Annals of Information Systems 21

Amit V. Deokar
Ashish Gupta
Lakshmi S. Iyer
Mary C. Jones *Editors*

Analytics and Data Science

Advances in Research and Pedagogy

Springer

Annals of Information Systems

Volume 21

More information about this series at http://www.springer.com/series/7573

Amit V. Deokar • Ashish Gupta
Lakshmi S. Iyer • Mary C. Jones
Editors

Analytics and Data Science

Advances in Research and Pedagogy

Springer

Editors

Amit V. Deokar
Robert J. Manning School of Business
University of Massachusetts Lowell
Lowell, MA, USA

Ashish Gupta
Raymond J. Harbert College of Business
Auburn University
Auburn, AL, USA

Lakshmi S. Iyer
Walker College of Business
Appalachian State University
Boone, NC, USA

Mary C. Jones
College of Business
University of North Texas
Denton, TX, USA

ISSN 1934-3221 ISSN 1934-3213 (electronic)
Annals of Information Systems
ISBN 978-3-319-58096-8 ISBN 978-3-319-58097-5 (eBook)
DOI 10.1007/978-3-319-58097-5

Library of Congress Control Number: 2017946842

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

About the Authors

Meng-Hsien (Jenny) Lin is an Assistant Professor of Marketing at California State University, Monterey Bay. Her research interests include studying various individual differences factors in the context of sensory marketing (influence of olfactory sensitivity on consumer behavior), advertising (gender differences and information processing in children), and focuses the mediating role of emotions on these relationships. She studies these topics using multi-methods, including behavioral experiments, survey research, neuroscience, and in-depth interviews. Her work also involves pedagogical research in marketing. Some of Dr. Lin's work has been published in *Neuroscience* and *Journal for the Advancement in Marketing Education*. Her research has implications for theory, public policy, and consumer well-being issues. Dr. Lin received her Ph.D. in marketing and an M.B.A. from Iowa State University.

William Jones is an Assistant Professor of Marketing at the Beacom School of Business at the University of South Dakota. Billy's work has explored consumers' use of numbers, behavioral pricing, individual differences in consumers' sensory processes, and issues in marketing education. Dr. Jones's work has been or will be published in *Biological Psychology, Journal for the Advancement of Marketing Education*, and *Psychology & Marketing* among other journals and presentations at national and international conferences. Dr. Jones received his Ph.D. in marketing from the University of Kentucky, an M.B.A. from Georgia Southern University, and a bachelor's degree in psychology from the University of Scranton.

Samantha N. N. Cross is an Associate Professor in Marketing in the College of Business at Iowa State University. Her research examines how diverse entities, identities, perspectives, beliefs, ways of sensing, and consuming co-exist in individuals, households, and society. Current research streams examine diverse cultural influences on decision-making, consumption, and innovation within the home; the impact of sensory influences on consumer identity and purchase behavior within the marketplace; and innovations in research methodology. She has received several awards for her research, including the Jane K. Fenyo Best Paper Award for Student Research, the ACR/Sheth Foundation Dissertation Award, and the Best Paper in Track Award at the American Marketing Association (AMA) Winter Conference. She has presented her work in several forums, both nationally and internationally. Her work has been accepted for publication in the *Journal of Marketing, the International Journal of Research in Marketing, Journal of Public Policy and Marketing, Journal of Business Research, Journal of Macromarketing, and Consumption, Markets and Culture*. Dr. Cross received her Ph.D. in marketing from the University of California, Irvine, her M.B.A. in international business from DePaul University, and a B.Sc. in management studies from the University of the West Indies.

Chapter 1
Exploring the Analytics Frontiers Through Research and Pedagogy

Amit V. Deokar, Ashish Gupta, Lakshmi S. Iyer, and Mary C. Jones

Abstract The 2015 Business Analytics Congress (BAC) brought together academic professionals and industry representatives who share a common passion for research and education innovation in the field of analytics. This event was organized by the Association for Information System's (AIS) Special Interest Group on Decision Support and Analytics (SIGDSA) and Teradata University Network (TUN) and held in conjunction with the International Conference on Information Systems (ICIS 2015) in Ft. Worth, Texas from December 12 to 16, 2015. The theme of BAC 2015 was *Exploring the Analytics Frontier* and was kept in alignment with the ICIS 2015 theme of *Exploring the Information Frontier.* In the spirit of open innovation, the goal of BAC 2015 was for the attendees to contribute their scientific and pedagogical contributions to the field of business analytics while brainstorming with the key industry and academic leaders for understanding latest innovation in business analytics as well as bridge industry-academic gap. This volume in the Annals of Information Systems reports the work originally reviewed for BAC 2015 and subsequently revised as chapters for this book.

Keywords Decision support • Business analytics • Congress • Business intelligence • Panels • Research • Pedagogy

A.V. Deokar (✉)
Robert J. Manning School of Business, University of Massachusetts Lowell, 72 University Ave, Pulichino Tong Business Center 436, Lowell, MA 01854, USA
e-mail: amit_deokar@uml.edu

A. Gupta
Raymond J. Harbert College of Business, Auburn University,
415 W. Magnolia Ave., 417 Lowder Hall, Auburn, AL 36849, USA
e-mail: ashish.gupta@auburn.edu

L.S. Iyer
Walker College of Business, Appalachian State University,
287 Rivers St, Boone, NC 28608, USA
e-mail: iyerLs@appstate.edu

M.C. Jones
College of Business, University of North Texas, 1155 Union Circle #305249,
Denton, TX 76203, USA
e-mail: mary.jones@unt.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_1

It has been a tradition for the AIS Special Interest Group on Decision Support and Analytics (SIGDSA) to organize the pre-International Conference on Information Systems (pre-ICIS) analytics workshop with the title of "Congress" when the event is held in the North American region. This "Congress" was the fourth such in its series that began in 2009. Planning for Business Analytics Congress held in December 2015 in Ft. Worth, Texas began in Fall 2014. The theme of Business Analytics Congress (BAC 2015) was decided as *Exploring the Analytics Frontiers* and was kept in alignment with the ICIS 2015 theme of *Exploring the Information Frontier*.

A major purpose of the Congress was to bring together a core group of leading researchers in the field to discuss the trends and future of business analytics in practice and education. This included discussion of the role of academicians in investigating and creating knowledge about applications of business analytics and its dissemination. This volume contributes to this purpose by striking a balance between investigating and disseminating what we know and helping to facilitate and catalyze movement forward in the field. This volume in the Annals of Information Systems includes papers that were originally reviewed for BAC 2015. These chapters were presented at BAC 2015 and subsequently revised for inclusion as chapters for this book.

BAC 2015 was sponsored by both industry and academia. The two main industry sponsors were Teradata University Network (TUN) and SAS, which in addition to providing financial support for the Congress, helped with bringing in distinguished speakers from industry. TUN also sponsored a reception for attendees the first evening of the event. Teradata University Network is a free, web-based portal that provides teaching and learning tools used by over 54,000 students and educators world-wide. These include majors as diverse as information systems, management, business analytics, data science, computer science, finance, accounting and marketing. The content provided by TUN supports instruction ranging from introductory information systems courses at the undergraduate level to graduate and executive level big data and business analytics classes. A key element of TUN success is that it is "led by academics to ensure the content will meet the needs of today's classrooms." SAS is a corporate leader in the provision of statistical and analytical software, services and support. SAS supports customers at over 80,000 sites around the world and provides several resources (www.sas.com/academic) for academics in support of their education needs. Academic sponsors included the University of Arkansas, University of North Carolina at Greensboro, University of North Texas, and University of Tennessee Chattanooga.

The day and a half BAC2015 event began on Saturday December 12th with several workshops. The first workshop was sponsored by SAS and focused on SAS® Visual Analytics and SAS® Visual Statistics. The workshop presented by Dr. Tom Bohannon focused on the basics of how to explore data and build reports using SAS Visual Analytics. It also covered topics on building predictive models in SAS Visual Statistics, such as decision tree, regression and general linear models.

The next workshop was sponsored by TUN and illustrated the vast academic resources available on TUN. It was presented by Drs. Barbara Wixom and Paul Cronan. The presenters discussed the rich repertoire of resources for faculty and students covering topics related to BI/Data Warehouse, database and analytics. Further, the talk session showcased software resources available from TUN and partnership with BI and Analytics companies such as MicroStrategy, SAS and Tableau that provide excellent resources to support analytics and visualization topics. The University of Arkansas is also a TUN partner and their resources were also discussed.

A workshop organized by Prof. Ramesh Sharda included Prof. Daniel Asamoah, Amir Hassan Zadeh, and Pankush Kalgotra and focused on pedagogical innovations related to delivering a Big Data Analytics course for MIS Programs. This session covered their experiences in offering a semester long course on Big Data technologies and included some hands-on demonstrations that they have used in their courses. Discussions also included the course outline and learning objectives followed by a description of various teaching modules, case studies, and exercises that they have developed or adapted.

The last session on Saturday was a panel on *Innovations in Healthcare: Actionable Insights from Analytics*. It was moderated and organized by Dr. Ashish Gupta. Panelists included Ms. Sherri Zink from BlueCross BlueShield of Tennessee, Ramesh Sharda from Oklahoma State University, David Lary from University of Texas Dallas and Ashish Gupta from Auburn University. This panel shared insights that have been derived using big data approaches, and how they have led to transformations in areas related to health. Example include analytics in insurance from consumer's perspective, sports, pollution and allergy management, utilizing disparate data using new data science paradigms such as deep learning framework and other enabling technologies.

The Sunday session began with an industry keynote by Ms. Sherri Zink, Senior VP, Chief Data and Engagement Officer, BlueCross BlueShield of Tennessee. The keynote address provided detailed insight into applications of analytics for empowering consumers, reducing redundant consumer touch points, optimal treatment plan based on information shared between provider and payer, informed decision making. Her talk provided an overview of how analytics could be used to develop a 360-degree view of consumers with the help of various approaches that foster the data integration, transformation & prediction, and eventually towards actionable insights. Key takeaways from the keynote address included a description of how clinical, life style and psychographic data could help develop a better understanding about consumer for stratification purposes using segmentation and clustering approaches. Such insights could help in developing better wellness programs and creating continuous feedback.

The keynote was followed by a panel entitled *AACSB Resources for Building a Business Analytics Program*. The panel was moderated by Dr. David Douglas and panelists included Drs. David Ahuja, Paul Cronan, Michael Goul, Eli Jones, Dan LeClair and Tom McDonald. The panel discussed AACSB's analytics initiative designed to help schools develop programs by providing a mix of curriculum content, pedagogy, and structure resources for schools contemplating development of or enhancement of Business Analytics. Panelists who were members of the AACSB Analytics Curriculum Advisory Group shared resources and encouraged interactive attendee discussion. Consistent with AACSB's goal of providing services to member schools across the globe, they shared information on initial analytics curriculum development seminars that are being be offered in the three cities that house AACSB's regional offices: Tampa (USA), Singapore, and Amsterdam.

Lunch and afternoon sessions focused on research presentations, both complete and research-in-progress, prototype and tutorials. The BAC 2015 event, for the first time, included a prototype presentation that highlighted various aspects of the prototype such as novelty, architecture, functioning, ongoing and future work, etc. Prototypes presented related to both teaching and research applications. Examples of such prototypes included original web applications, mobile apps, functional analytics models, devices (such as IoT) connected to data science applications, and teaching games that have an integrated study or analytics component. A report on one such prototype, *Say It Right: IS Prototype to Enable Evidence-Based Communication Using Big Data*, by Simon Alfano is included as a chapter in the book.

In keeping with the *Exploring the Analytics Frontier* theme of BAC 2015, the research track sought forward-thinking research in the areas of analytics and business intelligence, with special focus on the role of business intelligence and analytics in the creation, spread, and use of information. The research track was co-chaired by Drs. Barbara Dinter, Babita Gupta, and Anna Sidorova.

Likewise, the teaching track also aligned their call with the theme of the congress and sought pedagogical research contributions, teaching materials, and pedagogical practices/cases that address acquisition, application, and continued development of the knowledge and skills required in the usage of business analytics in the classroom, with emphasis on business intelligence, social media analytics, big data analytics, high performance analytics, data science, visualization, and other emerging analytic technologies. The teaching track was co-chaired by Drs. Sule Balkan, Joseph Clark, and Nick Evangelopoulos.

The later chapters in the book provide many of the research and teaching track papers, along with a prototype report, and a tutorial report.

Biographies

Amit V. Deokar is an Assistant Professor of Management Information Systems in the Robert J. Manning School of Business at the University of Massachusetts Lowell. Dr. Deokar received his Ph.D. in Management Information Systems from the University of Arizona. He also earned a M.S. in Industrial Engineering from the University of Arizona and a B.E. in Mechanical Engineering from VJTI, University of Mumbai. His research interests include data analytics, enterprise data management, business intelligence, business process management, and collaboration processes. His work has been published in journals such as *Journal of Management Information Systems*, *Decision Support Systems* (*DSS*), *The DATA BASE for Advances in Information Systems*, *Information Systems Frontiers* (*ISF*), *Business Process Management Journal* (*BPMJ*) and *IEEE Transactions*. He is currently a member of the editorial board of DSS, ISF, and BPMJ journals. He has been serving as the *Decision Support and Analytics* Track Chair at the international AMCIS 2014–17 conferences, and is currently the Chair of the AIS Special Interest Group on Decision Support and Analytics (*SIGDSA*). He was recognized with the 2014 IBM Faculty Award for his research and teaching in the areas of analytics and big data.

Ashish Gupta is an Associate Professor of Analytics in Raymond J. Harbert College of Business at the Auburn University. Prior to this, he served as the (founding) director of Analytics Research Center and an Associate Professor of Analytics & IS in the College of Business at the University of Tennessee Chattanooga. He has been a Visiting Research Scientist at the Mayo Clinic Rochester, Visiting Associate Professor in Biomedical Informatics at the Arizona State University and research affiliate with University of Tennessee Health Science Center in Memphis. He has a Ph.D. in MSIS from Spears School of Business at Oklahoma State University. Dr. Gupta's research interests are in the areas of data analytics, healthcare informatics, sports analytics, organizational and individual performance. His recent articles have appeared in journals such as *MIT Sloan Management Review*, *Journal of Biomedical Informatics*, *IEEE Transactions*, *Information Systems Journal*, *European Journal of Information Systems*, *Decision Support Systems*, *Information Systems Frontiers*, and *Communications of the Association for Information Systems*. His research has been funded by several agencies and private enterprises. He has published four edited books.

Lakshmi S. Iyer is Professor of Information Systems and Director of the Master's in Applied Data Analytics Graduate Programs at the Walker College of Business, Appalachian State University. Her research interests are in the area of business analytics, knowledge management, emerging technologies & its impact on organizations and users, and social inclusion in computing. Her research work has been published in or forthcoming in *Communications of the AIS*, *Journal of Association for Information Systems*, *European Journal of Information Systems*, *Communications of the ACM*, *Decision Support Systems*, *eService Journal*, *Journal of Electronic Commerce Research*, *International Journal of Business Intelligence Research*, *Information Systems Management*, *Journal of Global Information Technology and Management*, and others. She is a Board member of Teradata University Network, recent past-chair of the Special Interest Group in Decision Support and Analytics (*SIGDSA*, formerly *SIGDSS*). She has served as a Guest Editor for *Communications of the ACM*, and the *Journal of Electronic Commerce Research*. She is also co-editor of *Annals of Information Systems* Special Issue on "Reshaping Society through Analytics, Collaboration, and Decision Support: Role of BI and Social Media," from the 2013 pre-ICIS workshop in Milan, Italy.

Mary C. Jones is Professor of information systems and Chair of the Information Technology and Decision Sciences Department at the University of North Texas. She received her doctorate from the University of Oklahoma in 1990. Her work appears in numerous journals including *MIS Quarterly*, *European Journal of Information Systems*, *Behavioral Science*, *Decision Support Systems*, *System Dynamics Review*, and *Information and Management*. Her research interests are primarily in the impact on organizations of large scale, organizational spanning information systems such as ERP or business intelligence systems. She teaches a variety of courses including Enterprise Applications of Business Intelligence, IT Project Management, and a doctoral seminar in General Systems Theory.

Chapter 2
Introduction: Research and Research-in-Progress

Anna Sidorova, Babita Gupta, and Barbara Dinter

Abstract Inspired by the theme "Exploring the Information Frontier" of the ICIS 2015 conference, the Pre-ICIS Business Analytics Congress workshop sought forward-thinking research in the areas of data science, business intelligence, analytics, and decision support with a special focus on the state of business analytics from the perspectives of organizations, faculty, and students. The research track aimed to promote comprehensive research or research-in-progress on the role of business intelligence and analytics in the creation, spread, and use of information. This work has been summarized in this chapter.

Keywords Business intelligence • Analytics • Data science • Big data • Social media analytics • Decision support systems • Curriculum design • Pedagogy

2.1 Introduction

Business Intelligence and Analytics (BI&A) have become core to many businesses as they try to derive value from data. Although addressed by research in the past few years, these domains are still evolving. For instance, the explosive growth in big data and social media analytics requires examination of the impact of these

A. Sidorova (✉)
University of North Texas, 365D Business Leadership Building, 1307 West Highland Street,
Denton, TX 76201, USA
e-mail: anna.sidorova@unt.edu

B. Gupta
California State University Monterey Bay, Room 326, Gambord BIT Building,
100 Campus Center, Seaside, CA 93955, USA
e-mail: bgupta@csumb.edu

B. Dinter
Faculty of Economics and Business Administration, Chemnitz University of Technology,
Chemnitz, Germany
e-mail: barbara.dinter@wirtschaft.tu-chemnitz.de

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_2

technologies and applications on business and society. As organizations in various sectors formulate IT strategies and investments, it is imperative to understand how various technologies and applications under the BI&A umbrella such as business intelligence, data warehousing, big data and big data analytics, decision support, and data visualization contribute to organizational information processing, and ultimately organizational success.

In the rest of this editorial we introduce the papers included in this chapter. The papers address three broad issues: (1) business intelligence and analytics capabilities and organizational impact, (2) social media analytics, and (3) individual, organizational and societal implications of big data. The remainder of this editorial is structured around these themes.

2.2 Organizational Use and Impact of Business Intelligence and Analytics

As organizations invest heavily in BI&A in hopes to improve their competitive stance, researchers seek to develop theoretical frameworks that explain the strategic role of BI&A, and help better understand the key factors associated with successful organizational implementation and usage of BI&A. The papers presented here build on several theoretical perspectives, including the capabilities view of the firm, value chain model, and the IS success model.

A paper titled *Business Intelligence Capabilities* (Ramakrishnan et al. 2018) proposes a theoretical framework for understanding core business intelligence capabilities. As BI systems become an integral part of value delivery in modern organizations, organizations need to go beyond the view of BI as a tool or artifact, and focus on developing BI capabilities. The authors draw on the IT capabilities framework and propose three categories of BI capabilities, BI innovation infrastructure capability, BI process capability and BI integration capability, which contribute to the organizational success. The proposed taxonomy can be used to inform practitioners engaged in building BI capabilities in their organizations. It also represents an important step in developing comprehensive nomological network of BI capabilities.

A (RIP) paper titled *Big Data Capabilities: An Organizational Information Processing Perspective* (Isik 2018) presents a theoretical model of big data capabilities that is inspired by the organizational information processing perspective. The paper argues that the realization of value from big data depends on an adequate fit between big data processing requirements and big data processing capabilities. Another (RIP) paper titled *Business Analytics Capabilities and Use: A Value Chain Perspective* (Bedeley et al. 2018) proposes a value chain based approach for analyzing business analytics (BA) capabilities of a firm. The authors analyze extant academic and practitioner literature and identify and describe how descriptive, predictive, and prescriptive analytics is used in primary and supporting activities of Porter's (2001) value chain. The literature analysis suggests that organizations focus on building BA capabilities in value chain activities where they measure the outcome of BA use in terms of the firm value.

Building on the IS success model, a paper titled *Critical Value Factors in Business Intelligence Systems Implementations* (Dooley et al. 2018), proposes and empirically tests a theoretical model on business intelligence system success. The paper extends Delone and McLean's model of IS success (Delone and McLean 2003) by relating critical success factors identified in extant BI&A research perceived information quality and perceived system quality. Through the use of survey methodology, the study finds empirical support for the relationships among critical success factors, perceived information quality, perceived system quality and user satisfaction with the system and with the information provided by the system.

Song et al. (2018) present in their paper *Business Intelligence Systems Use in Chinese Organizations* an international perspective on BI&A systems by investigating the impact of natural culture, in particular of Guanxi, a universal and unique Chinese cultural form. The authors have conducted a series of interviews in two indigenous Chinese organizations (including Alibaba) in order to test previously identified research constructs. Based on the results five propositions of BI systems use in Chinese organizations have been formulated, introducing a Guanxi perspective in BI use theories. Their results confirm that national culture has a significant impact on BI&A usage in China. Future research should be guided by these insights given the high relevance and influence of Chinese firms worldwide.

2.3 Social Media Analytics

Web 2.0 and social media facilitate the creation of vast amounts of digital content that represents a valuable data source for researchers and companies alike. Social media analytics relies on new and established statistical and machine learning techniques to derive meaning from large amounts of textual and numeric data. In this section we present several papers that seek to advance social media analytics methods and to demonstrate how social media analytics can be applied in a variety of contexts to deliver useful insight.

The first paper in this category, titled *The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps* (Qiao et al. 2018) addresses a research field with promising opportunities—analyzing Web 2.0 data to foster innovation. The article examines the role played by customer reviews in influencing product innovations in the context of mobile applications. In particular, the authors verify the impact of online mobile app reviews on developers´ product innovation decisions and identify the characteristics of such reviews that increase the likelihood of future app updates. The findings suggest that it is important to explore user generated reviews in the context of customer-centered product innovation.

The paper *Whispering on Social Media* (Zhang 2018) examines the role of information circulated on social media in influencing stock performance during the so-called "quiet period" before an initial public offering (IPO). During such quiet periods organizations are not allowed to disclose any information that might influence investors´ decisions. Nevertheless, people discuss and comment about upcoming IPS's in social media. The author finds in her research that the number of IPO-related tweets (and re-tweets) have significant positive correlation with the IPO's first-day return, liquidity and volatility.

The next contribution in this category presents another interesting use case for social media analytics. The paper titled *Does Social Media Reflect Metropolitan Attractiveness? Behavioral Information from Twitter Activity in Urban Areas* (Bendler et al. 2018) describes how the analysis of social media activities can generate insights for urban planning. When tweets are combined with other data such as the temporal information, spatial coordinates, appended images, videos, or linked places, a variety of applications can be supported, for example city planning, city safety, and investment decisions. For these purposes, the paper presents methods and measures for identifying the places of interest.

The paper titled *The Competitive Landscape of Mobile Communications Industry in Canada—Predictive Analytic Modeling with Google Trends and Twitter* (Szczech and Turetken 2018) describes how social media and Google Trends can be analyzed to predict competitive performance. Their predictive model builds on the previous studies that use Google Trends for predicting economic and consumer behavior trends in a particular business or industry. The authors improve these existing models by adding competition variables and incorporate Twitter Sentiment scores into their models to discover if Twitter sentiment scores modify some of the variance in the dependent variable that is not already explained by Google Trends data.

The research-in-progress paper titled *Scale Development Using Twitter Data: Applying Contemporary Natural Language Processing Methods in IS Research* (Agogo and Hess 2018) illustrates the use of Twitter data analytics for scale development. With the rise in social media communication, these data are becoming an important source to understand consumer behavior. However, challenges abound in transitioning the traditional measurement scales into social media data such as tweets. This paper uses natural language processing methods to develop measurement scales using big data such as tweets. They present a new scale called the technology hassles and delights scale (THDS) to show how the content validity of the scale can be improved by using a syntax aware filtering process that identifies relevant information from analyzing 146 million tweets.

2.4 Individual, Organizational and Societal Implications of Big Data

The rise of big data and associated analytical techniques has important implication not only for organizational, but for the society in general.

The research-in-progress paper titled *Information Privacy on Online Social Networks: Illusion-in-Progress in the Age of Big Data?* (Sharma and Gupta 2018) focusses on the issues of privacy and information disclosure on social media. They present a research model that draws together concepts from behavioral economic theory, the prospect theory which is an extension of expected utility hypothesis, and the rational apathy theory, which is derived from the public choice theory in social psychology. The research methodology investigates why people choose to disclose vast amounts of personal information voluntarily on Online Social Networks (OSN). The proposed research model considers the effect of situational factors such as the information control, ownership of personal information, and apathy towards privacy concern of users on OSN. The article proposes value to practitioners in many different ways, the OSN providers and third parties could better understand how consumer's information disclosure behavior works and we could better understand why people tend to disclose too much of their personal information on OSN.

The research article, titled *Online Information Processing of Scent-Related Words and Implications for Decision Making* (Lin et al. 2018) takes a broader view of human information processing by examining the role of olfactory information in decision making. The authors propose a methodology to examine emotions triggered by olfactory-related information and how these could be simulated using visual cues in the context of consumer decision-making online. The methodology combines approaches from neuroscience with behavioral experiments. Their work studies the effectiveness of triggering olfactory emotions using sensory congruent brand names in online ads and also examines the influence on the consumers' attitudes and intentions towards brand and purchases. Results show that individual differences in olfactory sensitivity moderate the effects on cognitive and emotional processes. This work has implications for online advertising and marketing decisions made by the consumers.

2.5 Conclusion

The research work presented at the Special Interest Group on Decision Support and Analytics (SIGDSA) Workshop held on Dec 12, 2015, Fort Worth, TX was of considerable variety in addressing the issues facing the researchers in the business intelligence and analytics area. The research work included here represents some of the innovations taking place in the analytics, combining theories from not only information systems but also diverse fields such as neuroscience, psychology, behavioral economics, and social sciences. Future research promises to exciting with opportunities to extend literature and methodologies presented here to further the field of decision support systems in the context of business intelligence.

Biographies

Anna Sidorova is an Associate Professor in the Department of Information Technology and Decision Sciences at the University of North Texas. She received his Ph.D. in information systems from Washington State University. Her current research interests include strategic use of business intelligence and analytics, business process management, information sharing on social media, and adoption and use of intelligent systems. Her articles appear in MIS Quarterly, Journal of Management Information Systems, Journal of the AIS, Decision Support Systems, Information & Management, Communications in Statistics, Computational Statistics & Data Analysis, and others.

Babita Gupta is a Professor of Information Systems at the College of Business, California State University Monterey Bay. She received his Ph.D. from the University of Georgia. Her research interests are in the areas of business intelligence, online security and privacy, adoption of ICTs in government, and role of culture in IT. She has published in Communications of the Association for Information Systems, Journal of Electronic Commerce Research, the Journal of Strategic Information Systems, the Communications of the ACM, the Journal of Industrial Management and Data System, and the Journal of Information Technology Cases and Applications and others. She also serves on the Advisory Board of the Teradata University Network.

Barbara Dinter is Professor and Chair of Business Information Systems at Chemnitz University of Technology, Germany. She holds a Ph.D. from the Technische Universität München, Germany, where she previously earned a master's degree in computer science. She has worked for several years at University of St. Gallen, Switzerland as a Post-Doc and project manager. In her role as an IT consultant, she worked with a variety of organizations. Her research interests include business intelligence and analytics, big data, data driven innovation, and information management. She has published in renowned journals such as Decision Support Systems, Journal of Database Management, and Journal of Decision Systems, and on conferences such as ICIS, ECIS, and HICSS.

References

Agogo D, Hess TJ (2018) Scale development using Twitter data: applying contemporary natural language processing methods in IS research. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 163–178. http://www.springer.com/series/7573

Bedeley RT, Ghoshal T, Iyer LS, Bhadury J (2018) Business analytics capabilities and use: a value chain perspective. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 41–54. http://www.springer.com/series/7573

Bendler J, Brandt T, Neumann D (2018) Does social media reflect metropolitan attractiveness? Behavioral information from Twitter activity in urban areas. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 119–142. http://www.springer.com/series/7573

DeLone WH, McLean ER (2003) The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst 19(4):9–30

Dooley PP, Levy Y, Hackney RA, Parrish JL (2018) Critical value factors in business intelligence systems implementations. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 55–78. http://www.springer.com/series/7573

Isik O (2018) Big data capabilities: an organizational information processing perspective. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 29–40. http://www.springer.com/series/7573

Lin M-H, Cross SNN, Jones WJ, Childers TL (2018) Online information processing of scent-related words and implications for decision making. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 197–216. http://www.springer.com/series/7573

Porter ME (2001) Strategy and the Internet, Harvard Business Review (79:3). Harvard Business School Publication Corp, pp 62–78

Qiao Z, Wang A, Zhou M, Fan W (2018) The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 95–110. http://www.springer.com/series/7573

Ramakrishnan T, Khuntia J, Saldanha T, Kathuria A (2018) Business Intelligence Capabilities. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 15–27. http://www.springer.com/series/7573

Sharma S, Gupta B (2018) Information privacy on online social networks: illusion-in-progress in the age of big data? In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 179–196. http://www.springer.com/series/7573

Song Y, Arnott D, Gao S (2018) Business intelligence system use in Chinese organizations. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 79–94. http://www.springer.com/series/7573

Szczech M, Turetken O (2018) The competitive landscape of mobile communications industry in Canada—predictive analytic modeling with Google Trends and Twitter. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 143–162. http://www.springer.com/series/7573

Zhang J (2018) Whispering on social media. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series. 111–118. http://www.springer.com/series/7573

Chapter 3
Business Intelligence Capabilities

Thiagarajan Ramakrishnan, Jiban Khuntia, Abhishek Kathuria, and Terence J.V. Saldanha

Abstract Business intelligence (BI) is emerging as a critical area of expertise for firms' value proposition. Firms are trying to leverage BI as an inherent capability to create value. Considering an organizational systems view, BI extends beyond a tool or artifact to include a number of capabilities. We draw on IT capabilities and prior research on BI to uncover potential capabilities that BI bestows to an organization. A three category BI capability classification is suggested: BI innovation infrastructure capability, BI process capability and BI integration capability. We discuss the attributes of these three BI capabilities to provide insights into how the capabilities help organizations. This taxonomy will help decision-makers take informed decisions on how to effectively implement BI within their organization to improve performance.

Keywords BI capabilities • BI innovation infrastructure capability • BI process capability • BI integration capability • IT capability

T. Ramakrishnan (✉)
College of Business, Prairie View A&M University, 805 A.G. Cleaver St.,
Agriculture/Business Multipurpose Building, Room 447, Prairie View, TX 77446, USA
e-mail: Ram@pvamu.edu

J. Khuntia
Business School, University of Colorado Denver,
1475 Lawrence Street, Denver, CO 80202, USA
e-mail: jiban.khuntia@ucdenver.edu

A. Kathuria
Faculty of Business & Economics, The University of Hong Kong, Pokfulam, Hong Kong
e-mail: kathuria@hku.hk

T.J.V. Saldanha
Carson College of Business, Washington State University,
Todd Hall, Pullman, WA 99164, USA
e-mail: terence.saldanha@wsu.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_3

3.1 Introduction

Business Intelligence (BI) is referred to the techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions (Ramakrishnan et al. 2012). BI helps transform large amount of data from disparate sources into meaningful information to support decision making. BI investment is estimated to grow from $54.5 billion in the year 2012 to $96.9 billion in the year 2016 (Tabbitt 2013). BI is being used in almost all industry sectors and is a top priority for organizations (Isik et al. 2013). The opportunities associated with business analytics in different organizations have helped generate significant interest in BI. In addition to the underlying data processing and analytical technologies, BI includes business-centric practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, e-government, healthcare, and security applications.

The evolution of business intelligence has its roots in artificial intelligence and business analytics, and has entered into mainstream business and IT communities since the 2000s (Davenport 2006). Further, the database related technologies advanced avenues for data collection, extraction, and analysis in the business intelligence areas (Chaudhuri et al. 2011; Turban et al. 2008; Watson and Wixom 2007). Currently, BI involves both structured and unstructured (big) data analysis and intelligence gleaning. Very large (from terabytes to exabytes), real time (feeds and tweets) and complex (from sensor to social media) data is emerging central tenet to recent BI developments. In addition, BI involves analytical techniques in applications that require advanced and unique data storage, management, analysis, and visualization technologies.

Recent developments in the internet, web, social media and mobile systems have offered unique data collection and analytical abilities to BI area. Large amounts of company, industry, product, and customer information can be gathered from the web and organized and visualized through various text and web mining techniques. Web analytics tools can gather customer clickstream data logs. Social media data analytics presents a unique opportunity for businesses to treat markets as avenues of business-customer relationship based co-creation (Lusch et al. 2010). Furthermore, mobile applications ranging from information advisories and ecommerce infomediaries and aggregators to gaming systems, often with billions of users, are changing the way intelligence and analytics fields are helping businesses and societal developments. It is noteworthy to mention that along with businesses, sectors such as healthcare, education, and governments have been benefitted a lot from business intelligence area. Emerging technologies and developments regarding Internet of Things (sensors, RFIDs, barcodes, tags), or drone based surveillance or monitoring systems are providing conduits for highly mobile, location-aware, person-centered, and context-relevant operations and transaction data. Indeed, many agree that both practice and academic communities face unique challenges and opportunities in understanding, developing, researching and educating the next generation BI students, researchers and professionals (Chen 2011).

Notwithstanding the increasing trend in BI adoption and implementation, the return on investments from BI remains a complex puzzle for many organizations. Some practitioners note that only around 20% firms have been able to convert BI to tangible benefits (Henshen 2008). Although it sounds simple, but initiation, implementation and development of a set of capabilities that can leverage on BI is not an easy task, and often needs integration of a set of distinctly different capabilities, ranging from information infrastructure to analytical mindsets. Once developed and used to its best extent, BI can be influential in organizations, and helpful in decision making or efficiency enhancements (Popovic et al. 2012; Wixom and Watson 2001). Furthermore, some even suggest that BI capabilities can be an important strategy within organizations forming their position in a competitive landscape (Thamir and Poulis 2015). Thus, given the wide applications and understanding of BI, it is important that BI capabilities be explicated in a simple yet holistic manner. In addition, given a firm wants to move towards BI implementations, managers should have an understanding on what capabilities need to be developed, or which directions need to be taken with an integrated perspective of BI capabilities.

The goal of this chapter is to highlight a typology of BI capabilities in organizations. We provide three categories of classification for BI capability in organizations: (1) *BI innovation infrastructure* consists of the foundational ability to mobilize and deploy BI functionalities to support innovation in the organization through infrastructure, culture and technological improvements; (2) *BI process capability* is the penetration of BI into the firm's customer centric and business-to-business (B2B) centric processes, and (3) *BI integration capability* refers to how the organization builds and integrates such capability and develops ways to acquire and convert business intelligence towards organizational improvement. Salient features and components of each type of BI capability are suggested to help in understanding in practice and further research prospects. The next section focuses on the understanding of BI capability, followed by the approach to develop the taxonomy and detailed description of each type of capability. Finally, suggestions on how to use the taxonomy, including managerial and research implications are discussed.

3.2 What is BI?

Several definitions of BI reflecting on different perspectives have been suggested. Moss and Atre (2007) define BI from a technological perspective as "an architecture and a collection of integrated operational as well as decision support applications and databases that provide the business community easy access to business data." However, Olszak and Ziemba (2003) define BI from an organizational perspective as "a set of concepts, methods and processes that aim at not only improving business decisions but also at supporting realization of an enterprises' strategy."

The effectiveness of BI is situated in its ability to support decision-making within an organization and providing decision-makers with timely and relevant information (Buchanan and O'Connell 2006; Massa and Testa 2005; Ramakrishnan et al. 2012).

Researchers have examined the benefits of implementing BI (Cooper et al. 2000; Watson et al. 2004), implementation factors (Hwang et al. 2004), and decision making (Park 2006). Organizations are struggling to make sense of the growing variety, velocity, and volume of data; demanding development of BI capabilities to deal with the data produced by internal and external sources, and leverage it to improve performance. Prior work on BI capabilities focuses mainly on the technical and organizational aspects of BI. For example, Sukumaran and Sureka (2006) examine BI capability as the ability of BI to manage quantitative and qualitative data. Similarly, BI capability has been seen in terms of a tool that can manage internal and external data (Harding 2003; Hostmann et al. 2007; Isik et al. 2013). From an organizational perspective BI capability has been examined as the ability of BI to provide support for decision making under conditions of uncertainty (Harding 2003; Gebauer and Schober 2006; Isik et al. 2013). An overarching view of the capabilities that BI endows in organizations in terms of supporting innovation, integration, and different process is still a gap in the BI literature; this chapter tries to fill in this gap.

3.3 Classification of BI Capabilities

We draw on prior work of IT capability and BI capability to propose that BI capabilities help orient a firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments; using business intelligence as a tool, artifact, and process level integrative capabilities. IT capability has its roots in the resource-based view that suggests that organizations' gain competitive advantage through the application of a combination of resources that are non-substitutable, scarce, difficult to imitate, and economically valuable (Barney 1991). Bharadwaj (2000) define IT capability as a firms' "ability to mobilize and deploy IT-based resources in combination or co-present with other resources and capabilities," (p. 171). Early studies with regards to IT capability started with viewing IT capability within single dimension in terms of either technological capability (Sabherwal and Kirs 1994) or managerial capability (Sambamurthy and Zmud 1997) and has now evolved to comprise three dimensions: technological dimension, human dimension, and organizational dimension (Kim et al. 2011; Schaefferling 2013). The technological dimension refers to the configuration and structure of all the technological elements in a firm such as hardware, software, networking and telecommunications, and different applications; the human dimension of IT capability discusses the knowledge and skill sets of the IT worker in a firm to manage and leverage IT to achieve a competitive advantage for the firms. Similarly, the organizational dimension examines the influence of organizational resources and the IT/business partnership that can provide the organization with a competitive advantage (Melville et al. 2004; Bhatt and Grover 2005; Rockmann et al. 2014).

BI as a capability is more so justified as a process or operational capability (Isik et al. 2013). Following prior work we conceptualize that BI capability overall is a culmination of different process or operational capabilities, and in addition, provides a second layer or integrative capability in the organization. This integrative capability is manifested through the three underlying three dimensions: (1) integrate BI within the organization (integration of data and intelligence), (2) align BI towards innovation (infrastructure frontier), and (3) use BI to improve customer centric and business partner centric processes (process orientation) (see Table 3.1). These three dimensions translate to the three BI capabilities: BI innovation infrastructure capability, BI process capability (consisting of customer centric and B2B centric process capabilities) and BI integration capability. We elaborate on these three dimensions further in the next sub-sections.

3.3.1 BI Innovation Infrastructure Capability

BI Innovation capability is the ability to marshal and use the functionalities of BI to sustain innovation in organizations through technological, cultural, and infrastructure improvements. In order to support BI technology the proper infrastructure and the right data collections strategy for BI is needed (Ramakrishnan et al. 2012). Further, in order to leverage BI technology, it is imperative to have the appropriate organizational structure that can facilitate sharing and collaboration. Along, the same lines, culture also plays an important role in facilitating sharing and leveraging of information generated by BI. BI technology plays a crucial role in supporting decision-making within any organization (Isik et al. 2013).

BI innovation infrastructure capability constitutes technical, structural and cultural elements. First, BI technology refers to the degree and extent of technological readiness to adopt BI in the organization. The technology dimension may also include business intelligence, collaboration, distributed learning, discovery, mapping, opportunity recognition and generation as well as aspects related to security and privacy of the data and analytics. The structural element of BI innovation infrastructure refers to the modular organizational design that helps facilitate the technical architecture and subsequent functions and innovations relevant to BI. BI culture facilitates a firm's ability to manage data, knowledge and intelligence; and espouses interaction between individuals and groups is a basis of the creation of new ideas and innovation.

Technical, structural and cultural elements associated with BI innovation infrastructure provide the abilities to a firm that help in managing data, knowledge and intelligence through embedded routines and processes of the organization. Technology plays an important role in the structural dimension needed to capture, store, and analyzed data in a firm. The various communication systems and information systems can be linked in an organization to integrate the previously fragmented flow of data and information (Teece et al. 1997). These linkages can eradicate the hurdle to communication between different business units and enable collaboration among them. Further, BI technology can endow firms with the ability to engender information and knowledge regarding their external fiscal environment and their competition (Gold et al. 2001). Effective utilization of BI technology can help organizations deal with competitive and institutional pressures that firms face within an industry (Ramakrishnan et al. 2012).

Table 3.1 Conceptualization of BI capabilities and dimensions

Category	Core	Description	References
Infrastructure Frontier	Codification, connectivity and flow of data and information to derive intelligence	• Codification of specialized data and information from different organizational elements to be used by qualified staff or personnel • Detection, classification and planning of organizational data and information to be accessed by others • Provider formal access and provision to staff and employees to contextually use the data and information • Creating a culture or practice of intelligence based decision making.	Sukumaran and Sureka (2006), Parikh and Haddad (2008), Hostmann et al. (2007), Harding (2003).
Process Orientation	Exploitation of infrastructure and integration for crating organizational value through workflow and process coordination levels	• Conceptualize and execute BI as essential dimensions at each and every process and workflow levels • Realization that the actions related to BI can be used to create and facilitate economic and strategic values for the organization • Exploitation of BI as a capability-asset to produce income and maximize profit	Li et al. (2008), Sahay and Ranjan (2008), Elbashir et al. (2008), Isik et al. (2013), Wixom et al. (2011)
Integration of Data and Intelligence	Design and integration of spaces, practices and connectivity to foster the activities around data, information and intelligence gathering and conversion	• Design and use of organizational structures or networks • to acquire expertise and skills for intelligence generation • to acquire data and information from external sources • to convert the data and information to intelligence by using the gathered expertise and skills • seamless integration of the acquisition and conversion process within the organization	White (2005), Hostmann et al. (2007), Gebauer and Schober (2006), Petrini and Pozzebon (2009)

BI structure establishes an organizational framework and readiness to accommodate and leverage this foundation, while BI technology provides the foundation. Structure examines the distribution of tasks, coordination, flow of information, and decision-making rights within an organization (Pugh 1990). Further, firms with rigid structure may have the unintended effect of inhibiting the sharing of information and knowledge across internal boundaries (Gold et al. 2001), rather than enabling communication and collaboration. Therefore, we argue that in order to leverage BI technology it is important to have BI structure in place that encourages the sharing and exchange of information and intelligence. Organizations need to promote collective intelligence rather than individualistic acumen. Firms need to facilitate the transfer of intelligence across internal boundaries. Thus, BI structure plays an important role in supporting BI technology, and hence, is an important element in BI capabilities taxonomy.

Finally, BI culture espouses interactions between individuals and groups as a basis of the creation of new ideas and innovation. Thus, a more interactive and collaborative culture is a precursor for converting the data or fact based tacit information to more explicit intelligence, and move it from an individual to an organizational level. Employees in such a cultural glue within the organization can develop an ability to self-organize their knowledge and practices to facilitate solutions to new or existing problems.

To establish the value proposition of BI innovation infrastructure capability, we suggest that a firm can foster innovation using the technical, structural and cultural elements of BI capabilities. Structural element of innovation infrastructure will allow data and information to be exchanged seamlessly between different business units, thus improving the effectiveness of BI towards higher performance. Further, having a culture that will facilitate interaction between individuals and groups to exchange information and intelligence generated by BI to come up with new innovative ideas will make the BI more effective.

3.3.2 BI Process Capabilities

BI process capabilities is the ability of BI to penetrate into the firms' business processes. This capability examines the functionalities of BI that can sustain both B2B centric and customer centric activities. We argue that BI helps organizations by supporting the business processes that give a firm a competitive advantage. Business processes in a firm help orient its activities towards value creation. To create value, a firm needs to do at least three activities; first, operations that can convert goods to products or services (i.e., operations); second, relationship with other firms who supply materials and products to the firm (e.g., firms in the supply chain), and third, orienting its operations to deliver products and services to the customers (i.e., customer oriented activities). As noted previously in this paper, the operational BI capabilities are embedded within infrastructural development related to BI, or, in other words, the infrastructural BI development caters to the operations. On the contrary, supply chain and customer oriented BI activities need to be explicitly developed; and included in the firm's value chains as two sets of capabilities to cater to the two ends of the value chain, i.e., supply chain partners and customers. Based on these concepts in the existing literature, we propose that for an organization to achieve competitive advantage, two explicit BI capabilities need to either exist or be developed—the customer centric and a business to business (B2B) process related BI capabilities elements. Although BI adoption and implementation is oriented predominantly towards customer centric data-information-knowledge-intelligence paradigm, similar process oriented approach of BI can also be found in leveraging B2B relationships or supply chain visibility areas. For example, BI in the B2B or supply chain can eliminate waste by providing demand aggregation or reducing the 'bullwhip effect' associated with distribution.

Because processes are not unilateral directives in an organization, and often consist of a multitude of orientations, we conceptualize the two dimensions of BI process oriented capabilities as multi-dimensional constructs. For example, customer centric BI process capability consists of the way BI is oriented to meet the firms' customer needs and serve them, elements that enhance customer satisfaction and loyalty by providing insights regarding customers' long term goals and requirements, and ability to absorb customer oriented information/intelligence into the organization using BI. Similarly, B2B centric BI process capability consists of BI applications related to supply chain integration, engage new partners and improve coordination with existing partners, and using BI for process coordination and operational improvements. Inherently, these dimensions relate to and influence organizational performance due to responsiveness to customer needs, awareness of customer goals and the ability to learn from information generated during customer interactions. Furthermore, B2B centric BI process capability aids activities with B2B partners due to insights through visibility of goods and information, business level integration, and process-level coordination across channels. Together, BI process capabilities provide firms with the capacity to derive analytical insights in its business processes which in turn enhance organizational effectiveness.

3.3.3 BI Integration Capability

Prior studies recognize BI integration to be very important and critical for the successful utilization of BI (Isik et al. 2013). Integration refers to combining different types of explicit data and information into novel patterns and relations (Herschel and Jones 2005). Based on the existing literature, we posit that organizations need to develop ways to acquire and convert business intelligence towards organizational performance.

We argue that BI integration capability has two dimensions that are effective towards organizational performance, albeit in an interconnected manner. First, BI acquisition consists of gathering data from different types of sources across the organization and beyond, in addition to data aggregation, rollup and partitioning. Data extracted from operational systems need to be cleansed and transformed in order to make it suitable for use without errors (Ramakrishnan et al. 2012). Second, the data need to be converted to usable patterns and schemas to help an organization to glean more insights from the data. Thus, BI Integration consists of the acquisition of data from various sources, followed by the conversion of data to the right format and quality in order to be used effectively in the organization.

As much as the acquisition and integration of business intelligence from various sources is a prerequisite for the utilization BI capabilities; the outcome of the acquisition and conversion through integration helps to achieve higher organizational performance. For instance, customer centric activities require acquisition of business intelligence regarding customer behavior and experience, which in turn provide insights regarding goals and requirements. Second, the gathering and aggregation of data from different types of sources across the organization and beyond enables the organization to leverage BI to adequately respond to market and environmental changes. Hence BI can provide insights regarding the nature of change to which the organization needs to adapt, as well as the internal changes required to do so. Third, aggregation, cleansing and transformation of this data can make this data more substantive and insightful, thereby making subsequent decisions faster and more effective. Thus, integration capability of BI that facilitates the gathering and cleaning of data from disparate data sources and providing the decision-makers with timely and usable information will make the BI more effective.

3.4 Using the Taxonomy

With the advent of business intelligence, organizations are somewhat moved by a ‘fad’ effect around this tool. In practice, while the buzz around BI is very high, BI perspectives and viewpoints vary across firms, with differing concepts, definitions and applications. While eliminating the differences would be a herculean task, integrating BI perspectives into holistic models can certainly be a fruitful approach. The intention of this chapter is to provide such a holistic view around BI integration in an organization, albeit with a bias towards a capability perspective. Taking the capabilities perspective helps to highlight the fact the BI is ‘not just a fad’ or buzz’ in the practice and academic discourse, but it can be helpful in garnering higher organizational performance. Indeed, the theoretical concepts and model developed in this chapter are oriented towards establishing the relationships between different dimensions of BI capabilities, their integrated schema, and the influence of these dimensions on organizational performance.

The classification schema can be helpful in further research. A line of research that can be pursued is relating the BI capabilities in a causal way. For example, a relationship model can test whether a BI innovation infrastructure capability can lead to a higher BI integration capability, and a higher BI integration capacity can lead to a higher process capability as suggested in a conceptual diagram in Fig. 3.1. Further research can explore relationships between capability types and organizational performance or BI effectiveness.

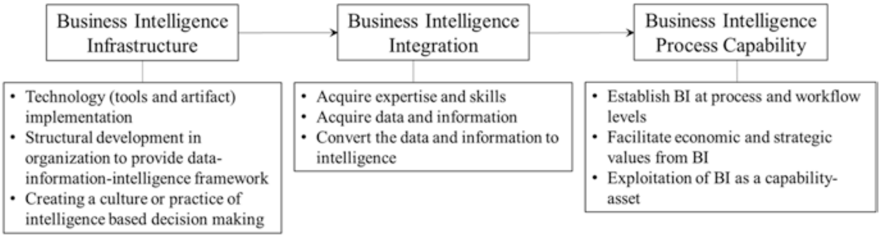

Business Intelligence Infrastructure
Business Intelligence Integration
Business Intelligence Process Capability
• Technology (tools and artifact) implementation
• Structural development in organization to provide data-information-intelligence framework
• Creating a culture or practice of intelligence based decision making
• Acquire expertise and skills
• Acquire data and information
• Convert the data and information to intelligence
• Establish BI at process and workflow levels
• Facilitate economic and strategic values from BI
• Exploitation of BI as a capability-asset

Fig. 3.1 A suggested relationship model between types of BI capabilities

The integrative model provides two theoretical contributions. Although existing studies note that BI helps in improving organizational decision making, the proposed model goes a step beyond to disentangle various dimensions of BI enabled capabilities. As a result, granular insights into BI orientation in a firm in relation to improving its capabilities are drawn. In addition, it is suggested here that BI is not a single directional or unilateral tool or perspective that just helps in siloed decision making; instead BI can be taken as a process-integrative organization wide framework that helps in improving firm performance.

The BI integrative model and dimensions have implications for managerial practice. The model provides directions to a step wise approach starting from acquisition to conversion and process integration for BI tools and applications. Furthermore, focusing on BI capabilities and integrating them into the functionalities of the organization may help in improving performance.

A number of factors might be unintentionally missing in the integrative model discussed here. For example, different industrial sectors might be leveraging on BI differently; indicating a variance in the model. In addition, early adopters and laggards may show variations with respect to BI capability and performance relationships. Additional research is warranted to extend this body of knowledge and related relationships. Future studies may focus on extending the model to other contexts by developing specific testable hypothesis on particular settings.

In conclusion, this chapter takes an integrative approach to BI. The central tenet of the chapter focusses on three dimensions of BI capability and relates it to organizational performance. The concepts proposed here are expected to provide a capability-integrative framework for BI implementation and motivate managers to see BI from organizational performance improvement perspective.

Biographies

Thiagarajan Ramakrishnan (Ram@pvamu.edu) is an Assistant Professor in the College of Business at Prairie View A&M University. He received his doctorate from the University of North Texas, Texas in 2010. His research interests include business intelligence and data mining, Information Systems discipline, disaster management, and Ecommerce. His work appears in such journals as *MIS Quarterly, Decision Support Systems*, and *Computers in Human Behavior*.

Jiban Khuntia (jiban.khuntia@ucdenver.edu) is an Assistant Professor in the Business School at University of Colorado, Denver. His research is in the areas of digital service innovation and digital health. He has published in *Decision Support Systems, Communications of the Association for Information Systems, Health Policy and Technology* and *International Journal of E-Business Research*. He received his Ph.D. from University of Maryland, College Park in 2013.

Abhishek Kathuria (kathuria@hku.hk) is an Assistant Professor of Innovation and Information Management at The University of Hong Kong. His research, teaching and consulting interests lie in the areas of business performance improvement, business & IT strategy, emerging technologies and technology entrepreneurship, with a focus on emerging economies. His work has been published in the Communications of the Association for Information Systems and nominated for and received best paper awards at the International Conference on Information Systems and the Annual Meeting of the Academy of Management. Abhishek received his Ph.D. from Emory University and a graduate degree in management from the Indian Institute of Management Indore.

Terence J.V. Saldanha (Terence.saldanha@wsu.edu) is Assistant Professor of Information Systems at the Carson College of Business at Washington State University, Pullman. He received his Ph.D. in Information Systems from the University of Michigan in 2012. His research interests include the business value of Information Systems (IS) and the role of IS in innovation. His research has appeared or is forthcoming in *MIS Quarterly, Journal of Operations Management, Journal of Organizational Computing and Electronic Commerce, Journal of Information Technology Theory and Application*, and in various academic conference proceedings. Prior degrees include an M.B.A. from S.P. Jain Institute of Management (Mumbai, India), and a Bachelor's of Engineering from University of Mumbai (India). Prior to his graduate studies, he spent about 4 years working in the IT industry.

References

Barney J (1991) Firm resources and sustained competitive advantage. J Manag 17:99–120

Bharadwaj AS (2000) A resource-based perspective on information technology capability and firm performance: an empirical investigation. MIS Q 24(1):169–196

Bhatt GD, Grover V (2005) Types of information technology capabilities and their role in competitive advantage: an empirical study. J Manag Inform Syst 22(2):253–277

Buchanan L, O'Connell A (2006) A brief history of decision making. Harv Bus Rev 84(1):32–40

Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Commun ACM 54(8):88–98

Chen H (2011) Design science, grand challenges, and societal impacts. ACM Trans Manag Inform Syst 2(1):1:1–1:10

Cooper BL, Watson HJ, Wixom BH, Goodhue DL (2000) Data warehousing supports corporate strategy at First American Corporation. MIS Q 24(4):547–567

Davenport TH (2006) Competing on analytics. Harv Bus Rev

Elbashir MZ, Collier PA, Davern MJ (2008) Measuring the effects of business intelligent systems: the relationship between business process and organizational performance. Int J Account Inf Syst 9:135–153

Gebauer J, Schober F (2006) Information system flexibility and the cost efficiency of business processes. J Assoc Inform Syst 7(3):122–145

Gold AH, Malhotra A, Segars AH (2001) Knowledge management: an organizational capabilities perspective. J Manag Inform Syst 18(1):185–214

Harding W (2003) BI crucial to making the right decisions. Financ Exec 19(2):256–268

Henshen D (2008) Special report: business intelligence gets smart. Inf Week

Herschel RT, Jones NE (2005) Knowledge management and business intelligence: the importance of integration. J Knowl Manag 9(4):45–55

Hostmann B, Herschel G, Rayner N (2007) The evolution of business intelligence: the four worlds. Gartner report. http://www.gartner.com/DisplayDocument?id=509002

Hwang H, Ku C, Yen DC, Cheng C, Critical Factors C (2004) Influencing the adoption of data warehouse technology: a study of banking industry in Taiwan. Decision Support Syst 37:1–21

Isik O, Jones MC, Sidorova A (2013) Business intelligence success: the roles of BI capabilities and decision environments. Inf Manag 50:13–23

Kim G, Shin B, Kim KK, Lee HG (2011) IT capabilities, process-oriented dynamic capabilities, and firm financial performance. J Assoc Inform Syst 12(7):487–517

Li S, Shue L, Lee S (2008) Business intelligence approach to supporting strategy-making of ISP service management. Exp Syst Appl 35:739–754

Lusch RF, Liu Y, Chen Y (2010) The phase transition of markets and organizations: the new intelligence and entrepreneurial frontier. IEEE Intell Syst 25(1):71–75

Massa S, Testa S (2005) Data warehouse-in-practice: exploring the function of expectations in organizational outcomes. Inf Manag 42:709–718

Melville N, Kraemer K, Gurbaxani V (2004) Information technology and organizational performance: an integrative model of IT business value. MIS Q 28(2):283–322

Moss LT, Atre S (2007) Business intelligence roadmap. Pearson Education Inc., Boston

Olszak CM, Ziemba E (2003) Business intelligence as a key to management of an enterprise. In: Proceedings of Informing Science and IT Education, Santa Rosa, CA

Parikh AA, Haddad J (2008) Right-time information for the real-time enterprise. DM review. http://www.dmreview.com/dmdirect/2008_92/10002003-1.html?portal=data_quality

Park Y (2006) An empirical investigation of the effects of data warehousing on decision performance. Inf Manag 43(1):51–61

Petrini M, Pozzebon M (2009) Managing sustainability with the support of business intelligence: integrating socio-environmental indicators and organizational context. J Strateg Inf Syst 18:178–191

Popovic A, Hackney R, Coelho PS, Jaklic J (2012) Towards business intelligence systems success: effects of maturity and culture on analytical decision making. Decision Support Syst 54:729–739

Pugh DS (1990) Organization theory: selected readings. Penguin, Harmondsworth

Ramakrishnan T, Jones MC, Sidorova A (2012) Factors influencing business intelligence (BI) data collection strategies: an empirical investigation. Decision Support Syst 52:486–496

Rockmann R, Weeger A, Gewald H (2014) Identifying organizational capabilities for the enterprise-wide usage of cloud computing. In: Proceedings of the Pacific Asia Conference on Information Systems

Sabherwal R, Kirs P (1994) The alignment between organizational critical success factors and information technology capability in academic institutions. Decis Sci 25(2):301–330

Sahay BS, Ranjan J (2008) Real time business intelligence in supply chain analytics. Inf Manag Comput Secur 16(1):28–48

Sambamurthy V, Zmud R (1997) At the heart of success: organizational wide management competencies. In: Sauer C, Yetton P, Alexander L (eds) Steps to the future: fresh thinking on the management of IT-based organizational transformation. Jossey-Bass, San Francisco, pp 143–163

Schaefferling A (2013) Determinants and consequences of IT capability: review and synthesis of the literature. In: Proceedings of the nineteenth American conference on information systems, Chicago, IL

Sukumaran S, Sureka A (2006) Integrating structured and unstructured data using text tagging and annotation. Bus Intell J 11(2):8–17

Tabbitt S (2013) BI services market predicted to double by 2016. Information Week

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. Strat Manag J 18(7):509–533

Thamir A, Poulis E (2015) Business intelligence capabilities and implementation strategies. Int J Glob Bus 8(1):34–45

Turban E, Sharda R, Aronson JE, King D (2008) Business intelligence: a managerial approach. Prentice Hall, Upper Saddle River

Watson HJ, Wixom H (2007) Enterprise agility and mature BI capabilities. Bus Intell J 12(3):13–28

Watson HJ, Abraham D, Chen D, Preston D, Thomas D (2004) Data warehousing ROI: justifying and assessing a data warehouse. Bus Intell J:6–17

White, C (2005) The next generation of business intelligence: operational BI. Information Management Magazine. http://www.information-management.com/issues/20050501/1026064-1.html

Wixom BH, Watson HJ (2001) An empirical investigation of the factors affecting data warehousing. MIS Q 25(1):17–41

Wixom BH, Watson HJ, Werner T (2011) Developing an enterprise business intelligence capability: the Norfolk southern journey. MIS Q 10(2):61–71

Chapter 4
Big Data Capabilities: An Organizational Information Processing Perspective

Öykü Isik

Abstract Big data is at the pinnacle of its hype cycle, offering big promise. Everyone wants a piece of the pie, yet not many know how to start and get the most out of their big data initiatives. We suggest that realizing benefits with big data depends on having the right capabilities for the right problems. When there is a discrepancy between these, organizations struggle to make sense of their data. Based on information processing theory, in this research-in-progress we suggest that there needs to be a fit between big data processing requirements and big data processing capabilities, so that organizations can realize value from their big data initiative.

Keywords Big Data • Organizational information processing • Capabilities

4.1 Introduction

At the World Economic forum in Davos every year, many public figures, politicians and brightest minds come together to discuss world's most important new developments. Since 2012, data has been discussed as a "critical new form of economic currency" (World Economic Forum Briefing 2012). Thanks to the availability of new data sources, the hyper-connectivity through social media and the internet of things, and digitalization of our business processes, big data is revolutionizing the way we interact with not only the businesses around us, but also the governments. Next to the mind-boggling increase in data amounts, the rise of analytics and quantification has also motivated organizations to leverage data for competitive

Ö. Isik (✉)
Information Systems Management, Vlerick Business School,
Vlamingenstraat 83, Leuven 3000, Belgium
e-mail: oyku.isik@vlerick.com

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_4

advantage. As many organizations try and fail, it became evident that our organizations do not have not only the technological means, but also the necessary organizational capabilities to process this voluminous and differently structured strategic resource.

Every now and then we hear success stories, such as how Macy's can optimize pricing of their 73 million items for sale in near-real time (Davenport and Dyché 2013), and how Tesco can do proactive maintenance by running analytics on 70 million refrigerator data points coming off its units (Goodwin 2013), or how Netflix managed to defeat its biggest competitor, Blockbuster, with only an algorithm and petabytes of data (Madrigal 2014). But, organizations cannot expect to 'go from zero to Netflix overnight' (Simon 2014). First, they need to figure out what business uncertainties they desire to address by processing big data. Then, it takes a well-planned process to organize the big data initiative; from developing the business case to understanding the business requirements, and to figuring out the necessary capabilities to address those requirements. When there is a discrepancy between these requirements and capabilities, organizations struggle to make sense of their data. Hence, we suggest that there needs to be a fit between big data processing requirements and big data processing capabilities, so that organizations can realize value from their big data initiative. To achieve the objective of finding support for our arguments, three research questions were formulated: (1) What elements constitute big data processing capabilities? (2) What elements constitute big data processing requirements? (3) How does the fit between big data processing capabilities and requirements impact value realization from the big data initiatives?

We intend to measure the research questions through a quantitative analysis of survey data. After refining the research model based on expert interviews, an instrument will be developed. Following the survey-based data collection phase, quantitative data analysis will be conducted. As a result, we expect to observe different levels of big data processing requirements as well as capabilities that can be clustered in four groups, based on high and low capabilities, as well as high and low levels of uncertainty. We expect to observe different levels of performance among these configurations.

This study not only addresses a very relevant topic, but does so rigorously. It contributes to the literature by improving the current understanding around what capabilities organizations need to have for big data success. Thus far, business literature has merely pointed to a number of variables that can play an important role for big data (e.g. hiring the right data scientist (Davenport and Patil 2012), top management support (McAfee and Brynjolfsson 2012), right IT infrastructure (LaValle et al. 2011), however, empirical validation of these variables has been limited, if not non-existent.

We design this research to empirically test big data capabilities, big data requirements and their fit by collecting quantitative data through an online survey. We use the established theory of organizational information processing as the foundation of our work. Besides testing the impact of fit on big data value realization, we also suggest and assess several elements that may contribute to big data processing capabilities as well as requirements.

4.2 Literature Review and Research Model

Big data definition has been evolving along with the technologies enabling as well as the expectations surrounding the concept. While the earlier definitions focused on volume by emphasizing "data sets that can no longer be easily managed or analyzed with traditional data management tools, methods and infrastructures" (Rogers 2011), later on the velocity of data (i.e., speed of data) and the variety of data (i.e., the format and structure of the data) were also added to the discussion and together, they are referred to as the '3 V's' of big data. More recently, value (i.e. extraction of benefits from data) and veracity (i.e., data and data source quality) were included and the definition was extended to '5 V's' (Fosso Wamba et al. 2015). We adopt Fosso Wamba et al.'s (2015) definition and suggest that 'big data' is a holistic approach to manage, process and analyze 5 V's in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantage.

Big data, coming from social media streams, banking transactions, sensors, GPS signals and countless other sources, may create new business opportunities in almost every industry (Gobble 2013). Yet, according to Gartner Group, big data is now in the "trough of disillusionment" phase of their hype cycle (Sicular 2013). This means that many organizations, even though they may have great ideas and opportunities, are disappointed with the difficulty of figuring out how to organize for their initiatives as well as the lack of reliable solutions that go beyond traditional vendor offerings. Therefore, a negative hype surrounding the topic is extant. Organizations are also grappling with cultural issues. For example, a retail organization heavily invested in new models and tools to optimize their returns on advertising, only to find out none of the frontline marketers were using them because they did not understand how the model worked and didn't believe in its results (Barton and Court 2012). Several other suggestions have been made with regards to why organizations struggle in their big data initiative; such as the lack of transparency between teams working on big data (Perrey and Arikr 2014), lack of qualified data scientists in the organization (Forsyth et al. 2014; Perrey and Arikr 2014) and the necessity of defining a realistic business case before delving into the data (Menon 2013).

These suggestions imply that there are factors not well understood to drive big data projects to success, and that we still do not have a clear approach that organizations can take for better performance with big data. Hence, we should look deeper into organizations that have started gaining positive value out of their big data initiatives, and understand how they're doing it. Current success stories imply that these organizations think differently about their data management methods as well as information processing capabilities to take advantage of this new resource (LaValle et al. 2011; McAfee and Brynjolfsson 2012). Yet, academic literature has yet to document the critical elements that may make or break an organization's big data initiative.

Several BI and analytics success models might be relevant for this research. For instance, Dinter et al. (2011)'s model build on the Delone and McLean IS success model to suggest a model for BI success, yet they have a distinct implementation success focus. On the other hand, Isik et al. (2013) have approached BI success from a capabilities perspective and suggested that, depending on the different nature of the decision to be made, certain BI capabilities may be more important than others. Another capabilities approach was used by Sidorova and Torres (2014), where the authors have distinguished between internal and external data, and suggested that capability building is key to success with BA. Yet none of these models incorporate uncertainty as a factor, they also do not assess whether these capabilities are being utilized for the right reasons. That is why a fit perspective is necessary. Using organizational information processing theory can help us make the case for certain capabilities being more critical for certain situations.

The power of big data depends heavily upon the context in which it's used (McAfee and Brynjolfsson 2012), and one key to success may be to have the right capabilities in place for that specific context (Davenport et al. 2012). The capabilities of the organization should be sufficient to meet the requirements of the business case put forward for big data. One lens that can be used to examine this match between capabilities and requirements is the Organizational Information Processing Theory (OIPT). OIPT emerged as a result of an increasing understanding that information is possibly the most important element of today's organizations (Fairbank et al. 2006; Galbraith 1977). OIPT focuses on information processing requirements (IPR), information processing capability (IPC), and the fit between them to obtain the best possible performance in an organization (Premkumar et al. 2005). In this context, information processing is defined as the gathering, analysis and synthesis of data for decision making (Tushman and Nadler 1978), and IPR are the means to reduce uncertainty (Daft and Lengel 1986). Uncertainty is the difference between information acquired and information needed to complete a task (Galbraith 1977; Premkumar et al. 2005; Tushman and Nadler 1978). Organizations that face uncertainty must acquire more information to learn more about their environment (Daft and Lengel 1986). When tasks are non-routine or highly complex (as mostly in the case of big data) uncertainty is high; hence IPR are greater for effective performance (Daft and Lengel 1986). Not surprisingly, when it comes to discovery and experimentation with big data, uncertainty increases significantly. Hence, it is critical that organizations build the right capabilities to minimize big data related uncertainty.

Organizations can only benefit from big data if they can manage to process, analyze and to turn it into useful knowledge, therefore it makes sense to study big data from an OIPT perspective. Although there is research using OIPT to explain various IS phenomena (e.g. Fairbank et al. 2006; Premkumar et al. 2005), there is very little focusing on business intelligence and analytics (e.g. Cao et al. 2015), and none on big data. This research suggests that the performance of big data initiatives significantly depend on the fit between IPC and IPR of the organization, specifically within the context of their big data initiatives (see Fig. 4.1 for the research model).

We posit that the value realized in big data initiatives depend on how well uncertainty is minimized by the IPC of the organization. In line with the OIPT literature, we suggest environmental uncertainty is an important factor contributing to processing requirements for big data as organizations in high uncertainty environments

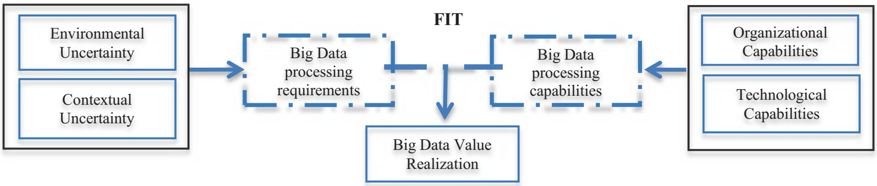

Environmental Uncertainty
Contextual Uncertainty
Big Data processing requirements
FIT
Big Data processing capabilities
Organizational Capabilities
Technological Capabilities
Big Data Value Realization

Fig. 4.1 Proposed research model

require more information processing, and, in turn, need more data to reduce uncertainty (Daft and Lengel 1986; Karimi et al. 2004; Premkumar et al. 2005). We also suggest contextual uncertainty as a source of uncertainty pertaining to big data; it can be defined as the potential biases, ambiguities, and inaccuracies in the data which need to be identified and accounted for to improve the accuracy of generated insights (Lukoianova and Rubin 2013; Schroeck et al. 2012). IBM suggests that in 2015, 80% of all available data is uncertain and will continue increasing (Claverie-Berge 2012). While data quality is a significant portion of this issue, enterprise data that can be subjected to data quality improvement (such as enterprise data quality solutions) forms only a fraction of the total data enterprises analyse. Most organizations include external data sources in their big data initiatives, such as social media accounts. These external data sources significantly increase data uncertainty, both in terms of content and expression. The ambiguity and lack of verifiability of these data increases contextual uncertainty. Environmental uncertainty and contextual uncertainty together influence the amount of processing an organization needs to do for value generation with big data.

Organizations may target to achieve different benefits from their big data initiatives. An organization may target using big data to change its products or the way it competes, this makes big data a strategic initiative. How Netflix competes based on its Cinematch algorithm is a good example for this. On the other hand, the purpose of the big data projects could be to cut costs and improve operational efficiency of the organization; such as how Tesco cut its annual refrigeration cooling costs by %20 across UK and Ireland by analyzing gigabytes of refrigeration data (Goodwin 2013)—this makes the big data project a transactional initiative. Finally, it can be all about bringing hidden information to light and to actually realize the knowledge residing in the organization. This would indicate an informational initiative. Such as LinkedIn, which developed several products, including People You May Know and Who's Viewed My Profile, based on the data they have already been collecting (Davenport and Dyché 2013). Acknowledging these different types of benefits an organization may realize through their big data projects, we prefer a wider definition of benefit realization and adopt the net benefits concept (DeLone and McLean 2003) which represents the individual and/or organizational impacts a certain IS investment has.

We suggest that the processing capabilities depend on the technological as well as organizational capabilities of the organization (Barton and Court 2012). Technological capabilities include the hardware and software that is being utilized for big data analytics; these capabilities should be sufficient enough to handle the '5 V's' of big data mentioned earlier. Organizational capabilities refer to the availability of the right skill set for big data analytics (Viaene and Van den Bunder 2011), analytical decision-making culture (Popovič et al. 2012), and top management support and championship of the big data initiative (Barton and Court 2012), which refers to the extent to which top management believes in the value of and actively participates in the efforts related to big data initiatives (Liang et al. 2007). Prior research has confirmed the importance of some these organizational elements in BI and analytics environments (Isik et al. 2013), but their impact on big data initiative are yet to be confirmed.

To obtain value from a big data initiative, big data processing requirements (BDPR) should match the big data processing capabilities (BDPC) of an organization. We posit that if organizations do not purposefully match their IPR with the IPC, configurations of misfit will occur and performance standards will be lower. For instance, if an organization is interested in strategic benefits from Big Data, it is more likely that they will include more data from a variety of sources, and deal with high uncertainty not in the environment but also within the data itself. To be able to manage these data, the technological as well as organizational capabilities should be on par. If not, less than optimal capabilities will render big data analytics ineffective. On the other hand, if the organization is interested in a rather isolated application of big data, using only internal sources, they would need to deal with less uncertainty compared to the previous example. Yet, having high level of IPC in this situation would be resource overkill (Mani et al. 2010). This is the case not only because unused resources will lead to inefficiency but also their management would be unnecessarily costly. So, while low uncertainty configurations may lead to benefits with low level of capabilities, high level of uncertainty would require high levels of capabilities. The other two configurations (low capability, high uncertainty and high capability, low uncertainty) represent misfit and will not realize benefits as much as the fit configurations.

4.3 Methodology

Our approach to empirically evaluating the research model consists of four phases: (1) research model fine-tuning, (2) instrument design, (3) pilot testing, and (4) final data collection and analysis.

4.3.1 Research Model Fine-Tuning

In the first phase of this research, an extensive literature review is carried out to fine-tune the conceptual model, using academic literature as well as industry outputs. Specific attention is paid to success and failure studies, in order to pinpoint organizational capabilities, whose existence or lack of contribute to big data success or failure. Following the literature review, interviews with big data experts, from academia as well as industry, who are responsible for or involved in big data projects, is conducted via a semi-structured process (currently on-going). The purpose is to understand the organizational requirements for big data, and whether big data projects are deemed value-adding or not. This interview process also provides information on the specific organizational context and its potential impact on value realization with big data, which may not be recognized during literature review. The deliverable of this phase should show whether the model needs revision, in terms of factors contributing to BDPR and BDPC. Another outcome might be the realization of additional factors that impact the suggested model (potential moderators or mediators). More detail about the interviews is provided in the Preliminary Findings section later in the paper.

4.3.2 Research Design and Measures

The required data for this study will be collected through an online survey. The instrument will measure variables using multiple indicator items derived from validated instruments used in prior research where possible. For variables that have no validated measures, items will be developed based on the comprehensive literature review and the interview round mentioned in the first phase above. These instruments will be tested for clarity of content, scope as well as purpose (representing content validity) by academicians and practitioners active within the field of big data.

4.3.2.1 The Conceptualization and Measurement of 'Fit'

Operationalization of the concept of fit has long been a topic of discussion in academia. As Galbraith and Nathanson (1979) observed, "although the concept of fit is a useful one, it lacks the precise definition needed to test and recognize whether an organization has it or not" (p. 266). As a lack of correspondence between the fit concept and how it is statistically formulated may lead to inconsistent research results (Venkatmaran 1989), it is critical to pay adequate attention to its formulation. Venkatmaran's (1989) seminal work on the fit concept in strategy literature provides a useful classification scheme that groups fit studies under six perspectives; fit as moderation, fit a mediation, fit as matching, fit as covariation, fit as gestalts and fit as profile deviation; each with distinct theoretical implications and specific analytical methods.

This research conceptualizes fit as matching, and evaluates the impact of fit on big data value realization through the impact of the interaction effect between BDPR and BDPC. This is a proper approach because fit is specified with a reference to a criterion variable, specifically 'big data value realization' in this research

(Umanath 2003). Even though there are various objections to multiplicative models that measure interaction, this form of operationalization has been shown to be reasonably robust (Kim and Umanath 1992; Umanath 2003; Venkatmaran 1989). The fit between BDPR and BDPC will be examined by evaluating four clusters formed by the interaction of these two variables. The number of clusters may be adapted in case additional constructs are added to the model based on the outcome of the research model fine-tuning phase. The interaction of BDPR and BDPC will be captured by the interaction effects of the main factors in ANOVA (Premkumar et al. 2005). The interaction effects are expected to have a greater impact on the value realization compared to the main effects (Premkumar et al. 2005; Venkatmaran 1989), because model posits that it is the notion of fit, and not the BDPR or BDPC individual effects, which results in better value realization.

4.3.2.2 Measures

Big Data Information Processing Requirements. This study suggests environmental and contextual uncertainty as the basis for BDPR. Environmental uncertainty has been extensively studied as a significant element of the OIPT and literature offers several validated measures for it. Karimi et al.'s (2004) questionnaire will be used to measure environmental uncertainty with three dimensions; dynamism (three items), hostility (three items) and heterogeneity (one item). Because contextual uncertainty is a concept unique to current model, a validated measure does not exist yet. A multi-item scale will be developed based on the literature as mentioned earlier.

Big Data Information Processing Capabilities. This study suggests technological and organizational capabilities as the basis for BDPC. Technological capabilities include the hardware and software necessary to deal with high volume, high velocity and high variety of data. Multi-item scales will be looked for in the literature and if necessary developed for this construct. For organizational capabilities, validated measures already exist. This research considers the availability of the right skill set for analytics, analytical decision-making culture, and top management support for the big data initiative as critical capabilities. Several studies, targeting to identify the right skill set for big data analytics, offer measures for this construct, e.g. Wixom et al. (2014), Liberatore and Luo (2013). Through a literature review, a list of such studies will be finalized and a comprehensive list of items will be generated. The analytical decision-making culture will be measure by three items adopted from Popovič et al. (2012). The level of top management support and championship of the big data initiative will be measured by seven items validated by Bajwa et al. (1998).

Big Data Value Realization. How much value an organization gains from its big data depends on the benefits an organization obtains from this initiative. In this research, the net benefits construct will be used to operationalize the value realized with big data. The net benefits concept, first introduced by DeLone and McLean (2003) in their widely studied IS success model, represents the individual and/or organizational impacts a certain IS has, representing its success level. Given that this research has an organizational level of analysis, the instrument developed by Mirani and Lederer (1998) will be adapted to measure the organizational benefits derived from IS projects.

4.3.3 Pilot Testing

After designing the instrument, a pilot test will be carried out for further refinement, validity and reliability checks. Data collection for this phase will be completed via our personal contacts in the industry, using a snowball sampling. Based on the findings, the instrument will be updated.

4.3.4 Data Collection

After finalization of the survey, the required data for this study will be collected by reaching out to small to large organizations that have big data initiatives, varying in scope. Our institution's research partnerships with several European financial services institutions will be leveraged for data collection, this way we will also be controlling for the context/industry.

4.4 Current State of the Research and Preliminary Findings

Currently we are in the process of conducting interviews. We have already conducted ten interviews with senior level experts, both from academia (4) and industry (6), who are immersed in the financial services industry yet work with big data on a daily basis. Interviews were conducted between February and May 2015, included open-ended questions and lasted around 1.5 h each. They were conducted by the researchers face to face where possible, and if not over the phone. Interviews were recorded and then transcribed by the researchers.

Preliminary interview results indicate a clear distinction between technological and organizational capabilities, as expected. The fact that better tools give traction to the analytics was commonly acknowledged. All interviewees suggested that technological capabilities are easier to develop, yet organizational capabilities, especially the talent side, is difficult to build. This is also in line with current academic knowledge. Partnership building (e.g. sharing and/or combining data sources, or using analytical capabilities of a partner in exchange of another service) was emphasized by almost all interviewees, making it another potential organizational capability for our research model. Another frequently emphasized factor was increasing regulations and/or increasing pressure from regulatory bodies. Given the fact that regulations form the boundaries of the environment these organizations operate in, the frequent emphasis of this as a factor impacting big data analytics confirms the importance of including environmental uncertainty in our model.

Currently, next to analyzing the above mentioned interviews, we are in the phase of scheduling interviews within the healthcare industry, with senior decision makers who are involved in big data projects. We target ten interviews. The intention with this research is to control for and to represent a wide variety of industries to increase generalizability of our findings.

4.5 Conclusion

The initial projects that have leveraged big data have provided some organizations with big returns and enabled them to disrupt their markets. For instance, shipping companies improving their fleets' on-time performance by using specialized weather forecast data and real-time information on port availability (Barton and Court 2012) and airlines improving their estimated time of arrivals by combining publicly available weather data and flight schedules and other proprietary data (McAfee and Brynjolfsson 2012) can serve as evidence that when done right, big data can transform the way organizations do business. Hence, this study has high practical relevance; many C-level executives are placing the topic on top of their agendas. Instead of sharing only high level and common best practices with executives, findings from this research will enable us to talk about specific capabilities they need to develop and environmental challenges they need to be careful about.

They key contribution of this research is that is suggests the alignment of capabilities and needs is key for benefit realization. This is especially important to justify given that today many organizations are trying to do too much without a clear strategy, just to not to miss the big data bandwagon. We believe this study will also contribute to the IS literature by (1) providing a comprehensive overview of factors, internal and external, influencing the success of big data initiatives, (2) examining both technological and organizational capabilities necessary to obtain value with big data, and by (3) being the first study looking into the big data phenomena with an OIPT lens.

Biography

Öykü Isik is an Assistant Professor of Information Systems Management at Vlerick Business School. She holds a Ph.D. degree in Business Computer Information Systems (2010) from the University of North Texas. Her research interests include business intelligence & analytics, business process management and information privacy. Her work appears in such Journals as *Information & Management*, *International Journal of Information Management* and *Business Process Management Journal.*

References

Bajwa DS, Rai A, Brennan I (1998) Key antecedents of executive information system success: a path analytic approach. Decis Support Syst 22(1):31–43

Barton D, Court D (2012) Making advanced analytics work for you. Harv Bus Rev 90(10):78–83

Cao G, Duan Y, Li G (2015) Linking business analytics to decision making effectiveness: a path model analysis. Trans Eng Manage 62(3):384–395

Claverie-Berge I (2012) Solutions Big Data IBM. IBM Presentation, 13 Mar 2012. http://www-05.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf

Daft RL, Lengel RH (1986) Organizational information requirements, media richness and structural design. Manag Sci 32(5):554–571

Davenport TH, Barth P, Bean R (2012) How 'Big Data' is different. MIT Sloan Manag Rev 54(1):21–24

Davenport TH, Dyché J (2013) Big Data in big companies. SAS International Institute for Analytics Report. http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. Harv Bus Rev 90(10):70–76

DeLone WH, McLean ER (2003) The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst 19(4):9–30

Dinter B, Schieder C, Gluchowski P (2011) Towards a Life Cycle Oriented Business Intelligence Success Model. AMCIS 2011 Proceedings

Fairbank JF, Labianca G, Steensma HK, Metters RD (2006) Information processing design choices, strategy and risk management performance. J Manag Inf Syst 23(1):293–319

Forsyth J, Moorman C, Spittaels S (2014) Recruit better data analysts. HBR Blog Network, 14 Feb 2014. http://blogs.hbr.org/2014/02/recruit-better-data-analysts/

Galbraith J (1977) Organizational design. Addison-Wesley, Reading, MA

Galbraith JR, Nathanson D (1979) The role of organizational structure and process in strategy implementation. In: Schendel D, Hofer CW (eds) Strategic management: a new view of business policy and planning. Little, Brown, Boston, pp 249–283

Gobble MM (2013) Big Data: the next big thing in innovation. Res Technol Manag 56(1):64–67

Goodwin B (2013) Tesco uses Big Data to cut cooling costs by up to €20m. Computer Weekly, 22 May 2013. http://www.computerweekly.com/news/2240184482/Tesco-uses-big-data-to-cut-cooling-costs-by-up-to-20m

Isik O, Jones M, Sidorova A (2013) Business intelligence success: the roles of BI capabilities and decision environments. Inf Manag 50(1):13–23

Karimi J, Somers TM, Gupta YP (2004) Impact of environmental uncertainty and task characteristics on user satisfaction with data. Inf Syst Res 15(2):175–193

Kim K, Umanath NS (1992) Structure and perceived effectiveness of software development units: a task contingency analysis. J Manag Inf Syst 9(3):157–181

LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big Data, analytics and the path from insights to value. MIT Sloan Manag Rev 52(2):21–31

Liang H, Saraf H, Hu Q, Xue Y (2007) Assimilation of enterprise systems: the effect of institutional pressures and the mediating role of top management. MIS Q 31(1):59–87

Liberatore M, Luo W (2013) ASP, the art and science of practice: a comparison of technical and soft skill requirements for analytics and OR professionals. Interfaces 43(2):194–197

Lukoianova T, Rubin VL (2013) Veracity roadmap: is Big Data objective, truthful and credible? Adv Class Res Online 24(1):4–15

Madrigal AC (2014) How Netflix reverse engineered hollywood. The Atlantic, 2 Jan. http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

McAfee A, Brynjolfsson E (2012) Big Data: the management revolution. Harv Bus Rev 90(10):60–68

Menon S (2013) Stop assuming your data will bring you riches. HBR Blog Network, 20 Sept. http://blogs.hbr.org/2013/09/stop-assuming-your-data-will-bring-you-riches/

Mirani R, Lederer AL (1998) An instrument for assessing the organizational benefits of IS projects. Decis Sci 29(4):803–838

Perrey J, Arikr M (2014) CMOs and CIOs need to get along to make big data work. HBR Blog Network, 4 Feb. http://blogs.hbr.org/2014/02/cmos-and-cios-need-to-get-along-to-make-big-data-work/

Popovič A, Hackney R, Coelho PS, Jaklič J (2012) Towards business intelligence systems success: effects of maturity and culture on analytical decision making. Decis Support Syst 54(1):729–739

Premkumar G, Ramamurthy K, Saunders CS (2005) Information processing view of organizations: an exploratory examination of fit in the context of interorganizational relationships. J Manag Inf Syst 22(1):257–294

Rogers S (2011) Big Data is scaling BI and analytics. information management, 01 Sept. http://www.information-management.com/issues/21_5/big-data-is-scaling-bi-and-analytics-10021093-1.html

Schroeck M, Shockley R, Smart J Romero-Morales D, Tufano P (2012) Analytics: the real-world use of Big Data. How innovative enterprises extract value from uncertain data. IBM Global Business Services Business Analytics and Optimization Executive Report. http://www.stthomas.edu/gradsoftware/files/BigData_RealWorldUse.pdf

Sicular S (2013) Big Data is falling into the trough of disillusionment. Gartner Blog Network, 22 Jan. http://blogs.gartner.com/svetlana-sicular/big-data-is-falling-into-the-trough-of-disillusionment/

Sidorova A, Torres RR (2014) Business Intelligence and Analytics: A Capabilities Dynamization View. AMCIS 2014 Proceedings

Simon P (2014) How to get over your inaction on Big Data. HBR Blog Network, February 24. http://blogs.hbr.org/2014/02/how-to-get-over-your-inaction-on-big-data-2/

Tushman ML, Nadler DA (1978) Information processing as an integrating concept in organizational design. Acad Manag Rev 3(3):613–624

Umanath NS (2003) The concept of contingency beyond 'It Depends': illustrations from IS research stream. Inf Manag 40(6):551–562

Venkatmaran N (1989) The concept of fit in strategy research: toward verbal and statistical correspondence. Acad Manag Rev 14(3):423–444

Viaene S, Van den Bunder A (2011) The secrets to managing business analytics projects. MIT Sloan Manag Rev 53(1):65–69

Wamba SF, Akter S, Edwards AJ, Chopin G, Gnanzou D (2015) How 'Big Data' can make Big Impact: findings from a systematic review and a longitudinal case study. Int J Prod Econ 165:234–246

Wixom B, Ariyachandra T, Douglas D, Goul M, Gupta B, Iyer L, Kulkarni U, Mooney JG, Phillips-Wren G, Turetken O (2014) The current state of business intelligence in academia: the arrival of Big Data. Commun Assoc Inf Syst 34(1):1–13

World Economic Forum Briefing (2012). Big Data, Big Impact: new possibilities for international development, January. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf

Chapter 5
Business Analytics Capabilities and Use: A Value Chain Perspective

Torupallab Ghoshal, Rudolph T. Bedeley, Lakshmi S. Iyer, and Joyendu Bhadury

Abstract This paper presents a mapping of the business analytics (BA) capabilities of a firm from a value chain lens similar to Porter's (Harv Bus Rev 79:62–78, 2001) internet capabilities framework. The generally accepted classification of analytics: descriptive, predictive and prescriptive, is used as basis for mapping BA capabilities. Using an extensive search of the academic and practitioner literature, analytics applications were analyzed and mapped onto the value chain framework. Given the increased interest and investment in BA, it is important to have a good understanding of what analytics capabilities firms use to enhance value through its value chain activities. We illustrate exemplar uses of BA applications, tools and technologies used by firms. Preliminary results suggest that organizations are focusing on

T. Ghoshal (✉)
Naveen Jindal School of Management, University of Texas at Dallas,
800 West Campbell Road, Dallas 75080, TX, USA

Department of Operations and Information Management, Isenberg School of Management,
University of Massachusetts at Amherst, Amherst, MA, USA
e-mail: txg140830@utdallas.edu

R.T. Bedeley
Department of Operations and Information Management, Isenberg School of Management,
University of Massachusetts at Amherst, Amherst, MA, USA

UMass, Amherst, 315 Isenberg Building, 121 Presidents Drive, Amherst, MA 01003, USA
e-mail: rbedeley@isenberg.umass.edu

L.S. Iyer
Walker College of Business, Appalachian State University,
287 Rivers St, Boone, NC 28608, USA
e-mail: iyerLs@appstate.edu

J. Bhadury
School of Business Administration and Economics, The College at Brockport,
350 New Campus Drive, Brockport, NY 14420, USA
e-mail: jbhadury@brockport.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_5

application of analytics where the outcome is easily measurable compared to application of analytics in other activities where it is harder to measure a direct value.

Keywords Business analytics • Capabilities • Descriptive analytics • Prescriptive analytics • Predictive analytics • Porter's value chain • Framework

5.1 Introduction

Value chain of an organization is the set of activities through which a product or service is created and delivered to customers (Porter 1985). According to Porter's framework (1985), an organization can create value by complementary sets of primary and secondary activities. Porter (2001) subsequently illustrated how internet-based technologies contribute to the firm's value chain. Since the introduction of this value chain model by Porter (2001), another significant trend has been the ubiquity of information technology and continuous increase of data availability and storage that has led businesses and researchers to address the issue of Business Intelligence and Analytics (BI&A) with more importance than ever before. According to studies by Gartner (2012, 2014), Business Intelligence and Analytics is the top priority of chief information officers and primary area of technology investment in most organizations. Owing in large part to such attention, BI&A has now become an important inclusion to increase value chain capabilities of many business organizations.

Hitherto, most prior literature has focused on the contribution of Business Analytics (BA) technologies within the context of their functional areas of application. However, it is our contention that the true value of such technology is not evident if the study focus is restricted to their functional area of application; instead, they should be viewed through the lens of how they contribute to the overall value chain of a firm. It is our premise that understanding the same is important for both academics and practitioners. From an academic standpoint, the development of such a framework is a novel contribution to the existing literature on BI&A in that it allows future researchers to study and evaluate BI&A systems not only from traditional standpoints (e.g., their contribution to inter and intra-organizational integration, user-friendliness or effectiveness) but from a new viewpoint of exactly where and how a BI&A system fits within the overall value creation by the organization. For practitioners, this research helps, first by cataloging existing applications of BI&A. But beyond that, the classification of these systems presented in the paper is not just by their functional areas of application but by their overall placement in the value chain of the organization. This potentially allows a manager to understand and evaluate BI&A systems in his/her organization by going beyond the usual cost-benefit analysis and asking the higher-level question "*Relative to other systems, how well does a given BI&A system contribute to the key strengths of my firm in creating value for the customer?*". Thus, developing such a framework that maps BA applications to the value chain of a firm is both the point of departure and the primary contribution of this paper.

Several current studies in the literature have proposed models, typologies and domains of BA in organizations (Chen et al. 2012; Holsapple et al. 2014; Wixom et al. 2013). Other studies focus on the supply chain analytics capabilities (Chae et al. 2014) of organization from a resource-based view (Barney 1991) and dynamic capabilities (Eisenhardt and Martin 2000) perspectives (Chae and Olson 2013). To understand the value created by BA in organizations, one important strand of research is to investigate how the value chain activities and processes of firms can be improved by the inclusion of BA. However, based on our extant review, no study yet has focused on the framework of analytics capabilities for the entire value chain and on empirically testing the application of different types of analytics capabilities in different activities of the value chain. In an effort to address this gap, this research accomplishes the first step by providing a framework to map business analytics capabilities and applications to Porter's value chain perspective (1985) of a firm.

When viewed from the standpoint of applications, analytics can be classified into three major categories (Davenport 2013): descriptive analytics, predictive analytics and prescriptive analytics. Descriptive analytics is used to answer 'what has happened?', predictive analytics to answer 'what could happen?', and prescriptive analytics to answer 'what should happen?' (Heching et al. 2013). Descriptive analytics can provide information of value chain components, predictive analytics can help in managerial planning, designing and value chain management, prescriptive analytics can be used to provide decision support tools and optimization based on the outcomes of descriptive and predictive analytics. Business organizations may use different combinations of the three types of analytics for deployment in their value chain. Hence, this study's research question: ***What are the prominent analytics capabilities that can be applied in different value chain activities of a firm?***

In pursuit of the same, the objective of this study is to develop a framework to classify different types of analytics capabilities from a value chain perspective. To accomplish this, a review of both academic and practitioner literature to uncover documented cases of BA applications in different industry/firms has been conducted. Thereafter, each application has been analyzed separately from two distinct standpoints: its categorical BA classification and its primary placement within the context of Porter's (1985) value chain framework. This was then used to develop the mapping given in the results section of the paper. Based on the findings, our future research will focus on developing a process model for how effective application of analytics can add value to organizations. In the following sections we provide the theoretical background and motivation; research methodology; preliminary analysis and results; and conclusion.

5.2 Background and Related Literature

At first a brief description of Porter's (1985) value chain activities and BA classification are provided, and then followed by literature conducted to review the application of analytics in different value chain activities that are present in extant literature.

5.2.1 Porter's Value Chain

Porter (1985) introduced the concept of value chain in his book with an idea that every organization has two distinct sets of activities to create value for the organization. One activity set is the primary activities organizations employ in creating physical product or service, marketing and delivery of the product or service, and support and after-sale service for that product or service. Another set is the supporting activities of the organization. The supporting activities are composed of internal activities of the organization which provides inputs and infrastructure to support the primary activities of the organization. Porter (1985) describes five primary activities as generic supply chain activities of organizations' value chain: Inbound logistics, operations, outbound logistics, marketing & sales, and after-sales service. The four supporting activities of organizations' value chain are: procurement, technology development, human resource management and firm infrastructure.

5.2.1.1 Analytics Capabilities of Organization

Analytics capabilities can generally be classified into three different categories. These are descriptive analytics, predictive analytics and prescriptive analytics (Davenport 2013). Following is a brief discussion on these different analytics capabilities:

- **Descriptive Analytics:** Descriptive analytics is the type of statistics that provide descriptive analysis of what is evident from the data. Based on this type of analysis it is possible to find out current trends, statistics of available data. This type of analytics tries to answer the question of what has happened.
- **Predictive Analytics:** Predictive analytics is the type of analytics where future of a process, product or activity can be predicted based on the result of the descriptive analytics. This type of analytics tries to answer the question of what could happen.
- **Prescriptive Analytics:** Prescriptive analytics is the most active type of analytics where the optimum output can be prescribed based on results of descriptive and predictive analytics. This type of analytics tries to answer the question of what should happen.

5.3 Methodology

To achieve the research objective, a review of analytics applications, technologies and tools used for different levels of analytics from a value chain perspective from both academic and practitioner sources was conducted. The purpose of literature review was not to exhaust all possible applications of analytics but rather focus on prominent exemplars of the usage of analytics in different value chain activities and processes. Review of academic and practice literature to accomplish similar purposes is common is IS literature (Zafar et al. 2014; Zhao and Zhu 2012; Phillips-Wren et al. 2015). For the literature search, keywords included 'analytics', 'value chain', and combinations of different levels of analytics (descriptive, predictive and prescriptive) with different value chain activities.

The data source included academic journals, practitioners' journals, academic and practitioners' conference publications, publications from research organizations, white papers, periodicals, etc. These included journals like *MIS Quarterly (MISQ)*, *Information Systems Research* (ISR), *Journal of Management Information Systems (JMIS)*, *Journal of the Association for Information Systems (JAIS)*, *Communications of the Association for Information Systems (CAIS)*, *MISQ Executive*, *Harvard Business Review (HBR), Sloan Management Review (SMR), etc. Conference publications include International Conference on Information Systems (ICIS), Americas Conference on Information Systems (AMCIS), Hawaii International Conference on System Sciences (HICSS)*, and *INFORMS Conference on Business Analytics & Operations Research*. White paper sources included The Data Warehouse Institute (TDWI), Teradata University Network (TUN), Booz Allen Hamilton (BAH), BeyeNETWORK, etc. Other than these sources, various periodicals, newspapers, analytics vendors' websites were explored. After identifying the different analytics capabilities in different value chain activities and processes, analytics capabilities from a value chain framework is presented.

5.4 Preliminary Analysis and Results

The results of the literature search are summarized in Tables 5.1 and 5.2. As indicated therein, most organizations are focusing on blends of descriptive, predictive and prescriptive analytics sets in their value chain activities. In Tables 5.1 and 5.2, different analytics capabilities are presented in primary and supporting activities of value chain respectively. Descriptive analytics is the first level of analytics conducted in any firm before predictive and prescriptive analytics (P&G 2012; Watson 2014). Given that all predictive and prescriptive analytics have underlying descriptive analytics (Davenport 2013), to highlight the prominent applications of analytics, Tables 5.1 and 5.2 are focused primarily on identifying prescriptive and predictive analytics.

Owing largely to their relative ease of implementation, long-standing and widespread availability of managerial statistics software, adoption of descriptive analytics by businesses is well-established over the past few decades. However, the same cannot be said of predictive and prescriptive analytics in today's firms. Nonetheless, in general the most noteworthy applications of analytics in recent times in terms of value creation has been in the use of predictive and prescriptive analytics. Therefore, while Table 5.2 above lists applications of all the three different types of analytics, *the remaining focus is on illustrating examples in the use of predictive and prescriptive analytics by organizations*.

Table 5.1 Analytics capabilities in primary activities of a value chain

Activity	Descriptive analytics	Predictive analytics	Prescriptive analytics
Inbound Logistics	• *Ad hoc* query and search-based BI (Gifford 2013). • Interactive visualization of goods in transit (Chen et al. 2012).	• Dynamic inventory balancing model to adjust inventories based on demand prediction (Happonen 2012). • Warehouse location planning based on capacity and demand analysis (Nelson 2012).	• Inventory optimization using tools like IBM-ILOG Inventory and Product Flow Analyst (Nelson 2012). • Telematics technology to optimize transportation. For example, UPS uses telematics in their logistics (Levis 2011).
Operations	• Multi-objective analysis for warehouse location selection (Nelson 2012). • Pattern recognition for supply chain performance improvement (Chae and Olson 2013). • Interactive visualization of processes and transformation of goods (Chen et al. 2012).	• Radio frequency identification (RFID) analytics to improve inventory management by predicting inventory count (Bertolucci 2014). • Neural network for various scheduling and planning tasks (Smith and Gupta 2000). • Clustering techniques such as k-means algorithm to find out causes of faults and process variations in production systems (Chien et al. 2007).	• Network optimization, and sourcing optimization using tools e.g. IBM ILOG LogicNET Plus XE (Nelson 2012). • Tabu search for optimization of distribution network (Gündüz 2015). • Genetic algorithm for assembly-line balancing problem, production scheduling (Che and Chiang 2012). • Cloud analytics to improve faster processing by providing capabilities to analyze large volume of data (BAH 2012).
Outbound Logistics	• Cluster analysis for network design (Gifford 2013). • Dispatch tool interface (Gifford 2013). • Delivery Information Acquisition Device (DIAD) of UPS for efficient parcel delivery (Levis 2011).	• Network flow model for vehicle routing (Gifford 2013). • Fleet progress prediction using interactive dashboards (Gifford 2013). • Capacity analysis for prediction of segmentation and delivery of orders (Gifford 2013).	• Transportation optimization using analytics e.g. ORION software of UPS (Rosenbush and Stevens 2014). • Self-organizing map clustering to determine the best ways of delivering products in terms of profitability and optimization of supply chain (Chae and Olson 2013).

Marketing & Sales	• Data mining of sales invoices to gain pattern information such as, Association Rule Mining for pattern identification of products bought together (Bhattacharjee 2012). • Customer Lifetime Value (CLV) analytics using stochastic models (Schweidel 2013). • Consumer heterogeneity analysis using neural networks (Hayashi et al. 2010).	• Machine Learning techniques such as Partial Recurrent Neural Network for sales forecasting (Müller-Navarra et al. 2015). • Customer churn analytics and upsell prediction model using logistic regression, random forests, and decision trees (Gillespie 2012). • Predictive modeling techniques to identify and retain the most profitable customers (nGenera 2008). • Profitability prediction using activity-based analytics (Sadovy 2010).	• Price optimization using revenue analytics (Koushik et al. 2012). • Use of analytics in creating profit optimizing search engine advertising tool such as, PROSAD (Skiera and Nabout 2012). • Integrated web data analytics to customize product offerings to customers (Franks 2011). • Use of analytics in agent-based modeling for marketing optimization (Ragusa 2013). • Landing page optimization using analytics tools such as Google website optimizer (King 2008).
After-sales Service	• Speech analytics in call centers to identify customer concerns (ITS 2015). • Location intelligence can help to provide real-time customer service by incorporating geospatial data with business data. (Steiner 2015). • Web-based, unstructured content such as information retrieval and extraction, opinion mining, web intelligence & analytics, social media analytics, social network analysis, spatial-temporal analysis (Chen et al. 2012).	• Text mining and Sentiment analysis to understand consumers' attitude and feedback (Gan and Yu 2015; Mayes 2015). • Customer propensity modeling using decision tree, logistic regression, neural network, and support vector machine (Leventhal and Langdell 2013).	• Real time analytics for faster customer service (HBR 2014). • Customer relationship management using logistic regression, decision trees, and neural network (Leventhal and Langdell 2013). • Consumer-centric website development using analytics (Albert et al. 2004). • Social media and other web 2.0 analytics to understand customer concern and act accordingly (Watson 2014).

Table 5.2 Analytics capabilities in supporting activities of value chain

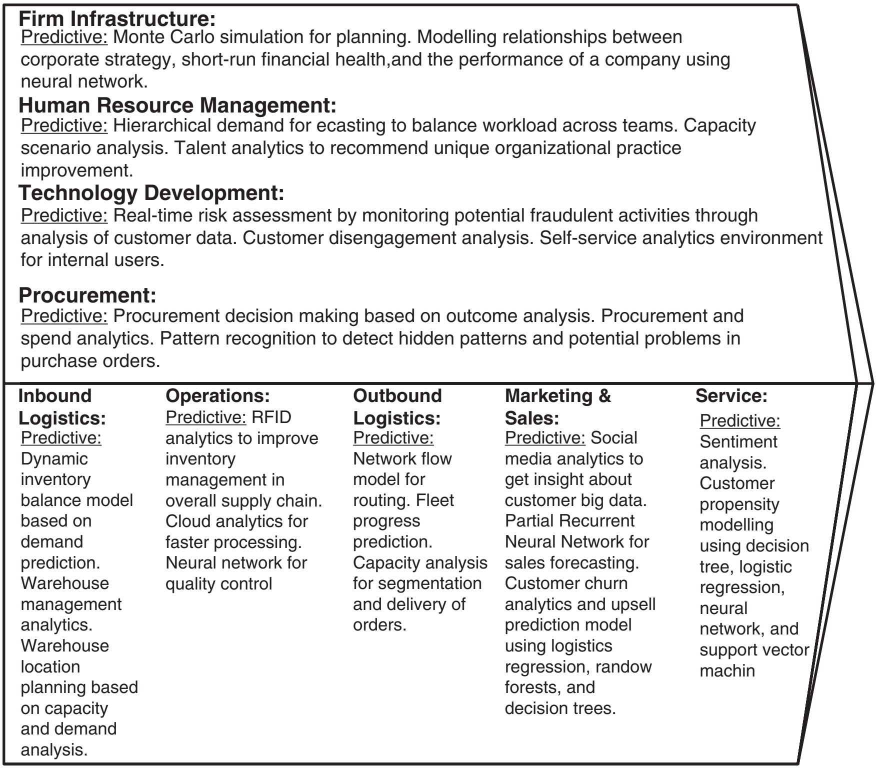
Activity	Descriptive analytics	Predictive analytics	Prescriptive analytics
Firm Infrastructure	• Financial analytics to provide organization better visibility into the factors that drive revenues, costs, and shareholder value (Oracle 2011). • Fuzzy analytic network process to assess ERP post-implementation success (Moalagh and Ravasan 2013). • Relational-DBMS and data warehousing. • ETL & OLAP (Chen et al. 2012).	• Monte Carlo simulation technique can be used to better understand all possible scenarios for planning, making decisions and mitigating risk (Underwood 2014). • Modeling predictive relationships between corporate strategy, short-run financial health, and the performance of a company using neural network (Smith and Gupta 2000).	• Energy informatics to go green. UPS used this technology to reduce CO_2 emission (Levis 2011). • Optimization modeling for financial planning, budgeting (Underwood 2013).
Human Resource Management	• Human resource dashboard/scorecard (Watson 2009, p. 496). • Mobile and sensor-based content analytics such as location-aware analysis, person-centered analysis, context-relevant analysis, and mobile visualization (Chen et al. 2012).	• Hierarchical demand forecasting to help balancing workload across teams (Heching et al. 2013). • Talent analytics to recommend unique organizational practice improvement (Davenport et al. 2010).	• Attrition modeling and compensation optimization (Mojsilovic 2013). • Capacity scenario analysis to optimize skill capacity considering constraints on skill requirements and agent utilization and service level agreements (Heching et al. 2013). • Optimatch technology used by IBM to match tasks with professionals (Mojsilovic 2013).
Technological Development	• Heat map to identify potential technological problems across the organization (ITS 2015). • Data mining workbenches (Chen et al. 2012).	• Real-time risk prediction by monitoring potential fraudulent activities through analysis of customer data (SAS 2012). • Real-time predictive analytics to improve business processes (Gartner 2013). • Text analytics to gain novel insights in various business processes (Mcneill 2015).	• Social media analytics to design products features tailored to customers' choices (Chau and Xu 2012). • Content and text analytics for faster processing of large datasets (Chen et al. 2012).
Procurement	• Spend analysis to make intelligent classification on spending, risk identification in supply base (Aberdeen Group 2014). • Interactive visualization analytics (Thomas and Cook 2006). • Web intelligence and analytics such as social media analytics (Chen et al. 2012)	• Procurement and spend analytics to develop strategies to rationalize supply base and to compare organization's practice with that of industry's (Aberdeen Group 2014) • Pattern recognition to find hidden pattern and potential problems in purchase orders (Chae and Olson 2013).	• Adaptive modeling for real-time bidding based on logistic regression (Kumar 2013). • Search engine advertisement optimization (Skiera and Nabout 2012) • Supplier/firm and firm/customer relational analytics (Chen et al. 2012)

5.4.1 Discussion of Results

Based on the preliminary findings, it is evident that analytics is mostly used in the primary activities of the value chain. Building on Porter's value chain, a framework that illustrates how the different types of analytics are being applied in the various stages of the value chain is now proposed (Figs. 5.1 and 5.2); the examples shown therein are generalizations from the specific applications cited in Tables 5.1 and 5.2.

5.5 Conclusion and Future Research

Based on the preliminary data analysis, we can convincingly argue that most organizations are using analytics in their primary activities of the value chain than in their supporting activities. The reason perhaps is that the outputs generated from



Firm Infrastructure:
Predictive: Monte Carlo simulation for planning. Modelling relationships between corporate strategy, short-run financial health,and the performance of a company using neural network.

Human Resource Management:
Predictive: Hierarchical demand for ecasting to balance workload across teams. Capacity scenario analysis. Talent analytics to recommend unique organizational practice improvement.

Technology Development:
Predictive: Real-time risk assessment by monitoring potential fraudulent activities through analysis of customer data. Customer disengagement analysis. Self-service analytics environment for internal users.

Procurement:
Predictive: Procurement decision making based on outcome analysis. Procurement and spend analytics. Pattern recognition to detect hidden patterns and potential problems in purchase orders.

Inbound Logistics:
Predictive: Dynamic inventory balance model based on demand prediction. Warehouse management analytics. Warehouse location planning based on capacity and demand analysis.

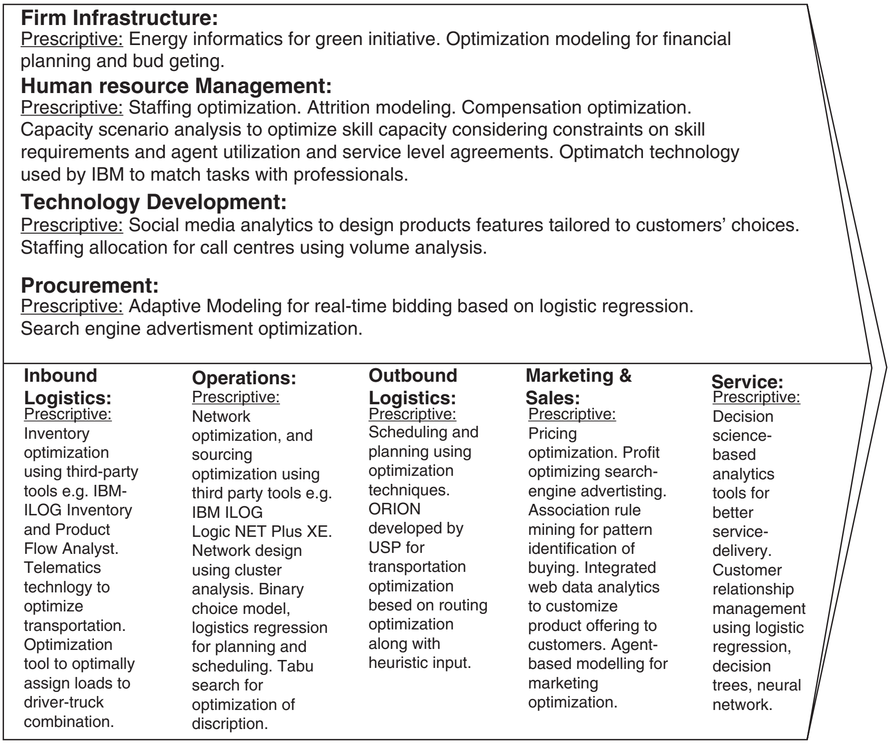
Operations:
Predictive: RFID analytics to improve inventory management in overall supply chain. Cloud analytics for faster processing. Neural network for quality control

Outbound Logistics:
Predictive: Network flow model for routing. Fleet progress prediction. Capacity analysis for segmentation and delivery of orders.

Marketing & Sales:
Predictive: Social media analytics to get insight about customer big data. Partial Recurrent Neural Network for sales forecasting. Customer churn analytics and upsell prediction model using logistics regression, randow forests, and decision trees.

Service:
Predictive: Sentiment analysis. Customer propensity modelling using decision tree, logistic regression, neural network, and support vector machin

Fig. 5.1 Application of predictive analytics tools and techniques in a firm's value chain



Firm Infrastructure:
Prescriptive: Energy informatics for green initiative. Optimization modeling for financial planning and bud geting.

Human resource Management:
Prescriptive: Staffing optimization. Attrition modeling. Compensation optimization. Capacity scenario analysis to optimize skill capacity considering constraints on skill requirements and agent utilization and service level agreements. Optimatch technology used by IBM to match tasks with professionals.

Technology Development:
Prescriptive: Social media analytics to design products features tailored to customers' choices. Staffing allocation for call centres using volume analysis.

Procurement:
Prescriptive: Adaptive Modeling for real-time bidding based on logistic regression. Search engine advertisment optimization.

Inbound Logistics:
Prescriptive: Inventory optimization using third-party tools e.g. IBM-ILOG Inventory and Product Flow Analyst. Telematics technlogy to optimize transportation. Optimization tool to optimally assign loads to driver-truck combination.

Operations:
Prescriptive: Network optimization, and sourcing optimization using third party tools e.g. IBM ILOG Logic NET Plus XE. Network design using cluster analysis. Binary choice model, logistics regression for planning and scheduling. Tabu search for optimization of discription.

Outbound Logistics:
Prescriptive: Scheduling and planning using optimization techniques. ORION developed by USP for transportation optimization besed on routing optimization along with heuristic input.

Marketing & Sales:
Prescriptive: Pricing optimization. Profit optimizing search-engine advertisting. Association rule mining for pattern identification of buying. Integrated web data analytics to customize product offering to customers. Agent-based modelling for marketing optimization.

Service:
Prescriptive: Decision science-based analytics tools for better service-delivery. Customer relationship management using logistic regression, decision trees, neural network.

Fig. 5.2 Application of prescriptive analytics tools and techniques in a firm's value chain

primary activities are easy to measure and quantify. Future research can examine the reason behind the highlighted gaps in BA use. For practitioners this framework helps not only to catalog the BI&A systems according to different value chain activities but also to allow managers to evaluate how different BI&A systems in different value chain activities can impact and create overall value in their organizations.

In future, the framework will be evaluated empirically to identify actual usage of analytics in organization's value chain. This should begin with studies that focus on delineating the specific organizational cultural and analytic infrastructural and capabilities variables that are germane to the different parts of the value chain (Marketing & sales, After sales service, Firm infrastructure, etc.). For example, it is quite conceivable that infrastructural/capability variables such as "Skills" will be different for "Marketing and sales" than for another aspect of the value chain such as "Technological development"; and exploration of the same for each part of the value chain will substantially enhance the research on the key success factors for BI&A in terms of contribution to the value chain of an organization.

Biographies

Torupallab Ghoshal (txg140830@utdallas.edu) is currently pursuing his Ph.D. degree in the Management Sciences with a concentration in Information Systems at the Naveen Jindal School of Management at the University of Texas, Dallas.His research interests include natural language processing, text analytics, machine learning, healthcare information technology, IT security, business intelligence, and big data analytics techniques. His research work appeared in various journals including Communications of the Association for Information Systems (CAIS), Information and Management (I&M), and Journal of Computer Information Systems (JCIS). His research work has been presented in several conferencs including Pre-ICIS Business Analytics Congress (BAC) and European Conference of Operations Research.

Rudolph T. Bedeley (Ph.D.) (rbedeley@isenberg.umass.edu) is an Assistant Professor in the Department of Information Systems and Operations Management in the Isenberg School of Management at UMass, Amherst. He teaches courses in the areas of Analytics and Business Intelligence, Database and Programming. His research interests encompass healthcare IT, big data analytics, business analytics and intelligence, and social inclusion. His research work has been presented in several conferences such as Pre-ICIS Business Analytic Conference (BAC), AMCIS, and Southern AIS. He was adjudged the best Graduate Teaching Assistant for the 2016/2017 academic year in the Bryan School of Business and Economics at UNC-Greensboro. He also won the best student research paper award at the 2017 Southern AIS conference held in Saint Simons Island, GA.

Lakshmi S. Iyer (Ph.D.) (Lsiyer@uncg.edu) is a Professor of Information Systems and Program Director of Applied Data Analytics in the Computer Information Systems and Supply Chain Management Department at Appalachian State University, Boone, NC. Her research interests are in the area of emerging technologies: business analytics, knowledge management, social computing; and social inclusion in IS: women in IS and technologies for disabled users. Her research work has been published in Communications of the AIS, Journal of Association for Information Systems, European Journal of Information Systems, Communications of the ACM, Decision Support Systems, eService Journal, Journal of Electronic Commerce Research, and others.

Joyendu Bhadury (Ph.D.) (jbhadury@brockport.edu) is a Professor and Dean in the School of Business Administration and Economics at the College of Brockport, State University of New York. His research is in the area of Management and Supply Chain and he has published his research in several journals including information system. His research work mostly focus on scheduling and optimization.

References

Aberdeen Group (2014) State of marketing automation 2014: processes that produce, A research report, Sept 19, 2014

Albert BTC, Goes PB, Gupta A (2004) Gist: a model for design and management of content and interactivity of customer-centric web sites. MIS Q 28(2):161–182

Barney J (1991) Firm resources and sustained competitive advantage. J Manag 17(1):99–120

Bertolucci J (2014) Radio tags generate vast quantities of information, but enterprises need to find ways to ingest, analyze, and archive that data. http://www.informationweek.com/bigdata/. Accessed 30 Apr 2015

BAH (2012) Improving intelligence analysis through cloud analytics. http://www.boozallen.com/media/file/Improving-Intelligence-Analysis-Through-Cloud-Analytics-fs.pdf. Retrieved May 1, 2015

Bhattacharjee S (2012) Quantifying growth and product assortment decisions across multiple retail stores: combining data analytics and optimization to connect global patterns with local constraints. In: Klampfl E (ed), Proceedings of INFORMS conference on business analytics & operations research. Huntington Beach, CA, pp 1–27

Chae B, Olson DL (2013) Business analytics for supply chain: a dynamic-capabilities framework. Int J Inf Technol Decis Mak 12(1):9–26

Chae B, Olson D, Sheu C (2014) The impact of supply chain analytics on operational performance: a resource-based view. Int J Prod Res 52(16):4695–4710

Chau M, Xu J (2012) Business intelligence in blogs: understanding consumer interactions and communities. MIS Q 36(4):1189–1216

Che ZH, Chiang T-A (2012) Designing a collaborative supply-chain plan using the analytic hierarchy process and genetic algorithm with cycle-time estimation. Int J Prod Res 50(16):4426–4443

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from Big Data to Big Impact. MIS Q 36(4):1165–1188

Chien C-F, Wang W-C, Cheng J-C (2007) Data mining for yield enhancement in semiconductor manufacturing and an empirical study. Expert Syst Appl 33(1):192–198

Davenport TH (2013) Analytics 3.0. Harv Bus Rev 91(12):1–14

Davenport TH, Harris J, Shapiro J (2010) Competing on talent analytics. Harv Bus Rev 88(10):1–13

Eisenhardt KM, Martin JA (2000) Dynamic capabilities: what are they. Strateg Manag J 21(1):1105–1121

Franks B (2011) Optimizing customer analytics: how customer level web data can help. In: Leonhardi DR (ed) Proceedings of INFORMS conference on business analytics & operations research. Chicago, IL, pp 1–35

Gan Q, Yu Y (2015) Restaurant rating: industrial standard and word-of-mouth a text mining and multi-dimensional sentiment analysis. In: Sprague RH Jr, Bui T, Laney S (eds) Proceedings of the 48th annual Hawaii international conference on system sciences. Kauai, HI, pp 1332–1340

Gartner Report (2012) Gartner says worldwide business intelligence, analytics and performance management software market surpassed the $12 Billion mark in 2011. http://www.gartner.com/newsroom/id/1971516. Accessed 1 May 2015

Gartner (2013) Gartner says by 2016, 70 percent of the most profitable companies will manage their business processes using real-time predictive analytics or extreme collaboration. http://www.gartner.com/newsroom/id/2349215. Retrieved October 25, 2015

Gartner Report (2014) Gartner says advanced analytics is a top business priority. http://www.gartner.com/newsroom/id/2881218. Accessed 1 May 2015

Gifford T (2013) Integrated analytics in transportation and logistics. In: Williams JT (ed) Proceedings of INFORMS conference on business analytics & operations research. San Antonio, TX, pp 1–39

Gillespie J (2012) Understanding customer behavior through marketing analytics: case studies from online gaming and chain restaurants. In: Klampfl E (ed) Proceedings of INFORMS conference on business analytics & operations research. Huntington Beach, CA, pp 1–36

Gündüz HI (2015) Optimization of a two-stage distribution network with route planning and time restrictions. In: Sprague RH Jr, Bui T, Laney S (eds) Proceedings of the 48th annual Hawaii international conference on system sciences. Kauai, HI, pp 1088–1097

Happonen A (2012) Adjusting inventories based on demand prediction using dynamic inventory balancing model. In: Proceedings of technology management for emerging technologies (PICMET), Proceedings of PICMET'12: IEEE, pp 3549–3565

Hayashi Y, Hsieh M-H, Setiono R (2010) Understanding consumer heterogeneity: a business intelligence application of neural networks. Knowl-Based Syst 23(8):856–863

HBR (2014) The new path to customer engagement: real-time analytics. https://hbr.org/resources/pdfs/comm/sap/18764_HBR_SAP_Telcom_July_14.pdf

Heching A, Lin P, Pratsini E (2013) Smarter workforce analytics for customer fulfillment transaction centers. In: INFORMS conference on business analytics & operations research. San Antonio, TX, April 7–9, 2013

Holsapple C, Lee-Post A, Pakath R (2014) A unified foundation for business analytics. Decis Support Syst 64:130–141

ITS (2015) Analytics application in information technology services department. A focus group Interview at a large US Public University

King A (2008) Website optimization Nutshell handbook. O'Reilly Media. https://books.google.com/books?id=f8-7pWbn9KEC

Koushik D, Higbie JA, Eister C (2012) Retail price optimization at intercontinental hotels group. Interfaces 42(1):45–57

Kumar M (2013) Predictive analytics in social media and online display advertising. In: Williams JT (ed) Proceedings of INFORMS conference on business analytics & operations research, San Antonio, TX, pp 1–37

Leventhal B, Langdell S (2013) Adding value to business applications with embedded advanced analytics. J Market Anal 1(2):64–70

Levis J (2011) Brown turning green delivering sustainability through data and technology. In Leonhardi DR (ed) Proceedings of INFORMS conference on business analytics & operations research. Chicago, IL, pp 1–58

Mayes M (2015) Sentiment analysis: more than a good idea monitoring and managing perceptions in social media, pp 1–3. http://www.b-eye-network.com/view/11395

Mcneill F (2015) Text analytics: deriving meaning from the deluge of documents and purging content chaos how organizations turn text into gold three keys to success, pp 1–4. http://www.b-eye-network.com/view/14437

Moalagh M, Ravasan AZ (2013) Developing a practical framework for assessing ERP post-implementation success using fuzzy analytic network process. Int J Prod Res 51(4):1–22

Mojsilovic A (2013) Smarter workforce changing the landscape of workforce management. In: Williams JT (ed) Proceedings of INFORMS conference on business analytics & operations research. San Antonio, TX, pp 1–23

Müller-Navarra M, Lessmann S, Voß S (2015) Sales forecasting with partial recurrent neural networks: empirical insights and benchmarking results. In: Sprague RH Jr, Bui T, Laney S (eds) Processing of the 48th annual Hawaii international conference on system sciences, Kauai, HI, pp 1108–1116

Nelson D (2012) Multi-objective optimization for strategic network design. In: Klampfl E (ed) Proceedings of INFORMS conference on business analytics & operations research. Huntington Beach, CA, pp 1–18

nGenera (2008) Business analytics six questions to ask about information and competition. nGenera Corporation report. http://www.sas.com/resources/whitepaper/wp_5483.pdf

Oracle (2011) Oracle knowledge for web self service. pp 1–4. http://www.oracle.com/us/products/applications/knowledgemanagement/oracle-knowl-web-self-service-2168374.pdf. Retrieved May 1, 2015

P&G (2012) Driving competitive advantage with OR. In: Klampfl E (ed) Proceedings of INFORMS conference on business analytics & operations research. Huntington Beach, CA, pp 1–35

Phillips-Wren G, Iyer LS, Kulkarni U, Ariyachandra T (2015) Business analytics in the context of big data: a roadmap for research. Comm AIS 37:23

Porter ME (1985) Competitive advantage. The Free Press, New York

Porter ME (2001) Strategy and the Internet. Harv Bus Rev 79(3):62–78

Ragusa D (2013) Bringing the consumer in the mix: using agent-based modeling to power marketing mix optimization. In: Williams JT (ed) Proceedings of INFORMS conference on business analytics & operations research. San Antonio, TX, pp 1–38

Rosenbush S, Stevens L (2014) At UPS, the algorithm is the driver—turn right, turn left, turn right: inside orion, the 10-year effort to squeeze every penny from delivery routes. Wall Street J: 14–17

Sadovy L (2010) Better decisions through profitability analysis, pp 1–4. http://www.b-eye-network.com/view/12489

SAS (2012) Banks, big data and high-performance analytics. http://www.teradatauniversitynetwork.com/assetmanagement/DownloadAsset.aspx?ID=475e4f52-caba-4ca7-acd1-51abe13c70d1&version=f6b950445fed4b46ab3aa5cd5eaf8c4e1.pdf. Retrieved May 1, 2015

Schweidel DA (2013) Stochastic models for customer analytics. In: Williams JT (ed) Proceedings of INFORMS conference on business analytics & operations research. San Antonio, TX, pp 1–37

Skiera B, Nabout NA (2012) PROSAD: a bidding decision support system for profit optimizing search engine advertising. In: Klampfl E (ed) Proceedings of INFORMS conference on business analytics & operations research. Huntington Beach, CA, pp 1–32

Smith KA, Gupta JND (2000) Neural networks in business: techniques and applications for the operations researcher. Comput Oper Res 27(11–12):1023–1044

Steiner J (2015) Business intelligence and GIS, systems within systems, and ubiquity how the world becomes part of every application in the 21st century, pp 1–4. http://www.b-eye-network.com/view/7956

Thomas JJ, Cook K (2006) A visual analytics agenda. IEEE Comput Graph Appl 26(1):10–13

Underwood J (2013) Beginning prescriptive analytics with optimization modeling. http://www.b-eye-network.com/view/17152. Accessed 1 May 2015

Underwood J (2014) Prescriptive analytics: making better decisions with simulation, pp 4–9. http://www.b-eye-network.com/view/17224. Accessed 1 May 2015

Watson H (2009) Tutorial: business intelligence—past, present, and future. Commun Assoc Inf Syst: 487–510

Watson HJ (2014) Tutorial: Big Data analytics: concepts, technologies, and applications tutorial: Big Data analytics: concepts, technologies, and applications. Commun Assoc Inf Syst 34:1246–1269

Wixom BH, Yen B, Relich M, Wixom BH, YenB RM (2013) Maximizing value from business analytics. MIS Q Exec 12(2):111–123

Zafar H, Ko MS, Clark JG (2014) Security risk management in healthcare: a case study. Commun Assoc Inf Syst 34(1):737–750

Zhao Y, Zhu Q (2012) Evaluation on crowdsourcing research: current status and future direction. Inf Syst Front 1(18):417–434

Chapter 6
Critical Value Factors in Business Intelligence Systems Implementations

Paul P. Dooley, Yair Levy, Raymond A. Hackney, and James L. Parrish

Abstract Business Intelligence (BI) systems have been rated as a leading technology for the last several years. However, organizations have struggled to ensure that high quality information is provided to and from BI systems. This suggests that organizations have recognized the value of information and the potential opportunities available but are challenged by the lack of success in Business Intelligence Systems Implementation (BISI). Therefore, our research addresses the preponderance of failed BI system projects, promulgated by a lack of attention to Systems Quality (SQ) and Information Quality (IQ) in BISI. The main purpose of this study is to determine how an organization may gain benefits by uncovering the antecedents and critical value factors (CVFs) of SQ and IQ necessary to derive greater BISI success. We approached these issues through adopting 'critical value factors' (CVF) as a conceptual 'lens'. Following an initial pilot study, we undertook an empirical

P.P. Dooley (✉)
College of Engineering and Computing, Nova Southeastern University,
Fort Lauderdale, FL, USA

330 East 39th St. #27M, New York, NY 10016, USA
e-mail: pd344@nova.edu

Y. Levy
College of Engineering and Computing, Nova Southeastern University,
Fort Lauderdale, FL, USA

College of Business, Arts and Social Sciences, Business School, Brunel University London,
Kingston Lane, Uxbridge, London, UB8 3PH UK
e-mail: levyy@nova.edu

R.A. Hackney
College of Business, Arts and Social Sciences, Business School, Brunel University London,
Kingston Lane, Uxbridge, London UB8 3PH, UK
e-mail: ray.hackney@brunel.ac.uk

J.L. Parrish
College of Engineering and Computing, Nova Southeastern University,
Fort Lauderdale, FL, USA

Department of Information Systems and Cybersecurity, College of Engineering and
Computing, 3301 College Avenue, Fort Lauderdale, FL 33314, USA
e-mail: jlparrish@nova.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_6

analysis of 1300 survey invitations to BI analysts. We used exploratory factor analysis (EFA) techniques to uncover the CVFs of SQ and IQ of BISI. Our study demonstrates that there is a significant effect in the relationships of perceived IQ of BISI to perceived user information satisfaction thereby confirming the importance BI system users place on information and the output produced. Our study also reported that there is a significant effect in the relationships of perceived IQ of BISI to perceived user system satisfaction thereby confirming the importance BI system users place on system output. We believe our research will be of benefit to both academics and practitioners in attempting to ensure BI systems implementation success.

Keywords Business intelligence • Business intelligence systems implementations • Information quality • Information systems success • Systems quality • Value theory

6.1 Introduction

Research evidence shows that spending on business intelligence (BI) systems has comprised one of the largest and fastest growing areas of information technology (IT) expenditures (Luftman and Ben-Zvi 2010). In spite of these investments, only 24% of 513 companies surveyed in a study conducted by Howson (2008), considered their BI implementations to be very successful. Furthermore, Marshall and de la Harpe (2009), noted 80% of the time spent in BI support involves investigating and resolving information quality (IQ) issues which if inadequately addressed, will severely affect organizations through decreased productivity, regulatory problems, and reputational issues.

It is apparent that pre-implementation activities for BI projects, particularly addressing system quality (SQ) and information quality (IQ) requirements are of paramount importance to business intelligence systems implementations (BISI) success (Howson 2008; Marshall and de la Harpe 2009; Negash and Gray 2008; Power 2008; Watson et al. 2002). Moreover, there has been a significant body of research that seeks to determine the role of SQ and IQ in information systems (IS) success (DeLone and McLean 2003; Petter and McLean 2009). However, very little attention has been given in the literature to addressing the role of SQ and IQ in the success of BISI (Arnott and Prevan 2008; Nelson et al. 2005; Ryu et al. 2006). Also, little attention has been given to the user's perceived value of SQ and IQ characteristics that have an impact on BISI success (Nelson et al. 2005; Popovic et al. 2012). Nelson et al. (2005) acknowledged the importance of identifying the appropriate SQ and IQ factors for BI success and indicated that some factors may not be stable across technologies or applications. Researchers in BI success have also suggested constructs and associated measurement items that consider the decision support environment and its maturity in BISI success (Dinter et al. 2011; Isik et al. 2013). However, few empirical studies have sought to uncover SQ and IQ characteristics that are of value to users of BI analytical systems, as measured by user satisfaction from BISI.

The relationships between the constructs of user perceived value (level of importance) and user satisfaction in the context of understanding the SQ and IQ necessary for BISI success have also received little attention in the literature. Research has been limited to studies that rely only on specific SQ and IQ factors for BI that are based on prior research, not on the universal set of antecedents for SQ and IQ that had been subjected to empirical analysis (Nelson et al. 2005). Thus, in the context of emerging technologies such as BI, it is important to focus on objectives and decisions that are of value, often requiring the exposure to underlying or hidden values that allow researchers and practitioners to be proactive and hence create more alternatives instead of being limited by available choices (Dhillon et al. 2002; Keeney 1999). According to Sheng, Siau, and Nah (2005), it is important to elicit and organize values in "developing constructs in relatively new and under-studied areas" (p. 40). Therefore, our research addresses the preponderance of failed BI system projects, promulgated by a lack of attention to SQ and IQ in BISI (Arnott and Prevan 2008; Jourdan et al. 2008). The main purpose of this study is to determine how an organization may gain benefits in the context of BISI by uncovering the antecedents and critical value factors (CVFs) of SQ and IQ necessary to derive greater BISI success. Furthermore, this study will empirically assess the cross-over relationships between the perceived SQ and IQ of BISI and perceived system and information quality to address any ambiguity in BI analytical user perceptions in distinguishing between the system (SQ) and the output (IQ) of the BI system as identified by Nelson et al. (2005).

6.2 Theoretical Background

6.2.1 Value Theory

In cognitive value theory, value refers to the individual's perceived level of importance (Rokeach 1969). According to Rokeach (1973), a value is "an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence" (p. 5). The concept of value is often referenced in various fields of social research but mainly in the context of economic value, thereby neglecting the applications of user perceived cognitive value (Levy 2006). According to Levy (2008), "several scholars have suggested that although it is important to investigate the nature of attitudes and opinions, it is more fundamental to investigate the nature of value since attitudes and opinions can often change based on experience, while value remains relatively stable over time" (p. 161). Keeney (1992) stated that values are what one desires to achieve. Bailey and Pearson (1983) measured the value (or level of importance) of information system (IS) characteristics using a scale featuring the semantic differential pair, important to unimportant (Levy 2003). These measures provided a deeper understanding of satisfaction with the IS (Etezandi-Amoli and Farhoomand 2011; Levy 2003; Sethi and King 1999). Levy (2009) defined user perceived value as a "belief about the level of importance that users hold for IS characteristics" (p. 94).

In the context of BI, as a large number of projects are considered to be failures because organizations do not see tangible business value, it is necessary to understand the value factors that are needed to benefit from BI investments (Todd 2009). Value based exploration techniques have been applied in many research areas such as value-focused assessment of privacy and security (Dhillon and Torkzadeh 2001; Dhillon et al. 2002), value-focused assessment of trust in mobile commerce (Siau et al. 2004), and assessing the values of mobile applications (Nah et al. 2005; Sheng et al. 2005). Levy (2008), in a study of online learning activities, used critical value factors (CVFs) to investigate and uncover issues related to learners' perceived value. Value theory has been established to uncover hidden attributes that users find important to IS success (Dhillon et al. 2002; Keeney 1999; Sheng et al. 2010). However, there has been little attention paid to ask the questions regarding what characteristics users find important in BISI. Furthermore, less is known about the CVFs that may lead to BISI success. Therefore, this study investigated issues related to the perceived value of IQ and SQ in BISI by uncovering CVFs as identified by users of BI analytical systems. CVFs are the factors that organizations should pay attention to in order to increase the BI systems perceived value, which in turn may lead to improved BISI success. Alternatively, IS success has also been assessed using the Critical Success Factor (CSF) methodology. The CSF methodology, however, has limited capacity to accommodate complexity and may produce models that do not accurately represent the actual environment (Boynton and Zmud 1984). Thus, this study used value theory as the basis to uncover the CVFs of SQ and IQ necessary to achieve BISI success. In this context, the CVFs for SQ and IQ have been identified and discovered using a process whereby SQ and IQ characteristics form clusters that provide an understanding of the factors that users of BI analytical systems find important or of value in BISI (Mertler and Vannatta 2001). This approach is particularly beneficial in performing research in an emerging technology such as BI where it is not a conventional application-based IT project but a complex undertaking (Yeoh and Koronios 2010).

6.2.1.1 IS and BI Success Theory

The measurement of IS success has been a top concern of researchers and practitioners for some time. Several models have been proposed to define and identify the causes of IS success. However, a universally agreed definition of IS success has not emerged due to differences in the needs of stakeholders who assess IS success in an organization (Urbach et al. 2009). For the purposes of this study, IS success is defined as a multi-dimensional phenomenon comprised of the technical, semantic, and effectiveness levels. Based on this definition, research models applicable to the specific requirements of a corresponding problem domain may be devised. The need for a general but comprehensive definition of IS success was recognized by DeLone and McLean (1992) in their review of existing definitions of IS success and their associated measures. This led to the multidimensional and interdependent model that classified the six major categories of system quality, information quality, user satisfaction, use, individual impact, and organizational impact. Since the publication of the DeLone and McLean (1992) IS success model, many researchers

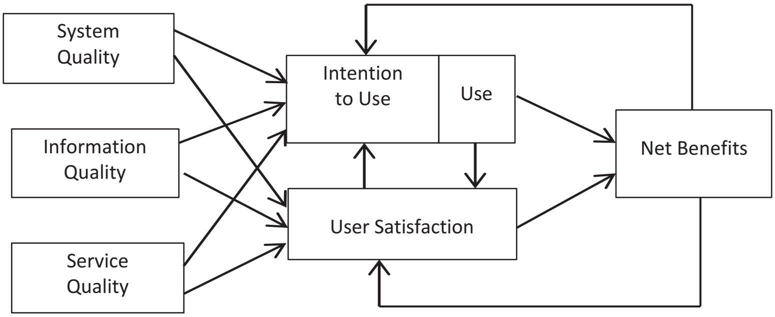


Fig. 6.1 DeLone and McLean (2003) IS success model

have treated IS success as a multidimensional construct (Urbach et al. 2009). Subsequent to the publication of the original DeLone and McLean (1992) IS success model, many researchers had suggested that it be extended or re-specified to include additional dimensions (Seddon 1997). As a result, DeLone and McLean (2003) published an updated IS success model to include the addition of service quality and intention to use as constructs, as depicted in Fig. 6.1. They also collapsed the individual and organizational impact constructs into the net benefits construct to measure the positive and negative influence of user satisfaction and use on an IS.

According to Urbach et al. (2009) “the majority of studies of IS success use the DeLone and McLean IS success model (1992, 2003) in combination with other theoretical models as a basis for deriving new research models that are applicable to the specific requirements of the corresponding problem domains” (p. 9). However, researchers have argued that certain constructs of the DeLone and McLean model do not significantly correlate with IS effectiveness. According to Levy (2009), “IS usage has been demonstrated to have mixed results as a predictor of IS effectiveness” (p. 99). Moreover, according to Petter et al. (2013), “while the service quality construct in the updated DeLone and McLean IS success model is an important dimension of IS success, they did not find any studies that considered the determinants of this construct. The few studies that did identify the determinants of service quality considered the overall quality of service provided by the IS department for all applications and services rather than for a specific IS” (p. 30). Furthermore, there is mixed support for the determinants of the construct ‘Intension to Use’ as an insufficient number of studies have investigated the relationship to IS success (Petter et al. 2013). Gatian (1994), in a study of 39 organizations found that there was a close relationship between user satisfaction, decision performance, and user efficiency. However, researchers had also recognized the complicated nature of establishing the dependent variable in IS success (DeLone and McLean 1992, 2003; Iivari 2005; Rai et al. 2002; Seddon 1997). According to Seddon “in the long run, it is people’s observations of the outcomes of use and the impacts that determine their satisfaction with the system” (p. 243).

Clark et al. (2007) followed the guidance of the DeLone and McLean IS success model (1992, 2003) to study the underlying threads of commonality with BISI success. Their study suggested that BISI success was theoretically grounded in IS success research. While much attention has been paid to IQ, SQ, and user satisfaction in IS success literature, little research has focused on the constructs of IS success in the domain of BISI. This may be related to a lack of understanding of BI technologies caused, in part, by the multifaceted nature of BI which combines a nonconventional application-based set of systems with infrastructure related projects (e.g. ERP and CRM) in an analytical user based decision support environment. Dinter et al. (2011) suggested alternatives for establishing BI specific success models to assist organizations in understanding the maturity of their BI decision environment by taking into consideration their BI capacity and capabilities. For instance, an organization may use the report writing and query capability of the BI implementation more than the analytical functionality in their implementation while another organization may use the analytical features of the BI system, such as predictive analytics, as their primary reason for implementing BI systems. In essence, BI success will be measured differently depending on the BI maturity level of the organization. Recognizing the differences in BI system maturity, Dinter et al. (2011) adopted and extended the updated DeLone and McLean (2003) IS success model in the BI domain thereby broadening its scope by adding additional constructs and items that have a causal relationship to the existing constructs in the BI decision environment.

Isik et al. (2013) examined the maturity of the required decision environment of BI to assess what capabilities are necessary to achieve success. They suggested technical, functional, and organizational elements of the decision environment that could lead to BISI success. Moreover, Isik et al. (2013) concluded that while the technical capabilities of the BI system represented a necessary foundation for BI success organizational capabilities that support flexibility in decision making should also be managed in relation to the decision environment in which the BI is employed.

Nelson et al. (2005) addressed a gap in the literature involving confusion in differentiating between SQ and IQ factors in the context of user satisfaction when using BI analytical tools in a data warehouse environment. Their model, which extended the DeLone and McLean (1992, 2003) success model, studied factors of SQ and IQ identified in the literature and their relationships with the constructs of system satisfaction and information satisfaction. The results of the Nelson et al. (2005) study suggested that "crossover or interaction effects may exist between the two constructs" (p. 207). They found that while the crossover effects of SQ on information satisfaction was significant within the context of BI analysis tools, the path leading from IQ to information satisfaction in the same context was surprisingly not significant. They concluded that future research was necessary to understand the characteristics of BI that led to the user perception that IQ did not strongly influence information satisfaction in the BI analytics domain. Nelson et al. (2005) expressed concern regarding this finding and offered the explanation that, from the user's perspective, it may be difficult to differentiate the BI system from the output it produces, leading to potential over-reliance on the system for IQ while ignoring the responsibility for user interaction with the interface and the generation of output.

While there is no commonly agreed definition of BI and BISI, there is some agreement in the literature on categorizing BI using the process, technological, and product perspectives. As a specific IS problem domain, BISI success falls within the context of key business analytics and processes that lead to decisions and actions that result in improved business performance. According to Chee et al. (2009), the organizational, functional, and process perspectives of BI focuses on the gathering of data from internal and external sources, followed by the generation of relevant information for decision making. BI success relies on multi-dimensional factors which also include those related to technology. Chee et al. (2009), acknowledged the similarities and differences in the interpretations of BI success and suggested that the technological aspect of BI be considered as a *BI system*, where the process perspective be regarded as the implementation of *BI systems*. Moreover, the product perspective is considered to be associated with the requirement for actionable information with established tools. For the purpose of this study we assessed BI success with the understanding that the process lifecycle approach addresses success in the implementation of BI systems following the premise that BI success has roots in technical and process capabilities in a decision environment.

Nelson et al. (2005) derived a model, depicted in Fig. 6.2, that identified, integrated, and assessed the dimensions of SQ and IQ as antecedents of the constructs of perceived user systems satisfaction and perceived user information satisfaction in their model titled "Determinants of information and system quality" (p. 208). Their model assumed that user satisfaction may be a reasonably good surrogate for net benefits if measures are confined to decision performance (Iivari 2005). Therefore, in this study the underlying theory of the DeLone and McLean (2003) model was explored with emphasis on the user satisfaction construct as the dependent variable for success (Iivari 2005). Furthermore, the BISI was considered effective when users perceived the characteristics of SQ and IQ to be of value or highly important and were also highly satisfied with these same characteristics. This study also uncovered the SQ and IQ characteristics that are of value in BISI as measured by user satisfaction. Participants in the study implemented BI analytical systems which represent a higher level of organizational BI system maturity in comparison to those who primarily perform report and query generation. The model expanded the user satisfaction construct and suggested that user perceived system satisfaction and user perceived information satisfaction could be considered as dependent variables and as a combined surrogate for user satisfaction. In essence, this study tested a proposed BI SQ and IQ research model which was based on the DeLone and McLean (1992, 2003) IS success model as extended by Nelson et al. (2005) and specifically tested the influence of the CVFs of SQ and IQ in BISI with user satisfaction from BISI in a decision support environment that leveraged BI analytics to improve and optimize decisions.

Various frameworks have been developed for categorizing and measuring IQ, SQ, and user satisfaction leading to IS success. The framework for IQ developed by

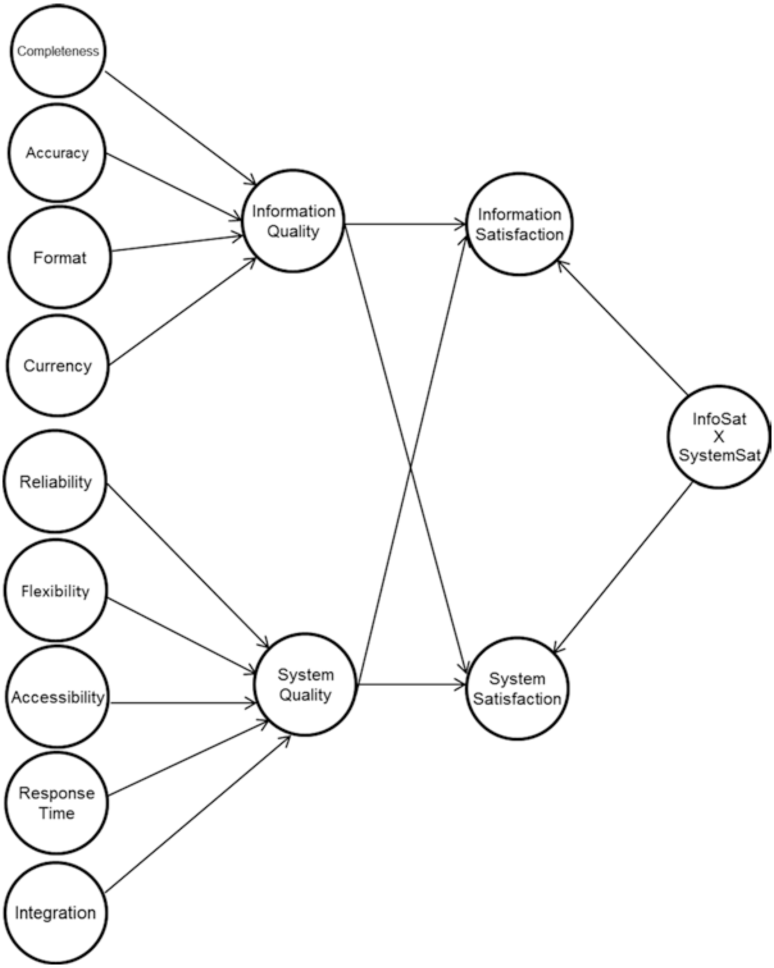


Fig. 6.2 Nelson et al. (2005) Determinants of information and system quality

Lee et al. (2002), for instance, provided four different categories used to assess IQ in IS. These categories were based on an empirical study of characteristics of a group of conventional IS. Moreover, Nelson et al. (2005) suggested a framework for the measurement of SQ for BI system satisfaction based on five dimensions of system quality.

Past confusion in differentiating SQ from IQ factors in BISI success suggested that crossover or interaction effects may exist between the two constructs leading Nelson et al. (2005) to explore the possibility that more complex quality/satisfaction relationships may exist. Thus, Nelson et al. (2005) studied the determinants of SQ and IQ which included the study of crossover relationships from quality (information and systems) to satisfaction (systems and information) as well as the interaction effect of information satisfaction and systems satisfaction. They suggested that future research should explore the relationship of SQ, IQ and perceived user satisfaction in the context of BI analytical systems to address the surprising results of their empirical analysis that indicated that the influence of SQ on user perceived IQ satisfaction was stronger than the influence of IQ on user perceived IQ satisfaction. It was, therefore, necessary to understand what dominant SQ and IQ characteristics are deemed important in BI to guide the design of BI systems and distinguish the system from its output. To address the surprising results of Nelson et al. (2005), the universal set of antecedents were empirically studied and once identified, data were gathered to determine what BI analysts valued in BI analytical systems with the expectation that the proposed CVFs of BISI could change after exploratory factor analysis (EFT) when subjected to confirmatory factor analysis (CFA). Therefore, this study used the BI SQ and IQ research model of Nelson et al. (2005) with the proposed CVFs of BISI depicted in Fig. 6.3.

The first specific goal of our research, following Keeney's (1992) methodology, was to gather a list of user perceived SQ and IQ characteristics from literature and augment it with input from an expert panel. The second research aim was to use the SQ and IQ characteristics to uncover the CVFs of SQ and IQ associated with BISI. The third specific goal of this research was to test the impact of the CVFs of SQ on perceived SQ of BISI and the CVFs of IQ on perceived IQ of BISI. The fourth research goal was to test the impact of perceived SQ of BISI on perceived user system satisfaction from BISI and perceived SQ of BISI on perceived user information satisfaction from BISI. The impact of perceived IQ of BISI on perceived user information satisfaction and perceived IQ of BISI on perceived user system satisfaction from BISI was also tested using the BI SQ and IQ research model based on the DeLone and McLean (1992, 2003) model for IS success as extended by Nelson et al. (2005).

The main research questions addressed in this study were:

RQ1: What SQ characteristics are valued in BISI by users? What IQ characteristics are valued in BISI by users?

RQ2: What are the CVFs for SQ that users' value in BISI? What are the CVFs for IQ that users' value in BISI?

Stemming from the research questions, this study then addressed the following specific hypotheses:

H1a–d: The CVFs of SQ will have a positive significant impact on perceived SQ of BISI.

H2a–d: The CVFs of IQ will have a positive significant impact on perceived IQ of BISI.

H3: The perceived SQ of BISI will have a positive significant impact on perceived user system satisfaction from BISI.

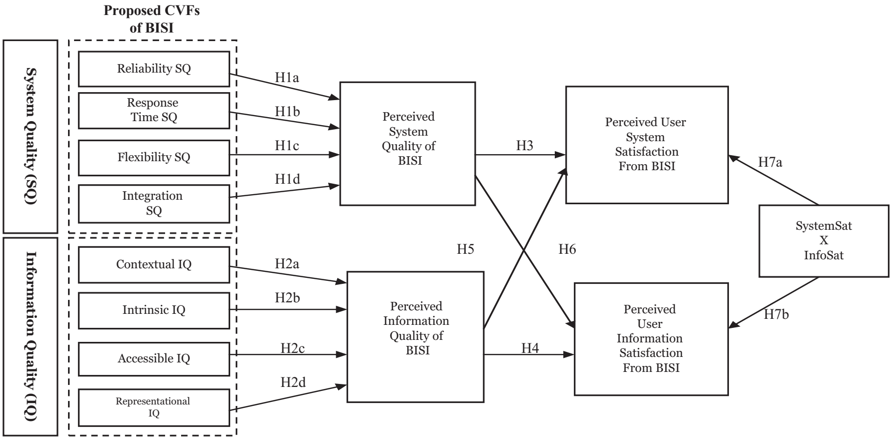


Fig. 6.3 BI SQ and IQ research model based on DeLone and McLean (1992) IS Success Model as extended by Nelson et al. (2005)

H4: The perceived IQ of BISI will have a positive significant impact on perceived user information satisfaction from BISI.

H5: The perceived SQ of BISI will have a positive significant impact on perceived user information satisfaction from BISI.

H6: The perceived IQ of BISI will have a positive significant impact on perceived user system satisfaction from BISI.

H7a: The interactions of perceived user system satisfaction from BISI and the perceived user information satisfaction from BISI will have a positive significant impact on perceived user system satisfaction from BISI.

H7b: The interactions of perceived user system satisfaction from BISI and the perceived user information satisfaction from BISI will have a positive significant impact on perceived user information satisfaction from BISI.

6.3 Methodology

Our study used a mixed method approach following the work of Keeney (1999), utilizing both qualitative and quantitative research methods. Using value theory and IS success theory, the study validated empirically a model for IS success that investigated how an organization may gain user satisfaction in the context of BISI by uncovering the CVFs of SQ and IQ necessary to derive BISI success. Hanson et al. (2005) stated that quantitative and qualitative data could be complementary when variances are uncovered that would not have been found by a single method.

Qualitative research could be used to discover and uncover evidence, while quantitative methods are often used to verify the results, thereby improving the integrity of the findings of the study (Shank 2006). Additionally, both qualitative and quantitative methods each carry their own capabilities to uncover the underlying meaning of phenomena in research (Straub 1989).

6.3.1 Phase I: Expert Panel and Open-Ended Questionnaire

The qualitative process (Phase I) began with the creation and distribution of an open-ended questionnaire designed to elicit SQ and IQ characteristics considered to be important in BISI. Development of the instrument followed the process proposed by Straub (1989). The open-ended questionnaire was developed to uncover new characteristics of SQ and IQ for BISI. An expert panel was formed, consisting of a small group of six individuals with experience in business analytics. The expert panel members had an average of 20 years' experience implementing business analytics systems in large organizations. Four experts were Business Analysts with leading financial institutions in banking, pension finance, and brokerage services. Two of these experts have also managed departments devoted to analytics. The remaining two experts, in addition to implementing business analytics systems were also responsible for BI system infrastructure and implementation services for organizations providing systems services. All experts have performed business analyst functions and have been responsible for decision making using BI system output. SQ and IQ characteristics drawn from the expert panel's responses to the open-ended questionnaire and the literature review of validated sources (Arazy and Kopak 2011; Goodhue 1995; Jarke and Vassiliou 1997; Lee et al. 2002; Nelson et al. 2005; Wand and Wang 1996; Wang and Strong 1996) were analyzed using Keeney's (1999) approach. Similar SQ and IQ characteristics identified from literature as well as responses from the expert panel were grouped into the four main proposed SQ categories of reliability SQ, response time SQ, flexibility SQ, and integration SQ, as well as the proposed four high level IQ categories of intrinsic IQ, contextual IQ, representational IQ, and accessibility IQ. These SQ and IQ characteristics were evaluated for inclusion in an updated list of SQ and IQ items. Items that did not appear to relate to any category were investigated for inclusion in a new SQ or IQ category. After considering the grouping of similar responses as well as the feedback from the expert panel using Keeney's (1999) approach there were 33 SQ and IQ characteristics identified, consisting of 16 SQ items and 17 IQ items identified and grouped under the appropriate SQ and IQ category. This included nine SQ and IQ items identified by the expert panel that did not correspond with any of the initial sources of BI success identified in the literature. As a result, the following nine measurement items were added to the survey instrument: functionality and features of the BI system are dependable, frequency of data generation and refresh in the BI system are flexible, the BI system accommodates remote access, the BI system is scalable, the BI system has an intuitive user interface, the BI system provides appropriate navigation to obtainable information, the BI system provides portability of data and data sources including import and export features, the source of BI information is traceable and verifiable, information is reproducible in the BISI.

6.3.2 Phase II: Instrument, Data Collection, and Exploratory Factor Analysis (EFA)

The quantitative process (Phase II) began with the development of a two part quantitative survey instrument to collect data. This preliminary survey instrument was based on the results of phase I. The quantitative assessment of the SQ and IQ characteristics found in literature, augmented by additional SQ and IQ characteristics uncovered in phase I of the study was performed using value theory under Keeney's (1999) methodology. After a further review by the expert panel, an instrument was developed that had content validity, construct validity, and reliability. The feedback from the expert panel was used to adjust the proposed instrument and included the removal of unnecessary items and the modification of questions, language, and the layout of the instrument (Straub 1989). The final survey instrument emerged from this process which was distributed to a larger group of users of BI systems to assess the perceived value attributed to the items using a 7-point Likert scale ranging from not important to highly important. Our study used the revised quantitative survey instrument to collect data in order to empirically determine the CVFs of SQ and IQ for BISI success. Hair, Anderson, and Tatham (1994) suggested 15–20 observations for each variable for the results of a study to be generalizable. This study targeted 250 participants as an appropriate sample size (Schumacker and Lomax 2010). Approximately 1300 survey invitations were sent to analysts through a service of SurveyMonkey to achieve the response rate necessary to reach the targeted sample size of 250 participants. After completion of pre-analysis data screening, 257 responses were available for analysis for a 20.8% response rate with 176 or 68.5% completed by females and 31.5% completed by males. Analysis of the ages of respondents indicated that 217 or 84.4% were above the age of 30. Additionally, 55 or 21.4% of the respondents considered themselves novices in the use of BI systems, 115 or 44.7% considered themselves average users, 77 or 30% considered themselves advanced users and only 10 or 3.9% considered themselves expert users. Respondents with graduate degrees comprised 35% of the subject population. Overall, 198 respondents or 77% had a university degree.

The study used EFA techniques to uncover the CVFs of SQ and IQ of BISI. Factorial validity assessed whether the measurement items corresponded to the theoretically anticipated CVFs of SQ and IQ in a successful BISI. Principal component analysis (PCA) was used as the extraction method to provide variances of underlying factors (Mertler and Vannatta 2001). The perceived SQ and IQ CVFs of BISI were identified by conducting EFA via PCA using Varimax rotation. PCA was used to extract as many factors as indicated by the data.

6.3.3 Phase III: Confirmatory Factor Analysis (CFA)

In phase III, hypotheses were tested to validate the proposed BI SQ and IQ research model based on IS success theory and the DeLone and McLean (1992, 2003) IS success model as extended by Nelson et al. (2005). This study then gathered data regarding the perceived SQ and IQ of BISI as it relates to perceived user system satisfaction and perceived user information satisfaction from BISI. Since SQ and IQ can separately influence user satisfaction, after determining the CVFs for SQ and IQ of BISI, this study tested each construct of the proposed BI SQ and IQ research model for reliability followed by the testing of the entire model. In addition to the data analysis performed in phase II of the study that established the CVFs for SQ and IQ of BISI, data was also analyzed in Phase III for the conceptual model constructs of perceived SQ of BISI, perceived IQ of BISI, perceived user system satisfaction from BISI, and perceived user information satisfaction from BISI.

6.4 Data Analysis and Results

6.4.1 SQ: Exploratory Factor Analysis—PCA

After conducting EFA via PCA using Varimax rotation, the Kaiser criteria was applied to the SQ factor analysis. Based on the Kaiser criterion, the results of the PCA factor analysis suggested that two SQ factors with a cumulative variance of 61.9% should be retained. Using the factor loadings, survey items were scrutinized for low loadings (<0.4) or for medium to high loadings (~0.4 to 0.6) on more than one factor. The results of this review indicated that five items could be eliminated from further analysis. Furthermore, the Cronbach Alpha analysis indicated that all remaining items supported the reliability of the items and the factors. Moreover, the Cronbach's Alpha of each factor was 0.83 or higher, indicating very high reliability. As a further test of reliability, the Cronbach's Alpha "if item is deleted" was calculated to test the reliability of the items for all SQ factors. Based on an analysis of the results it was concluded that the appropriate number of SQ factors for extraction were two as represented in Table 6.1 and were comprised of 12 items.

As a result of the analysis, integration flexibility SQ was found to explain the largest variance in the SQ data collected and consisted of characteristics that addressed the ability of the BI system to combine information using compatible systems that supported integrated communications and transmissions among a variety of systems and the associated data in various functional areas. The new factor of integration flexibility SQ was also comprised of the BISI SQ characteristics of extendibility, expandability, modularity, and configurability, as well as adaptability and scalability with an intuitive user interface. In particular the characteristic of data portability was considered to be very important to BI users. It is clear that flexibility in integrated systems is important to BISI success. Reliability SQ explained

Table 6.1 SQ CVFs of BISI resulting from PCA

Factor Name	Item	1	2	Factor's Alpha if Item is Deleted
Integration Flexibility SQ	SQI3	.797	.060	.888
	SQI1	.770	.291	.879
	SQI2	.758	.260	.883
	SQF2	.730	.348	.878
	SQF3	.707	.356	.881
	SQI4	.662	.295	.889
	SQF4	.621	.318	.891
	SQF1	.610	.369	.889
Reliability SQ	SQR2	.203	.851	.765
	SQR3	.328	.795	.761
	SQR1	.217	.735	.827
	SQR4	.376	.663	.814
Cronbach's Alpha		.898	.837	

the remaining variance in the data collected and represented a combination of the characteristics of system dependability, recoverability, and low downtime. In essence, BI users found the technical quality of the system to be important. The list of SQ characteristics of BISI is provided in Table 6.2.

6.4.2 *IQ: Exploratory Factor Analysis—PCA*

The results of the IQ EFA under PCA using Varimax rotation and the Kaiser criteria suggested that three IQ factors with a cumulative variance of 75.3% should be retained. Using the factor loadings, survey items were scrutinized for low loadings (<0.4) or for medium to high loadings (~0.4 to 0.6) on more than one factor. The results of this review indicated that three items could be eliminated from further analysis. The Cronbach's Alpha's of the individual factors provided high reliability: representation IQ—0.896, intrinsic IQ—0.957, accessibility IQ—0.852. Based on an analysis of the results it was concluded that the appropriate number of SQ factors for extraction were three, as represented in Table 6.3 and were comprised of 14 items.

Representation IQ was found to explain the largest variance in the IQ data collected and consisted of characteristics that addressed the representation of information in BI systems which rely on the user to ensure that IQ is retained as information from various sources are joined, aggregated, updated, configured, manipulated, and mapped into suitable representations and formats. The item IQC4 "traceability and verifiability of the source of information in BISI" loaded high on the CVF of representation IQ. Accessibility IQ explained the next largest variance in the data collected and included items representing a combination of ease of access to locatable, obtainable, and searchable information. In essence, BI users

Table 6.2 SQ characteristics of BISI

Item	CVF	Perceived SQ Items
SQI3	Integration flexibility SQ	The ability of the BI system to communicate and transmit a variety of data between other systems servicing different functional areas.
SQI1		The ability of the BI system to combine information with other information and deliver to the user.
SQI2		The compatibility of BI system software with other software and hardware
SQF2		The BI system is extendible, expandable, modular, and configurable
SQF3		The BI system is scalable (e.g. hardware, software, memory)
SQI4		The BI system provides portability of data and data sources including import and export features
SQF4		The BI system has an intuitive user interface (UI)
SQF1		The BI system is adaptable to user needs
SQR2	Reliability SQ	The BI system has a low percentage of hardware and software downtime.
SQR3		The BI system can easily recover from malfunctioning equipment and restore data
SQR1		The functionality and features of the BI system are dependable
SQR4		The BI system is of high technical quality

Table 6.3 IQ CVFs of BISI resulting from PCA

Factor Name	Item	1	2	3	Factor's Alpha if Item is Deleted
Representation IQ	IQR3	. 848	. 171	. 144	. 873
	IQR4	. 798	. 296	. 002	. 883
	IQR5	. 733	. 143	. 335	. 876
	IQR1	. 703	. 290	. 381	. 871
	IQR2	.693	.078	.400	.883
	IQC4	.604	.320	.334	.884
Intrinsic IQ	IQI1	. 176	. 914	. 196	. 937
	IQI3	. 223	. 905	.231	. 932
	IQI4	. 211	. 877	. 214	. 949
	IQI2	.249	.864	.178	.953
Accessibility IQ	IQA3	. 358	. 255	. 765	. 772
	IQA2	. 048	. 304	. 764	. 873
	IQA4	.476	.158	.720	. 784
	IQA1	. 527	. 160	. 615	. 816
Cronbach's Alpha		.896	.957	. 852	

found interactive information access for the purpose of improving information content quality important in their BI IQ work. The IQ CVF of BISI with the third highest variance belonged to intrinsic IQ and consisted of the items of information accuracy, consistency, reliability, and correctness. The list of IQ characteristics of BISI is provided in Table 6.4.

Table 6.4 IQ characteristics of BISI

Item	CVF	IQ Items
IQR3	Representation IQ	Information is easily joined, aggregated, updated, configured, and manipulated in BISI
IQR4		Information is reproducible in the BISI
IQR5		Information is mapped into suitable representations at the user level in the BISI
IQR1		Understandability of Information in BISI
IQR2		Format of information in BISI
IQC4		Traceability and verifiability of the source of information in BISI
IQI1	Intrinsic IQ	Accuracy of information in BISI
IQI3		Reliability of information in BISI
IQI4		Correctness of information in BISI
IQI2		Consistency of information in BISI
IQA3	Accessibility IQ	Accessibility to locatable and searchable information in BISI
IQA2		Security of accessed information in BISI
IQA4		Appropriate navigation to obtainable information in BISI
IQA1		Ease of accessing information in BISI

6.4.3 Confirmatory Factor Analysis (CFA)

The strength and direction of the hypothesized relationships (Fig. 6.4) in the conceptual model were validated using the partial least squares (PLS) method, a subtype of structured equation modeling (SEM) used in performing CFA. The bootstrapping resampling method (5000 samples) was also employed. As a result of Phase II factor analysis, the model was revised to replace the proposed theoretically anticipated CVFs of BISI with the empirically determined CVFs of BISI. The paths from the two empirically assessed CVFs of SQ to the perceived SQ of BISI have been named H1.1 and H1.2. Likewise, the paths from the three empirically assessed CVFs of IQ to the perceived IQ of BISI have been named H2.1, H2.2, and H2.3. The paths from user perceived SQ and user perceived IQ of BISI to perceived user system satisfaction and perceived user information satisfaction from BISI as hypothesized in the proposed BI SQ and IQ research model, based on the Delone and McLean IS success model (2003) as extended by Nelson et al. (2005) were tested in the overall context of BISI success.

The PLS generated loadings for the SQ and IQ selected items and CVFs from phase II EFA were again considered high with the lowest item loading at 0.675. Moreover, the SQ CVF's of reliability-SQ and integration flexibility-SQ loaded at 0.836 and 0.898 respectively. The CVF's of intrinsic-IQ loaded at 0.957, accessibility-IQ at 0.868, and representational-IQ at 0.897. PLS was then used to empirically

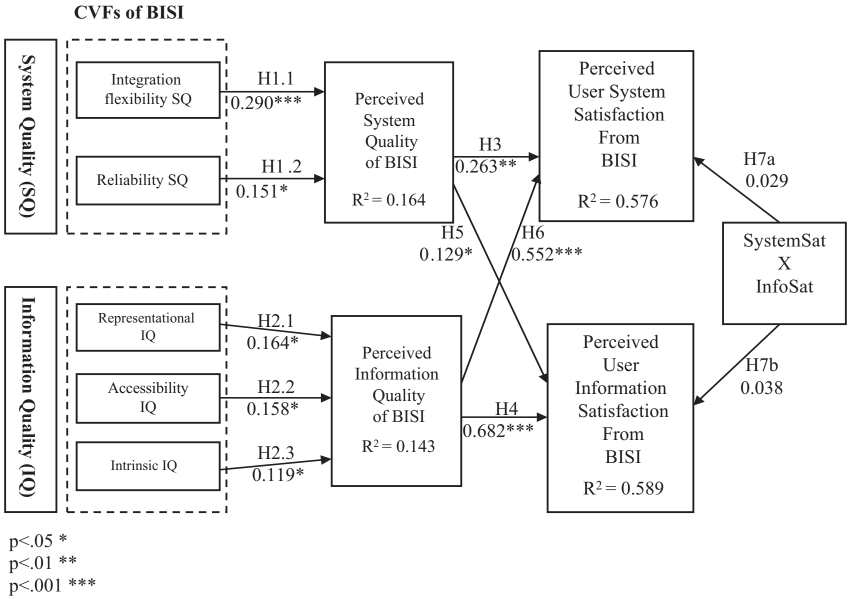


Fig. 6.4 Structural equation model testing results of conceptual model. $p<0.05$*, $p<0.01$**, $p<0.001$***

test the conceptual model path coefficients to determine the significance of the relationships. As indicated in the conceptual model in Fig. 6.4, all CVFs of BISI for SQ and IQ had significant positive impacts on the perceived SQ and IQ of BISI.

6.5 Findings

The results of the testing of the hypotheses clearly indicated support for the empirically determined CVFs of SQ and IQ of BISI as depicted in Table 6.5. Moreover, these results provided evidence that many of the antecedents uncovered in the literature and by the expert panel in the qualitative phase of the study were highly valued by BI users and contributed to the strength of the relationships between the CVFs of BISI and perceived SQ and IQ of BISI. Furthermore, seven of nine items recommended for inclusion in the survey by the expert panel were reliable and grouped accordingly within the retained CVFs.

The results confirm that there is a significant positive impact between perceived SQ and perceived user system satisfaction as well as a significant positive impact between perceived IQ and perceived user information satisfaction. The results also provided confirmation that there is a significant positive impact in the crossover relationships between the perceived SQ and IQ of BISI and the perceived user system satisfaction and perceived information satisfaction from BISI. It is also noted

Table 6.5 Summary of hypotheses results

Hypotheses	Results
H1.1 and H1.2: The CVFs of integration flexibility SQ and reliability SQ will have a positive significant impact on SQ for BISI success.	Supported
H2.1-3: The CVFs of representational IQ, accessibility IQ, and intrinsic IQ will have a positive significant impact on IQ for BISI success.	Supported
H3: The perceived SQ of BISI will have a positive significant impact on perceived user system satisfaction from BISI.	Supported
H4: The perceived IQ of BISI will have a positive significant impact on perceived user information satisfaction from BISI.	Supported
H5: The perceived SQ of BISI will have a positive significant impact on perceived user information satisfaction from BISI.	Supported
H6: The perceived IQ of BISI will have a positive significant impact on perceived user system satisfaction from BISI.	Supported
H7a: The interactions of perceived user system satisfaction from BISI and the perceived user information satisfaction from BISI will have a positive significant impact on perceived user system satisfaction from BISI.	Not Supported
H7b: The interactions of perceived user system satisfaction from BISI and the perceived user information satisfaction from BISI will have a positive significant impact on perceived user information satisfaction.	Not Supported

that the interaction effect did not have a significant positive impact on either perceived user information satisfaction from BISI or perceived user system satisfaction from BISI. These results were shared with members of the expert panel who expressed their agreement and support of the findings.

6.6 Discussion

The main goal of this study was to validate empirically a model for IS success that investigated user satisfaction in the context of BISI by uncovering the CVFs of SQ and IQ necessary to derive BISI success. The study found that a BISI project should place emphasis on the CVFs of integration flexibility SQ and reliability SQ as the primary drivers for SQ of BISI success. Emphasis should also be placed on the CVFs for IQ of representational IQ, intrinsic IQ, and accessible IQ, as the primary drivers for IQ of BISI success.

The CVF of integration flexibility SQ had the most significant effect on the SQ of BISI as greater emphasis was placed on the capability of the BI system to easily combine information from multiple sources while retaining compatibility with other software and hardware. This is important to users of BI analytics as the ability of the BI system to communicate and transmit a variety of data between other systems supporting different functional areas is necessary for BISI success. This had previously been understood to be merely a relevant attribute and expected in BI systems that leveraged data warehouse technologies (Nelson et al. 2005). The results of this study also confirm the importance of integration flexibility SQ to facilitate integration of changing information from various sources to support business decisions. The system must be flexible in supporting ad hoc and unplanned requests for information in various representations. Reliability SQ was also considered as an important CVF as system dependability, recoverability, and low downtime are valued by BI users. On the other hand, the SQ CVF of response time SQ was not a reliable CVF in BISI success. It may be that response time for BISI was considered less important as a separate CVF but was assumed to be available in reliable and flexible BI systems. It might also be possible that due to the analytical nature of BI systems, response time does not carry the same level of importance as would be necessary in a transaction based system.

The CVF of representation IQ had the most significant effect on IQ as the representation of information in BI analytical systems, as with most analytical based applications, relies on the user to ensure that IQ is retained as information from various users and sources are joined, aggregated, updated, configured, manipulated, and mapped into suitable representations and formats. Of particular interest was the high level of importance placed on traceability, verifiability, and the ability to reproduce information in BISI. This may point to user recognition of the need for accountability for the output produced by the user in BI analytical systems. The CVF of accessibility IQ was also considered important in successful BISI as emphasis was placed on the importance of ease of access to locatable, obtainable and searchable information as well as the security of the accessed information and the ability to navigate within the BI system. Intrinsic IQ was also a reliable CVF as information accuracy, consistency, reliability, and correctness have generally been a cornerstone to BI success. The CVF of contextual IQ, however, was not a reliable CVF of perceived IQ of BISI. This may be due to the nature of BI systems which often rely on historical data to perform analytics and, as with response time expectations and assumptions, the contextual characteristics of currency, timeliness, sufficiency, and relevancy of information may be assumed to be of less importance than in systems that are more time dependent and transaction oriented.

The effects of perceived SQ and IQ of BISI on perceived user system and information satisfaction from BISI were also of particular interest in the study. The perceived IQ of BISI had a significant positive impact on perceived user information satisfaction from BISI. Perceived IQ of BISI also had a significant positive impact on perceived user system satisfaction from BISI. While the perceived SQ of BISI had a significant positive impact on perceived user system satisfaction from BISI there was less of an impact on perceived user information satisfaction from BISI, thereby highlighting the differences between the BI system and the information produced. It is apparent that BI analytical systems provide advanced interfacing capabilities that may influence the users' perception that the interaction with the interface has an impact on the output produced thereby making it difficult to differentiate between the system interface and the user's responsibility for the quality of the output. This study also confirms that while the empirically determined CVFs of SQ and IQ of BISI and their crossover effects are perceived to be important to user perceived SQ and IQ user satisfaction from BISI, the strength of the impact of IQ on the system corresponds to the importance users place on the output in analytical BISI. Moreover, this finding emphasizes the differences between the BI system tools and the output that is produced as well as the need for BI system implementers to accept responsibility for IQ. The results of this study and particularly the crossover effects found in the research model shed light on our understanding of quality and highlight a continuum of interactivity in BISI that distinguishes SQ and IQ characteristics and their effect on output and user perceived satisfaction.

6.7 Contributions of the Study

Our study has several implications in the field of BI, particularly for practitioners. First, it contributes to the body of knowledge by empirically identifying the CVFs and characteristics of SQ and IQ that users find important in successful BISI. Secondly, this study empirically addressed the relationship between the quality of the BI system (SQ) and its output (IQ). The study determined that there was a significant positive impact from perceived SQ and IQ of BISI on perceived user system and information satisfaction from BISI. Previous studies in BISI placed emphasis on the use of a data warehouse within the BISI domain. However, while DW can be used with varying levels of importance in BI systems, they can also exist without a DW. There had also been some ambiguity between the system (SQ) and its output (IQ) whereby the strength of the relationship between SQ and information satisfaction was stronger than the relationship between IQ and information satisfaction. The empirically developed findings of this study are in line with expectations for system success as theorized in the BI SQ and IQ research model, based on the DeLone and McLean IS success model (1992, 2003) as extended by Nelson et al. (2005). Lastly, this study identified characteristics of SQ and IQ that are valued or important in BISI, thereby assisting practitioners in determining the best areas of focus for BISI success. This study provided compelling evidence that the antecedents and CVFs of integration flexibility SQ and reliability SQ are important to BISI success. Moreover, this study also provided compelling evidence that the antecedents and CVFs of representation IQ, accessibility IQ, and intrinsic IQ are important to successful BISIs. This study represents the first empirical analysis of CVFs that affects SQ and IQ for BISI success and has uncovered important factors and characteristics for BISI success that will enable BI stakeholders to better optimize scarce resources.

6.8 Limitations and Suggestions for Future Research

The primary limitation of this study surrounds the possibility that participants may have varying degrees of exposure to analytical BI systems. While BI systems are associated with decision making, the complexity of the implemented system and the interpretation of its output could require skill levels that may not be consistent among all participants. It is, therefore, assumed for the purposes of this study that participants had, at a minimum, BI or analytical system implementation experience. The gender differences among BI users may also be examined more closely, as there were twice as many females that participated in the survey than males. Another limitation surrounds the lack of consistency in the BI technologies used. For example, one participant may have experienced BI using the IBM Cognos tool. Another participant may have experienced BI using systems that were integrated in an ERP system.

Our study provided a solid theoretical foundation from which future studies can originate. Firstly, it was designed to empirically validate a model for IS success for user satisfaction in the context of BISI and although the individual CVFs of SQ and IQ necessary to derive BISI success were significant, future studies may be warranted to examine and assess other constructs and items that are important to BI systems users that lead to BISI success such as governance and service quality. Moreover, BI systems are expected to accommodate the big data phenomenon which represents additional, unusual, and complex sources of data in BISI (Wixom et al. 2014). Furthermore, future research could assess the needs of BISI in a big data environment whereby information is often unstructured. With more attempts to manipulate input streams, many issues have been raised in the field of big data, accompanied by a wide variety of potential failures. There have been few attempts to actually apply big data analytics to the validation of big data, particularly in the analysis of data streams (Wixom et al. 2014). Social media for instance is open to a wider range of validation techniques. This could explain, in part, the high degree of importance placed by BI users in this study on validity of data sources. This finding may also point to the need to establish tailored systems development methodologies with emphasis on testing and verification for the delivery of BI systems in the future.

6.9 Conclusion

This study provided further evidence that the antecedents of integration flexibility SQ and reliability SQ are important to BISI success. Moreover, it demonstrated compelling evidence that the antecedents and CVFs of representation IQ, accessibility IQ, and intrinsic IQ are important to successful BISI. These findings confirm the widely held view that BISI is not a conventional application-based IT project but a complex undertaking requiring an appropriate infrastructure over a lengthy period of time. The findings also confirm that successful BISI require a robust and easy to use interface for user-driven information representation in an analytical user-based decision support system context from multiple integrated heterogeneous sources (Goodhue and Thompson 1995; Yeoh and Koronios 2010). Our study also reported that there is a significant effect in the relationships of perceived IQ of BISI to perceived user information and system satisfaction thereby confirming the importance BI system users place on information and the system output produced.

Biographies

Paul P. Dooley is a lecturer in the analytics program at Northeastern University. He has also held various executive management positions in business and information management for large international financial institutions. His research interests include data science and business intelligence with specific attention to quality factors affecting successful implementations.

Yair Levy is a Professor of Information Systems and Cybersecurity at the College of Engineering and Computing at Nova Southeastern University, the Director of the Center for e-Learning Security Research, and chair of the Information Security Faculty Group at the college along with serving as the director of the M.S. and Ph.D. programs in Cybersecurity and Information Assurance. He heads the Levy CyLab (http://CyLab.nova.edu/), which conducts innovative research from the human-centric lens of four key research areas Cybersecurity, User-authentication, Privacy, and Skills, as well as their interconnections. Levy authored one book, three book chapters, and numerous peer-reviewed journal as well as conference proceedings publications. His scholarly research have cited over 1400 times. Dr. Levy has been an active member of the US Secret Service (USSS)'s—Miami Electronic Crimes Task Force (MECTF) and The South Florida Cybercrime Working Group (SFCWG). He was trained by the Federal Bureau of Investigation (FBI) on various topics, and actively serves as a member on of the FBI/InfraGard, and consults federal agencies, state and local government groups on cybersecurity topics. He is also a frequent invited keynote speaker at national and international meetings, as well as regular media interviews as a Subject Matter Expert (SME) on cybersecurity topics. Read more about Dr. Levy via: http://cec.nova.edu/~levyy/.

Raymond A. Hackney is Chair in Business Systems within Brunel University London, Business School, UK. His research interests are the management of information systems with a speciality in government sectors including leadership, environmental sustainability, Big Data analytics and transformational digital inclusion. He is a former President of the Information Resource Management Association, USA, served on the Board of UK AIS and the EDAMBA Executive.

James L. Parrish is the Chair of the Department of Information Systems and Cybersecurity at Nova Southeastern University's College of Engineering & Computing. He received his Ph.D. from the University of Central Florida in 2008 and has several publications in the area of cybersecurity and knowledge management. In addition to his role at NSU, Dr. Parrish also serves on the boards and leadership teams of multiple academic and industry organizations.

References

Arazy O, Kopak R (2011) On the measurability of information quality. J Am Soc Inf Sci Technol 62(1):89–99

Arnott D, Prevan G (2008) Eight key issues for the decision support systems discipline. Decis Support Syst 44(3):657–672

Bailey JE, Pearson SW (1983) Development of a tool for measuring and analyzing computer user satisfaction. Manag Sci 29(5):530–545

Boynton AC, Zmud RW (1984) An assessment of critical success factors. Sloan Manag Rev 25(4):17–27

Chee T, Chan L-K, Chuah M-H, Tan C-S, Wong S-F, Yeoh W (2009) Business intelligence systems: state-of-the-art review and contemporary applications. In: Symposium on progress in information & communication technology, p 96–101

Clark TD, Jones MC, Armstrong CP (2007) The dynamic structure of management support systems: theory development, research focus, and direction. MIS Q 31(3):579–615

DeLone WH, McLean ER (1992) Information systems success: the quest for the dependent variable. Inf Syst Res 3(1):60–95

DeLone WH, McLean ER (2003) The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst 19(4):9–30

Dhillon G, Bardacino J, Hackney R (2002) Value-focused assessment of individual privacy concerns for internet commerce. In: Proceedings of the Twenty-Third international conference on information systems, p 705–709

Dhillon G, Torkzadeh G (2001) Value-focused assessment of information system security in organizations. In: Proceedings of the twenty-second international conference on information systems, p 561–565

Dinter B, Schieder C, Gluchowski P (2011) Towards a life cycle oriented business intelligence success model. In: Proceedings of the Americas conference on information systems

Etezadi-Amoli J, Farhoomand AF (2011) On end-user computing satisfaction. MIS Q 15(1):1–5

Gatian AW (1994) IS user satisfaction: a valid measure of system effectiveness? Inf Manag 26(1):119–131

Goodhue DL (1995) Understanding user evaluations of information systems. Manag Sci 41(12): 1827–1844

Goodhue DL, Thompson RL (1995) Task-technology fit and individual performance. MIS Q 19(2):213–236

Hair JF, Anderson RE, Tatham RL, Black WC (1994) Multivariate data analysis. Prentice Hall, Upper Saddle River, NJ

Hanson WE, Plano-Clark VL, Petska KS, Creswell JW, Creswell JD (2005) Mixed methods research designs in counseling psychology. J Counsel Psychol 52(2):224–235

Howson C (2008) Successful business intelligence: secrets to making BI a killer application. McGraw-Hill, New York

Iivari J (2005) An empirical test of the DeLone-McLean model of information system success. ACM SIGMIS Database 36(2):8–27

Isik O, Jones MC, Sidorova A (2013) Business intelligence success: the roles of BI capabilities and decision environments. Inf Manag 50(1):13–23

Jarke M, Vassiliou Y (1997) Data warehouse quality: a review of the DWQ project. In: Proceedings of the conference on information quality, p 299–313

Jourdan Z, Kelly RK, Marshall TE (2008) Business intelligence: an analysis of the literature. Inf Syst Manag 25(2):121–131

Keeney RL (1992) Value-focused thinking. Harvard University Press, Cambridge, MA

Keeney RL (1999) The value of internet commerce to the customer. Manag Sci 45(4):533–542

Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inf Manag 40(1):133–146

Levy Y (2003) A study of learner's perceived value and satisfaction for implied effectiveness of online learning systems. Diss Abstr Int A65(03):1014

Levy Y (2006) Assessing the value of e-learning systems. Information Science, Hershey, PA

Levy Y (2008) An empirical development of critical value factors (CVF) of online learning activities: an application of activity theory and cognitive value theory. Comput Educ 51(4):1664–1675

Levy Y (2009) A value-satisfaction taxonomy of IS effectiveness (VSTISE): a case study of user satisfaction with IS and user-perceived value of IS. Int J Inform Sys Service Sect 1(1):93–118

Luftman J, Ben-Zvi T (2010) Key issues for IT executives 2009: difficult economy's impact on IT. MIS Q Exec 9(1):49–59

Marshall L, de la Harpe R (2009) Decision making in the context of business intelligence and data quality. SA J Inform Manage 11(2):1–15

Mertler CA, Vannatta RA (2001) Advanced and multivariate statistical methods: practical application and interpretation. Pyrczak, Los Angeles, CA

Nah F, Siau H, Sheng H (2005) The value of mobile applications: a study on a public utility company. Commun ACM 48(2):85–90

Negash S, Gray P (2008) Business intelligence, handbook on decision support. C.W. Holsapple, Berlin

Nelson RR, Todd PA, Wixom BA (2005) Antecedents of information and system quality: an empirical examination within the context of data warehousing. J Manag Inf Syst 21(4):199–235

Petter S, McLean E (2009) A meta-analytic assessment of the DeLone and McLean IS success model: an examination of IS success at the individual level. Inf Manag 46(3):159–166

Petter S, DeLone W, McLean E (2013) Information systems success: the quest for the independent variables. J Manag Inf Syst 29(4):7–61

Popovic A, Hackney R, Coelho PS, Jacklic J (2012) Towards business intelligence systems success: effects of maturity and culture on analytical decision making. Decis Support Syst 54(1):729–739

Power DJ (2008) Understanding data-driven decision support systems. Inf Syst Manag 25(2): 149–154

Rai A, Lang SS, Welker RB (2002) Assessing the validity of IS success models: an empirical test and theoretical analysis. Inf Syst Res 13(1):50–69

Rokeach MJ (1969) Beliefs, attitudes, and values. Jossey-Bass, San Francisco, CA

Rokeach MJ (1973) Nature of human values. The Free Press, New York, NY

Ryu KS, Park JS, Park JH (2006) A data quality management maturity model. ETRI J 28(2):191–204

Schumacker RE, Lomax RG (2010) A beginner's guide to structural equation modeling. Routledge, New York, NY

Seddon PB (1997) A respecification and extension of the DeLone and McLean model of IS success. Inf Syst Res 8(3):240–253

Sethi V, King RC (1999) Nonlinear and noncompensatory models in user information satisfaction measurement. Inf Syst Res 10(1):87–96

Shank G (2006) Six alternatives to mixed methods in qualitative research. Qual Res Psychol 3(4):346–356

Sheng H, Nah K, Siau K (2005) Strategic implications of mobile technology: a case study using value-focused thinking. J Assoc Inf Syst 9(6):344–376

Sheng H, Siau K, Nah FF (2010) Understanding the values of mobile technology in education: a value-focused thinking approach. ACM SIGMIS Database 41(2):25–44

Siau K, Nah F, Sheng H (2004) Value of m-Commerce to customers. In: Proceedings of the tenth Americas conference on information systems, p 2811–2815

Straub D (1989) Validating instruments in MIS research. MIS Q 13(2):147–169

Todd G (2009) The imperative of analytics. Inf Manag 19(2):44–47

Urbach N, Smolnik S, Riempp G (2009) The state of research on information systems success: a review of existing multidimensional approaches. Busin Inf Syst Eng 1(4):315–325

Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. Commun ACM 39(11):86–95

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12(4):5–34

Watson HJ, Goodhue DL, Wixom BH (2002) The benefits of data warehousing: why some organizations realize exceptional payoffs. Inf Manag 39(1):491–502

Wixom BH, Ariyachandra T, Douglas D, Goul M, Gupta B, Iyer L, Kulkarni U, Mooney JG, Phillips-Wren G, Turetken O (2014) The current state of business intelligence in academia: the arrival of Big Data. Commun Assoc Inf Syst 34(1):1–13

Yeoh W, Koronios A (2010) Critical success factors for business intelligence systems. J Comput Inf Syst 50(3):23–32

Chapter 7
Business Intelligence System Use in Chinese Organizations

Yutong Song, David Arnott, and Shijia Gao

Abstract Chinese business has developed exponentially in the last few decades and Chinese firms are highly influential in world trade. Business intelligence (BI) systems are large-scale decision support systems (DSS) that analyze enterprise data to generate business insights. BI was developed in the West and is integral to contemporary Western management practices. It is generally assumed that western BI systems are useable and effective in a Chinese context. No study has been undertaken to investigate the use behavior of large-scale DSS in Chinese organizations. We conducted two exploratory case studies in large indigenous Chinese organizations. The case analysis shows that a complex cultural factor (provisionally termed Factor X) affects BI systems use in China. A set of propositions are formulated from the analysis. They will be used as a foundation for future research on Chinese BI.

Keywords Business intelligence • System use • China • Guanxi • Exploratory case study

7.1 Introduction

Guanxi is a universal and unique Chinese cultural norm (CN). Guanxi refers to 'a whole complex of social practices, strategies and ethics of the exchange and reciprocity of gifts, favors and banquets' (Davis 2005, p. 232). It has been practised for centuries and remains highly relevant in Chinese society today. Guanxi has affected the evolution of Chinese society, economy, and business environment. Guanxi is not as simple as Western relationships; Western relationships grow out of deals, while Chinese deals grow out of relationships. Gu et al. (2008) identified eight key differences from Western perspectives when considering guanxi as a construct and governance strategy in business environments. Guanxi has attracted research attention.

Y. Song (✉) • D. Arnott • S. Gao
Faculty of Information Technology, Monash University,
197, Caulfield East, VIC 3145, Australia
e-mail: yuri.song@monash.edu; David.Arnott@monash.edu; Caddie.Gao@monash.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_7

Many studies have been conducted in foreign companies, which operate their business in China (for example, Millington et al. 2005).

Business intelligence (BI) is a large-scale decision support systems (DSS) approach that analyzes enterprise data using specialized techniques to generate business insights. Managers, at different levels, may leverage these business insights to make decisions in order to improve organization performance. The "mega BI vendors" (Microsoft, Oracle, IBM, and SAP) have increased their investments in BI over recent years. Gartner Inc. reported that BI and analytics is the top technology investment priority for CIOs, and has been in the top five technology priorities over the last decade (Gartner 2015). The Chinese economy attracts considerable attention from IT vendors that wish to sell contemporary Western technology, such as BI. The assumption is that Western IT is appropriate to Chinese uses. However, research has lagged the vendor efforts and the nature of the use of BI systems in large indigenous Chinese organizations remains unknown. This is the research gap addressed by this project.

This chapter describes an exploratory study that develops a set of propositions of BI systems use in indigenous Chinese organizations. It is explicitly concerned with Chinese CN, the nature of Chinese organizations, and Chinese business decision-making. The remainder of this chapter is structured as follows: the next section outlines relevant literature, and identifies the research constructs and concepts that will inform the empirical research. This is followed by the research design discussion, which in turn is followed by the case study analysis and results. This analysis leads to the formulation of the Chinese BI systems use propositions.

7.2 Theoretical Background

The aim of this section is to identify research constructs and concepts that may be relevant to Chinese BI systems use. The section discusses literature that is relevant to the project in three subsections: IS and BI research in China, Chinese CN, and research constructs and concepts.

7.2.1 IS and BI Research in China

The literature review scope was broadened to information systems (IS) in general, and was restricted to mainland Chinese IS studies. Davison and Martinsons (2016) argued that particularism of research design is critical to the validity of research design. Studies that were conducted of the Chinese diaspora, that is where data were collected in Hong Kong, Macau, and Taiwan, were excluded from the sample due to variations in their CN.

A keyword search was conducted in the 'basket of eight' journals. Two elite IS journals *Information & Management*, and *Decision Support Systems* were then added to the sample because of their A* status by Australian Business Deans Council, and their relevance to DSS. Within the sample of 162 published papers, only eight papers (5.4%) focused on DSS. The only Chinese BI study, Li et al. (2013), focused on intrinsic and extrinsic motivation as a lens to investigate both routine and innovative use of BI systems in the post-acceptance period. Cultural aspects of BI systems use were not investigated in this study. Two of the Chinese DSS papers, Zhang et al. (2007) and Lowry et al. (2010), employed Hofstede's culture dimensions in their investigation on how national cultural differences affect group decision-making. Hofstede's theory (1980) concerns national comparisons and is not widely used outside of IT research where this theory was formulated between 1967 and 1973 for IBM. Hofstede's theory is not appropriate for this project as its focus is on cultural comparisons between countries. This project concerns one culture and country—People's Republic of China.

7.2.2 Guanxi and Other Chinese Cultural Norms

There have been many debates about applying theories, guidelines, and other forms of research outcomes that have been derived from studying Western business to Eastern contexts. In China, IS use is constantly formed by Chinese business culture (Martinsons and Westwood 1997). Cultures determine how people react and behave, and research results should only be applied to other cultural contexts with caution (Mao and Palvia 2008). Western IS theories may or may not be applicable to Chinese practice.

Researchers have been aware of cultural impacts in research conducted in China. Many researchers briefly mentioned cultural impacts in their research limitations discussion; examples include Liu et al. (2013) and Wang et al. (2013). A small number of researchers have considered culture as a factor in their studies (14 of 162 or 8.64%). Two of these 14 papers studied guanxi as a research concept: Shin et al. (2007) pointed out that guanxi and collectivism had a stronger influence on in-group than on external information sharing, while Confucian dynamism appeared to have less effect on in-group information sharing. Davison et al. (2013) argued that guanxi strengthened transactive memory networks, and facilitated knowledge sharing. Nevertheless, Guanxi has not been investigated in the terms of BI or the wider DSS discipline.

7.2.3 Research Constructs and Concepts

Technology adoption and use represent different phases of the utilization of technology, and are affected by different factors. Karahanna et al. (1999) concluded that adoption and use are sequential phases of technology acceptance. Adoption is determined by normative considerations, while continued usage is determined by attitudinal factors (Karahanna et al. 1999). It is consistent with Deng and Chi's (2012) identification of different BI use patterns between the initial usage phase and the continued usage phase. Researchers have used the Technology Acceptance Model (TAM) (Davis 1989) and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al. 2003) to investigate not only technology acceptance but also use behaviors in China. For example, Liu and Forsythe (2011) used these theories to examine the continuous use of online shopping channels and Zhou (2011) investigated the use of mobile Internet in China. Further, using constructs from TAM and UTAUT to investigate BI systems use is informed by the adaptation of TAM and UTAUT for IT continuance by Bhattacherjee and Lin (2015) and the adaptation of UTAUT for clinical DSS by Shibl et al. (2012).

In UTAUT, use behavior (UB) is the actual use of a system and, behavioral intention (BeI) is an individual's attitudes towards using the system. Early research indicated that BeI is a major determinant of UB (Davis et al. 1989). Social influence focused on the extent to which important others believe the person should use or not use a particular technology. This factor involves investigating how CN directs social interactions between people, and how people will react to these social interactions. CN is a subjective factor that could influence both BeI and UB. Leidner and Kayworth (2006) reviewed 82 IS journal articles, and concluded that CN has significant impacts, especially in IS development and use. In China, CN consists of many dimensions but is especially informed by guanxi.

Perceived usefulness (PU) and perceived ease of use (PEOU) and their effect on BeI has been extensively researched in IS (Taylor and Todd 1995). Perceived facilitating conditions (FC) refers to the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system (Venkatesh et al. 2003). Gogindarajan and Trimble (2012) argued that the degree of IT innovation is higher in emerging economies than in Western countries. This is because organizations in an emerging economy do not have extensive legacy systems. This means that there are less constraints on the level of IS innovation. As a result, FC may have a higher influence in a Chinese context than in the West.

Gender and age are self-explanatory constructs. Generational differences are derived from, and shaped by, political, socioeconomic, and cultural events. Unlike the west, the major events that caused generational difference in China were the foundation of the PRC in 1949, the Great Leap Forward in the 1960s, the Cultural Revolution in the 1970s, and the One Child and Open Policies in the 1980s. Erickson's (2009) framework of classifying Chinese generations has been adopted in this research, namely Traditionalists (1928–1945), Boomers (1946–1964), Generation X (1965–1979), and Generation Y (1980–1995). Age affects an individual's status in an organization, as respecting one's elders is a key moral principle in China.

Emerging economies do not follow Western trajectories of economic development because of their different infrastructures, geographies, cultures, languages, and governments (Gogindarajan and Trimble 2012). This means that managers who have worked or trained in the West can have difficulty in adapting to current Chinese circumstances. Therefore in this research, experience will have at least two dimensions: category (work or education) and place (domestic or overseas).

Voluntariness of use refers to the degree to which BI users are forced to use the system. Chinese managers may have less discretion in their use of BI systems than those in the West. In addition, managers who experienced the Cultural Revolution may lack technology and mathematics literacy. They may use assistants or intermediaries to use BI systems at a higher rate than Western managers.

Table 7.1 summarizes the research constructs and definitions that were discussed.

Table 7.1 Definitions of research constructs and concepts

Constructs & Concept	Original Definitions	Definition Adopted for this Project
Behavioral Intention (BeI)	Attitude toward using technology is defined as an individual's positive or negative feeling about performing the target behavior (Davis et al. 1989; Venkatesh et al. 2003).	Behavioral intention refers to an individual's positive and negative attitudes towards using the BI system to perform the target behavior.
Use Behavior (UB)	Not defined explicitly (appears in Venkatesh et al. 2003, 2012).	Use behavior refers to the actual use of a BI system.
Guanxi	Social influence is defined as 'the degree to which an individual perceives that important others believe he or she should use the new system' (Venkatesh et al. 2003)	Guanxi refers to a whole complex of social practices, strategies and ethics of the exchange and reciprocity of gifts, favors and banquets.
Perceived Usefulness (PU)	Perceived usefulness is defined as 'the prospective user's subjective probability that using a specific application system will increase his or her job performance within an organizational context' (Davis et al. 1989).	Perceived usefulness refers to the user's subjective probability that using a specific BI system will increase their job performance within an organizational context.
Perceived Ease of Use (PEOU)	Perceived ease of use is defined as 'the degree to which the prospective user expects the target system to be free of effort' (Davis et al. 1989).	Perceived ease of use refers to the degree to which the real user experiences the target BI system to be free of effort.
Perceived Facilitating Conditions (FC)	Facilitating conditions are defined as 'the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system' (Venkatesh et al. 2003).	Perceived facilitating conditions is a subjective factor and refers to the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the BI system.
Gender	Self-explanatory	Self-explanatory
Age	Self-explanatory	Age refers to different generations. These generations are defined by Erickson (2009) for a Chinese context.
Experience	Not defined explicitly (appears in Venkatesh et al. 2003, 2012).	Experience refers to education and work experience, and whether the experience was gained overseas or domestically.
Voluntariness of Use	Not defined explicitly (appears in Venkatesh et al. 2003, 2012).	Voluntariness of use refers to the degree to which individuals perceive that they have freedom to choose whether or not they can use the BI system.

7.3 Research Method and Design

The research scope excludes Hong Kong, Macau, and Taiwan. Taking Hong Kong as an example, executives from Hong Kong may be considered as Chinese insiders by Westerners but they will be treated as outsiders by indigenous Chinese business people (Fock and Woo 1998). As a result, this research focuses only on large business organizations in mainland China. Further, these organizations are indigenous to China and are not the Chinese subsidiaries of foreign multinational corporations. The majority of general IS research that has been conducted in China has used a survey method. However, survey data is not rich enough for an exploratory study of BI systems use in a Chinese context. A multiple case study research design was adopted from Yin (2014) with semi-structured interviews as the data collection technique.

The case selection criteria were based on the above discussion. First, the organizations involved were indigenous mainland Chinese. Second, the size of the case organizations was large as BI systems provide large-scale decision support. Large organizations are more likely to have a complex managerial hierarchy, which will help to understand BI systems use behavior across different managerial levels. Third, researchers had to obtain access and receive sufficient support from high-level management in the case organizations. There were two sets of participants: developers who could provide sufficient details about BI systems, and managers and senior professionals who use BI systems to support their decision tasks.

Based on the identified research constructs and concepts, interview protocols were developed in English and translated into Chinese for data collection. This research adapted a back-translation technique based on Brislin (1970). These adaptations were inspired by Jones et al. (2001) and Sousa and Rojjanasrirat (2011). A pilot study to pre-test the interview protocols was conducted with professionals and academics who have extensive experience in the BI field.

7.3.1 Case Study Sites

Two large indigenous Chinese organizations were selected for this research. The first case is the Chinese Insurance Company (CIC). The name of the company has been disguised for ethics approval reasons. CIC has more than 5,000 employees. Though CIC was founded by a local Chinese group with foreign investment in 2002, CIC transferred to total Chinese ownership in 2010. Most employees are mainland Chinese and do not have overseas qualifications or work experience. CIC sells insurance policies all over China via traditional physical sales offices in many provinces and online transactions.

The second case is Alibaba Group (AG), which is the largest Internet business in China. AG is representative of a group of newly established Internet companies in China. AG has a Western-like appearance that is significantly different to traditional Chinese companies like CIC. AG had about 35,000 employees in 2015. AG was founded by a local Chinese group led by Jack Ma in 1998; it floated on the New York

Stock Exchange in 2014. AG operates in both domestic and international markets via different portals. Its main business operations are located in mainland China.

Both CIC and AG fitted the case selection criteria. CIC represents traditional Chinese organizational structures and values, while AG represents modern Chinese business. This means that the research project is informed by a range of Chinese organization styles. CIC and AG were founded around the same economic development era.

7.3.2 Data Collection and Analysis Method

Empirical data collection was carried out from July to September, 2014. Data was collected in three branches of CIC, located in two cities, and three campuses of AG that are located in two cities. No incentive was offered to any participant. Detailed notes were taken during all interview sessions at CIC, while at AG, in addition to notes, all interview sessions were audio recorded. All field notes were recorded, stored, and sorted by interview session, and all audio recordings were transcribed. This research adopted recommendations from Miles et al. (2013) to guide data analysis.

All field notes and transcripts were loaded into a qualitative data analysis software package—NVivo. The first cycle of coding used techniques from Miles et al. (2013), namely provisional coding (an exploratory method) and simultaneous coding (a grammatical method). Provisional codes were generated from the literature review. This set was later expanded to include new codes. Under simultaneous coding one piece of a transcript could be coded under multiple codes. These techniques assisted in creating codes from emerging themes, and for the second cycle of coding.

7.4 Preliminary Results and Discussion

This section discusses the removal of some constructs from further consideration, refining relevant constructs, and adding emerging constructs to the propositions. This section also identifies relationships among constructs according to their strength and importance. It is important to mention that only constructs and concepts were taken from the literature review, relationships between constructs were developed from the BI case study data.

7.4.1 Changes to the Research Construct Set

After the case study analysis, Gender, voluntariness of use, FC, and BeI have been removed from the proposition development on the BI systems use. No significant difference was found between male and female use patterns in Chinese BI systems. Most BI users stated that the use of BI was mandatory, while voluntary use was typical for managers and senior professionals who did not have extensive analysis needs. All users confirmed that managers were in favor of staff using BI systems, and both organizations provided sufficient support to BI systems development and enhancement. There was no budget constraint on BI systems in either case. UB in Chinese BI systems does not depend on, or is affected by, BeI. This is partly a consequence of the lack of voluntary usage patterns. Based on the data analysis, gender, voluntariness of use, FC, and BeI did not appear to affect the relationships among important research constructs.

7.4.2 Propositions about Chinese BI Systems Use

The case data analysis yielded five propositions to guide further investigation on Chinese BI use.

Proposition 1. BI systems are developed for supporting specific decision tasks in the organizations. This proposition involves two important emerging constructs from the data analysis. A decision is a commitment to an action. A decision task is part of a manager's job that requires decision-making. The nature of a decision task (NT) has been categorized differently in DSS research with the most influential classifications being Simon (1960) and Anthony (1965). Simon's decision types are based on the levels of understanding or structure that is perceived in a decision situation. Anthony's managerial activity categories help to classify decision tasks in terms of operational, tactical, and strategic management activities. Gorry and Scott Morton (1971), the seminal article for the DSS field, proposed a two-dimensional IS task framework based on the Simon and Anthony decision typologies.

The decision tasks of CIC and AG managers were applied to the Gorry and Scott Morton (1971) decision task framework. In the case studies no BI systems were used to support unstructured or strategic tasks. Semi-structured tactical tasks were the most commonly reported decisions in both CIC and AG. The low proportion of structured tasks may be due to the seniority of the participants, but it also may be the situation that structured tasks are supported by other IT systems. Examples of decision tasks include, a structured operational decision task—after-sale support in order to respond to clients' further enquiries about purchased insurance policies (CIC); a structured tactical decision task—to identify which developmental phase of electronic business has more laws and regulations established, and this analysis may be used in later analysis for strategic decisions (AG); a semi-structured operational decision task—financial data analysis support of consolidating data from alternative source systems (CIC); and, a semi-structured tactical decision task—identifying and consolidating real life business cases and providing these cases to external researchers (AG).

The nature of a BI system (NS) refers to the overall characteristics of a BI system, which are determined by the technology environment, data quality, and system quality. Data with high quality are current, well maintained, and at the appropriate level of detail. High system quality occurs when BI systems are highly reliable and dependable, and available on the platforms that users need. In both CIC and AG it was often difficult to locate where and when the needed data were available. Understanding NS is essential in identifying its impacts on the alignment of decision tasks with the BI system (TSA) and PEOU.

CIC has several generations of DSS/BI that were developed in house, including the Management Information Systems (MIS), Key Performance Indicator (KPI), Customer Management Systems (CMS), and the new core BI system. MIS was copied directly from a Western product and was developed for regulatory reporting to the Insurance Supervision Department. CIC built most decision support for other decision tasks on the MIS, and the system became inefficient. The two most mentioned BI applications are KPI and CMS, where KPI mainly focuses on audit operations and summary purposes, and CMS is mainly used to investigate the details of policies. These two systems were built to solve the shortcomings of MIS. KPI and CMS also assisted with CIC's compliance with industrial standards. Data stored in these systems need to be manually exported and consolidated, and inconsistencies were frequently reported due to different extract, transform, and load (ETL) process and data definitions among systems, branches, and reports. In October 2014, CIC tested a new system that will replace all existing BI systems in late 2015.

AG's technological environment is unique. Data are often only collected and used within the same business unit (BU) to maintain a high level of data security, though AG has a consolidated data platform in operation. If anyone requests to review data from another BU, then the request needs to be assessed by many layers of management. As a result, different BUs have developed their own BI systems based on their data analysis needs. At AG, all developed platforms, applications, tools and systems are assigned to individual system managers who take responsibilities of monitoring the status of the systems, processing requests by users, and refining the systems. The most common decision tasks were summarizing, monitoring, and predicting business operation performance.

In both case studies BI systems were developed for one fundamental reason—supporting specific decision tasks. This could be a point of difference with western BI. The data from the cases show that NT determines NS.

Proposition 2. Decision task—BI system alignment is a superior research construct than perceived usefulness in Chinese BI use. According to past research, PU is important in determining BeI and therefore UB (Taylor and Todd 1995). PU refers to an actual user's perception of subjective probability that using a specific application system will increase their performance. PU is one of the key factors in utilization-focused IS models (e.g. TAM). BI systems were used for multiple purposes at CIC and AG. Internal users focused on comparing, managing, reporting, and analysis. For external users, BI systems were used to report on industrial standards, respond to data requests by government officials, and to assist business partners.

TAM was built by testing word processor use and TTF was built by testing different applications in industry in 1990s. The technology has advanced significantly, and contemporary technology BI use will be different from the TAM and TTF era. Besides, TAM is a utilization focus model, while TTF is a fit focus model. A utilization focus model has two major limitations: First of all, system use is not always voluntary. System use was more about how a job was designed than the usefulness of the systems or the attitude of the user (Goodhue and Thompson 1995). Secondly, more utilization does not necessary lead to better performance. Increasing use of a poor system may even produce negative impacts on organizations. TAM and TTF have significant overlapping constructs. The dependent constructs of both models are related to the actual use of IT, and the aims of both models concern understanding users' choices and evaluation of information systems.

Many users conveyed that it was not important how useful BI systems were but whether BI systems could support completion of the decision tasks. Goodhue and Thompson (1995) proposed the task technology fit model to explain how task performance and technology performance interact with each other, and the individual's role in using technology to perform tasks. More importantly, the expression "decision task—BI system alignment" was repeatedly mentioned by interviewees at both CIC and AG, even though the interviewer had not explicitly mentioned the phrase or the concept of alignment during any interview session. The emergent construct, Decision task—BI system alignment (TSA), refers to the degree to which a system assists an individual in performing his or her portfolio of tasks. This definition is adapted from Goodhue and Thompson's (1995) task technology fit idea. Therefore, replacing PU with TSA offers a more comprehensive way to explore, explain, and predict UB in Chinese organizations. Importantly, using TSA, as a construct, overcomes the limitations of utilization-focused and fit-focused models in this project.

Proposition 3. BI system characteristics affect level of perceived ease of use. Perceived BI system ease of use (PEOU) refers to the degree to which the actual user experiences the BI system to be free of effort, during the process of gathering system requirements, developing and maintaining, learning and training, using, and understanding and communicating. BI system development was not the focus of this research and a simpler conception of PEOU emerged. Many users discussed BI ease of use in terms of learning, training, understanding, and communicating. This is due to the intuitive nature of the BI systems interfaces, as well as data and system quality. Hence, the case studies indicate that NS affects levels of PEOU.

Proposition 4. Use behavior is affected by decision task—BI system alignment and perceived ease of use. BI system use behavior (UB) refers to the actual hands-on use of a BI system. Interview data helped to tease out the UB construct in terms of user types, usage, and use satisfaction. Most users fell into the category of direct users, that is, they had direct access to the BI systems. Assisted users are often senior managers in the organization, and they do not personally access BI systems. The data analysis required and inadequate technical skills were often the reasons for receiving assistance in using BI systems. There were much fewer assisted users than direct users at both organizations, but all assisted users could have direct access to BI systems if they wished.

For semi-structured and tactical decisions, users did not feel that they were adequately supported by BI. Users encountered technical as well as business operation issues. When users met technical issues, they contacted the person who is in charge of managing a particular BI system. When users met business operation issues, they would communicate with BI team members in order to find data that would support solving those problems. Other use issues include inconsistent data feeds among different systems and departments, different BI systems had repeated functions, some decisions required more advanced analytic support, and users required better performance from the systems. All of these use issues were caused by lower levels of TSA or lower levels of PEOU.

Usage was measured by the frequency and duration of sessions with the BI systems. All CIC interviewees, except five developers, reported daily usage of BI systems. Most AG interviewees reported daily use. Three AG managers revealed less use frequency; the nature of their tasks did not require a significant amount of data. Spending additional time using BI systems did not necessary lead to higher quality information being discovered or sourced. Finally, most CIC and AG users appraised medium to high satisfaction of BI systems use, in terms of sufficient data feeds, adequately formatted reports, improved decision logic, and overall satisfaction.

Proposition 5. Factor X, a composite of trust, closeness, experience, and generations, affects decision task—BI system alignment and perceived ease of use. Higher levels of FX lead to higher decision task—BI system alignment and perceived ease of use. A number of concepts relating to Chinese CN emerged from the analysis of the case study data and the literature review. This group of concepts is best conceived as a multi-attribute construct. There is no obvious name for this factor and to avoid biasing the analysis it was given a provisional title of "Factor X" or FX. This is a similar neutral naming approach to the System 1/System 2 terminology of cognitive systems in behavioral economics (Kahneman 2011).

Guanxi may influence users' decisions to use BI systems and affect their system use patterns. Social cognitive theory (SCT) suggests that future behaviors are shaped by past behavior and beliefs about ability and the environment. Applying SCT to the post-adoption use of IS, users' beliefs are more likely to be shaped by repeatedly presented opportunity and the outcomes of using IS (Craig et al. 2010). Unlike guanxi in business negotiations, guanxi inside an organization is bounded by hierarchical positions and tasks. These activities, such as collaboration, often initiate, maintain, and utilize a guanxi dyad. Chinese managers make decisions according to specific circumstances that are determined by the people involved, the occasion of the event, and place where the event takes place (Fu et al. 2006).

Relationships are based on the interaction between guanxi hu (two individuals in one guanxi dyad), and trust is one of the common measures of guanxi quality. For instance, people in general perceive a lower level of trust when they work for different organizations. For example, an AG manager assisted external researchers to conduct research by providing them with data from AG. However, he expressed that many researchers requested sensitive data that he cannot offer due to security and ethical concerns. In this case TSA is low because one AG manager had lower trust with external researchers although NT could be supported by NS. Another AG manager reported a high level of trust with his subordinates when acquiring extra resources. This manager requested extra technical and analysis support from the BI team to assist his subordinates' use of the BI system. This means that PEOU might be low for this manager's subordinate but trust helped to increase PEOU although NS was constant. Trust is therefore important in investigating BI systems use.

The decision tasks in the case studies usually required at least two users to work in a collaborative manner. This is not a common use pattern in Western organizations. Closeness is a concept that was introduced by multiple interviewees who were responding to follow-up questions related to xinren (trust) in describing their relationships with their colleagues. For example, an AG analyst's superior was in charge of all data products in an AG BU. Due to a shortage of resources, his superior had to prioritize development tasks. The analyst asked his superior for help when he needed his superior to promote a particular development task. The analyst declared that his supervisor trusts him with development decisions, because they have worked closely together for some years.

Often, subordinates reported their obedience to their supervisors. However many interviewees believe this obedience does not lead to a closer relationship or a higher level of trust. Guanxi closeness comprises two components, trust and feelings, where trust is more cognitive based and feeling is more affect based (Chen and Peng 2008). A higher level of trust may motivate a closer relationship, while the degree of closeness may impact on the level of trust. One AG analyst described that he was close to his immediate supervisor, so he trusted that his supervisor would provide additional resources when required. In this research closeness refers to the distance of relationship a person feels between them and another. Therefore, closeness is important in understanding and assessing the level of trust, and as a result has been added to FX.

None of the TAM or UTAUT articles explicitly define experience. Presumably experience means the users' experience with using a particular technology. A decision maker is born with a natural endowment, and through dint of practice, learning, and experience they develop their endowment into a mature skill (Simon 1960). This means that, to some extent, decision-making skills can be learned through education and work experience. The majority of employees from CIC and AG possessed bachelor degrees. AG had more employees with graduate degrees (two with PhDs), while CIC had more employees with diploma education and below. Some employees had studied graduate degrees part-time while working. From the case analysis, the level of qualification did not explicitly affect their competence in completing assigned decision tasks. However, the fields of their degree major and their work experience did lead to different ways of looking at, and analyzing, data. For example, some managers only considered BI as a data extraction tool, while others believed that BI systems were essential in assisting and supporting their decision-making process. In this research, experience consists of education and work experience. Both dimensions of experience are critical to TSA and PEOU, and therefore UB.

The different generations in China have received considerably different education. None of the interviewees from CIC and AG were Traditionalists or Boomers. Generation X was the first generation to receive high quality education after the Chinese education system recovered from the Cultural Revolution. Generation Y was born when China opened trade with international companies in the 1980s. Modern business IT was introduced around the same time. Generation Y was able to be educated and adopt innovative technology whereas Generation X did not have the same opportunities. Both CIC and AG have more Generation Y than Generation

X employees; 69% of CIC interviewees and 86% of AG interviewees were from Generation Y. The nature of AG as an Internet company means that more recent graduates are hired than at CIC. More Generation Y employees have studied IT than Generation X employees. In general, Generation Y reported higher TSA and PEOU than Generation X.

Factor X is summarized in Eq. 7.1.

$$FX = f\left(trust, closeness, experience, generation\right) \tag{7.1}$$

7.5 Working Conclusion

In recent years, the Chinese economy has grown to be one of the largest and most influential in the world. BI systems contribute to overall organization performance and may lead to competitive advantage. However, no indigenous Chinese organization has been investigated regarding BI systems use nor has any study considered CN and BI use. This project attempts to fill this research gap.

The contribution to knowledge of the project is a set of five research propositions based on empirical research of BI systems use in Chinese organizations. These propositions introduce guanxi into BI use theories. This set of propositions is a foundation for future BI research in China. For practitioners, these propositions contribute to the understanding of Chinese management and decision support. This improved understanding may help achieve more effective use of large-scale DSS, and in turn lead to higher organization performance.

This research is subject to the normal limitations of exploratory research. The BI use propositions were based on two case studies, and the selected organizations operated in quite specific industrial environments. It is not possible to generalize the research outcomes to all Chinese organizations. Another common concern raised about case study research is the lack of objective interviews (Eisenhardt 1989). This project adopted the most rigorous possible case study methods and techniques.

Factor X is a complex attribute construct. It is an umbrella construct that represents a group of important concepts that have influence from, or on, guanxi. Analysis suggested there are impacts from this group of concepts on TSA and PEOU. In addition, these concepts have effects on each other. For example, the level of closeness may be interrelated with the level of trust. The case study data also suggested that there may be a complex interaction between two groups of attributes (generation and experience, and closeness and trust). However, these complex interactions remain ambiguous.

The next stage of this research project will have two aims. The first is to investigate the detailed nature of FX. The second aim is to formulate a model of BI use in Chinese organizations. It is planned to return to CIC and AG to conduct further investigations.

Biographies

Yutong Song (Yuri.Song@monash.edu) is currently a doctoral student at Monash University, Australia. She graduated with first class honors in the Master of Business Information Systems from the same university in 2012. Her research interests are the development and use of business intelligence systems, and the cross cultural use of IT. She has published in *Decision Support Systems*.

David Arnott (David.Arnott@monash.edu) is Emeritus Professor of Information Systems at Monash University, Melbourne, Australia. His research areas are behavioral economics, personal decision support systems, and business intelligence. He is the author of many papers in the decision support area, including papers in the *European Journal of Information Systems*, *Information Systems Journal, Journal of the Association of Information Systems, Decision Support Systems*, and the *Journal of Information Technology*. A pioneer in management support education, he first taught graduate courses in the area in 1980. He is a Fellow of the Australian Computer Society and Senior Editor for the journal *Decision Support Systems*.

Shijia (Caddie) Gao (Caddie.Gao@monash.edu) is a lecturer at the Faculty of Information Technology, Monash University. She received her Ph.D. degree in business information systems from the University of Queensland Business School. Her research interests include business intelligence, decision support systems, decision theory, risk management, financial information systems, business process management, and knowledge management. Her research has appeared in *Decision Support Systems*, *Journal of Knowledge Management*, *Journal of Decision Systems*, *Journal of Database Management*, *Expert Systems with Applications*, among others.

References

Anthony RN (1965) Planning and control systems: a framework for analysis. Harvard University, Boston, MA

Bhattacherjee A, Lin CP (2015) A unified model of IT continuance: three complementary perspectives and crossover effects. Eur J Inf Syst 24(4):364–373

Brislin RW (1970) Back-translation for cross-cultural research. J Cross-Cult Psychol 1(3):185–216

Chen XP, Peng S (2008) Guanxi dynamics: shifts in the closeness of ties between Chinese coworkers. Manag Organ Rev 4(1):63–80

Craig K, Tams S, Clay P, & Thatcher J (2010) Integrating trust in technology and computer self-efficacy within the post-adoption context: an empirical examination, In: Proceedings of Americas Conference on Information Systems

Davis ELE (2005) Encyclopedia of contemporary Chinese culture. Routledge, New York, NY

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340

Davis FD, Bagozzi RP, Warshaw PR (1989) User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 35(8):982–1003

Davison RM, Martinsons MG (2016) Context is King! considering particularism in research design and reporting. J Inform Tech 31(3):241–249

Davison RM, Ou CXJ, Martinsons MG (2013) Information technology to support informal knowledge sharing. Inf Syst J 23(1):89–109

Deng X, Chi L (2012) Understanding postadoptive behaviors in information systems use: a longitudinal analysis of system use problems in the business intelligence context. J Manag Inf Syst 29(3):291–326

Eisenhardt KM (1989) Building theories from case study research. Acad Manag Rev 14(4):532–550

Erickson T (2009) Generations in China. http://blogs.hbr.org/2009/03/generations-in-china/

Fock HK, Woo K (1998) The China market: strategic implications of Guanxi. Bus Strateg Rev 9(3):33–43

Fu PP, Tsui AS, Dess GG (2006) The dynamics of Guanxi in Chinese hightech firms: implications for knowledge management and decision making. Manag Int Rev 46(3):277–305

Gartner (2015) Flipping to digital leadership. http://www.gartner.com/imagesrv/cio/pdf/cio_agenda_insights2015.pdf

Gogindarajan V, Trimble C (2012) Reverse innovation: create far from home, win everywhere. Harvard Business Review Press, Boston, MA

Goodhue DL, Thompson RL (1995) Task-technology fit and individual performance. MIS Q 19(2):213–236

Gorry GA, Scott-Morton MS (1971) A framework for management information systems. Massachusetts Institute of Technology, Cambridge, MA

Gu FF, Hung K, Tse DK (2008) When does Guanxi matter? Issues of capitalization and its dark sides. J Mark 72(4):12–28

Hofstede G (1980) Culture's consequences: international differences in work-related values. Sage, Newbury Park, CA

Jones PS, Lee JW, Phillips LR, Zhang XE, Jaceldo KB (2001) An adaptation of Brislin's translation model for cross-cultural research. Nurs Res 50(5):300–304

Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York, NY

Karahanna E, Straub DW, Chervany NL (1999) Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Q 23(2):183–213

Leidner DE, Kayworth T (2006) A review of culture in information systems research: toward a theory of information technology culture conflict. MIS Q 30(2):357–399

Li X, Hsieh JJPA, Rai A (2013) Motivational differences across post-acceptance information usage behaviors: an investigation in the business intelligence systems context. Inf Syst J 24(3):659–682

Liu C, Forsythe S (2011) Examining drivers of online purchase intensity: moderating role of adoption duration IIN sustaining post-adoption online shopping. J Retail Consum Serv 18(1):101–109

Liu H, Ke W, Wei KK, Hua Z (2013) The impact of IT capabilities on firm performance: the mediating roles of absorptive capacity and supply chain agility. Decis Support Syst 54(3):1452–1462

Lowry PB, Zhang D, Zhou L, Fu X (2010) Effects of culture, social presence, and group composition on trust in technology-supported decision-making groups. Inf Syst J 20(3):297–315

Mao E, Palvia P (2008) Exploring the effects of direct experience on IT use: an organizational field study. Inf Manag 45(4):249–256

Martinsons MG, Westwood RI (1997) Management information systems in the Chinese business culture: an explanatory theory. Inf Manag 32(5):215–228

Miles MB, Huberman AM, Saldaña J (2013) Qualitative data analysis: a methods sourcebook, 3rd edn. Sage, Thousand Oaks, CA

Millington A, Eberhardt M, Wilkinson B (2005) Gift giving, Guanxi and illicit payments in buyer–supplier relations in China: analysing the experience of UK companies. J Bus Ethics 57(3):255–268

Shibl R, Lawley M, Debuse J (2012) Factors influencing decision support system acceptance. Decis Support Syst 54(2):953–961

Shin SK, Ishman M, Sanders LG (2007) An empirical investigation of socio-cultural factors of information sharing in China. Inf Manag 44(2):165–174

Simon HA (1960) The new science of management decision. Harper & Row, New York, NY

Sousa VD, Rojjanasrirat W (2011) Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract 17(2):268–274

Taylor S, Todd PA (1995) Assessing IT usage: the role of prior experience. MIS Q 19(4):561–570

Venkatesh V, Morris MG, Gordon BD, Davis FD (2003) User acceptance of information technology: toward a unified view. MIS Q 27(3):425–478

Venkatesh V, Thong JYL, Xu W (2012) Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q 36(1):157–178

Wang Y, Wang S, Fang Y, Chau PYK (2013) Store survival in online marketplace: an empirical investigation. Decis Support Syst 56(12):482–493

Yin RK (2014) Case study research: design and method, 5th edn. Sage, Los Angeles, CA

Zhang DS, Lowry P, Zhou L, Fu X (2007) The impact of individualism-collectivism, social presence, and group diversity on group decision making under majority influence. J Manag Inf Sys 23(4):53–80

Zhou T (2011) Understanding mobile internet continuance usage from the perspectives of UTAUT and flow. Inf Dev 27(3):207–218

Chapter 8
The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps

Zhilei Qiao, G. Alan Wang, Mi Zhou, and Weiguo Fan

Abstract Product innovation is important for firms to gain competitive advantages in a dynamic business environment. Traditionally, customers are not very much involved in product innovation processes. With the technology of Web 2.0, online users are enabled and motivated to provide reviews and discussions about product features and use experiences. User generated product reviews have been found to have a word-of-mouth effect as a new element of marketing communication. However, their implication on improving product innovation cycles have not been studied before. Guided by a persuasion theory, we extracted the central and peripheral persuasion cues from user generated reviews and examined their impact on mobile app developers' product innovation decisions. Using data collected from the Google App store, our empirical study shows that long and easy-to-read user reviews with mildly negative reviews can increase the likelihood of a future mobile app update. Our findings highlight the need for researchers to explore user generated reviews in the context of customer-centered product innovation.

Keywords Product innovation • Text mining • Mobile apps • Survival analysis • Persuasion

8.1 Introduction

A product innovation strategy is critical for firms to survive and prosper in a dynamic business environment (Alegre and Chiva 2008; Holahan et al. 2014; Wales et al. 2013). Dunk (2011) defines product innovation as an innovation process that

Z. Qiao (✉) • G.A. Wang
Department of Business Information Technology, Pamplin College of Business,
Virginia Tech 1007, Blacksburg, VA 24061, USA
e-mail: qzhilei@vt.edu

M. Zhou • W. Fan
Department of Accounting and Information Systems, Pamplin College of Business,
Virginia Tech 1007, Blacksburg, VA 24061, USA

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_8

conceives new and better products, which are unique or different in some ways from existing products (Nakata and Sivakumar 1996). The ability of firms to develop innovative products is key to their competitive advantages (Cankurtaran et al. 2013; Jayaram et al. 2014). Evidence suggests that product innovation can assist firms to enter an emerging industry and strengthen their competitiveness in the corresponding market (Keupp et al. 2012; Kotabe et al. 2011). Therefore, product innovation is critically important to a firm's performance (Prajogo and Ahmed 2006; Yao et al. 2013).

Existing innovation literature suggests several different channels of information acquisition for the product innovation process (von Hippel 1998; O'Hern and Rindfleisch 2008; Ramaswamy and Prahalad 2004). The traditional perspective suggests that firms dominate the product innovation decisions (Porter 1980). It views product innovation as a firm-centric activity, with most information flowing one way from the firm to its customers (Ramaswamy and Prahalad 2004). While customers are considered as the passive recipients of product innovation, firms have very limited understanding of customers' perception and opinions before the release of a new product. Firms only target the "right" customers and cannot accurately capture their customers' needs. Recent studies show that customers have been more involved in product co-creation processes. For example, the 3M company took advantage of identifying lead users before creating breakthrough products in order to avoid a market decline (Von Hippel et al. 1999). In addition, Cohen et al. (2002) show evidence that customers can offer useful ideas for new R&D projects and contribute substantially to the improvement of existing R&D projects. However, firms, still taking a dominant position in the innovation process, present their ideas to customers and gather customers' needs and feedback from only a small fraction of customers (Di Gangi and Wasko 2009; von Hippel and Katz 2002). Firms tend to be biased towards listening to their current customers, and even among these, to their most important customers or those who speak the most (Sawhney et al. 2005).

Recent literature shows that the product innovation process is shifting from a firm-centric view to customer-driven perspective. While customers are considered as the passive recipients, firms have very limited understanding of customers' perception and opinions before the release of a new product. Due to the limitations of developing new knowledge internally, integrating and using external knowledge is critical. While most existing literature searches external knowledge from other companies and alliances and finds evidence that sourcing this knowledge is beneficial to the firm's product innovation decisions, some other scholars identify customers as the most important source of information for new product development. Also, customer oriented innovation literature illuminates why and how external knowledge is significant and potentially valuable. O'Hern and Rindfleisch (2008) consider users as being central and vital participants in the product innovation process. Particularly, with the fast development and proliferation of online customer review communities, customers today willingly contribute and share their thoughts and opinions online. Zhang et al. (2013) show that innovative users share common interests and ideas in online communities. Since product innovation aims to provide higher quality products and give higher benefit to users. Therefore, customer-driven product innovation, enabled by the Web 2.0 technologies, is getting more and more attention from researchers and practitioners.

Online customer reviews have become an important new channel to acquire customers' feedback about product features and potential product defects (Abrahams et al. 2013; Lee and Seo 2013; Mudambi and Schuff 2010). Existing studies show that online customer reviews have a significant impact on other customers' adoption decision and firms' sales performance due to the word-of-mouth (WOM) effect (Chen and Xie 2008; Duan et al. 2008; Ghose and Ipeirotis 2011). In addition to being useful for customers and marketing purposes, online customer reviews also contain useful information for product development and improvement. Jin et al. (2015) find that online customer reviews are an important information source for collecting customer feedback and new requirements for product developers or designers. Some researchers have developed text analysis methods in order to extract and measure aggregated customers' preferences and feedback on product features (Decker and Trusov 2010; Xiao et al. 2015). However, to the best of our knowledge, we have not found any literature that show empirical evidence about the impact of online customer reviews on product innovation decisions. Our study aims to find empirical evidence that online product reviews can affect product developers' innovation decisions.

The mobile app industry provides a perfect research context for our research question. First, unlike physical products or enterprise software products, mobile apps are updated much more frequently (Syer et al. 2013). It provides more observation instances for product innovation activities than traditionally developed products that usually have a much longer development cycle. Second, mobile apps have a large user base through which a large number of user reviews have been generated through online app stores. Our research can potentially help improve the communication between app users and developers. Knowing the impact of user reviews on developers, the app users will be more motivated to contribute reviews. App developers can quickly identify those reviews that have high impact on their innovation decisions. Our study can also benefit app store providers such as the Google Play Store by making a better ecosystem in support of product innovation for mobile apps.

This study contributes to literature in several ways. First, our research enriches existing customer-driven product innovation literature. Prior studies suggest that firms have to design the right toolkits in order to get users involved in the product innovation process. Our study shows that product developers can learn from the widely available online customer reviews without developing specialized tools. Specifically, we reveal how online customer reviews can affect the product innovation cycles. By analyzing online customer reviews, product developers can learn customer feedback and feature requests in their complete view compared to traditional ways of collecting customer feedback such as surveys. The second contribution of our study is to complement the online customer reviews literature, which mainly show the impact of online customer reviews on the perception and purchasing decisions of future customers (Chen and Xie 2008; Duan et al. 2008; Ghose and Ipeirotis 2011). Our study will be the first to empirically show the impact of online customer reviews on product developers and designers. The third major contribution lies in a deeper understanding in how innovation works in the emerging mobile apps industry. According to the Silvias (2014), the global mobile app market will reach $187 billion in 2017. Examining the innovation activities in this emerging industry will be economically significant.

From a managerial perspective, our study underscores the business strategy value of online customer reviews that executives struggle to quantify. Our results indicate that investment made on analyzing online customer reviews would pay off over time in terms of better product quality and a higher customer retention rate. In addition, based on marketing literature, it is important to understand customers' needs so that product managers can allocate resources to more productive and promising product innovation activities.

The rest of the chapter is organized as follows. We first review the Elaboration Likelihood Model, a persuasion theory that we use to guide our research design. We then develop our research hypotheses followed by our empirical study. We provide conclusions, discussions, and future directions at the end.

8.2 A Persuasion Theory—Elaboration Likelihood Model

Existing literature indicates that online customer reviews provide important external knowledge for product developers to identify new user requirements, detect product defects, and incorporate user solutions (Abrahams et al. 2013; Lee and Seo 2013; Mudambi and Schuff 2010). Therefore, online customer reviews have not only a word-of-mouth (WOM) effect for fellow customers, but also an implicit persuasion effect on product designers and developers.

The Elaboration Likelihood Model (ELM) has been commonly used to explain how a message can possibly change the perception of the message recipient. The theory suggests that a message recipient has a continuum of elaboration methods to deal with persuasive messages (Tam and Ho 2005). The essence of elaboration processing goes beyond simply focusing on comprehending the arguments embedded in the text content of the received message. When a message recipient does not have the motivation or ability to read and understand the arguments in a received message, persuasion is made through the peripheral route rather than the central route or argument quality, according to the ELM model. However, in most cases, both central and peripheral routes work collectively in persuading message recipients' decisions.

The central route of persuasion requires a message recipient to carefully scrutinize the arguments in a received message, thus the recipient's cognitive efforts on argument processing determines its influence (Zhang 1996). Existing studies have found that argument quality, such as information completeness and accuracy, has a significant impact on the message recipient's perception on information usefulness and willingness to adopt the message (Sussman and Siegal 2003).

The peripheral route relies on simple cues that are content-irrelevant indicators reflecting a recipient's perception of the credibility of the message source (Chaiken 1980). ELM researchers find that source credibility becomes an important predictor of the recipient's attitude change especially when the recipient cannot comprehend the arguments embedded in the received message (Petty et al. 1981). When a message recipient cannot or is not willing to scrutinize the message arguments, he or she will access the expertise, knowledgeability, reliability, and trustworthiness of the message source (Wu and Shaffer 1987). In addition, mood reflected from message content is also considered as a peripheral route that can affect message recipients' decisions (Batra and Stayman 1990; Payne-James and Khawaja 1993).

In our research context, we consider mobile app reviews generated by app users as persuasion messages, with which app users try to influence app developers in their product design and improvement. App users may provide arguments promoting certain features or demoting features to be improved or abandoned. App developers, as message recipients, are likely to scrutinize the arguments in each review and assess the peripheral cues about the reviewer in order to prioritize feature requests and make product release decisions. Guided by the ELM theory, we develop our research hypotheses about the impact of central and peripheral routes on app developers' product innovation decisions.

8.3 Research Hypotheses

Our research model, as depicted in Fig. 8.1, illustrates how we hypothesize central route and peripheral route would affect app developers' product innovation decisions. The central route constructs include the amount of information and readability, reflecting the argument quality of each app review. Both have been used to assess the argument quality of text messages (Zhou et al. 2015). The peripheral route constructs include review sentiment and sentiment strength. They reflect nothing about the arguments but the general mood of the reviewer. Classic peripheral cues such as expertise and trustworthiness do not apply in our research context

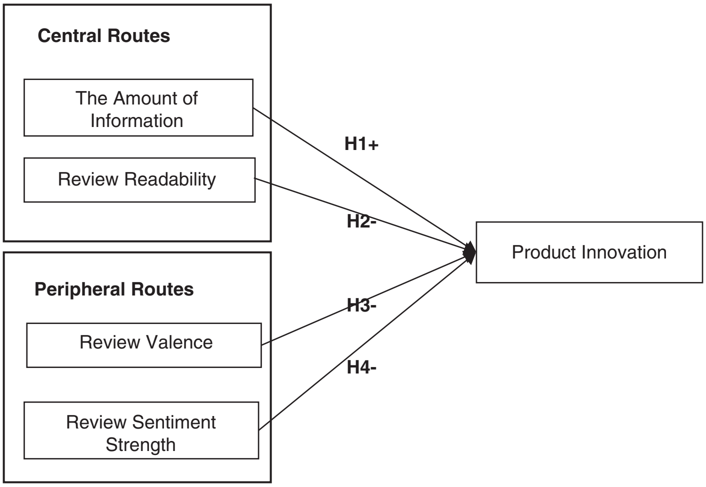


Fig. 8.1 Research model

because many user reviews are anonymous. As we discussed earlier, positive sentiment or mood conveyed through messages may positively influence the attitude of message recipients (Petty et al. 1993), i.e., the app developers in our context. In the rest of this section, we describe our research hypotheses in details.

8.3.1 The Amount of Information

The amount of information directly influences the capability of the message recipient to scrutinize the arguments embedded in a message. When the amount of information in a message is low, the recipient has few opportunities to elaborate because the motivation to elaborate is low (Palmer and Griffith 1998). Previous literature shows that the amount of information in product reviews, which is measured by the review length, has a positive influence on customers' adoption decisions due to the WOM effect (Chen and Turut 2013; Duan et al. 2008). Mudambi and Schuff (2010) also suggest that long reviews are perceived as being more useful than short ones because readers consider the word count as the depth of information usefulness and comprehensiveness.

A greater amount of information in app reviews presumably has more value to app developers. Lee (2007) suggests that customer reviews reflect customer needs. Some studies show that customer reviews contain critique about existing product features and suggestions about new product features (Mudambi and Schuff 2010; Troy et al. 2001). Customer feedback and suggestions on product features can help app developers reduce the uncertainty in customers' perception on new products, increase their confidence in product innovation decisions, and increase the frequency of product innovation (Dougherty and Dunne 2011; Zirger and Maidique 1990). Therefore, we propose:

Hypothesis 1: Mobile apps that have received a higher amount of information in its user reviews are more likely to have a new product release.

8.3.2 Review Readability

In addition to the amount of information, the persuasion effect of argument quality is also related to the willingness of message recipients exerting their mental efforts. Obscure words will inhibit the message recipient's willingness to make sense of the message content. Readability measures the effort it takes for a message recipient to comprehend a text message. It relates to the linguistic complexity of the text, in particular to the semantic and syntactic dimensions of the text. Text readability affects the message recipient's ability to cognitively process the arguments in a message (Lehavy et al. 2011). Bloomfield (2002) shows that less readable text requires investors to devote more time and effort to identify and extract relevant information. By contrast, easy-to-read text improves the message recipient's reading speed, comprehension, and memory retention (Ghose and Ipeirotis 2011). Existing studies have shown that the readability of product reviews can be used to predict the usefulness and impact of the reviews (Ghose and Ipeirotis 2011). Similarly, we propose:

Hypothesis 2: *Mobile apps that have received reviews with higher readability are more likely to have a new product release.*

8.3.3 Review Sentiment

Review sentiment, that includes both sentiment valence and extremity, reflects the subjectivity of the reviewers. Although it is derived from the message content, sentiment is often considered as a peripheral cue because it reveals the mood or affection status of the author (Bardzil and Rosenberger 1996). Past research shows that review sentiment can affect the perceived value or usefulness of product reviews. For example, Schindler and Bickart (2012) find that a moderate proportion of positive evaluative statements in product reviews positively relates to consumers' perceived helpfulness. Sen and Lerman (2007) find that review readers are more likely to consider negative opinions as being helpful for utilitarian products. Cheung et al. (2012) concludes that fair reviews are perceived more favorably when they cover both positive and negative aspects of the reviewed product. Existing studies do not provide consistent conclusions for the impact of review sentiment because of the moderation effects of product category and different message recipients. In our research, we aim to study the effect of review sentiment on the perceived usefulness of product reviews for app developers, not for fellow consumers. We use the SentiStrength method proposed by Thelwall et al. (2010) to automatically identify and classify the emotional information of customers. SentiStrength estimates the strength of positive and negative sentiment in informal short text messages using sentiment word dictionaries. We consider negative app reviews to be the major concerns for app developers due to the negative word-of-mouth (nWOM) effect, which has shown to have both short-term and long-term effects on firms' financial performance (Luo 2009). Moreover, Schindler and Bickart (2012) find that a product review with more descriptive statements is considered as being more helpful. Reviews with extreme sentiment do not have increased value and may decrease the readers' perceptions about its helpfulness. According to review sentiment strength in the text, this will make reviews lean toward neutral polarity. Therefore, we hypothesize the following:

Hypothesis 3: *Mobile apps that show negative sentiment in their reviews are more likely to have a new product release.*

Hypothesis 4: Mobile apps with a lower review sentiment strength are more likely to have a new product release.

8.4 Research Methodology

8.4.1 The Stratified Cox Proportional Hazard Model

Our primary interest is to study the impact of app user reviews on product innovation, i.e., the probability of having a new product release. More specifically, we would like to know if app reviews might shorten the time to the next product release. If we define an app update to be an event, we can use survival analysis to model this "time to event" data. On the other hand, survival analysis is capable of incorporating time-independent explanatory variables, which fits our scenario since our hypothesized explanatory variables are time invariant. Moreover, mobile apps usually have several updates over time. Therefore, events are recurrent and event time order matters (i.e., for each mobile app, an update at time t1 is different from that at time t2). We choose the Stratified Cox Proportional Hazard (SCPH) model (Cox 1972) for our empirical analysis, which makes no assumption about the form of the baseline hazard function. The SCPH model does not depend on distributional assumptions of survival time and defines the hazard ratio as the relative risk based on comparison of event rates. Thus, we employ the SCPH model to examine the relative association between the effects of independent variables (i.e., amount of information, review readability and review sentiment) and a subsequent product release event.

The hazard function, $h(t)$, represents the occurrence rate of a product per unit time (t). We use T to denote the time to event. The hazard function has the following form:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t} \tag{8.1}$$

The SCPH model assumes that the elapsed time to event T is conditional on the independent variables ($X_1, X_2, \ldots, X_j$). In our study, T measures the time between the product launch date or the previous product update date and the date of the event of interest—a new product release—or the end of the observation period. Thus, our hazard ratio represents the "risk" of having a new product release within a time unit (where the time is measured in days). The SCPH model is expressed as:

$$h_g(t, X) = h_{0g}(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j} \tag{8.2}$$

where $\beta_1, \beta_2, \cdots, \beta_j$ is a vector of regression parameters to be estimated. The baseline hazard function $h_{0g}(t)$ corresponds to the case where $x_j = 0$, involving time but not independent variables at each stratum with $g = 1, \cdots, k^*$. The second component is the exponential functions with the sum of $\beta_j X_j$, which involves independent variables' effects but not time.

$$\frac{h_g\left(t, X_i + 1, |X_{j(J \neq i)}\right)}{h_g\left(t, X_i, |X_{j(J \neq i)}\right)} = e^{\beta_i} \tag{8.3}$$

8.4.2 Data

As of Q1 of 2015, Google Play store was the largest mobile app provider in terms of number of app downloads, surpassing the first mobile app store-the Apple App store. We have collected data for 1215 mobile apps from the Google Play store between April 8, 2014 and April 5, 2015. All apps have appeared at least once in the top charts of new game apps. Because our study is focused on text analysis, we dropped 63 apps that did not receive any user review or have empty reviews during the data collection period. The final dataset contains 281,202 customer reviews for 1152 mobile apps. Data collected include basic app attributes, the business model (free or paid apps), app release/update dates, user ratings, user reviews, and developer information. Table 8.1 shows the basic summary statistics for collected data.

8.4.3 Variables

The dependent variable. We retrieved and analyzed the reviews that app users posted for each mobile app before the app's next update. We used the variable *Hazard Rate* to represent whether the app update event happened and how long the update interval (i.e., time between the previous update or the initial release date and its next update date). If a mobile app has not been updated, *Hazard Rate* represents the (instantaneous) rate of update for the apps to some time point during the next instant of time. Some mobile apps were not updated at all during the observation period. Therefore, the right censoring problem occurs in our data set. To solve the problem, we used an *event* variable to indicate whether an observation is censored (i.e., event is 1 for a complete observation and 0 for a censored one).

Table 8.1 The summary statistics of collected data

Measure	Value
Time period	April 8, 2014–April 5, 2015
Number of mobile apps	1152 (171 paid apps; 981 free apps)
Number of app reviews	281,202
Number of app updates	3307

Independent variables. *Amount of Information.* As suggested by Mudambi and Schuff (2010), we used review length to measure the amount of information in user reviews. We calculated the average review length and the total number of words for the reviews that we collected during each app update cycle.

Review Readability. The Fog index is commonly used to computationally measure text readability (Li 2008). It estimates the number of years of formal education that a reader of average intelligence would need to understand the text. It is built on the premise that complex words and long sentences are difficult to understand. A word is considered as a complex word if it contains three or more syllables. The larger the Fog index, the more difficult it is to understand the text. The Fog index is calculated as follows:

$$Fog = \left(Words\_Per\_Sentence + Percent\_of\_complex\_words\right) \times 0.4 \qquad (8.4)$$

Review Sentiment Strength. SentiStrength is a lexicon-based classifier that uses supplementary (non-lexical) linguistic information and rules to identify sentiment strength in short informal English text, which is perfect for analyzing text in user generated app reviews. For each text message, SentiStrength generates two integer values ranging from 1 to 5, one being the positive sentiment strength and the other the negative sentiment strength (Thelwall et al. 2010). The average review sentiment strength or extremity was calculated over all the reviews collected for each app update cycle.

Review Valence. Review valence indicates whether a positive or negative sentiment stands out in a text message. We used the difference between positive and negative sentiment strength values calculated by SentiStrength to measure review valence. If the difference is less than 0, the review valence is negative. When it is greater than 0, it is positive. The average review sentiment valence was calculated over all the reviews collected for each app update cycle.

Control Variables. We considered the mobile app business model (free 0 or paid 1) and competition intensity (strong or weak) to be important control variables in our study. Existing literature find that both can affect product innovation decisions (Goettler and Gordon 2011; Greenstein and Ramey 1998; Stewart and Zhao 2000). Casual, arcade, puzzle, and action game categories have the strongest competition among all game categories because each of those game app categories has at least 100 games.

8.4.4 Results

8.4.4.1 Descriptive Statistics

Table 8.2 shows the descriptive statistics and pairwise correlations of our variables. In our dataset, 97% mobile apps had released at least one update during the observation period. For those mobile apps where at least one update occurred, the average

Table 8.2 Descriptive statistics

	Min.	Max	Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Avg. review length	0.00	184.00	7.46	10.03	–	–	–	–	–	–	–
(2) Total no. of words	1	62,136	53.67	21.76	0.1	–	–	–	–	–	–
(3) Review readability	0.00	40.4	4.25	4.73	0.08	0.24	–	–	–	–	–
(4) Review sentiment strength	−3.00	4.00	0.63	0.63	0.04	0.27	0.38	–	–	–	–
(5) Review valence	0.00	4.00	0.48	0.76	0.02	0.15	0.27	0.68	–	–	–
(6) Business model	0.00	1.00	0.14	0.35	−0.08	0.11	−0.04	0.05	0.05	–	–
(7) Competition intensity	0.00	1.00	0.58	0.49	−0.02	−0.07	0.01	−0.04	−0.06	−0.14	–

Time to Update is 39.3 days. Fifty percent mobile app update cycles received user reviews before an update occurred. All pairwise correlations between independent variables are below 0.5 except the correlation between review valence and review sentiment strength. We also checked the variance inflation factor (VIF) values for all independent variables in our model. The result indicated that multicollinearity was not a concern (Zhang et al. 2013).

8.4.4.2 Hypotheses Testing Results

Table 8.3 presents the estimates of our research model. Review length was found to positively influence the likelihood that a future mobile app update would occur (β = 0.0083, $p < 0.01$). This suggests that mobile apps receiving longer user reviews on average are more likely to receive a new update. However, we found that the total number of words was negatively related to the possibility of having a future mobile app update. The hypothesis 1 is only partially supported. We need to conduct further analysis on this hypothesis in the future.

The Fog index, used to indicate review readability, was found to be negatively associated with a future app update ($\beta = -0.038$, $p < 0.01$). Mobile apps that receive user reviews with a high Fog index (i.e., more difficult to read) are less likely to receive a new update. The observation supports Hypothesis 2.

As predicted by Hypothesis 3, mobile apps that had received positive user reviews were less likely to receive a new update ($\beta = -0.10$, $p < 0.1$). Mobile apps that have received user reviews with extreme sentiment were less likely to receive a future update ($\beta = -0.084$, $p < 0.05$). Therefore, Hypothesis 4 is also supported.

Table 8.3 Results of hypothesis testing

	Coefficients	Hazard ratios
Central routes		
1. The amount of information (avg. review length)	0.0083***	1.0083***
2. The amount of information (total no. of words)	−1.1e-05***	1.0***
3. Review readability	−0.038***	0.96***
Periphera routes		
4. Sentiment strength	−0.084*	0.92*
5. Review valence	−0.10**	0.90**
Control variables		
6. Business model	−0.48***	0.62***
7. Competition intensity	−0.046	0.95
Wald χ^2	165.5	
Likelihood ratio test	197.3	

*Note. Significance levels: *p < 0.1, **p < 0.05, ***p < 0.01*

8.5 Discussion and Conclusions

User generated product reviews have been found to have a word-of-mouth effect as a new element of marketing communication. However, their implication on improving product innovation cycles have not been studied before. Guided by the ELM persuasion theory, we examined the central and peripheral cues of online mobile app reviews and their impact on app developers' product innovation decisions. Our empirical study shows that easy-to-ready user reviews with high average review length and mildly negative reviews can increase the likelihood of a future app update. Our findings highlight the need for researchers to explore user generated reviews in the context of customer-centered product innovation.

Our work has theoretical and practical implications. First, our research enriches customer-centered product innovation literature and is the first paper to empirically examine the impact of online product reviews on product innovation in the mobile app industry. Second, our research can benefit different stakeholders in the mobile app industry. For customers, our research encourages them to continue contributing reviews because those reviews do matter in getting better products in return. Moreover, our study provides specific guidelines for writing online product reviews that can be better perceived by app developers. For app developers, our study can be used to automatically process user reviews and extract useful information content in their product innovation processes. Lastly, our study can also benefit mobile app platform providers such as the Google Play Store and Apple App Store by promoting useful user reviews and making a better ecosystem for product innovation.

Our work has also several limitations. First, our data set may contain mobile apps that was only updated once during our observation period. That will introduce anomalies in our analysis. Second, our model can be improved by including important control variables such as mobile app rank, app category, and app tenure. Third, our findings can only be applied to the data set that we collected. Additional analysis is necessary to improve the generalizability of our findings. We acknowledge that predictive analytics could be used to generalize our conclusions to other mobile app platforms such as Apple's App Store and Windows App Store.

Biographies

Zhilei Qiao is a third year Ph.D. student in Business Information Technology at Virginia Tech. He received Master degree in Computer Science from Tianjin Polytechnic University, P.R. China, in 2007, and a Bachelor degree in Computer Science and Technology from Shandong University of Science and Technology, P.R. China, in 2004. He has more than 6 years' work experience in IT companies (Infosys and DNV) as Software Engineer/Senior Software Engineer. His research interests include social media analysis, text mining, product innovation and decision support systems. He has published papers in a variety of conferences.

G. Alan Wang is an Associate Professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. He received a Ph.D. in Management Information Systems from the University of Arizona. His research interests include heterogeneous data management, data mining and knowledge discovery, and decision support systems. He has published in Production and Operations Management, Decision Support Systems, Communications of the ACM, IEEE Transactions of Systems, Man and Cybernetics (Part A), IEEE Computer, Group Decision and Negotiation, Journal of the American Society for Information Science and Technology, and Journal of Intelligence Community Research and Development.

Mi Zhou is a Ph.D. student in Accounting and Information Systems at Virginia Tech. His research interests include Big Data Analytics, Corporate Disclosures, and Capital Markets. He has published papers in a variety of journals and conferences. Jamie earned his master degrees concurrently in Accounting and Computer Science from UNC Charlotte. He started his career with IBM and AmerisourceBergen Corporation as a software engineer. He then worked as a consultant/Senior Consultant at Deloitte for 3 years.

Weiguo Fan is a Full Professor of Accounting and Information Systems and Full Professor of Computer Science (courtesy) at Virginia Tech. He is also a L. Mahlon Harrell Research Fellow. He received his Ph.D. in Business Administration from the Ross School of Business, University of Michigan, Ann Arbor, in 2002, a M.Sc. in Computer Science from the National University of Singapore in 1997, and a B. E. in Information and Control Engineering from the Xi'an Jiaotong University, P.R. China, in 1995. His research interests focus on the design and development of novel information technologies—information retrieval, data mining, text/web mining, business intelligence techniques—to support better business information management and decision making. He has published more than 140 refereed journal and conference papers. His research has appeared in journals such as Information Systems Research, Journal of Management Information Systems, Production and Operations Management, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Communications of the ACM, Journal of the American Society on Information Science and Technology, Information Processing and Management, Decision Support Systems, ACM Transactions on Internet Technology, Pattern Recognition, IEEE Intelligent Systems, Pattern Recognition Letters, International Journal of e-Collaboration, and International Journal of Electronic Business. His research has been cited more than 3600 times (h-index:33, i10-index:72) Google Scholar. His research has been funded by five NSF grants, one PWC grant, and one KPMG grant.

References

Abrahams AS, Jiao J, Fan W, Wang GA, Zhang Z (2013) What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. Decis Support Syst 55(4):871–882

Alan S. Dunk (2011) Product innovation, budgetary control, and the financial performance of firms. The British Accounting Review 43(2):102–111

Alegre J, Chiva R (2008) Assessing the impact of organizational learning capability on product innovation performance: an empirical test. Technovation 28(6):315–326

Bardzil J, Rosenberger P III (1996) Atmosphere: does it provide central or peripheral cues. Asia Pacific Adv Consum Res 2:73–79

Batra R, Stayman DM (1990) The role of mood in advertising effectiveness. J Consum Res 17:203–214

Bloomfield RJ (2002) The 'incomplete revelation hypothesis' and financial reporting. Account Horiz 16(3):233–243

Cankurtaran P, Langerak F, Griffin A (2013) Consequences of new product development speed: a meta-analysis. J Prod Innovat Manag 30(3):465–486

Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 39(5):752

Chen Y, Turut Ö (2013) Context-dependent preferences and innovation strategy. Manage Sci 59(12):2747–2765

Chen Y, Xie J (2008) Online consumer review: word-of-mouth as a new element of marketing communication mix. Manage Sci 54(3):477–491

Cheryl Nakata, K. Sivakumar (1996) National Culture and New Product Development: An Integrative Review. Journal of Marketing 60 (1):61

Cheung MY, Sia C-L, Kuan KKY (2012) Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective. J Assoc Inf Syst 13(8):618–635

Cohen WM, Nelson RR, Walsh JP (2002) Links and impacts: the influence of public research on industrial R&D. Manage Sci 48(1):1–23

Cox DR (1972) Regression models and life-tables. J R Stat Soc B Methodol 34:187–220

Decker R, Trusov M (2010) Estimating aggregate consumer preference from online product reviews. Int J Res Mark 27(4):293–307

Di Gangi PM, Wasko M (2009) Steal my idea! Organizational adoption of user innovations from a user innovation community: a case study of Dell IdeaStorm. Decis Support Syst 48(1):303–312

Dougherty D, Dunne DD (2011) Organizing ecologies of complex innovation. Organ Sci 22(5):1214–1223

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—An empirical investigation of panel data. Decis Support Syst 45(4):1007–1016

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. IEEE Trans Knowl Data Eng 23(10):1498–1512

Goettler RL, Gordon BR (2011) Does AMD spur intel to innovate more? J Polit Econ 119(6):1141–1200

Greenstein S, Ramey G (1998) Market structure, innovation and vertical product differentiation. Int J Ind Organ 16(3):285–311

Holahan PJ, Sullivan ZZ, Markham SK (2014) Product development as core competence: how formal product development practices differ for radical, more innovative, and incremental product innovations. J Prod Innov Manag 31(2):329–345

Jayaram J, Oke A, Prajogo D (2014) The antecedents and consequences of product and process innovation strategy implementation in Australian manufacturing firms. Int J Prod Res 52(15):4424–4439

Jin J, Ji P, Kwong CK (2015) What makes consumers unsatisfied with your products: review analysis at a fine-grained level. Eng Appl Artif Intel 47:38–48

Jonathan W. Palmer, David A. Griffith (1998) An emerging model of Web site design for marketing. Communications of the ACM 41 (3):44-51Keupp MM, Palmié M, Gassmann O (2012) The strategic management of innovation: a systematic review and paths for future research. Int J Manag Rev 14(4):367–390

Kotabe M, Jiang CX, Murray JY (2011) Managerial ties, knowledge acquisition, realized absorptive capacity and new product market performance of emerging multinational companies: a case of China. J World Bus 46(2):166–176

Lee TY (2007) Needs-based analysis of online customer reviews. In: Proceedings of the ninth international conference on electronic commerce, ACM, New York, pp 311–318

Lee H, Seo S (2013) What determines an agreeable and adoptable idea? A study of user ideas on MyStarbucksIdea Com. In: System sciences (HICSS), 2013 46th Hawaii international conference, IEEE, Hawaii, pp 3207–3217

Lehavy R, Li F, Merkley K (2011) The effect of annual report readability on analyst following and the properties of their earnings forecasts. Account Rev 86(3):1087–1115

Li F (2008) Annual report readability, current earnings, and earnings persistence. J Account Econ 45(2–3):221–247

Luo X (2009) Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. Mark Sci 28(1):148–165

Mudambi SM, Schuff D (2010) What makes a helpful review? A study of customer reviews on Amazon. Com. MIS Q 34(1):185–200

O'Hern M, Rindfleisch A (2008) Customer co-creation: a typology and research agenda. Rev Mark Res 4:84–106

Payne-James JJ, Khawaja HT (1993) Review: first choice for total parenteral nutrition: the peripheral route. J Parenter Enteral Nutr 17(5):468–478

Petty RE, Cacioppo JT, Goldman R (1981) Personal involvement as a determinant of argument-based persuasion. J Pers Soc Psychol 41(5):847

Petty RE, Schumann DW, Richman SA, Strathman AJ (1993) Positive mood and persuasion: different roles for affect under high-and low-elaboration conditions. J Pers Soc Psychol 64(1):5

Porter ME (1980) Competitive strategy: techniques for analyzing industries and competition. The Free Press, New York

Prajogo DI, Ahmed PK (2006) Relationships between innovation stimulus, innovation capacity, and innovation performance. R&D Manag 36(5):499–515

Ramaswamy V, Prahalad CK (2004) Co-creation experiences: the next ractice in value creation. J Interact Mark 18(3):5–14

Sawhney M, Verona G, Prandelli E (2005) Collaborating to create: the internet as a platform for customer engagement in product innovation. J Interact Mark 19(4):4–34

Schindler RM, Bickart B (2012) Perceived helpfulness of online consumer reviews: the role of message content and style. J Consum Behaviour 11(3):234–243

Sen S, Lerman D (2007) Why are you telling me this? An examination into negative consumer reviews on the web. J Interact Mark 21(4):76–94

Silvias (2014) Mobile Load: Performance Testing for Mobile Applications. https://community.saas.hpe.com/t5/LoadRunner-and-Performance/Mobile-Load-Performance-Testing-for-Mobile-Applications/ba-p/273396#.WUjnXzOB000

Stewart DW, Zhao Q (2000) Internet marketing, business models, and public policy. J Public Policy Mark 19(2):287–296

Sussman SW, Siegal WS (2003) Informational influence in organizations: an integrated approach to knowledge adoption. Inf Syst Res 14(1):47–65

Syer MD, Nagappan M, Hassan A E, Adams B (2013) Revisiting prior empirical findings for mobile apps: an empirical case study on the 15 most popular open-source android apps. In: Proceedings of the 2013 conference of the Center for Advanced Studies on Collaborative Research (CASCON), pp. 283–297

Tam KY, Ho SY (2005) Web personalization as a persuasion strategy: an elaboration likelihood model perspective. Inf Syst Res 16(3):271–291

Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 61(12):2544–2558

Troy LC, Szymanski DM, Varadarajan PR (2001) Generating new product ideas: an initial investigation of the role of market Information and organizational characteristics. J Acad Mark Sci 29(1):89–101

Von Hippel E (1998) Economics of product development by users: the impact of ‘sticky’ local information. Manage Sci 44(5):629–644

Von Hippel E, Katz R (2002) Shifting innovation to users via toolkits. Manage Sci 48(7):821–833

Von Hippel E, Thomke S, Sonnack M (1999) Creating breakthroughs at 3M. Harv Bus Rev 77:47–57

Wales WJ, Parida V, Patel PC (2013) Too much of a good thing? Absorptive capacity, firm performance, and the moderating role of entrepreneurial orientation. Strat Manag J 34:622–633

Wu C, Shaffer DR (1987) Susceptibility to persuasive appeals as a function of source credibility and prior experience with the attitude object. J Pers Soc Psychol 52(4):677

Xiao S, Wei C-P, Dong M (2015) Crowd intelligence: analyzing online product reviews for preference measurement. Inf Manag 53(2):169–182

Yao Z, Yang Z, Fisher GJ, Ma C, Fang EE (2013) Knowledge complementarity, knowledge absorption effectiveness, and new product performance: the exploration of international joint ventures in China. Int Bus Rev 22(1):216–227

Zhang Y (1996) Responses to humorous advertising: the moderating effect of need for cognition. J Advert 25(1):15–32

Zhang C, Hahn J, De P (2013) Continued participation in online innovation communities: does community response matter equally for everyone? Inf Syst Res 24(4):1112–1130

Zhou M, Du Q, Fan W, Qiao Z, Wang G, Zhang X (2015) Money talks: a predictive model on crowdfunding success using project description. In: Twenty-first Americas conference on information systems, Puerto Rico

Zirger BJ, Maidique MA (1990) A model of new product development: an empirical test. Manag Sci 36(7):867–883

Chapter 9
Whispering on Social Media

Juheng Zhang

Abstract Using Twitter as the primary social media platform, we study the predictive relationship of social media buzz in quiet periods and the IPO's first-day return, liquidity, and volatility. We compare social media buzz with conventional press news coverage and show that social media buzz is stronger at predicting the first-day returns than conventional press news.

Keywords IPO • Initial returns • Cocial media analytics

9.1 Introduction

With the ease and speed of the Internet, individual investors, also called retail or unsophisticated investors, have been attracted to financial markets. New technologies have streamlined the procedures to trade stocks and have provided investors with an improved information environment. Individuals can access information about companies online much more easily now than they could in the era when the mainstream communication channels were television and radio broadcasting only.

Current information technologies also provide individuals with social media platforms (e.g., Twitter, Facebook) that enable them to be socially connected and to exchange opinions and information. Social media as a means of disseminating information differ from traditional media in several ways: cost, speed, impact, and reach. For instance, tweets can reach a large number of people almost immediately and at a negligible cost. Social media have become a major platform for individuals to access news and for investors to learn about investing opportunities. Social media is changing how people find and interpret news, and has been identified as the most powerful outlet of information (Greenslade 2014). The study (Bennett 2013) found that 73% of online users are active on at least one social network. Through social media, people can get the latest corporate news, market trends, investment information, etc. They can also use social media to discuss stocks and the markets with other investors and to research information about companies and brokers.

J. Zhang (✉)
Department of Operations and Information Systems, University of Massachusetts Lowell,
One University Ave., Lowell, MA 01854, USA
e-mail: juheng_zhang@uml.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_9

We know that economic transactions are often embedded in social relationships (Granovetter 1985). Individuals increasingly rely on the online opinions and comments of others when making purchase decisions (Zhang and Zhang 2015). When making investment decisions, individual investors are probably influenced by other investors in social media. Unlike sophisticated investors (e.g., institutional shareholders), who review the critical information contained in the statutory prospectus, unsophisticated investors fail to read the detailed prospectus or cannot absorb most of the information even if they do read it (Newkirk 1991). Unsophisticated investors turn to online chat rooms or social media when faced with having to make investment decisions under conditions of uncertainty (Newkirk 1991). Investors may adopt other investors' recommendations and decisions that appear on social media in the belief that the people they are following online have more accurate information about an issuing company than they do. When investors follow others' behaviors, information cascades occur.

The quiet period rules were adopted by the SEC decades before the advent of the Internet and of personal computers. In this rich information world, Twitter users tweet their favorite IPOs and may influence other users to acquire the same portfolios; bloggers forecast the outcome of IPOs and their blogs may sway people's decisions; newspapers choose certain IPOs to cover and that news coverage can influence retail investors' opinions of the IPOs; retail investors chat and discuss in social media and their investment decisions may be influenced by the opinions of people of whom they know little or nothing. The influence of public information disseminated through the media outlets should never be neglected.

In this paper, we consider buzz in social media and news in conventional media. We study the relationship between the buzz and news about IPOs and the IPOs' stock performance. In Sect. 9.2 of this paper, we review relevant studies related to our study, including research work in finance field and in Information Systems field. In Sect. 9.3, we formulate our research questions, and conjecture our hypotheses. In Sect. 9.4, we describe our data collection processes and provide simple statistics on the data variables used in this study. In Sect. 9.5, we analyze the first-day performances of IPOs with reference to quiet period social media content and news coverage. Section 9.6 concludes our study.

9.2 Literature Review

Research related to our study is found in the IPO literature. We refer readers to the review paper (Ritter and Welch 2002), which overviews various theories and reasons for going on public and includes a detailed discussion of the two stylized characteristics of IPOs: first-day return (underpricing) and long-run underperformance, both of which are related to the behaviors of retail investors. Liu et al. (2014) study media coverage of IPOs. They examine conventional media coverage (press news) prior to an IPO to predict the IPO firm's long-run liquidity and its following analysts and institutional investors. They measure the pre-IPO news coverage of a company by the number of articles mentioning the company name during the 30 days prior to the IPO date, and find a positive correlation between the pre-IPO press coverage, the firm's long-run following analysts, institutional investors, and the stock's liquidity. Da et al. (2011) use Google search trend as the index of retail investors' attention in predicting stock returns. Our paper differs from these studies in that we focus on social media content rather than on conventional press news and in that we use social media content to study both the IPO's first-day performance and its long-term underperformance rather than the long-term attention from analysts and institutional investors.

In the information systems (IS) literature, social media content has been examined (e.g., Bollen et al. 2011; Luo et al. 2013; Zhang 2014, 2015a, b; Zhang et al. 2015). Bollen et al. (2011) use the mood of daily Twitter feeds to predict the Dow Jones Industrial Average (DJIA) over time. Two mood-tracking tools, OpinionFinder and Google Profile of Mood States, are used to detect the mood of daily tweets. They find that the inclusion of public mood can improve the ability to predict stock prices. Luo et al. (2013) use sentiment analysis to study the impact of the sentiment of social media content towards a company on the company's equity value. These IS studies focus on whether a message about a company is positive or negative and whether the sentiment affects the firm's stock price. They use sentiment analysis to study the impact of positive and negative social media content on firm performance.

Cook et al. (2006) suggest that 99% of the news about IPOs is non-negative, so we use the simple count of tweets mentioning a stock as the measure of the volume of social media buzz about the stock, and then study the impact of that buzz amount on the IPO's first-day return, liquidity, and volatility. As for the content of tweets, the informedness and consensus measure the value of information release (Holthausen and Verrecchia 1990). We capture the informedness and consensus of each tweet by the number of favorites and retweets of the tweet jointly, and study their impact on IPOs' initial returns.

9.3 Research Questions

We used three categories of media content in the quiet period—press news, business tweets, and microbloggers' tweets—to study the first-day return, liquidity, and volatility. Individual investors are often influenced by press news about IPOs or recommendations available on social media platforms. The volume of social media buzz accumulated during quiet period may have an impact on the IPOs' first day return, trading volume, and bid-ask spread, and it may even lead to unjustified upward price pressure on IPOs. Hence, we formulate our hypothesis as the following.

Hypothesis 1: *Social media buzz about an IPO during quiet period is positively correlated with the IPOs' first day return, volatility, and liquidity.*

Twitter users read and evaluate feeds posted on Twitter, and the number of retweets and favorites indicate the agreement among users on the information contained in the tweet and the degree of its information value. If a Twitter user retweeted or favorited a tweet, s/he often found that the tweet contained useful information or could be helpful to others, and also somewhat agreed with the information conveyed in the tweet. The number of retweets and favorites indicate the agreement among users on the information contained in the tweet and the degree of its information value. We use the number of retweets and favorites of a tweet to jointly capture the informedness and consensus of the tweet. The informedness and consensus of the tweets about IPOs during quiet period may have an impact on investors' investment decisions and further accelerate the impact of the tweets on IPOs' initial return, volatility, and liquidity. Therefore, we conjecture the following hypothesis:

Hypothesis 2: *The information value of tweets about an IPO during quiet period are positively correlated with the IPO's first day return, volatility, and liquidity.*

9.4 Data Description

We began with the list of IPOs in the year 2014 and use them as our company samples. We downloaded the offer prices, open prices, and closing prices of the IPOs from IPOScoop.com. We retrieved the accounting numbers prior to the IPO companies from the Compustat database. Around twenty of the companies have missing data. We provide the basic statistics of the IPOs in Table 9.1.

We downloaded tweets by searching either ticker symbol or company name of each IPO. Twitter users use variant company names when mentioning a company: for example, "Alibaba Group Holding Ltd.," "Alibaba Group Holding," or "Alibaba" for Alibaba company. We allowed variations of a company name when searching in Twitter, but refined the search keywords and removed noisy or ambiguous ones: for example, "King" for King Digital Entertainment PLC. We downloaded the tweets along with each tweet's posted date, number of times being retweeted, times being "favorited," and user account name.

Table 9.1 Descriptive statistics of IPOs in 2014 year

Variable	Mean	Std Dev	Minimum	Maximum	N
Offer price	15.16	7.80	4.00	91.00	288
Opening price	17.24	9.98	4.00	92.70	288
First day close	17.33	10.24	4.28	93.89	288
First day return	0.13	0.27	−0.35	2.07	288
Asset	2305.4	2306.4	2307.4	2308.4	266
Net income	21.88	363.42	−3426	3750.56	263
Liabilities	1912.12	11108.7	0.131	136,959	266
Intangible asset	216.35	776.52	0	7061	260

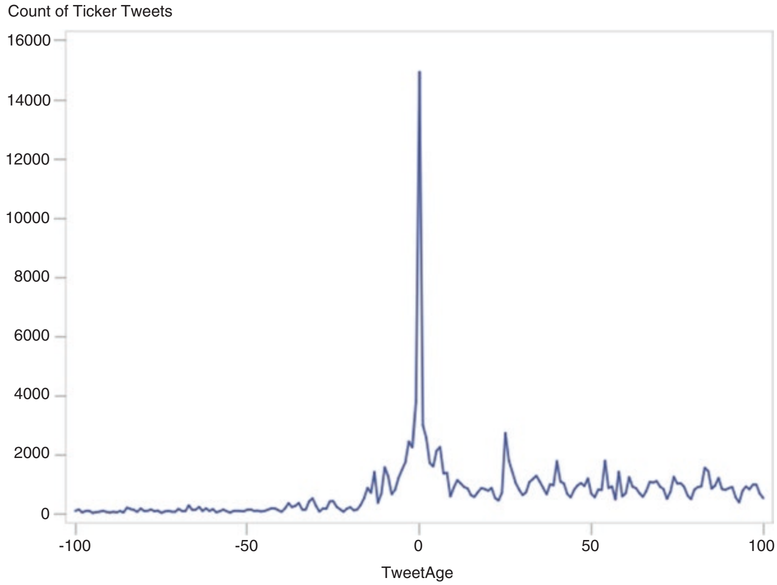


Fig. 9.1 Number of downloaded tweets over time

We checked the account names of downloaded tweets and found that the tweets were mainly posted by individual Twitter users. We separated those tweets from the ones posted by the companies themselves. We call the tweets by individual Twitter users as "microbloggers' tweets" and the tweets from businesses as "businesses' tweets" in our analysis. Following the IPO literature (e.g., Liu et al. 2014), we defined the 30 days prior to the IPO as the quiet period. We then used the volume of tweets in the quiet period to study the effect of social media buzz on an IPO's first-day performance.

We plot the number of downloaded tweets mentioning IPO ticker symbols in Fig. 9.1. The tweet age is the number of days between the tweet posted date and the tweeted stock's IPO date. Figure 9.1 shows that users start to tweet the IPOs actively around 30 days before the IPO date. The volume of tweets peaks on the IPO day, with 15,000 tweets mentioning the sampled IPOs in total. After the IPO date, the stocks are tweeted less frequently than the IPO date but more than during the quiet period, around 1000 tweets in total about the IPOs per day.

We collected detailed information about the IPO companies' twitter accounts. In Table 9.2, we provide the statistics for the twitter account information. 118 out of 288 IPO sample companies have a twitter account.

In addition to analyzing the collected Twitter data, we considered the news coverage in the traditional media. We downloaded the articles in the press about IPOs from the Lexis Nexis database. We used not only stock ticker symbols but also the variant company names as search keywords to find the IPO news of the companies during the period from January 1, 2013, to December 31, 2014.

Table 9.2 Descriptive statistics of twitter accounts of 2014 year IPOs

Variable	Mean	Std Dev	Minimum	Maximum	N
Twitter age	3.827569	2.1419	−1.31667	7.666667	118
No. tweets	5667.66	14693.11	0	96,155	118
No. followers	36537.58	156683.4	5	1,423,522	118
No. friends	2382.07	8807.85	0	81,508	118
Likes	640.75	2046.81	0	19,532	118

Table 9.3 Descriptive statistics of media activities in quiet period

	Microbloggers			Lexis Nexis	Business		
	Tweets	Retweets	Favorites	News	Tweets	Retweets	Favorites
Mean	28.8	40.66	31	4.58	78.75	129.81	130.94
Std Dev	46.05	150.59	106.49	20.73	125.	371.54	394.95
Min	1	0	0	1	1	0	0
Max	634	1598	1503	257	598	2453	2587
N	266	266	266	154	83	83	83

In Table 9.3, we summarize the number of traditional news articles, microbloggers' tweets, and business tweets during the quiet period. As shown in the table, on average each company had 4.58 items of press coverage, 28.89 microbloggers' tweets, and 78.75 business tweets per quiet period.

We collected other variables from standard data sources, such as the underwriter rankings found in Loughran and Ritter (2004) and market sentiment data from CNN (http://money.cnn.com/). Daily stock prices in the year of 2014 are from CRSP and in the first half of the year 2015 from Yahoo! Finance.

9.5 Empirical Results

We include the prediction variables for the first-day returns as suggested in the IPO literature (e.g., Da et al. 2011; Kim and Ritter 1999) to conduct the analysis of variance. The log of a company's total assets, the reputation of underwriters, and market sentiment are included. We also use the variables of media buzz, including the number of news articles, the log of the number of microbloggers' tweets (including retweets and favorites), the log of the number of company's followers, and the age of the Twitter account. Table 9.4 shows the variance using the stepwise selection of variables. The adjusted R-Square of the prediction model is 0.27.

To determine if first-day return is related to long-run underperformance, we looked at the cumulative return for the following time periods: 30 days, 45 days, 1 year, and 1 year after the IPO from 30 days after the IPO. Table 9.5 lists the correlation coefficients and t-values. As shown in Table 9.5, price reversion is observed

Table 9.4 Analysis of variance on the first-day return

Source	DF	Sum of squares	Mean square	F value	Pr > F
Model	7	1.93557	0.27651	5.07	0.0001
Error	70	3.819	0.05456		
Corrected total	77	5.75456			
Root MSE	0.23	R-square	0.3364		
Dependent mean	0.1612	Adj R-Sq	0.27		
Coeff var	144.8				
Parameter estimates					
Variable	DF	Parameter estimate	Standard error	t value	Pr > |t|
Intercept	1	−0.68	0.26206	−2.62	0.01
News	1	0.002	0.00984	0.2	0.84
Log tweets	1	0.21	0.11116	1.92	0.06
Log asset	1	−0.09	0.03253	−2.96	0.004
Underwriter ranking	1	0.06	0.02645	2.3	0.02
Market bearish	1	−0.01	0.51338	−0.02	0.98
Twitter age	1	0.02	0.01976	0.9	0.37
Log twitter followers	1	0.02	0.01633	1.07	0.29

Table 9.5 Correlation coefficients of returns

	First day	30 day	45 day	1 year	1 year from 30th day
First day	1	−0.191	−0.111	−0.136	−0.0471
		0.022	0.191	0.196	0.6571
30 day	−0.191	1	0.861***	0.408***	0.0521
	0.022		<.0001	<.0001	0.6238
45 day	−0.111	0.861***	1	0.464***	0.1302
	0.1916	<.0001		<.0001	0.2183
1 year	−0.136	0.408***	0.464***	1	0.91***
	0.1968	<.0001	<.0001		<.0001
1 year from	−0.0471	0.0521	0.1302	0.91***	1
the day 30th	0.6571	0.6238	0.218	<.0001	

Signif. codes: ***, **, *, for <0.0001, 0.01, and 0.1 respectively.

as early as 30 days after the IPO. The first-day return is significantly negatively correlated with the 30-day return, with a coefficient of −0.19. One percent of gain in stock return on the first day will be reversed with an 0.19% loss after the 30th day. The first-day return is also negatively correlated with the cumulative returns in the 45-day period, the 1-year period, and 1-year-after-30-day period. The finding is consistent with IPO underpricing and long run underperformance in existing studies (e.g., Loughran and Ritter 2004; Ritter and Welch 2002).

9.6 Conclusion

We find that the number of feeds tweeted by Twitter users on the IPO and the number of times that the tweets are retweeted or favorited during the quiet period are significantly positively correlated with the IPO's first-day return, trading volume, and bid/ask spread. The findings suggest that social media buzz in a quiet period is significantly correlated with an IPO's stock performance on the first day.

Biography

Juheng Zhang is an Associate Professor of Information Systems at University of Massachusetts, Lowell. She earned a Ph.D. in Management Information Systems from University of Florida in August 2011. Her research focuses on data analytics and data manipulation. Juheng Zhang has published in Information Systems Research, Decision Support Systems, European Journal of Operational Research, and other academic journals. She has taught Business Intelligence & Data Mining and other core courses in the field of Information Systems.

https://www.uml.edu/MSB/faculty/Zhang-Juheng.aspx

References

Bennett S (2013) 73% of online adults now use social media. http://www.mediabistro.com/alltwitter/pew-social-study_b53501. Accessed 30 Dec 2013

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2:1–9

Cook DO, Kieschnick R, Van Ness RA (2006) On the marketing of IPOs. J Financ Econ 82(1):35–61

Da Z, Engelberg J, Gao P (2011) In search of attention. J Financ 66(5):1461–1499

Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91(3):481–510

Greenslade R (2014) More digital disruption ahead for mainstream news groups, says survey. http://www.theguardian.com/media/greenslade/2014/jun/12/digital-media-social-media. Accessed 11 June 2014

Holthausen RW, Verrecchia RE (1990) The effect of informedness and consensus on price and volume behavior. J Account Rev 65:191–208

Kim M, Ritter JR (1999) Valuing IPOs. J Financ Econ 53(3):409–437

Liu LX, Sherman AE, Zhang Y (2014) The long-run role of the media: evidence from initial public offerings. Manag Sci 60(8):1945–1964

Loughran T, Ritter JR (2004) Why has IPO underpricing changed over time? Financ Manag 33(3):5–37

Luo X, Zhang J, Duan W (2013) Social media and firm equity value. Inf Syst Res 24(1):146–163

Newkirk RG (1991) Sufficient efficiency: fraud on the market in the initial public offering context. Univ Chicago Law Rev 58(4):1393–1422

Ritter JR, Welch I (2002) A review of IPO activity, pricing, and allocations. J Financ 57(4):1795–1828

Zhang J (2014) Information revelation and social learning. Int J Bus Soc Sci 5(2):115–125

Zhang J (2015a) Ensuring trust online through the wisdom of crowd. J Internet e-Bus Stud 2015:886172

Zhang J (2015b) Voluntary information disclosure on social media. Decis Support Syst 73(2015):28–36

Zhang XM, Zhang L (2015) How does the internet affect the financial market? An equilibrium model of internet-facilitated feedback. MIS Q 39(1):17–37

Zhang J, Khan RM, Shih D (2015) The rating determinants factored in decision-making for hotel selection. Int J Appl Manag Technol 14(1):1–20

Chapter 10
Does Social Media Reflect Metropolitan Attractiveness? Behavioral Information from Twitter Activity in Urban Areas

Johannes Bendler, Tobias Brandt, and Dirk Neumann

Abstract The rapid and ongoing evolution of mobile devices allows for increasing ubiquity of online handhelds, yet boosting the recent growth of social platforms. This development facilitates participation in social media for an enormous amount of individuals independently from time and location. When navigating through a city and especially when following activities worthy to be shared with others, people uncover their traces in both geographical and temporal dimension. Using these traces to spot popular areas in a metropolitan region is valuable to a broad variety of applications, reaching from city planning to venue recommendation and investment. We propose a density-based method to determine the attractiveness of areas based solely on spatial and content characteristics of Twitter activity. Furthermore, we show the relation of attached images, videos, or linked places to the activity users are engaged in and assess the explanatory power of Twitter messages in a geographical context.

Keywords Social media • Spatial statistics • Data analysis

10.1 Introduction

In recent years, the rapid evolution and development of mobile devices have heavily boosted the growth of social web services and their users' online activity. For instance, in March 2013, Flickr users uploaded more than 3.5 million new images per day (Jeffries 2013); during December 2013, Facebook had 757 million daily

J. Bendler • T. Brandt (✉)
Geospin GmbH, c/o University of Freiburg, Platz der Alten Synagoge,
79098 Freiburg, Germany

University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg, Germany
e-mail: tobias.brandt@is.uni-freiburg.de

D. Neumann
University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg, Germany

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_10

active users worldwide on average with 556 million daily active users accessing the service from their mobile devices (Facebook Inc. 2013). On an average day, more than 500 million Twitter messages are sent (Twitter Inc. 2014), 60 million photographs are shared on Instagram collecting 1.6 billion likes (Instagram 2014) and several million people share their locations by checking in on Foursquare (2014).

The ubiquity of smartphones and other devices that enable mobile internet access facilitates the access to social networks independent of time and location. Wherever users are, they are able to share their mood, feelings, activities, photographs, and places. Not simply limited to posting statuses, a wide variety of online platforms are present where people can additionally rate the place they are visiting. Recommendations and reviews from users cover any imaginable location, ranging from restaurants and bars, stores and malls to even beaches and other public areas, as well as companies. For instance, foursquare is an online service that allows the sharing and rating of arbitrary venues and collects several million user check-ins per day, aggregating an enormous crowd-sourced recommendation database. With all this available geo-referenced data (e.g. tagged photographs or recommended venues) trips to foreign cities have fundamentally changed. When searching for a trip, people can rely on hotels others have rated and commented on before. When looking for a round-trip, visitors can check tagged and rated photographs and interesting spots, ranging from unappealing areas that should be left out to must-see locations. Therefore, one could assume that visiting a foreign city is becoming less exciting, because there is nothing left to be explored. Contrastingly, travelling rather seems to require less preparation through the support of the social web and consulting mobile devices, and so offers new opportunities, by spontaneously looking up recommendations during the stay. The amounts of data generated by users in social services all over the world provides evidence for the key aspects of reference and recommendation from others that makes information valuable.

One major drawback in social media data concerning the analyzability is that users are aware of what they are doing. The disutility of this fact is twofold. On the one hand, people are able to significantly manipulate comments, ranks, and recommendations, whether maliciously or not, and thus can considerably skew the results, possibly even rendering the entire analysis useless. On the other hand, a noticeable dominance of extremes can be observed considering the online recommendations of arbitrary entities. People tend to comment and rank either if they are positively excited or if they are truly disappointed. Thus, recommendations and comments would be distinctively more reliable if users posted independent of their current mood and so had no opportunity to maliciously impact ratings on the social web. We consider the behavior of users who are not aware of what traces they leave behind to be more honest than publicly posted opinions where users know that they might have an impact on others' decisions.

In this research, we aim to explore the possibilities of identifying places of interest in urban environments based on user-generated mass data that is more robust and reliable than pure recommendations. Complete navigation tracking is impossible because people need the ability to opt-out if they don't want their geographical location to be disclosed. Thus, we focus on Twitter as the data source of choice. People use Twitter to show where they are, but also for sharing feelings, thoughts, and desires with friends and followers. From earlier findings, we have evidence that people use Twitter on an irregular basis, but status updates are particularly cumulated around places where there is something worth being shared, whether it is interesting, boring, exciting, or disgusting. Analysis of a social activity stream that comes along with geo-spatial and temporal information is a promising and auspicious option to collect robust data on urban hot-spots, though it may originally be designed for a different purpose than pure location sharing. From these preliminary thoughts, we derive our main research questions:

- Can general social media activity in an urban environment be utilized to spot socially attractive places?
- How does social media blend into already available criteria (e.g. natural or historical factors) to determine places of interest within a city?
- How do socially attractive places relate to the environmental conditions, such as known hot spots or establishments in the vicinity?

The remainder of this work is structured as follows. In the first section, we analyze relevant work from related research streams, namely geographical attractiveness, location-based recommendations, and event recognition. The insights found in related research are used to derive characteristics of social attractiveness in the subsequent section. We describe the characteristics of our Twitter data set and derive a model to measure attractiveness that solely relies on measurable attributes from Twitter messages. In the subsequent section we provide a regression analysis to support and verify our attractiveness model. The resulting findings then are described and visually outlined. This paper closes with a concluding section with a short summary and an outlook on potential future research.

10.2 Related Work

The research at hand studies attractiveness of places in urban environments, which is originally a tourism research subject. Our approach blends in the social media perspective, and draws a relation to Big Data analytics. In terms of this multi-discipline setup, corresponding literature that covers all such aspects is rare, if existent. The following literature review is split into the three main areas of research this approach relates to; (1) definition and measurement of geo-spatial attractiveness; (2) location-based recommendation systems; (3) recognition of (eventually unusual) events from social media streams. Based on these streams, we identify the research gap that is cleared in this work.

10.2.1 Definition and Measurement of Geo-spatial Attractiveness

The rating of attractiveness of geographical regions or spots is a controversial subject and thus is always contested. What makes certain nations, cities or regions appear more attractive than others to live in, visit, or invest in (Niedomysl 2010)? Factors that influence the attractiveness of places have been studied for years by tourism experts, business investors, and governments. However, to our best knowledge none of the publications reach a common consensus regarding relevant factors or their relative weighting to be included when measuring geographical attractiveness (Rogerson 1999).

In the area of tourism research, many different methods have been proposed and applied for measuring or rating attractiveness or competitiveness of places. While many of these approaches rely on questionnaires and the opinions of experts or vacationists, other publications compare different cities or areas from specific regions or countries with similar weather conditions or cultural circumstances (Kozak and Rimmington 1999). Other research carried out by Hu and Ritchie (1993) emphasizes the touristic attribute in contributing to the attractiveness of a destination from a more general perspective. For instance, in order to measure the performance of a destination, both quantitative performance data (*hard data*), as well as qualitative aspects of the regions' competitiveness (*soft data*) are taken into account. According to Hu and Ritchie (1993), hard data refers to explicitly measurable quantities, such as tourist arrivals and incomes from the tourism sector, while soft data can contain sources that are hard to measure, for example the environmental constitution or visitor satisfaction. In order to generate and establish a competitiveness set for Turkey, they delivered questionnaires to British vacationists visiting Turkey to obtain details of their travel motivation and levels of satisfaction according to various criteria.

Numerous papers perform research towards multi-dimensional attribute sets that together determine the attractiveness of a certain place or area. They "include not only the historical sites, amusement parks, and spectacular scenery, but also the services and facilities which cater to the everyday needs of tourists" (Lew 1987). Gearing et al. (1974) propose a grouping of these attributes into five major categories; (1) natural factors; (2) social factors; (3) historical factors; (4) recreational and shopping facilities; and (5) infrastructure, food, and shelter. The importance of each sub-attribute is quantitatively supported by questionnaire-based research. The method of classifying attributes into groups is the basis for numerous additional research attempts towards place attractiveness carried out in the tourism sector. Similar categories have been affirmed in later studies by, among others, Jansen-Verbeke (1986) and Enright and Newton (2004).

10.2.2 Location-Based Recommendation Systems

Geo-spatial data (e.g. trajectories of people navigating through a city) has always been employed to serve as input for location-based recommendation systems, but lately received a boost due to the emergence of location-based social networks (Hongzhi et al. 2013; Liu et al. 2013; Ye et al. 2011). Publications on the topic of point of interest recommendation systems are manifold and many of them even propose the implementation of a well-functioning recommender system. Towards understanding the general mobility pattern of individuals, González et al. (2008) carry out a study based on the trajectories of 100,000 mobile phone users, which allows for important insights in to human mobility that are useful when proposing location-based recommenders. The certain types and formats of input data that are fed into recommendation systems are manifold. These types contain, for instance, location data, trajectories or GPS coordinates, data on activities as well as from services, point of interest categories in the close vicinity, user profiles, personal preferences (on locations), or geo-tagged photographs (Ballatore et al. 2010; Bao et al. 2012; Waga et al. 2012; Zheng et al. 2010). As baseline approaches, the named researchers choose clusters, matrices, and collaborative filters. When places are scored for location recommendation, the systems rely on various metrics, some of them being users' ratings, experts' comments, photo tags, or the distance between a user and a service (Arase et al. 2010; Bao et al. 2012; Waga et al. 2012; Zheng et al. 2010). The context-aware recommendation system presented by Waga et al. (2012) is accessible via the web and recommends relevant locations based on user-generated data, as well as based on a dataset consisting of GPS routes, trusted services, and locations from geo-tagged photographs. Divided into three different databases, the proposed system scores locations based on the corresponding three sections service, photos, and routes.

A different approach for location-based recommendation is proposed by Arase et al. (2010). The authors focus on detecting patterns from users' trips and try to make suggestions on the travel route. Trip patterns are mined by analysis of geo-tagged photographs publicly posted on the social image sharing platform Flickr by other users who have travelled to the same geographical area. Not only the geo-tag itself is used by the system, but the authors additionally rely on each picture's title and tags. The large amounts of data backing the geographical scores increase the trip recommendations' accuracy and can thus be more convincing to users. The collaborative filtering approach proposed by Zheng et al. (2010) is another approach for creating geo-spatial ranks. Potentially interesting locations and activities worth being suggested to a user are mined by analyzing user locations and activity histories. In their approach, the authors generate further knowledge, such as location profiles or activity-activity relations from geographical databases and the web. Based on matrix factorization and grid-based clustering, Zheng et al. (2010) are able to identify regions with different characteristics according to people's behavior within a city.

In addition to the location category, Bao et al. (2012) further take user preferences and social opinions into account. In their approach, a user's preference is compared to the preferences of highly experienced users that are regarded as local experts. Ballatore et al. (2010) propose a geographical information system called "RecoMap" to provide personalized recommendations by monitoring social interaction and context. Yue et al. (2009) rely on user-generated GPS data in order to discover potentially interesting locations. Furthermore, the authors analyze areas where users travel and where they rest. The idea behind this is that places of interest can be inferred from clustering the pick-up and drop-off locations of taxi passengers. The research of Leung et al. (2011) proposes clustering and recommendation based on activities drawn from GPS log files. Yoon et al. (2012) analyze user-generated GPS trails to learn transmission routes from experts and residents in order to propose itineraries to first-time visitors.

10.2.3 Recognition of Events from Social Media Streams

Apart from the identification of location-bound activities and attractions that lie in the common interest of many social media users, there is an active field of research in recognition of global, regional, or even local events by analysis of social media data. For instance, events can be popular sports games, festivals, regional weather phenomena, epidemics, or natural disasters. The principle of studies by Lee et al. (2011) and Rattenbury et al. (2007) is to monitor the sudden increase or decrease of tweets within a short time period. If exceeding a regular amount of messages per day according to the geographical regularities, the boost is considered an event. Lee et al. (2011) monitor certain geographical areas and are able to identify unexpected as well as expected activities from Twitter messages. Contrastingly, Rattenbury et al. (2007) rely on an unstructured set of Flickr tags when extracting events and places. They focus on photographs from the San Francisco Bay Area and extract semantics from their assigned tags. The authors are able to detect important spots and events from within the city.

10.2.4 Research Gap

One aspect that most social-media related research threads have in common is their focus on the textual content of messages. The widely used abbreviations in internet speech, as well as the large number of different languages used in Twitter messages render textual analysis a demanding task. Since the exact semantics (i.e. not only semantic estimates or sentiment) are not easily extracted and usually contain heavy uncertainty given the style and pace at which Twitter messages are generated—even given today's tools and methods—we intend to explore the value of the faster and simpler geographic analysis. Furthermore, to be able to use any Tweet in any city all around the globe, no matter whether it contains information as such, we think that there is a need to rely solely on the geo-spatial and temporal dimensions of information provided by social media. Furthermore, in contrast to existing research, we aim to detect popular places by people's social activity without necessitating the reason "why" exactly. We argue that digging for certain environmental reasons for which people visit places may limit the findings of the research and analysis to pre-defined spots. To our best knowledge, there is no present research that identifies geographical areas by their social attractiveness rather than environmental conditions. Our proposed method is a more general approach, appropriate to serve for travel research, city planners, and real estate investments, for example.

10.3 Identifying Areas of Social Attractiveness

In order to fill the identified research gap, we propose a method to find hot spots within a city by mining the geo-spatial information from Twitter messages. In this section, we present categories of activities representing people's actions when visiting an area. We provide information on our dataset and draw evidence on the applicability of Twitter data for measuring attractiveness. Finally, we propose a mathematical formalization to estimate social attractiveness of areas and support it visually. The model then is validated in the subsequent section.

Wherever there is an increased number of people in a definable geographical vicinity, we expect something to be there that attracts their attention and thus lies in their common interest. Identifying the position of people using social web services with geo-tagging, we synonymize their common interest as the social attractiveness of the location. From a tourist's perspective, Gearing et al. (1974) and Enright and Newton (2004) identified five groups of factors that can be used to measure the touristic attractiveness of an area. Furthermore, 17 criteria were identified and assigned to the five major groups, as outlined in Table 10.1. The authors have laid out their work together with an assessment of the criteria and a calculation of weights to fit larger geographic areas that may or may not attract tourists, such as entire cities.

In this research, we focus on a finer grained resolution since we want to identify attractive places within single cities. Nevertheless, the findings of Gearing et al. (1974) are adoptable as a guideline. In order to measure not only the touristic attractiveness but to also judge the attractiveness for the social and daily lives of residents as well, the original assignments may be inappropriate. To adapt the findings of Gearing et al. (1974) to modern urban living, we rearrange the criteria to be geared towards civil living and businesses, in addition to the touristic alignment. This requires a more general definition of attractiveness that complies with the habits of social service users. According to the research carried out by Leung et al. (2011), activities can be the key aspect towards measuring location-based attractiveness. Where many people are engaged in activities in a certain area, the density of individuals can possibly reflect the attractiveness of that region and thus serve well as a

Table 10.1 Groups and criteria to judge touristic attractiveness according to Gearing et al. (1974)

Group	Criterion
(1) Natural factors	Natural beauty
	Climate
(2) Social factors	Artistic and architectural features
	Festivals
	Distinctive local features
	Fairs and exhibits
	Attitudes toward tourists
(3) Historical factors	Ancient ruins
	Religious significance
	Historical prominence
(4) Recreational and shopping facilities	Sports facilities
	Educational facilities
	Facilities conductive to health, rest, and tranquility
	Nighttime recreation
	Shopping facilities
(5) Infrastructure and food and shelter	Infrastructure above "minimal touristic quality"
	Food and lodging facilities above "minimal touristic quality"

Table 10.2 Mapping of tourist attractiveness groups to activity categories

Activity category	Tourist attractiveness group
(a) Sightseeing, culture, landmarks	Natural factors (1), historical factors (3)
(b) Nightlife, special events, entertainment	Social factors (2)
(c) Shopping, sports, business	Recreational and shopping facilities (4)
(d) Restaurant, accommodation	Infrastructure and food and shelter (5)
(e) Transportation	Infrastructure (5)

measure. Similar research is proposed by Jaffe et al. (2006), who use densities to build an algorithm that is able to cluster geo-referenced images into collections. Inspired by their research, our starting point is to model human attention and action using official Twitter data at hand. In order to utilize Twitter status updates as a proxy for attention and action, the following assumptions are to be made; (1) Twitter status messages are posted in real-time when a user experiences something he assumes to be attracting others' attention and interest; (2) Twitter status messages with geo-tags are especially posted near the locations that provide some important activity or landmark; (3) The density of tweets reflects to some extent the number of people moving around in a specific location, cf. Jaffe et al. (2006).

We refrain from questionnaire-based research as carried out by Gearing et al. (1974) and employ a quantitative methodology instead. Due to the tools and methods available today, we can efficiently mine large amounts of data gathered from Twitter, extract user activity and locations, and calculate densities on temporal and geographical bases. Adjusting the categories outlined above, we propose mapping according to Gearing et al. (1974), outlined in Table 10.2.

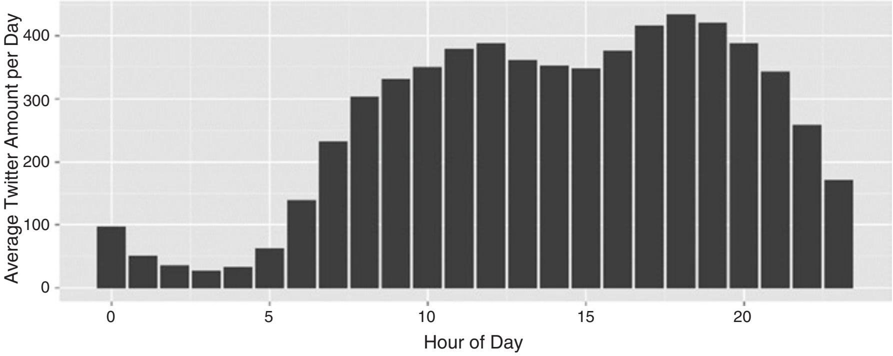


Fig. 10.1 Daily twitter pattern in San Francisco city

10.3.1 Twitter Data Characteristics

People are engaged in activities. Tourists on a city trip can for example follow their pleasure when sight-seeing, residents go about their everyday business. Independently from the certain activity, people encounter various circumstances and conditions on their routines in a city. Among those people, active Twitter users post status messages to inform their friends and followers about their experiences, thoughts, feelings, or impressions. Irrespective of the kind of status update people post, they reveal themselves in an activity at a certain location at that very time. In this research, we rely solely on geo-tagged tweets in order to use spatial information to determine patterns of activities. The patterns identified in the following provide evidence of Twitter being an appropriate data source to measure geographical attractiveness.

We have gathered more than 600,000 geo-tagged tweets from the city area of San Francisco that were posted from August to October 2013. The data has been directly obtained by Twitter and as such can be considered to represent the full extent of geo-tagged tweets from within the observation period. The data points at hand contain the tweet's user and text, the geographical location, as well as additional information, such as contained URLs and images. The Twitter messages from our three-month period shows a robust and stable pattern over a 24 h period, as depicted in Fig. 10.1. With generally low activity at around 50–100 geo-tagged tweets per hour, the night hours draw the baseline. The hourly volume of tweets increases during the morning hours and peaks around 12 a.m., followed by a slight drop until 3 p.m. The afternoon and early evening cover the most active phase of the day with a peak at around 6 p.m. After that peak, the tweet volume steadily drops until it reaches the nightly baseline. Given the steadiness of this pattern, we can assume Twitter users to act in daily patterns as well. As found by Bendler et al. (2014), there is a causal link between the time of day and the Twitter activity in the vicinity of certain points of interest. For example, Twitter activity is above average around restaurants in the evening, around bars and night clubs during late-night hours, and

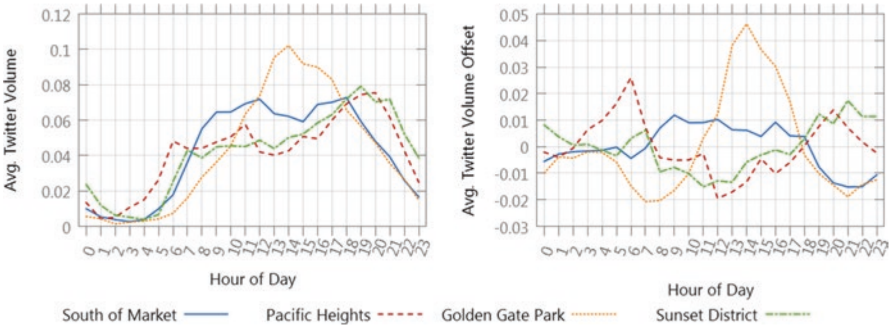


Fig. 10.2 Twitter pattern and offset in San Francisco districts

around cafés during forenoon. Thus, given the official Twitter data at hand, we are confident in finding the attractiveness measures for certain areas of a city, tentatively based on the hour of day.

Figure 10.2 illustrates the difference in averaged Twitter volume by hour of day in four different districts of San Francisco. On the left-hand panel, the absolute average is plotted for the administrative districts South of Market, Pacific Heights, the Golden Gate Park, and Sunset District. The right-hand panel shows the offset between each district and the average Tweet volume of the entire city. The districts are chosen in a manner to represent areas that possibly offer different activities and thus yield different social user behaviors at different times of day. South of Market, as a business district shows a Twitter volume pattern that is close to the entire average but still lies above average during daytime and below average in the evening hours. Pacific Heights and Sunset District are residential areas, which is most likely the cause of their similar patterning. We can identify an above-average volume during morning and evening hours and a volume below average during working hours. The Golden Gate Park, a recreational area, has an entirely different pattern. It lies below average during most time of the day but shows a high peak from noon to the early evening. These different patterns for different districts in San Francisco provide some evidence that Twitter data can represent societal habits and thus may be consulted as a source for measuring the attractiveness of urban areas.

10.3.2 Social Attractiveness

The Twitter data enables the identification of distinctive usage patterns with respect to city district and time of day. Based on this perception, we develop a model in this section that allows for estimating the social attractiveness of regions within a city. The outline of our approach is given in Fig. 10.3. Driven by the dataset of Twitter messages, we formulate estimations of popularity and activity based on a grid.

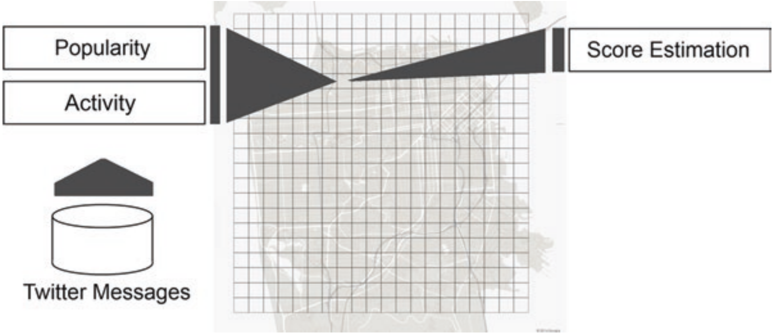


Fig. 10.3 Calculating scores from estimated popularity and attractiveness on a grid

These two estimated measures will then be combined into a score that indicates the social attractiveness of the area covered by the respective cell. Once performed for the entire grid, the method allows comparison of the estimated social attractiveness among grid cells, effectively identifying hot spots of a city.

The social attractiveness model proposed in this section targets the attractiveness estimation for regions based on Twitter activity. Our approach is different from recent relevant literature in various ways. Lee et al. (2011) attempt to detect unusual events by dividing an area into a grid and employ clustering methods to discover unusual regional social activities. The authors monitor the tweet count and Twitter users within a specific time horizon to spot exactly the unusual growth. Bao et al. (2012) employ a location-based social network to gather the location history data of users, combine it with rating information from local experts, and are therefore able to provide personalized recommendations. In contrast to these approaches, we neither rely on events that can be isolated in the temporal dimension, nor do we track location histories of certain users or consult experts. Conversely, we carry out analyses that rely solely on the social activity on Twitter using a grid that covers the city area of San Francisco. We further refine the characteristics of activities by calculation of additional key metrics, such as the number of unique users, for example.

Following Tobler's first law of geography "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970), we employ a grid-based approach for analysis of attractiveness within an urban environment, efficiently splitting the major problem into sub-problems of smaller regional diversity and complexity. As Lee et al. (2011) note, determining an adequate cell size is very difficult for grid-based approaches. If analyses are performed based on administrative regions, the results can be considered oversimplified because large regions most likely span multiple areas with different focuses of interest. Furthermore, accuracy and plausibility are questionable. The same applies to the case of choosing a grid with a very high resolution. A single area of interest could be split up into multiple grid cells yielding different results though they belong together from an activity point of view. Thus, we leave the grid size variable and carry out our analyses with different resolutions. The grid's definition is given in Eq. (10.1).

$$G\ (\lambda,\ \phi,\ \Delta\lambda,\ \Delta\phi,\ x\,,\ y) = \begin{pmatrix} g_{1,1} & g_{2,1} & \cdots & g_{x,1} \\ g_{1,2} & \square & \square & g_{x,2} \\ \vdots & \square & \ddots & \vdots \\ g_{1,y} & g_{2,y} & \cdots & g_{x,y} \end{pmatrix} \tag{10.1}$$

with

$x\,,y\in N$: grid dimensions
$\lambda\,,\phi$: longitude and latitude describing grid's origin
$\Delta\lambda\,,\Delta\phi$: longitude, latitude describing edge length of cells
$g_{i,j}\mapsto(\lambda_g,\phi_g)$: grid cells

Let a grid G be defined as a matrix of $x\times y$ grid cells $g_{i,j}$. Each grid cell g is defined as a tuple of longitude λ_g and latitude ϕ_g corresponding to the center of the respective cell. The area covered by a single cell can thus be described by $\lambda_g \pm \frac{\Delta\lambda}{2}, \phi_g \pm \frac{\Delta\phi}{2}$.

Since we attempt to measure social activity, we denote all Twitter messages a as representatives of the set of all activities A, as outlined in Eq. (10.2). Each Twitter message maps to an 8-tuple consisting of the point in time t the message has been published, longitude λ_a and latitude ϕ_a describing the exact location, a unique user identifier u, the message text m and sets I, V, P containing all images, videos, or places attached to the tweet, respectively.

$$A = \left\{a_1, a_2, \ldots a_{|A|}\right\}$$

$$a \in A \mapsto \left(t, \lambda_a, \phi_a, u, m, I, V, P\right) \tag{10.2}$$

with

t: point in time of publication
$\lambda_a\,,\phi_a$: longitude, latitude describing location of publication
u: unique user identifier
m: message text
$I=\{i_1\ldots i_{|I|}\}$: set of all images attached to the tweet
$V=\{v_1\ldots v_{|V|}\}$: set of all videos attached to the tweet
$P=\{p_1\ldots p_{|P|}\}$: set of all places attached to the tweet

Let the geographic reference between a Twitter message and a grid cell be denoted by the ~ operator. Following the principles of topological spaces in mathematics, if a is geographically enclosed, i.e. embedded in the area covered by g, then the unary operation ~*a* yields the respective cell *g* as defined in Eq. (10.3). Note that ~ does not represent proportionality in this context.

$$\sim: A \rightarrow G, a \mapsto g \in G \left| \left|\lambda_a - \lambda_g\right| \leq \frac{|\Delta\lambda|}{2} \wedge \left|\phi_a - \phi_g\right| \leq \frac{|\Delta\phi|}{2} \right. \tag{10.3}$$

For each cell *g* within a grid *G* a score value σ_g is calculated as outlined in Eq. (10.4). Our proposed score consists of two main parts; a popularity estimate φ_g reflecting the relative social activity in terms of published Twitter messages, and an activity estimate α_g that describes the likelihood of contents being shared via Twitter.

$$\sigma_g = \begin{cases} \varphi_g^{\frac{1}{\alpha_g}} & \text{if } \alpha_g > 0 \\ 0 & \text{else} \end{cases} \tag{10.4}$$

with

α_g: activity estimate, likeliness of contents being shared
φ_g: popularity estimate, relative social activity
$\sigma_g \mapsto R \geq 0, \alpha_g \mapsto R \geq 0, \varphi_g \mapsto R \in [0,1]$

Our baseline is to estimate the attractiveness of urban areas based on online social activity. Both estimates, activity and popularity, should have a positive impact on the score with growing values. A small value of α along with a small value of φ should yield a low score. Accordingly, both estimates in the upper region of their respective domains should result in a high score. Since both values can hardly be compared directly to each other due to their different domains and distributions, the popularity is risen to the power of the activity's inverse to obtain a score value. Note that it is crucial to employ the inverse due to the domain of α in order to preserve the monotonically increase of σ for increasing values in both α and φ. With increasing popularity value φ, the slope in dependency of α gets steeper. This means, that for areas with a higher popularity, a higher activity estimate leads to increased score values, whereas for lower popularity areas, a high activity estimate only yields lower scores. This behavior is intended to constrain high scores caused by single users in solitary areas. Eqs. (10.5) and (10.6) provide detailed descriptions on how φ and α are being calculated.

The popularity estimate φ_g of a specific grid cell as given in Eq. (10.5) is defined as the share of Twitter messages that relate to a certain cell from the total amount of Twitter messages available. Thus, the resulting value reflects the relative activity weight of the covered district, which can be directly compared among different cells. Whether people only traverse the cell and publish a tweet in the meantime or whether they actually follow some interest in the cell, the popularity value can be regarded as the density of activities.

$$\varphi_g = \frac{\left|\{a :\sim a = g\}\right|}{|A|} \tag{10.5}$$

with

$\varphi_g \mapsto R \in [0, 1]$

The activity estimate α_g is outlined in Eq. (10.6). It is defined as the sum of the Twitter messages' characteristics of unique users U_g, image attachments I_g, video attachments V_g, and linked places P_g, divided by the total amount of Twitter messages in the respective cell.

$$\alpha_g = \frac{U_g + I_g + V_g + P_g}{\left|\{a :\sim a = g\}\right|} \tag{10.6}$$

with

$U_g = |\{u_a : \sim a = g\}|$: the amount of distinct users within the cell
$I_g = \sum_{a: \sim a = g} |I_a|$: the amount of images posted from within the cell
$V_g = \sum_{a: \sim a = g} |V_a|$: the amount of videos posted from within the cell
$P_g = \sum_{a: \sim a = g} |P_a|$: the amount of places posted from within the cell
$\alpha_g \mapsto R \geq 0$

The value U_g reflects the number of unique users identified in the respective cell g, i.e. all distinct users that have posted at least one Twitter message in the cell. This rate supports the distinguishability between areas where people reside longer and those areas where people only stay for a short period of time, measured by their actual activity. If people remain in the same area for an extended time-span, then it is more likely that each unique user publishes more than only one Twitter status update. For example, a residential area that is home to several Twitter users will probably contain multiple Twitter activities by each user during the observation period. Spots with a higher fluctuation of unique users (e.g. airports) are potentially less likely to see the same user over and over again. The rate of Twitter activities that contain an image reference, denoted I_g, is essential for estimating the activity in sharing contents in an area. We expect users to be more likely to publish a tweet that contains a photograph if something in their vicinity exists that they think is worthy of being shared. For example, photographs of food and drinks are very common among Twitter users, as well as pictures of people in front of well-known landmarks. As a result, not only the contents of the photographs contain information on the activity the user is following, but the pure fact that a photograph or some other image is attached to the tweet underlines the importance of the activity. Though videos V_g possibly have a different impact on the attractiveness than images, their meta-information remains valuable. Finally, the amount of directly linked places P_g

is taken into account as well. Note that no textual analysis is performed and the places mentioned are not extracted from the Twitter message itself. Instead, links leading to websites that allow location-sharing are counted, such as Foursquare check-ins.

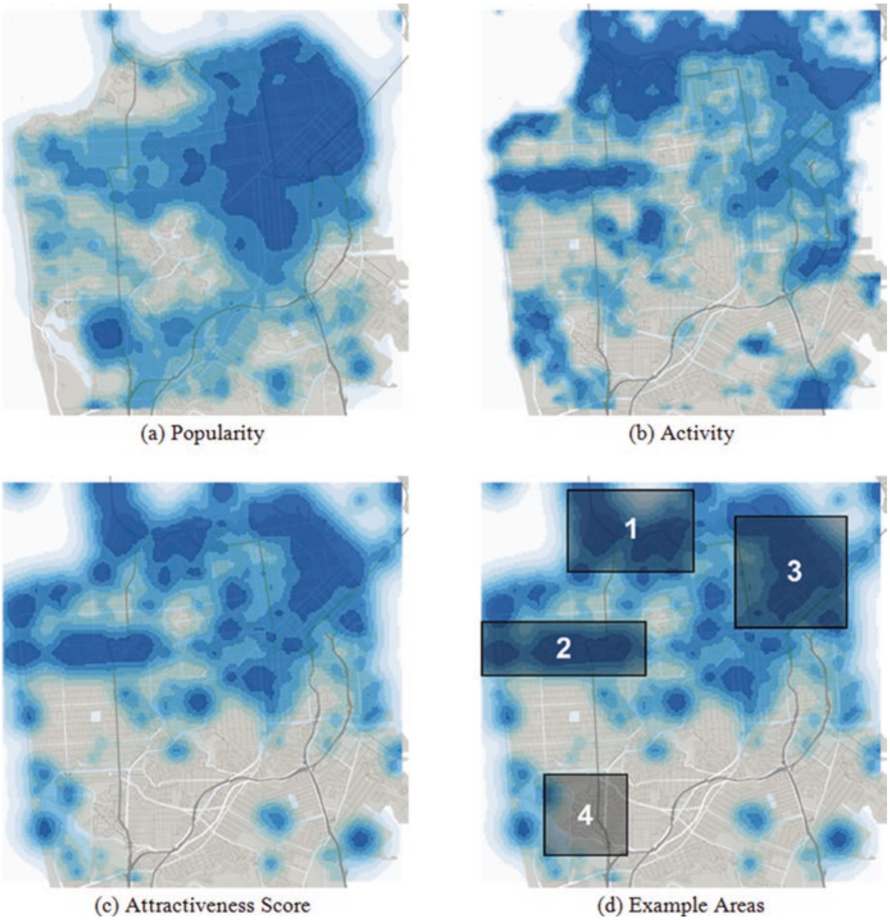
Generally, the popularity estimate represents a sole social activity while the activity estimate reflects the willingness to share additional content by many different users. The visualization of estimated popularity φ and activity α on a fine-grained grid covering San Francisco is shown in Fig. 10.4. Panel (**a**) depicts the popularity estimate. The city center can easily be identified, as well as the university campus in the south-west. While panel (**a**) does not reveal increased values along the coastlines, they are easily recognized in panel (**b**). Furthermore, panel (**b**) also highlights the Golden Gate Park and the entire pier area between the Golden Gate Bridge (north) and the San Francisco-Oakland Bay Bridge (north-east). Bringing these two estimates—popularity and activity—together by calculating a score as defined in Eq. (10.4), we obtain the distribution delineated in panel (**c**). Darker areas account for areas that are likely to attract more people, who in turn are more likely to attach additional contents to their Twitter messages.

10.4 Regression Analysis

Validity of our proposed model can be shown by application of statistical analysis. In this section, we carry out regression analyses to support the explanatory power of our approach. We test our identified activity categories for their certain impact on the measures obtained from the Twitter dataset. Furthermore, we describe to what extent multicollinearity exists in our model and how we cope with it. Our findings are condensed in the last subsection.

As first evidence, areas with high scores are highlighted on a map of San Francisco in panel (**d**) of Fig. 10.4. The first tinted area (1) is the San-Francisco-side of the Golden Gate Bridge and is a tourist hot-spot and important transportation route. Area (2) covers the Golden Gate Park and parts of the Pacific coastline, which are popular recreational areas. Area (3) is the pier area with a view of Alcatraz and Treasure Island and furthermore covers the Market Street and South of Market. Area (4) in the south-west covers the San Francisco State University campus and does not unveil increased attractiveness estimates even though it shows a high degree of popularity (cf. panel **a**). Table 10.3 shows the relevant activity categories (according to those defined in Table 10.2) for the identified areas of interest as framed on Fig. 10.4(**d**) previously.

The identified areas depicted on Fig. 10.4, panel (**d**), and the classification given in Table 10.3 provide some evidence for an actual correlation between activity categories and areas of increased scores within a city. The activity categories can be represented by venues as available from Google Maps, OpenStreetMap, or other online mapping providers. Google provides points of interest of more than 90 different categories, OpenStreetMap provides far more—their common ground is


(a) Popularity

(b) Activity

(c) Attractiveness Score

(d) Example Areas

Fig. 10.4 Visualization of estimated popularity, activity, and attractiveness

however that most of them are anchored to a very specific geographical location. In order to employ points of interest categories for cross-checking our scores, we assigned each of them to none, one, or more of our activity categories (a)–(e). When carrying out geo-spatial analyses, the spatial presence of the entity to be employed should be given in order to obtain valid results. Figure 10.5 shows the presence of all activity categories over the city of San Francisco.

Visual inspection of the categories' distribution in Fig. 10.5 suggests presence of strong multicollinearity. This problem is inherent to the approach and requires special consideration. The rationale for present multicollinearity is twofold. First, between some categories points of interest, there is a large overlap. This category overlap is yielded by the assignment of multiple categories to the same point of interest, such as, for instance, the additional labeling of restaurants with the category 'food'.

Table 10.3 Classification of areas by activity categories

#	Area	Activity categories				
		Sightseeing, landmarks	Nightlife, events, entertainment	Shopping, sports, business	Restaurant, accommodation	Transportation
1	Golden Gate Bridge	×				×
2	Golden Gate Park	×	×			
3	Piers, Market street, and south of Market	×	×	×	×	
4	San Francisco state university					

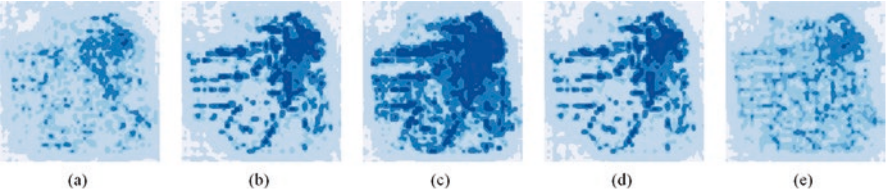


Fig. 10.5 Geo-spatial presence of points of interest for each activity category

The same holds for a souvenir shop, which can be tagged as ‘sight-seeing’ and ‘shop’ at the same time. The fact that categories are not disjoint inevitably leads to implicit correlation. The second reason for multicollinearity among points of interest is their natural geographical clustering. Even when a point of interest is described by a single category, it may naturally be more present in the proximity of other categories. For example considering shopping malls or pedestrian zones, different kinds of establishments often are located close to each other, which results in their densities behaving similarly in a spatial manner.

When dealing with multicollinearity, the most common measure is the variance inflation factor (VIF). However, O’Brien (2007) points out that researchers should be cautious and should not blindly follow VIFs. According to his remarks, VIF thresholds are somewhat arbitrary, and attempts to eliminate multicollinearity often result in more damage than was originally caused. The major effect of a likely multicollinearity in our model is that variances of coefficient estimates are likely to be increased, even though the coefficients themselves are generally unbiased.

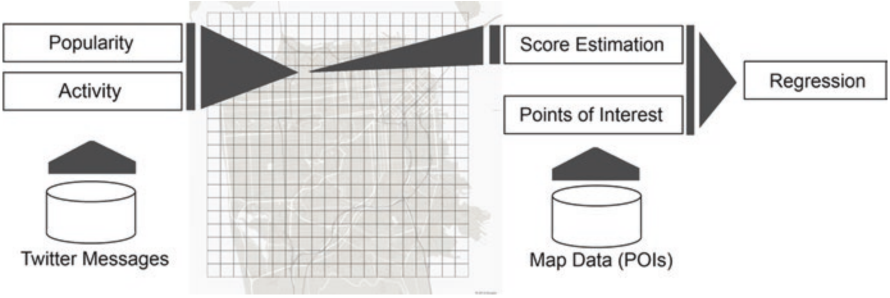


Fig. 10.6 Blending points of interest into the model for validation of results

For analysts, the determination of the source of insignificance is blurred between lack of influence and possibly present multicollinearity. This blur can lead to false refusal of variables that have a significant influence. We cannot avoid multicollinearity in our model and hence have to regard the model as a whole. Variables that are insignificant can become significant by discarding other variables due to multicollinearity. However, variables that show substantial influence truly are significant in face of multicollinearity (O'Brien 2007).

10.4.1 Assessing Explanatory Value of Twitter Measures

In order to support our estimated attractiveness scores in each grid cell, we test our measures against the activity categories using points of interest from map data as a second, independent data source. Despite present multicollinearity, we are able to employ linear regressions for identification of significant influence of variables (O'Brien 2007).

Regressions carried out in this context are based on our complete dataset consisting of more than 600,000 Twitter status messages and more than 60,000 points of interest from the city of San Francisco. The Twitter data spans the 3 months of August through to October 2013. The points of interest are available in many different categories and are reassigned to our five categories (a)–(e). Figure 10.6 outlines how the external data source for result validation blends into our research approach.

The classification of points of interest into our activity categories enables us to test the explanatory power of measures from Twitter messages per cell, such as unique users, the number of images or videos attached, and the number of places shared. Each of these measures is employed as the dependent variable in

Table 10.4 Regression results for twitter measures and activity categories

	Unique users	Images attached	Videos attached	Places shared
Intercept	26.46 (6.73)***	11.38 (5.27)***	0.29 (2.55)*	3.06 (1.73)
(a) Sightseeing, culture, landmarks	−1.40 (−0.98)	0.18 (0.22)	−0.01 (−0.33)	−1.29 (−2.00)*
(b) Nightlife, events, entertainment	36.82 (19.10)***	15.55 (14.68)***	0.17 (3.09)**	15.38 (17.69)***
(c) Shopping, sports, business	0.87 (4.20)***	−0.19 (−1.67)	0.01 (1.09)	0.52 (5.56)***
(d) Restaurant, accommodation	1.77 (2.46)*	1.70 (4.31)***	−0.001 (−0.06)	1.16 (3.591)***
(e) Transportation	−2.37 (−1.39)	−1.88 (−2.01)*	0.03 (0.57)	−1.22 (−1.59)
Adjusted R^2	0.5147	0.3375	0.0415	0.5205

Stated: OLS coefficients, t-statistics in parentheses, based on 2500 observations
Significance levels: * 0.05 ** 0.01 *** 0.001

a linear regression (cf. Eq. (10.7)) and tested for dependency of activity categories (a)–(e).

$$y \sim a + b + c + d + e + \hat{\epsilon} \tag{10.7}$$

with

y: dependent variable; unique users U, attached images I, attached videos V, or shared places P

$a \dots e$: independent variables, the amount of points of interest assigned to the respective category:

a: sightseeing, landmarks
b: nightlife, events, entertainment
c: shopping, sports, business
d: restaurant, accommodation
e: transportation

$\hat{\epsilon}$: the residual error

From the regression results shown in Table 10.4, we can identify different combinations of activity categories responsible for the different shapes in Twitter message characteristics. For instance, category (a) only has a significant influence on places shared. Category (b), most likely due to its entertainment characteristic, has a highly significant impact on all of the Twitter measures. The category (c) of shopping, sports, and business shows an increased impact on high user fluctuation and the likeliness of attaching a current location to a tweet. When in restaurants or at their hotel or hostel (d), people tend to post images as well as link their location, whereas posting videos from these venues is very uncommon. Finally, the transportation category (e) only has a slightly significant influence on posting images. These results support the selection of our measures since we obtain characteristic tweeting behavior based on the respective geographical environment. According to O'Brien (2007), we can interpret the significance of our regression results despite existing multicollinearity. Variables that show significant impact truly are significant, only their coefficient estimates are not directly comparable. We can identify distinctive combinations of categories being significant for each of our measures, effectively meaning that different conditions induce different Twitter behavior. From Tobler's first law of geography, we learned that the impact of geographic relation depends on distance, and from research by Lee et al. (2011) we know that determining an adequate cell size is difficult for geographical analyses based on a grid (Tobler 1970). We used a grid resolution of 50 by 50 cells to cover the city area of San Francisco. Low-resolution grids may blur results and oversimplify the problem, whereas a very high resolution can lead to inappropriate results due to over-fitting. The adequate grid resolution depends on the density of observations, points of interest, population, and presumably additional societal and environmental circumstances.

10.4.2 Findings

Answering the first research question posed at the outset, general social media activity can be utilized to spot socially attractive places. We propose a model that is capable of estimating the areal attractiveness within a city by taking solely Twitter data into account. Our model delivers a score that is composite of popularity of a region and visitor's activity density. From identified activity categories we can infer that social media data implicitly covers aspects of attractiveness delivered by other criteria known from literature. The cross-checking of our model with points of interest from map services renders distinctive combinations of activity categories to be responsible for Twitter characteristics. This observation reveals Twitter activity to be closely related to points of interest in the vicinity and furthermore supports Twitter as a suitable proxy for people's activity within a city. Figure 10.7 depicts our results for San Francisco after being smoothed by a Gaussian filter (3×3). The left-hand panel shows a 3D-plot, the right-hand panel depicts the contours of the same data. Areas of interest can clearly be spotted, while their intensity can be metered by the spike amplitudes. The contiguous areas from the contour plot can directly be applied by a broad variety of domains.

10.5 Concluding Remarks

Due to the ongoing increase in velocity and volume, social media becomes increasingly powerful. In this research, we aimed to explore the methods and measures of identifying places of interest within a city, solely relying on metrics from social media data.

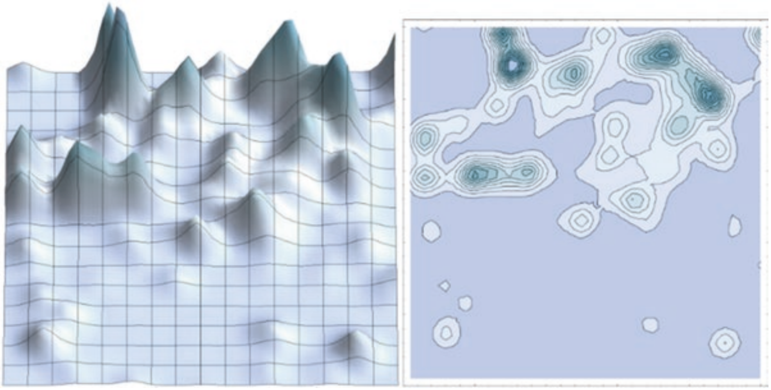


Fig. 10.7 Smoothed attractiveness scores σ in San Francisco

Whenever people generate data unconsciously, this data reflects their true activities and thus is of enormous value in the analysis of behavioral patterns. Furthermore, the spatial relationship between people's activities, their recent location, and points of interest in their vicinity can reveal detailed information on attractiveness of places within a city. Based on relevant related research, we have identified aspects that render urban locations attractive to people, whether they are tourists, residents, city planners, or investors. We identified five different categories describing activities in urban living and, furthermore, were able to prove their existence using a linear regression model on Twitter message characteristics after obtaining evidence from initial visual analysis. We developed a grid-based scoring approach to determine areas of interest within a city. Using points of interest from mapping services as an independent data source, we showed the validity of our scores and identified an appropriate grid resolution to work on when performing analyses on an inner-city level.

With respect to the research questions, we state that social media activity can be exploited to identify urban hot spots of various kinds. The spatial coordinates that are available as a part of the messages from social platforms and services reflect public activity. Additional data from social media messages, such as appended images, videos, or linked places allow inference on the type of activity dominant in certain areas within a city. According to our findings, social media activity in an urban environment can be utilized to spot socially attractive places. Furthermore, we have shown that social media correlates to certain point of interest categories and thus can serve as a proxy for environmental characteristics based on a grid. Socially attractive places relate strongly to environmental conditions, including tourist hot spots, attractions, restaurants, hotels, and many other establishments. Our contribution of spotting areas of interest solely based on social media data is a valuable addition to a broad variety of applications, covering for example city planning, disaster management, city safety, venue recommendation and trip advises, launching businesses, and investment strategies.

Our study and findings are limited by the small fraction of tweets providing geospatial information, since most users opt-out from disclosing their location when publishing a message. Only around 1% of all Twitter messages sent contain the exact geographical location of publication. Nonetheless, our results indicate a strong correlation to the urban environment. In addition, most of the points of interest available from various mapping APIs online, which were used for the validation of results, relate to a certain respective geographical point instead of reflecting detailed areal dimensions. The assignment of point of interest categories to our five activity categories is an aggregation that includes some tolerance. A strict rule-set for assignment between these two different types of categories would be preferable and could possibly lead to more stable and robust results.

In future research, we plan to extend our approach by applying Kernel Density Estimations (KDE) on our grid in order to obtain a valid interpolation that diminishes the explicit separation among grid cells. Additionally, we will analyze whether the Twitter measures can be refined by assigning weights to each of them in order to adapt the attractiveness estimation to individual needs. Further refinement of our results could be achieved by incorporating additional temporal coordinates available from Twitter messages. The social attractiveness of areas or places within a city is likely to vary depending on the time of day and day of week. The inclusion of a temporal dimension allows time-series that could yield valuable insights in to the navigational pattern of people within a city based on the current time. For assessing attractiveness, the general fluctuation of people between grid cells is of increased interest and can further help to spot cells users mainly traverse and those, where users stay for a longer period of time.

Biographies

Joahnnes Bendler (jbendler@geospin.de) is co-founder and CTO of Geospin, a startup company providing geospatial analytics services to corporate and governmental clients. He has received his B.Sc. and M.Sc. in Computer Science from DHBW Stuttgart in 2009 and the University of Freiburg in 2012, respectively. In 2015, he completed his Ph.D. with a dissertation on spatial analytics, with a particular focus on the explanation and prediction of criminal incidents.

http://www.geospin.de/en/about/

Tobias Brandt (tobias.brandt@is.uni-freiburg.de) heads the Smart Cities & Industries research group at the University of Freiburg, Germany, and is co-founder of Geospin. He studied economics in Freiburg and Bologna and completed his Ph.D. at the University of Freiburg in 2015. He has held visiting positions at Lawrence Berkeley National Laboratory in Berkeley, California, and at the University of Texas at Austin. His research on smart grids and smart cities has received best papers awards at ICIS and HICSS and his articles are published in JMIS, the European Journal of Operational Research, Omega, and Business & Information Systems Engineering.

http://www.is.uni-freiburg.de/mitarbeiter/team/tobiasbrandt

Dirk Neumann (dirk.neumann@is.uni-freiburg.de) is full Professor and Chair for Information Systems Research at the University of Freiburg, Germany. He holds degrees in economics from the University of Giessen, Germany, and the University of Wisconsin–Milwaukee, as well as a Ph.D. in economic and business sciences from the Karlsruhe Institute of Technology, Germany. His research interests include the design of electronic markets, novel methods for news and social media analysis, and issues concerning IS and sustainability. His articles have been published in JMIS, the European Journal of Operational Research, Decision Support Systems, ACM Transactions on Internet Technology, and Group Decision and Negotiation.

http://www.is.uni-freiburg.de/mitarbeiter/team/dirk-neumann

References

Arase Y, Xie X, Hara T, Nishio S (2010) Mining people's trips from large scale geo-tagged photos. In: Proceedings of the 18th ACM international conference on multimedia. ACM, New York, pp 133–142

Ballatore A, McArdle G, Kelly C, Bertolotto M (2010) RecoMap: an interactive and adaptive map-based recommender. In: Proceedings of the 2010 ACM symposium on applied computing. ACM, New York, pp 887–891

Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: Proceedings of the 20th international conference on advances in geographic information systems. ACM, New York, pp 199–208

Bendler J, Wagner S, Brandt T, Neumann D (2014) Taming uncertainty in big data. Bus Inf Syst Eng 6(5):279–288

Enright MJ, Newton J (2004) Tourism destination competitiveness: a quantitative approach. Tour Manag 25(6):777–788

Facebook Inc. (2013) Facebook form 10-K annual report

Foursquare (2014) About Foursquare. https://foursquare.com/about

Gearing CE, Swart WW, Var T (1974) Establishing a measure of touristic attractiveness. J Travel Res 12(4):1–8

González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782

Hongzhi Y, Yizhou S, Cui B, Zhiting H, Chen L (2013) LCARS: a location-content-aware recommender system. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 221–229

Hu Y, Ritchie JB (1993) Measuring destination attractiveness: a contextual approach. J Travel Res 32(2):25–34

Instagram (2014) Instagram Press News. http://instagram.com/press/

Jaffe A, Naaman M, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval. ACM, New York, pp 89–98

Jansen-Verbeke M (1986) Inner-city tourism: resources, tourists and promoters. Ann Tour Res 13(1):79–100

Jeffries A (2013) The man behind Flickr on making the service 'awesome again'. http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer

Kozak M, Rimmington M (1999) Measuring tourist destination competitiveness: conceptual considerations and empirical findings. Int J Hosp Manag 18(3):273–283

Lee R, Wakamiya S, Sumiya K (2011) Discovery of unusual regional social activities using geo-tagged microblogs. World Wide Web 14(4):321–349

Leung KW-T, Lee DL, Lee W-C (2011) CLR: a collaborative location recommendation framework based on co-clustering. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information. ACM, New York, pp 305–314

Lew AA (1987) A framework of tourist attraction research. Ann Tour Res 14(4):553–575

Liu B, Fu Y, Yao Z, Xiong H (2013) Learning geographical preferences for point-of-interest recommendation. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 1043–1051

Niedomysl T (2010) Towards a conceptual framework of place attractiveness: a migration perspective. Geogr Ann Ser B, Hum Geogr 92(1):97–109

O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quan 41(5):673–690

Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from Flickr tags. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 103–110

Rogerson RJ (1999) Quality of life and city competitiveness. Urban Stud 36(5–6):969–985

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234

Twitter Inc (2014) About Twitter. https://about.twitter.com/company

Waga K, Tabarcea A, Fränti P (2012) Recommendation of points of interest from user generated data collection. In: 8th IEEE international conference on collaborative computing: networking, applications and worksharing. IEEE, pp 550–555

Ye M, Yin P, Lee W-C, Lee D-L (2011) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 325–334

Yoon H, Zheng Y, Xie X, Woo W (2012) Social itinerary recommendation from user-generated digital trails. Pers Ubiquit Comput 16(5):469–484

Yue Y, Zhuang Y, Li Q, Mao Q (2009) Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: 17th international conference on geoinformatics. IEEE, pp 1–6

Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with GPS history data. In: Proceedings of the 19th international conference on world wide web. ACM, New York, pp 1029–1038

Chapter 11
The Competitive Landscape of Mobile Communications Industry in Canada: Predictive Analytic Modeling with Google Trends and Twitter

Michal Szczech and Ozgur Turetken

Abstract Google Trends, the service that illustrates the trends in Google search activity, has recently received attention form analytics researchers for the prediction of economic trends and consumer behavior. Previous studies used Google Trends to estimate consumption and sales for a particular business, or provide general trends for an economic sector or industry. This study reported here differs from these attempts as it aims to estimate the performance of a single player in an industry by not only trends related to that player, but also those of its competitors. Further, these trends have been modified by Twitter based sentiment scores. It is demonstrated that the incorporation of competitive factors results in better estimates by as much as 5% while the addition of a Twitter sentiment score is not beneficial. The Twitter related findings could be because the tweet volumes in the particular industry that was examined are low and volatile.

Keywords Predictive analytics • Social media data • Market model

11.1 Introduction

Modern day people turn to the Internet more and more often to find information for making decisions and educating themselves about various topics. At the end of 2014, about 42% of world's population had experienced at least one internet based

M. Szczech
CGI, 750 Hillman Crescent, Mississauga, ON, Canada, L4Y2J2

Ted Rogers School of Management, Ryerson University,
350 Victoria Street, Toronto, ON, Canada, M5B 2K3
e-mail: mszczech@gmail.com

O. Turetken (✉)
Ted Rogers School of Management, Ryerson University,
350 Victoria Street, Toronto, ON, Canada, M5B 2K3
e-mail: turetken@ryerson.ca

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information Systems 21, DOI 10.1007/978-3-319-58097-5_11

service. In North America, the penetration of the Internet is close to 87% of the population whereas in Canada the figure was even higher at 94% of the population in 2014 (Internet World Stats 2014). For most people, the Internet is practically synonymous with the Web. At the time of the writing, the number of active websites on the Internet is estimated to be over one billion (Internet Live Stats 2015). The sheer amount of web-based content emphasizes the importance of powerful and effective search engines. Over 65% of all searches on the Internet are conducted by Google, which amounts to over 3.5 billion searches per day. Like other search engines, Google stores the search queries that users submit along with geolocation information. Most of this data are available to the wide public through a service called Google Trends. The data are updated daily, even hourly, through a special service called Google Trends Hourly. Typically, whatever people search for in a given region of the world tends to somewhat reflect events in that region. This provides rich and timely data for predictive analytics, which many modelers have attempted to use.

Meanwhile, the typical Web-content that can be reached through a service like Google search (and hence Google trends) is static in the sense that it is not updated and shared as often as Web 2.0 content. Web 2.0, on the other hand, allows for the creation of dynamic, highly interactive and content rich applications that differ from static websites that constituted Web 1.0. These new Internet technologies gave rise to social networking and social media sites. The most popular social networking application in the world is Facebook, which currently has about 1.3 billion active monthly users (Statistic Brain 2015a). Other popular social networking and social media platforms are YouTube, Twitter, LinkedIn, Pinterest, Instagram, and Google+. Twitter allows posting short messages, not longer than 140 characters, called “tweets”. It has about 650 million users worldwide. On average, the world creates about 58 million tweets per day (Statistic Brain 2015b).

While these changes in technology have shaped how individuals find information and communicate it, the ability to store information about millions of users and their online interactions in searchable databases has significant influence on various aspects of business, especially for analysts who need market data, but suffer from the fact that market data that are collected through traditional methods arrive too late and at a high cost. To satisfy their need for timely market indicators and analysis, they utilize the above mentioned databases maintained by search engines such as Google and Yahoo and social networking platforms such as Twitter and Facebook. Google search volume data are available to everyone through the Google Trends interface (https://www.google.com/trends/) and Twitter data are publicly available to everyone through searchable databases holding almost the entirety of tweets dating many years back. These immense databases provide a versatile environment for many forms of descriptive and predictive analytics. As will be further detailed in the “Literature Review” section of the chapter, Google Trends and Twitter have been used to capture market trends and used for predicting the demand for various goods and services. However, somewhat surprisingly, there is little to no research that we have been able to encounter that shows the usefulness of web search and social media based analytics on providing a view of the competitive environment within a certain industry and geographical area. Our research aims to fill that void and determine if Google Trends and Twitter data can be effectively used to assess the competitive landscape of a certain industry within a given geographical area.

The specific context of our study is the mobile communication service provision market in Canada. Although each of the main Canadian mobile communication service providers knows the market share it holds relative to its competition for a given quarter, their expressed desire is to have insight as to how well they are performing relative to the competition on a more frequent (e.g. weekly or monthly) basis. To provide such insight for this market, we observe the following characteristics of the mobile telecommunications industry in Canada: (1) there exist a fairly stable and small number of competitors, and (2) the market is already mature and the overall consumer growth year by year is small. These characteristics help in the formulation of our predictive models. By using Google Trends, regression models for predicting the market share changes of each of the main mobile service providers in Canada are developed. We then attempt to improve the models by adding competition variables. Finally, we incorporate Twitter sentiment scores into our models to detect if those scores explain some of the variance in the dependent variable that is not already explained by Google Trends data.

Armed with this information, decision makers in mobile telecommunication service providers can assess the presence and importance of the threat their direct competition poses to their organization. In turn, they can adjust marketing and sales strategies by quickly assessing their operator's performance relative to the competition. Once the competitive threats are identified, a closer analysis of the competitors' activities may help identify the source of their success, which in turn helps strategic planning.

The rest of this chapter is organized as follows. In the next section, we provide a brief review of the relevant academic literature. Section 11.3 details our approach to the predictive modeling exercise. In Sect. 11.4, we present the results of the data analysis, which are discussed further in Sect. 11.5. Section 11.6 presents concluding remarks and directions for future research.

11.2 Literature Review

11.2.1 Consumer Related Research Involving Google Trends Data

Many studies have demonstrated a relationship between online search and offline sales. Chandukala et al. (2014) have shown that the market potential or demand, which is expressed by "latent interest", can be measured by analysing online search. Their findings have been drawn from relationships between product search and product sales, search for jobs and unemployment rates, search for a flu medicine and incidence of flu in a given region and time, and search for cancer and incidence of cancer.

In 2009, Hal Varian, then Google's Chief Economist, and Hyunyoung Choi showed that short-term forecasts of automotive and home sales could be significantly improved by using Google search data as provided by the Google Trends framework. Through this example, they demonstrated the potential of employing Google Trends search data in research. Later, in 2012, they extended their study and showed the potential of Google Trends in predicting the sales of vehicles and motor part sales (Choi and Varian 2012). It was demonstrated that it is possible to create Google Trends based consumer consumption indicators that outperform most commonly used survey-based consumer consumption indicators such as Michigan Consumer Sentiment Index (Vosen 2011). It has been confirmed that Google search data can improve forecasting models in general, and specifically that it can improve commercial real estate price and demand forecasts (Marian et al. 2014).

Google Search seems to reveal the intentions of user actions. For example, there is correlation between Google search volume for marijuana and consumption of marijuana by youth (Cavazos-Rehg et al. 2015). With the use of Google Trends it is also possible to predict, with reasonable margin of error, box office movie results days, or even weeks, before the showings (Goel et al. 2010). It was also shown that Google Search Volume Index (SVI) can measure the retail market investor desire to buy certain stocks (Da et al. 2011). Jun et al. (2014) have developed a model for forecasting the sales of Toyota Prius based on search traffic and some environmental variables. Our study expands these typical search-volume based models by incorporating search-volume based competitive variables. For example, to forecast Toyota Prius sales, it would be beneficial to see where the sales of direct competition such as Chevrolet Volt and Honda CR-Z are going. In general, such competitive factors tend to be omitted in most studies utilizing search traffic for consumer choice estimation and forecasting. This study addresses this gap by assessing the impact of competitive factors on the ability to forecast performance of three main competitors in an oligopolistic market. The introduction of a competitive factor is especially important in industries that are characterized as oligopolies, because oligopolies are characterized by a high level of mutual interdependence between firms (Goel 2007), thus changes in offering by one firm my affect the performance of the other. For example, if a particular business improves its offering or lowers price and the others do not, then that business increases its sales at the expense of the others. In the Canadian wireless communications industry there exist a limited number of competitors, and barriers to entry are high. In fact, about 90% of the market is controlled by three main players. Our models aim to capture the degree of interdependence between the provider under consideration and its direct competition by adding competition variables to the prediction. If there is a high degree of interdependence between the providers as expected, then the competitive variables should have significant impact on market performance.

One weakness of search volume analytics is the fact that we know the scale of the interest, but are unaware of the underlying sentiment. One solution to this problem would be analyzing the combination of search terms with sentiment indicators (Shawn and Stridsberg 2015). Such an approach can be used to predict financial market changes such as financial index moves or financial crises (Preis et al. 2013).

Similarly, Google Trends data were used to create an investor sentiment index that was able to closely reflect investor sentiment as obtained by traditional means (Beer et al. 2013). Another approach to estimate user sentiment related to their topics of search interest is the use of the collective wisdom available in social media. The next section reviews research on that latter stream.

11.2.2 Use of Social Media and Twitter in Predictive Models

Recently, there have been a number of big data predictive analytics studies that incorporated Twitter Analytics data. These include predictions of stock market moves, financial markets, election results, crime volumes, health outbreaks, or even unemployment. In 2013, Matthew S. Gerber, attempted to answer the question of whether Twitter data can be used to predict crime in a large US city (Gerber 2014). He argued that such information could assist decision making processes. That study was not able to identify a significant correlation, but when using Twitter data to support other crime predicting models, it was able to improve crime prediction in 19 out of 25 crime types. One of the main reasons why it was difficult to obtain an effective prediction framework could be the fact that tweets cannot be easily grouped by location on such a granular basis such as neighborhood-by-neighborhood in a given city. Possibly better results could be obtained if GPS-tagged tweets were used in the study.

Research aiming to predict election results from social media has indicated that the volume and sentiment of electoral party mentions on Twitter reflects the election result for that party (Tumasjan 2011). It has also been shown that the size of the social media network of an electoral candidate can be indicative of an election result if the election is “closely” contested (Cameron et al. 2015).

In the world of commerce, services and retail, it has been established that a particularly effective way of affecting sales is the word-of-mouth (WOM) (Engel et al. 1969). Not surprisingly, social networking sites such as Twitter and Facebook have been identified as possible environments where WOM could be shared. It was established that microblogging websites are widely used as a form of electronic word-of-mouth (eWOM) with regards to a brand, and they disseminate brand sentiments among the members of social networks (Jansen 2009). Thus, it can be expected that Twitter and other microblogging websites allow sharing of user sentiment about a certain product, and can influence user purchasing decisions. Not surprisingly, some studies try to predict sales based on Twitter data analysis. It was demonstrated that social media data can be used to predict quarterly sales of iPhones with average error of 5–10% (Lassen 2014). It is important to note that sentiment analysis tools and techniques are still evolving and currently there are no conclusive studies that would indicate which approach to sentiment analysis, especially for short texts such as tweets, is the most effective (Lak and Turetken 2014).

11.3 Predictive Modeling

To reiterate, the primary objective of our research is to show that the big data archived by Google Trends and Twitter can be utilized to give business managers an early indication of how well they are doing against their direct competition at any point in time. The Canadian wireless telecommunications industry is dominated (90% market share) by three major players, namely Rogers Wireless, Telus Mobility and Bell Mobility with over eight million subscribers each. The only independent company that tries to compete with the big three on a Canada-wide front, and is present in more than four provinces, is Wind Mobile Corporation. Its size is about a tenth of the size of a single provider from the big three, and currently sits just above 800 thousands subscribers. The other competitors are either constrained geographically (Sasktel, MTS), or too small to be considered serious contenders (Windmobile, Mobilicity and others). Therefore, in this study, we consider only Rogers Wireless, Telus Mobility, and Bell Mobility.

Because the barriers to entry are high, the market is already significantly penetrated, and there is a limited amount of new consumers, new subscribers that join one competitor are likely to be ones that leave another competitor. In this highly regulated landscape, the management for each wireless telecommunication company faces three main concerns:

- To retain as many of their existing subscribers as they can,
- To attract existing subscribers form competition, and to a lesser degree
- To attract first-time wireless subscribers.

Success, in all of the three objectives above, can be measured by the change in total subscribers in any given quarter. This indicator is called “net new subscribers” (NNS). Each of the main wireless services carriers in Canada publishes the number of net new subscribers they were able to attract quarterly. Therefore, each provider has the opportunity to compare their net new subscribers against the competition, more or less, every 3 months. Three months, in some cases, can prove to be too long. If a provider is not doing well compared to the competition, an early detection of a threat could lead to early intervention through marketing and service-offering adjustments. In addition, in this type of industry, what is important besides the growth of customers, or subscribers, is the actual market share. If a company has a decent growth in subscribers, but is losing its market share overall, this can hardly be considered a success. Therefore; it is beneficial for any provider to be able to determine how well it is doing when compared to the competition. It is not just its own NNS figure, but also its current position against the competition that is important. If a decision maker can identify which competitor has made the greatest progress in a given period of time (e.g. a month), then (s)he can try to understand what stands behind their success, and try to balance this with his/her company’s offering and marketing strategy in a more timely fashion.

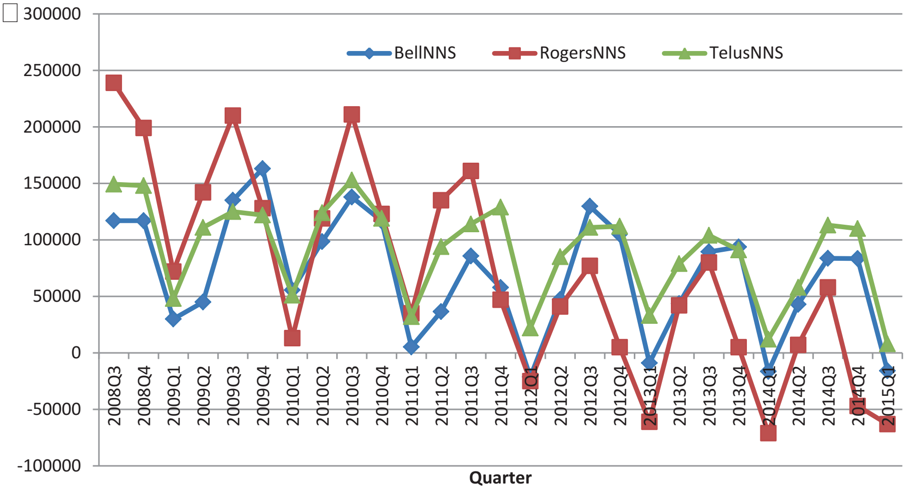


Fig. 11.1 NNS over time for Bell, Rogers and Telus

11.3.1 Market Data

To serve as the basis of historical values for the dependent variable, the quarterly results for net new subscribers, for each of the Canadian wireless carriers are widely available. The data for years 2006 through 2015 can be collected through Canadian Wireless Telecommunications Association (CWTA). Therefore, for each company, we can collect net new subscriber data for 37 quarters. The dependent variable that is used in this study is net new subscribers (NSS):

$$\text{NNS for a wireless provider} = \left(\#\text{of New Subscribers}\right) - \left(\#\text{of Lost Subscribers}\right) \tag{11.1}$$

This variable is commonly used in the wireless communications industry to measure its overall competitive standing, as it directly affects the overall market share a given provider holds for the industry in the specific geographical area. Figure 11.1 illustrates the NNS data over the 37 quarters.

Our first observation about the data is that NNS is highly seasonal, and has a general downward trend over time for all three competitors, with the highest rate of decrease experienced by Rogers. This can be partly because smaller competitors like Windmobile, Mobilicity, Shaw and Sasktel are becoming more and more aggressive in their efforts to capture market share that traditionally belonged to the top three providers, and that the size of the overall mobile service market in Canada is fairly stable. The reason for the particular decline in Rogers' NNS could be that,

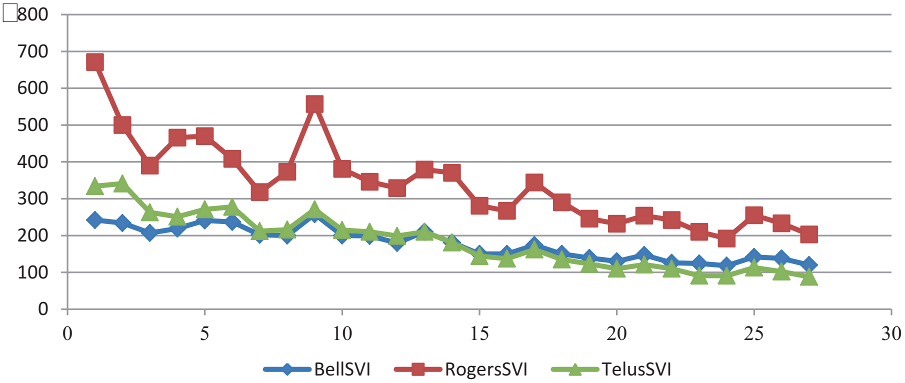


Fig. 11.2 SVI over time for Bell, Rogers and Telus

Rogers is, and historically has been, the largest wireless services provider in Canada. This may cause most of the competitors to aim their offering and marketing campaigns against the leader, resulting in a higher average customer loss for Rogers than the other two competitors.

One of the main premises of this study is that trends in market data can be modeled by Google Trend data as was done in previous literature. It is possible to obtain search volume data for each of the quarters, and for each of the competitors through Google Trends. Google Trends search volume data can be fine-tuned to a particular region and industry. One characteristic of the data is the fact that it is approximated relative to the collection of all searches. In general, what Google Trends returns is a search volume index (SVI) where it returns a number for a given time range, relative to the highest search volume for all the included search terms. Therefore, only one search volume result would have an SVI of 100, that is, the highest one. The rest of the SVIs would be between 0 and 100. For our research we selected Canada as the geographical location, "Internet and Telecom" as the category, and "Mobile and Wireless" as the subcategory. It is important to note that while Google does really well in splitting search queries into geographical location, the algorithm it uses for sorting the searches into various categories and subcategories is not guaranteed to be fully accurate as it infers the category based on the search terms the users enter. Finally we used "Bell", "Rogers" and "Telus" as search terms we wished to compare. Google Trends only provides weekly data; therefore we summed the results up and grouped them into quarters. Figure 11.2 displays SVI trends for each provider.

Based on previous literature, it is reasonable to expect that there will be a positive correlation between the number of searches that included a given provider name and its ability to attract net new subscribers. However as seen in Fig. 11.2, Google Trends data do not exhibit the same pattern of seasonality as the NNS data displayed in Fig. 11.1. This makes it essential to include a seasonality factor in the formulation of subsequent models. It is possible that other industries do not exhibit such seasonality, and a seasonal factor is not needed in a more general model whereas in this

Table 11.1 Average seasonal weight factor for each provider (Eq. 11.3)

	Bell	Rogers	Telus
ASWF.Q1	0.153	0.112	0.462
ASWF.Q2	0.757	1.016	1.051
ASWF.Q3	1.493	1.796	1.226
ASWF.Q4	1.596	1.076	1.261

industry, the first quarter of the year seems to be a low season for all the providers and quarters 3 and 4 are where the highest number of net new subscribers is registered. The fourth quarter includes December holidays, and is characterized by the highest levels of consumer spending, which translates to higher NNS in the wireless industry. Third quarter is the back-to-school and back to work quarter, which often leads to higher sales in retail and services industry. Quarter one is generally the slowest season for retail and services as the consumers pull back from the high spending that characterizes quarter four. In addition, quarter one may have the highest number of deactivations or cancellations due to some "over-purchasing" in December.

The Seasonal Weight Factor (SWF) was approximated by the averages for all the known historical quarters. The average seasonal weight ASW for each of the four quarters can be obtained by calculating the average NNS for each quarter, (ANNS(q)) and dividing it by the sum of averages for each quarter as follows:

$$\mathrm{ASW}\left(\mathrm{Qi}\right)=\frac{\mathrm{ANNS}\left(\mathrm{Qi}\right)}{\mathrm{ANNS}\left(\mathrm{Q1}\right)+\mathrm{ANNS}\left(\mathrm{Q2}\right)+\mathrm{ANNS}\left(\mathrm{Q3}\right)+\mathrm{ANNS}\left(\mathrm{Q4}\right)} \tag{11.2}$$

If there is no seasonality, that is if all four quarter ANNS(q) are equal, then the formula will always result in a value of 0.25. Therefore, the respective average seasonal weight factor (ASWF) can be calculated by dividing the ASW by 0.25:

$$\mathrm{ASWF}\left(\mathrm{q}\right)=\frac{\mathrm{ASW}\left(\mathrm{q}\right)}{0.25} \tag{11.3}$$

The average seasonal weight factors for each provider and for each quarter are shown in Table 11.1. Once again, if there was no seasonality, all of the values in the table would be equal to 1. Therefore, the average of the ASWF values over the four seasons for each competitor is 1.

To predict NNS by the seasonality factor alone we formulate our base model, ***Model 1:***

$$\mathrm{NNS}_{\mathrm{i}}\left(\mathrm{q}\right)=B_{\mathrm{i}}^{*}\mathrm{ASWF}_{\mathrm{i}}\left(\mathrm{q}\right)+A_{\mathrm{i}}+\varepsilon_{\mathrm{i}}$$

where $\mathrm{NNS}_{\mathrm{i}}$ (q) is the number of net new subscribers for a given provider i for quarter q, B_{i} is the coefficient for seasonality, $\mathrm{ASWF}_{\mathrm{i}}$(q) is the average seasonal weight factor for provider i for quarter q, A_{i} is the model constant, and ε_{i} is the error term.

As the next step, we formulate a model using both time series (ASWF) and causal (SVI) variables to attempt improving predictive power by capturing both the seasonality and the trend observed in NNS. Empirical results in previous literature suggest a linear relationship between SVI and NNS. There is no theoretical reason to argue that this relationship should be nonlinear either, therefore ***Model 2*** is formulated as a linear model as follows:

$$\mathrm{NNS_i}\left(\mathrm{q}\right)=B1_{\mathrm{i}}{}^{*}\mathrm{ASWF_i}\left(\mathrm{q}\right)+B2_{\mathrm{i}}{}^{*}\mathrm{SVI_i}\left(\mathrm{q}\right)+\mathrm{A_i}+\varepsilon_{\mathrm{i}}$$

where the added term $B2_{\mathrm{i}}$ is the coefficient for the search volume index, and $\mathrm{SVI_i}$ (q) is the Google search volume index for provider i for quarter q.

11.3.2 Competitor Effects

The next step is to add competitive factors to our model. In an industry like this where there are a fixed number of competitors and fairly stable target market, performance of one player is expected to be affected by the performance of the competition. This leads us to believe that there will be a correlation, very likely negative, between the number of Google searches for the target company's NNS and those of its competitors.

Following from our discussion of the market, each target provider (i) has two direct competitors (j, k). We model the relationship between a company's market movement in the presence of its competitors with ***Model 3***

$$\mathrm{NNS_i}\left(\mathrm{q}\right)=B1_{\mathrm{i}}{}^{*}\mathrm{ASWF_i}\left(\mathrm{q}\right)+B2_{\mathrm{i}}{}^{*}\mathrm{SVI_i}\left(\mathrm{q}\right)+B2_{\mathrm{j}}{}^{*}\,\mathrm{SVI_j}\left(\mathrm{q}\right)+B2_{\mathrm{k}}{}^{*}\,\mathrm{SVI_k}\left(\mathrm{q}\right)+\mathrm{A_{ijk}}+\varepsilon_{\mathrm{ijk}}$$

where the added term $B2_{\mathrm{j}}$ is the coefficient for the search volume index for provider j (competitor 1), $B2_{\mathrm{k}}$ is the coefficient for the search volume index for provider k (competitor 2), $\mathrm{SVI_j}$ (q) is the Google search volume index for provider j for quarter q, and $\mathrm{SVI_k}$ (q) is the Google search volume index for provider k for quarter q.

11.3.3 Effects of Sentiments and Twitter Data

The sentiment (ST) expressed by consumers about a provider in their online activity could indicate whether the volume of search expressed in the trends is a positive or negative indicator. By their very nature, web search tools are indexing engines: they only store pointers to content in their databases. As is more than typical, when web content changes (or is altogether removed) over time, so do the results to a query. As a result, it is impossible to replicate the results to a *historic* web query, which means one could not reproduce the content to which a consumer was exposed when they searched for a certain term in the past. Therefore; as it was noted in the literature review section, a shortcoming of using Google Search Volume (SV) as an indicator for sales or consumption is the fact that the volume information does not carry with it the sentiment of the search results. Even though the volume does reflect interest in a product or a brand, we miss information about the nature, *i.e.* the sentiment, of the interest.

Meanwhile, content on social media, especially on Twitter, is not overwritten unless a user specifically chooses to do so. Twitter Analytics database holds all the "non-deleted" tweets back to year 2006. Therefore past sentiments regarding a given topic can be identified by analyzing Twitter data. Historical data from Twitter is not freely available; however the Twitter website makes it possible to run search queries specifying dates and key search terms. This led us to collect historical tweets from the Twitter website through a series of individual searches. Although very time-consuming, this process proved to be viable for collection of tweets. One weakness of the Twitter data is the fact that very few of the historical tweets can be localized, which means the Twitter searches we performed were "global" in nature. To narrow down the results, full names of the companies were included in the search query instead of just the short name. For example, instead of just searching for "Bell" the search included "Bell Mobility" as a search term. After examining the returned tweets, it was verified, by examining the content and the authors of the tweets, that what is included in the result are tweets that are relevant, and were originated by Canadian users.

After the tweets for each of the three competitors for every quarter were collected, it was noted that the volume of tweets that were returned for years prior to quarter 3 (Q3) of 2008 were rather negligible. This could be due to the fact that Twitter adoption in Canada did not reach significant scale before this date. Therefore, the time span of the analysis was narrowed down to 27 quarters, from Q3 of 2008 to Q1 of 2015.

After the initial processing of the tweets as described, sentiment analysis of the tweet contents was performed. For the purposes of this study, a sentiment analysis tool called "SentiStrength" was used. SentiStrength returns two scores for each Tweet it analyzes: a negative score, a number between −1 and −5, and a positive score, a number between 1 and 5 where "–5" is the most negative score and "5" is the most positive score. For each tweet, these two scores were added to obtain a single sentiment score in the range of −4 to 4. About 1000 scores were reviewed manually and it was observed that scores between −1 and 1 were mostly associated with tweets that were perceived to be "neutral". Therefore, a decision was made to classify the tweets into Negative, Positive and Neutral categories. Any tweet with a score below −1 was considered Negative, a score between −1 and 1 (inclusive) was Neutral, and a score greater than 1 was Positive. Positivity to negativity ratio, which is simply the number of positive tweets divided by the number of negative tweets (Pos/Neg), is a commonly used measure of the general sentiment of a collection of tweets. Another way of tallying sentiment scores is to use the ratio of positive tweets to tweets with opinions. This approach ignores neutral tweets for sentiment measurement as those do not directly contribute to the shaping of the sentiment. In this study, we followed this latter approach to calculate sentiment scores for each of the providers:

$$\text{Sentiment ST} = \frac{\text{Pos}}{\text{Pos+Neg}} \tag{11.4}$$

In the last model we formulated, we included this sentiment variable in an additive fashion as none of the multiplicative models yielded desirable results. Formulating a linear model also simplifies the subsequent analyses and makes this model comparable to the previous three. The resulting model, ***Model 4***, is as follows:

$$\begin{aligned}\text{NNS}_{\text{i}}\left(\text{q}\right) = B1_{\text{i}}^{*}\,\text{ASWF}_{\text{i}}\left(\text{q}\right) + B2_{\text{i}}^{*}\,\text{SVI}_{\text{i}}\left(\text{q}\right) + B2_{\text{j}}^{*}\,\text{SVI}_{\text{j}}\left(\text{q}\right) + B2_{\text{k}}^{*}\,\text{SVI}_{\text{k}}\left(\text{q}\right) + \text{ST}_{\text{i}}\left(\text{q}\right)\\ + \text{A}_{\text{ijk}} + \varepsilon_{\text{ijk}}\end{aligned}$$

where the new term STi(q) is the Twitter sentiment for provider i in quarter q as displayed in Eq. (11.4).

11.4 Results

IBM SPSS Statistic software was used to analyze the data. Normal P-P plots showed that the distribution of the NNS variables is reasonably normal (Fig. 11.3). The models were estimated separately for all three wireless service providers: Bell, Rogers and Telus through multiple linear regression models based on historical quarterly NNS data. The relationships between these players are complex and diverse, which implies competitor one may be affecting competitor two in a completely different way than it affects competitor three. The predictive power of the

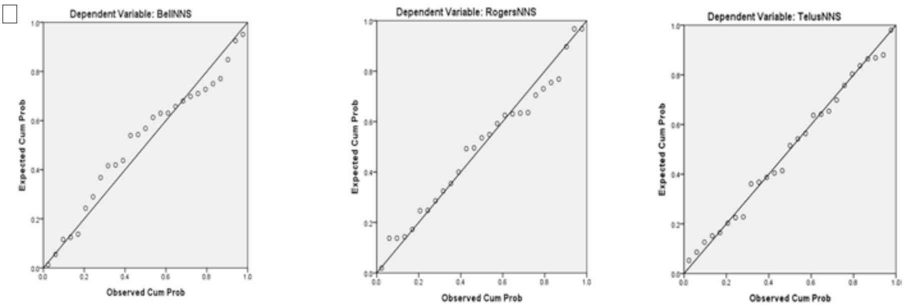


Fig. 11.3 Normal P-P plots

Table 11.2 Adjusted R^2 values for Bell

Model summary[a]				
Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.861[b]	0.741	0.730	27,035.0451
2	0.922[c]	0.850	0.838	20,966.8961
3	0.930[d]	0.865	0.840	20,825.0304
4	0.933[e]	0.870	0.839	20,867.0641

[a] Dependent Variable: BellNNS
[b] Predictors: (Constant), BellASWF
[c] Predictors: (Constant), BellASWF, BellSVI
[d] Predictors: (Constant), BellASWF, BellSVI, RogersSVI, TelusSVI
[e] Predictors: (Constant), BellASWF, BellSVI, RogersSVI, TelusSVI, BellST

Table 11.3 Adjusted R^2 values for Rogers

Model summary[a]				
Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.664[b]	0.441	0.419	68,135.1150
2	0.934[c]	0.872	0.861	33,296.5604
3	0.964[d]	0.929	0.916	25,894.8570
4	0.964[e]	0.930	0.913	26,338.3258

[a] Dependent Variable: RogersNNS
[b] Predictors: (Constant), RogersASWF
[c] Predictors: (Constant), RogersASWF, RogersSVI
[d] Predictors: (Constant), RogersASWF, RogersSVI, BellSVI, TelusSVI
[e] Predictors: (Constant), RogersASWF, RogersSVI, BellSVI, TelusSVI, RogersST

Table 11.4 Adjusted R^2 square values for Telus

Model summary[a]				
Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.901[b]	0.811	0.804	19,075.2100
2	0.974[c]	0.948	0.943	10,245.3655
3	0.979[d]	0.959	0.951	9489.6418
4	0.980[e]	0.960	0.951	9557.3387

[a] Dependent Variable: TelusNNS
[b] Predictors: (Constant), TelusASWF
[c] Predictors: (Constant), TelusASWF, TelusSVI
[d] Predictors: (Constant), TelusASWF, TelusSVI, RogersSVI, BellSVI
[e] Predictors: (Constant), TelusASWF, TelusSVI, RogersSVI, BellSVI, TelusST

models for Bell, Rogers and Telus are in Tables 11.2, 11.3 and 11.4 respectively, while Tables 11.5, 11.6, and 11.7 provide the regression analysis results. Analysis of the independent variables indicates a high level of collinearity between BellSVI, RogersSVI and TelusSVI. The collinearity of provider SVIs is not a significant issue for this study, because we are not trying to assess how much each of the variables

Table 11.5 Regression analysis results for Bell

Model		Unstandardized coefficients		Standardized coefficients		Collinearity statistics	
		B	Std. error	Beta	Sig.	Tolerance	VIF
1	(Constant)	−5768.789	10,241.676		0.578		
	BellASWF	73,892.696	8743.072	0.861	0.000	1.000	1.000
2	(Constant)	−71,090.456	17,493.243		0.000		
	BellASWF	65,679.966	7058.128	0.765	0.000	0.923	1.084
	BellSVI	414.133	98.814	0.345	0.000	0.923	1.084
3	(Constant)	−83,470.678	28,096.986		0.007		
	BellASWF	69,132.983	7624.626	0.805	0.000	0.780	1.282
	BellSVI	720.190	352.303	0.599	0.053	0.072	13.961
	RogersSVI	−151.910	108.492	−0.340	0.175	0.105	9.561
	TelusSVI	34.053	222.765	0.049	0.880	0.059	17.032
4	(Constant)	−49,595.000	45,295.369		0.286		
	BellASWF	69,505.554	7649.976	0.810	0.000	0.778	1.285
	BellSVI	601.912	374.122	0.501	0.123	0.064	15.681
	RogersSVI	−144.701	108.972	−0.324	0.198	0.104	9.607
	TelusSVI	67.035	225.872	0.097	0.770	0.057	17.440
	BellST	−57,506.738	60,235.214	−0.085	0.351	0.776	1.289

Table 11.6 Regression analysis results for Rogers

Model		Unstandardized coefficients		Standardized coefficients		Collinearity statistics	
		B	Std. error	Beta	Sig.	Tolerance	VIF
1	(Constant)	−25,980.580	25,210.662		0.313		
	RogersASWF	95,741.599	21,545.129	0.664	0.000	1.000	1.000
2	(Constant)	−17,1487.949	20,351.726		0.000		
	RogersASWF	52,511.594	11,576.575	0.364	0.000	0.827	1.209
	RogersSVI	554.068	61.683	0.722	0.000	0.827	1.209
3	(Constant)	−15,9847.668	35,614.074		0.000		
	RogersASWF	70,791.232	10,308.706	0.491	0.000	0.631	1.585
	RogersSVI	−13.828	143.913	−0.018	0.924	0.092	10.880
	TelusSVI	839.596	299.399	0.711	0.010	0.050	19.899
	BellSVI	54.810	447.959	0.027	0.904	0.068	14.599
4	(Constant)	−14,2219.908	49,830.147		0.010		
	RogersASWF	71,594.924	10,600.675	0.497	0.000	0.617	1.620
	RogersSVI	−29.773	149.614	−0.039	0.844	0.088	11.367
	TelusSVI	826.092	305.653	0.699	0.013	0.050	20.046
	BellSVI	80.034	458.254	0.039	0.863	0.068	14.767
	RogersST	−34,565.528	67,096.470	−0.035	0.612	0.708	1.412

Table 11.7 Regression analysis results for Telus

Model		Unstandardized coefficients		Standardized coefficients	Sig.	Collinearity statistics	
		B	Std. error	Beta		Tolerance	VIF
1	(Constant)	−25,259.464	11,805.033		0.042		
	TelusASWF	116,492.411	11,241.138	0.901	0.000	1.000	1.000
2	(Constant)	−54,642.892	7347.173		0.000		
	TelusASWF	10,6530.132	6167.428	0.824	0.000	0.958	1.043
	TelusSVI	214.726	27.126	0.377	0.000	0.958	1.043
3	(Constant)	−72,537.868	13,121.357		0.000		
	TelusASWF	10,0911.939	6222.900	0.780	0.000	0.808	1.238
	TelusSVI	−15.565	100.781	−0.027	0.879	0.060	16.788
	BellSVI	207.120	160.133	0.208	0.209	0.072	13.891
	RogersSVI	84.753	49.186	0.229	0.099	0.106	9.463
4	(Constant)	−87,720.657	22,560.777		0.001		
	TelusASWF	10,0171.674	6330.387	0.774	0.000	0.792	1.263
	TelusSVI	32.753	116.998	0.058	0.782	0.045	22.307
	BellSVI	125.393	188.938	0.126	0.514	0.052	19.065
	RogersSVI	104.197	54.793	0.282	0.071	0.086	11.578
	TelusST	28,873.751	34,773.994	0.070	0.416	0.270	3.707

affects the final output, but rather their combined impact on our ability to predict NNS for each quarter. In other words, our models are predictive rather than exploratory therefore collinearity is not a severe violation. However; this also implies that the regression coefficients for the competitor models (Models 3 and 4) should not be used to make any conclusions about the effect of competition in the market.

For Bell, as seen in Table 11.2, if only the seasonal weight factor is used as an independent variable, the adjusted R^2 is 0.73 indicating that seasonality plays a major role in determining market share. When we add Google Trends SVI to the model, the adjusted R^{2}increases to 0.838; this is a substantial change. When variables representing competition (TelusSVI and RogersSVI) are added, we see a minor increase in the adjusted R^2 to 0.84. Finally, the addition of the Twitter sentiment decreases the adjusted R^2 to 0.839.

Interestingly, for Rogers, the model with the seasonal factor only (RogersASWF) explains only about 0.419 of the variance (R^2) in NNS (Table 11.3). When Google Trends SVI for Rogers is added to the model, the adjusted R^2 is more than doubled to 0.861, and standard error of the estimate is more than halved to 33,297. After adding the variables representing competition to the model, we notice a sizeable improvement in adjusted R^2 to 0.916, and standard error of the estimate goes down to 25,894. Yet again, after adding the Twitter sentiment factors, the adjusted R^2 goes down to 0.913.

Finally, for Telus, it is observed that the Seasonal Weight Factor alone explains about 80% of the variance in the dependent variable (Table 11.4). Seasonality seems to play a very significant part in Telus's ability to attract new users and retain existing ones. Once again, we see that adding the Google Trends SVI for Telus improves the model and increases the adjusted R^2 to 0.943. The addition of competition variables, as expressed by RogersSVI and BellSVI, is able to improve the model, but only slightly, to an adjusted R^2 of 0.951. Finally the addition of Twitter Sentiment (TelusST) has no impact on the model.

11.5 Discussion

Our results show that Bell NNS performance is the least susceptible to competition variables or to the performance of the competition in general. Generally speaking, Bell is the oldest and most renowned telecommunications provider in Canada; therefore it has a solid base of loyal customers that are unlikely to switch to other competitors. In addition, Bell is the largest wired telecommunications provider in Canada, and it has the most advanced fiber optics base TV service. Hence customers can enjoy additional benefits when bundling their wireless and wired solutions with the same provider.

Rogers' performance, on the other hand, seems to be fairly sensitive to the performance of the other two providers. This can be explained by the fact that both Telus and Bell are aiming to capture the market share from the leader, which is Rogers. Their marketing efforts and diverse offers are often targeted at Rogers' customers, and leads to higher Rogers deactivation and churn rates. It is interesting to observe that the characteristics of the competitive landscape in the Canadian wireless telecommunications industry are reflected in the impact of the competitive factor on the prediction accuracy of our models.

The four models tested in this study are fairly robust and easy to justify. Yet, we are aware that there are many more variations of the presented models that could have been developed and compared. The argument for our choice of linear models was made before. Another variation of the models could be created by aggregating the time variant variables differently noting that there can possibly be a time lag between searching for information on the Web and deciding to choose a wireless service provider. To test whether this time lag had any effect on the models, we rebuilt the models with 4 weeks, 2 weeks and 1 week time lags between the independent and dependent variables. When we compare these results with the original (no time lag) results, we observe that longer time lags result in poorer predictive capability in terms of adjusted R^2. This suggests that the time lag between searching for information and choosing a wireless solution is less than a week (Table 11.8). It is not possible to apply time lags that are less than a week because Google search volume data is available in a week-by-week form. All of the three providers studied here maintain e-commerce sites through which users can subscribe to the service immediately after conducting an online search. We do not have the statistics on the

Table 11.8 Time-lagged adjusted R square for Bell, Roges, Telus (all four models)

Provider—Model	No time-lag Adjusted R square	1 week time-lag Adjusted R square	2 week time-lag Adjusted R square	4 week time-lag Adjusted R square
Bell—Model 2	0.838	0.835	0.834	0.829
Bell—Model 3	0.840	0.834	0.832	0.825
Bell—Model 4	0.839	0.834	0.833	0.827
Rogers—Model 2	0.861	0.853	0.836	0.797
Rogers—Model 3	0.916	0.913	0.912	0.898
Rogers—Model 4	0.913	0.911	0.910	0.898
Telus—Model 2	0.943	0.943	0.943	0.939
Telus—Model 3	0.951	0.949	0.944	0.935
Telus—Model 4	0.951	0.949	0.945	0.935

percentage of online phone activations, but we expect it to be substantial as all of the players in Canadian wireless market, including the average-sized ones, have invested in e-commerce sites that provide such capability. Generally speaking, consumers can make fast decisions with regards to wireless purchases, because top smartphone models such as iPhone or Samsung Galaxy are provided by virtually every competitor, and constitute majority of the new activations. In addition, a typical customer is very likely to have already performed research with regards to the product that they are interested in and decided on the model they wish to purchase before doing their research on, and making a decision about, the wireless provider. As a result, models built based on data with short (to none) Google search time-lags are remarkably accurate.

11.6 Conclusions

In this study, we studied the success of readily available web search metadata along with social media content in predicting the market share of the three major mobile service providers in Canada. The main contribution of this work is the use of these data both for the provider of interest (target company) and its competitiors in the process improving prediction accuracy. The impact of competition variables vary depending on the competitor. Nevertheless, the results suggest that when nowcasting or predicting operational results for a company with a known and limited set of direct competitors, it is beneficial to include competition variables based on Google Trends data. So far, all of the Google Trends studies that analyzed company performance such as sales in retail, real estate, or car sales, focused only on Google Trends or Twitter data for the target company, omitting the data for direct competition. This study demonstrates that when trying to predict sales of, for example, Toyota Dealerships, it would be beneficial to include Google Trends values for direct competition such us Honda, Hyundai and Nissan. We also observe that, for an industry with a limited and stable number of competitors, we could expect the market leader to be most influenced by the inclusion of competition variables in the Google Search based model. The results show that Rogers' performance for a quarter is more dependent on competition performance than, for example, Bell's performance.

It comes as a surprise to see that the inclusion of Twitter sentiments did not improve the performance of the models for any of the three competitors. The existing body of literature would suggest a strong correlation, but our data for the Canadian Telco sector indicate otherwise. Some of the possible reasons for this situation are discussed next.

Due to financial limitations, the historical data for Twitter were obtained by manually entering the queries on the Twitter website and by copying all of the resulting tweets to a spreadsheet. This is a very time-consuming process hence it could only be done once. Therefore, it was not possible to experiment with various sets of queries to select one that has the best fit for the model. With better access to Twitter data, this issue could be revisited in the future.

In addition, the Twitter data that were obtained through this laborious process seemed to have significant levels of variance, where in some quarters the number of returned tweets would be counted in hundreds, and in others, in thousands. It could be that the adoption levels of Twitter in Canada were not sufficient in some of the historical periods to provide an adequate dataset for modelling. Future research could further explore the impact of sentiments by incorporating additional sources of data such as Facebook posts, other microblogging website posts, or posts from forums with wireless provider reviews.

It would also be desirable to include more historical quarters into our model but the Twitter data prior to year 2008 were nearly non-existent for the queries we used in our study. It would be beneficial to update the model in the future with new quarterly data and to verify that the findings contained herein can be confirmed with a larger dataset. Ideally, we would wish to utilize monthly, rather than quarterly, data for the study, but monthly subscriber figures for each of the competitors are not publicly available. Therefore, for our Model 4 with five predictors, the findings could not be based on a dataset that is large enough. It is possible that the findings related to Twitter sentiments could be revised and reinterpreted after applying the model to a more sizeable dataset. Research on sentiment analysis suggests context or topic specific corpuses for automated sentiment analysis (Lak and Turetken 2014). In this study we used the general purpose sentiment analysis engine SentiStrength. In the future, the tool can be modified to better fit the specific industry being studied.

Another direction for future research is to confirm our findings in other similar industries around the globe before making any generalizations. There are a number of other industries that are focused on services and have oligopoly characteristics in certain geographical areas. For example, Canadian retail banking industry is dominated by top five players: RBC, TD, Scotiabank, CIBC and BMO. Likewise, the US wireless provider market is dominated by four top players: Verizon, AT&T, Sprint, and T-Mobile. Our models can be replicated for these markets to identify whether they are generalizable or have simply captured the idiosyncrasies about the context of the current study.

The purpose of our research was purely prediction; therefore we did not specifically emphasize each independent variable's contribution and significance in the model. This is mainly due to the fact that the independent variables in the study cannot be (substantially) manipulated by the managers of the companies explored. As such, collinearity and dependence of error terms was not a specific concern as long as the overall predictive power of the models was satisfactory. Future studies where some of the social media related variables can be significantly manipulated should ensure avoiding the violation of regression assumptions to be able to explain the individual factors' role in influencing the overall change in the dependent variables.

One advantage of using Google Trends as a surrogate for potential consumer interest is the fact that Google Trends captures nonlinearities in the consumer search patterns hence making simple linear models such as those presented in this chapter feasible. As a result, the predictive power of our models was very high. However, market share may not be as easily predictable in other contexts with longer time periods and market volatility. Further research should explore whether nonlinear time variant models such as Markov chains may be needed for web and social media based prediction of market success.

Biographies

Michal Szczech (mszczech@gmail.com). For past 10 years serving in the IT consulting field servicing mostly Canadian Telecommunications Industry including clients such as: Bell Canada, Rogers Communications, Shaw Communications and SaskTel. Michal graduated from Ted Rogers School of Management—Ryerson University MBA program with Information Technology Management specialization in 2015.

Ozgur Turetken (turetken@ryerson.ca). Before joining Ryerson, Professor Turetken (Ph.D., Oklahoma State University, M.B.A., B.Sc., METU, Ankara, Turkey) served on the faculty of Temple University's Fox School of Business and Management. Dr. Turetken's scholarly interests are mainly in applied (text) analytics especially in the context of individual decision making. His research on the organization and presentation of information as their relation to decision outcomes has been funded by NSERC and SSHRC among other agencies. Ozgur is also a recipient of Ryerson University's Scholarly Research and Creative award in 2010. He is active in various research communities as a reviewer, editor, and track chair.

http://www.ryerson.ca/itm/faculty/ozgur-Turetken.html

References

Beer F, Hervé F, Zouaoui M (2013) Is big brother watching us? Google, investor sentiment and the stock market. Econ Bull 33(1):454–466

Cameron MP, Barrett P, Stewardson B (2015) Can social media predict election results? Evidence from New Zealand. J Polit Market:1–17

Cavazos-Rehg P, Krauss M, Spitznagel E, Buckner-Petty S, Grucza R, Bierut L (2015) Monitoring marijuana use and risk perceptions with Google Trends data. Drug Alcohol Depend 146:e242–e243

Chandukala SR, Dotson JP, Liu Q, Conrady S (2014) Exploring the relationship between online search and offline sales for better "nowcasting". Cust Need Solut 1(3):176–187

Choi H, Varian H (2012) Predicting the present with Google Trends. Econ Rec 88(s1):2–9

Da Z, Engelberg J, Gao P (2011) In search of attention. J Financ 66(5):1461–1499

Engel JF, Blackwell RD, Kegerreis RJ (1969) How information is used to adopt an innovation. J Advert Res 9(4):3–8

Gerber MS (2014) Predicting crime using Twitter and Kernel density estimation. Decis Support Syst 61:115–125

Goel RK (2007) Oligopoly. In: Kaliski BS (ed) Encyclopedia of business and finance. Macmillan, Detroit, MI, pp 558–559

Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. Proc Natl Acad Sci U S A:17486–17490

Internet Live Stats (2015) Total number of websites. http://www.internetlivestats.com/total-number-of-websites

Internet World Stats (2014) World Internet users statistics. http://www.internetworldstats.com/stats.htm

Jansen BJ (2009) Twitter power: Tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169–2188

Jun S-P, Park D-H, Yeom J (2014) The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technol Forecast Soc Chang 86:237–253

Lak P, Turetken O (2014) Star ratings versus sentiment analysis—a comparison of explicit and implicit measures of opinions. In: Proceedings of the 47th Hawaii international conference on system sciences (HICSS). IEEE, pp 796–805

Lassen NB (2014) Predicting Iphone sales from Iphone tweets. In: IEEE 18th international enterprise distributed object computing conference. IEEE, pp 81–90

Marian AD, Nicole B, Wolfgang S (2014) Sentiment-based commercial real estate forecasting with Google search volume data. J Prop Invest Financ 32(6):540–569

Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using Google Trends. Sci Rep 3:1684

Shawn KJL, Stridsberg D (2015) Feeling the market's pulse with Google Trends. Int Fed Tech Anal J

Statistic Brain (2015a) Facebook statistics. http://www.statisticbrain.com/facebook-statistics

Statistic Brain (2015b) Twitter statistics. http://www.statisticbrain.com/twitter-statistics/

Tumasjan A (2011) Election forecasts with Twitter: how 140 characters reflect the political landscape. Soc Sci Comput Rev 29(4):402–418

Vosen S (2011) Forecasting private consumption: survey-based indicators vs. Google Trends. J Forecast 30(6):565–578

Chapter 12
Scale Development Using Twitter Data: Applying Contemporary Natural Language Processing Methods in IS Research

David Agogo and Traci J. Hess

Abstract The availability of big data sources and developments in computational linguistics present an opportunity for IS researchers to pursue new areas of inquiry and to tackle existing challenges with new methods. In this paper, a novel way of developing measurement scales using big data (i.e., tweets) and associated methods (i.e., natural language processing) is proposed and tested. The development of a new scale, the technology hassles and delights scale (THDS), is used to demonstrate how a syntax aware filtering process can identify relevant information from a large corpus of tweets to improve the content validity of a scale. In comparing themes generated from analyzing 146 million tweets, with themes generated from semi-structured interviews, a reasonable overlap is observed. Further, the potential for identifying even more relevant themes from within subsets of the tweet dataset is uncovered.

Keywords Scale development • Content validity • Tweets • Technology hassles and delights • Affective evaluation • Natural language processing • Computational social science

12.1 Background

In the past two decades, significant advances in computer science and related disciplines have unleashed the power of algorithms and advanced hardware on the classical problem of interpreting spoken and written text. Methods for machine translation, speech recognition and speech synthesis have been widely applied in consumer products such as spoken dialogue systems (SDS) (e.g., Apple's Siri, Amtrak's Julie, and Microsoft's Cortana), and a host of other technologies,

D. Agogo • T.J. Hess (✉)
Operations and Information Management Department, Isenberg School of Management,
University of Massachusetts Amherst, 121 President's Drive, Amherst, MA 01003, USA
e-mail: thess@isenberg.umass.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_12

including the mining of social media data for various purposes (Hirschberg and Manning 2015). There is also a vast amount of digital data now available for the purpose of understanding human behavior, an emerging area called computational social science (Lazer et al. 2009). This paper aims to demonstrate a potentially useful application of big data in information systems (IS) research, with applicability for social science research in general.

Like many other academic disciplines, the IS field has a growing interest in the use of big data (Agarwal and Dhar 2014). Further, the IS field is known for its best practices in the development and validation of measurement scales in empirical research (e.g., Boudreau et al. 2001; MacKenzie et al. 2011; Straub et al. 2004). The purpose of this paper is to demonstrate how big data can be used to develop better measurement scales using an IS scale as an example. While IS research has focused on scale validation issues such as construct validity and reliability, content validity has received little attention. Content validity refers to "*the degree to which items in an instrument reflect the content universe to which the instrument will be generalized*" (Straub et al. 2004, p. 424). Content validity is a property of a set of items or measures taken together (Anderson and Gerbing 1991), and can be difficult to establish due to challenges in sampling the domain of interest (Hinkin 1995, 1998; Rossiter 2002).

In this paper, natural language processing (NLP) methods are applied on data collected from the Microblogging website, Twitter, with the objective of identifying frequently occurring themes in affective evaluations of technology, in order to support the development of a content valid measurement scale. Both positive and negative evaluations are considered, thus we refer to this scale as the technology hassles and delights scale (THDS). Further, semi-structured interviews on affective evaluations of technology are also carried out to evaluate the relative performance of both approaches at identifying the most prevalent elements in this domain. By employing both Twitter data and interview data, separately and collectively, towards the development of a measurement scale, we hope to demonstrate the capacity for contemporary big data and big data methods to contribute to scale development practices in IS research. After all, as acknowledged by Agarwal and Dhar (2014, p. 445), the big data opportunity in IS research enables us to "address the same types of questions as we have in the past but with significantly richer data sets".

Beyond the methodological contributions of this paper, the IS scale being developed (THDS), is also important. This scale captures both positive and negative events experienced when interacting with computers and phones, and has the potential to improve our understanding of the process-based factors that lead to affective evaluations of technology (e.g. computer anxiety, technostress, flow and enjoyment) (Zhang 2013). THDS may also shed light on the user experiences that lead to switching from one platform to the other (e.g. from Windows to Mac or from Android to Apple iOS). In the remainder of the paper, a background on scale development and the social media data source are presented, after which the NLP methods used in this paper are introduced. The analysis and results are then reported, followed by a discussion of the next steps in this research project.

12.2 The State of Scale Development

The development of valid measurement scales is a prerequisite for carrying out quality empirical social science research. The earliest stages in the scale creation process include conceptual definition of the concept to be measured and item generation (Churchill 1979; Hinkin 1998; MacKenzie et al. 2011). While established methods exist for later stages of scale development (e.g. reliability analysis, factor analysis, etc.), there is more ambiguity regarding the earliest stages of the process as postulated by domain sampling theory. This theory states that it is not possible to measure the complete domain of interest, thus it is important that the sample of items adequately represents the construct under examination (Ghiselli et al. 1981; Hinkin 1998). Thus, the quality of any scale is highly dependent on the degree to which the items generated sample the domain being studied. This concept is referred to as content validity (Nunnally 1978; Straub et al. 2004).

Two broad approaches to scale development exist: inductive and deductive, with the former an iterative approach best suited for exploratory research and the latter more suitable for areas with well-established theory. An inductive approach refers to situations where a theoretical foundation is lacking, and scales are developed inductively by having a sample of participants describe their feelings and behaviors related to systems or organizations (Hinkin 1995). *"An example might be, 'describe how your organization supports your use of a system' or 'describe how you feel when you use the system.' Responses are then classified into a number of categories by content analysis based on key words or themes."* (Hinkin 1995, pp. 969–970). A deductive approach refers to situations where a review of the literature is conducted to provide a theoretically-grounded definition of the construct (Hinkin 1995). With this approach, a sample of participants is often used to determine critical incidents that can be used in the measurement scale. Thus, regardless of the scale development approach, samples of respondents are often consulted to obtain descriptions of the beliefs, feelings, and behaviors related to the construct of interest.

In these circumstances, item generation based on semi-structured interviews with a purposive sample (Hinkin 1995; Rossiter 2002) may be limited by validity threats. For instance, the quality of the questions created by the researcher can greatly influence the items generated (Clark and Watson 1995). Similarly, memory and response bias of participants may introduce unknown bias to the items generated in the first phase of scale creation. These factors may introduce restrictions to the scales created and ultimately bias research programs that adopt the scales. Thus, generating an inclusive and extensive pool of items is critical to developing valid scales. As articulated by Loevinger, "*The items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait*" (Loevinger 1957, p. 659). Given the popularity of social media and the willingness of individuals to share personal information and express their experience of daily situations using these media, it is worth exploring how data from these sources can be utilized in the scale development process.

12.2.1 Extracting Meaning from Social Media Data

This paper employs data obtained from a leading social networking and microblogging service, Twitter. Founded in 2006, Twitter is a service that allows users to post and read short messages (maximum of 140 characters). These "tweets", as they are called, are typically public and visible to anyone online. The success of Twitter can been attributed in part to its wide global reach and support for establishing weak ties, with the "promise of transcending distance, connecting everyone with anyone" (Takhteyev et al. 2012, p. 81). Beyond that, however, Twitter creates ambient awareness between people in social networks and provides a platform for virtual exhibitionism and voyeurism for active contributors and passive observers (Kaplan and Haenlein 2011). Ambient awareness is a form of awareness facilitated by the exchange of fragments of information online which can result in high levels of social presence and media richness between individuals on social media (Kaplan and Haenlein 2011). The support for ambient awareness and virtual exhibitionism/voyeurism make Twitter a unique source of data for researchers seeking an unfettered glimpse into the daily lives of people around the world.

The public nature of tweets enables users to engage in self-disclosure and self-presentation by sharing their experiences with the world (i.e., while users can restrict access to their tweets to those who follow them, this is not a common practice) (Kaplan and Haenlein 2011). These motivations have led to extensive use of Twitter to disclose information on personal experiences, attitudes, etc. One important aspect of Twitter is the length restriction imposed by the creators of the service. By restricting the length of a tweet, information shared by users is more likely to be focused on a single idea or topic, which enables easy interpretation by researchers. However, this also leads Twitter users to (1) compress sentences by omitting words and using unusual spellings to fit within character limits, and (2) send out multi-part tweets i.e. multiple tweets on a single subject. In addition, limited context information is made available, since tweets are generally meant to be read at that instant, and the sender assumes that the audience is already part of the conversation. In addition to these idiosyncrasies with Twitter data, there are other issues such as an abundance of spam, automated posts and the relative anonymity of users.

Nevertheless, tweets have been reliably used in aggregate to analyze and even predict events/trends such as weather (Lampos and Cristianini 2012), box-office revenues (Asur et al. 2010), consumer confidence and political polls (O'Connor et al. 2010), politics and election outcomes (Beauchamp 2013; Gayo-Avello 2013), public health concerns such as a flu pandemic (Lampos and Cristianini 2010), and unemployment (Llorente et al. 2015), as well as a myriad of applications in online marketing. On the individual level, tweets have been used to identify gender, age, regional origin (Rao et al. 2010), political affiliation (Pennacchiotti and Popescu 2011), and post-partum changes in affect (De Choudhury et al. 2013).

12.3 Natural Language Processing (NLP) Methods

NLP, also known as computational linguistics, involves the use of computational and statistical methods to learn, understand, and generate human language content. The fundamental building block of linguistic analysis is identifying patterns that occur in language use (Manning and Schütze 1999). Current NLP methods focus on more complex representations of linguistic patterns which are designed to transcend the count-based 'body of words' approaches common in early NLP (Hirschberg and Manning 2015). Currently, statistical and machine learning (ML) methods are being used to achieve deeper levels of linguistic analysis (Smith 2011). The NLP methods applied in this paper include tokenization, part-of-speech tagging, and n-gram sequences (as described in Table 12.1). Part of speech tags are assigned based on word definitions and rules devised to interpret the context in which a word appears (e.g. the relationship of a word with adjacent words) (Brill 2000). Several parts-of-speech taggers have been developed for use on different types of text with accuracies maxing out at about 97% (Manning 2011). However, these taggers perform poorly with the informal text obtained on Twitter, which has led to taggers developed specifically for conversational online text. TweetNLP is such a tagger and it achieves up to 93% accuracy when applied to tweets (Owoputi et al. 2013). An example of an actual tweet with tags is shown in Table 12.2, along with definitions of the tags produced by TweetNLP (See Owoputi et al. 2013 for a full list).

12.3.1 The NLP Approach: Syntax-Aware Phrase Extraction

By first assigning parts of speech to a tweet, it is possible to extract n-gram pairs containing relevant phrases which capture the core information in each tweet. Based on this, sequences of words that co-occur most often out of a large corpus of tweets can be expected to be indicative of the main themes within that corpus. Further, by combining syntax rules constituted from part-of-speech information, more nuanced

Table 12.1 Definition of NLP methods used in this project

NLP method	Definition	Relevance to this study
Part-of-speech tagging	Part-of-speech tagging (POST), is the process of marking up words in text as corresponding to a part of speech based on a set of rules	Grammatical syntax added to tweets provides filtering beyond 'bag of word' approaches
Tokenization	Given a character sequence, tokenization is the process of splitting into pieces called tokens, usually discarding content such as punctuation	Each tagged tweet is broken into individual (word, POS) pairs
N-gram sequences	N-gram refers to a sequence of n items or tokens based on syllables, letters, words, or more complex structures	Word tokens are combined into n-grams for syntax-aware filtering to occur

Table 12.2 Illustration of NLP methods used

<table>
<tr><th>Unit of analysis</th><th colspan="8">Contents</th></tr>
<tr><td>Tweet</td><td colspan="8">{My phone crashed the second it turned midnight}</td></tr>
<tr><td rowspan="2">POS Tagged Tweet</td><td>{My</td><td>phone</td><td>crashed</td><td>the</td><td>second</td><td>it</td><td>turned</td><td>midnight</td></tr>
<tr><td>D</td><td>N</td><td>V</td><td>D</td><td>A</td><td>O</td><td>V</td><td>N}</td></tr>
<tr><td>Tweet Tokens</td><td colspan="8">{My}, {phone}, {crashed}, {the}, {second}, {it}, {turned}, {midnight}</td></tr>
<tr><td>{Word, POS} Pair Tokens</td><td colspan="8">{My, D}, {phone, N}, {crashed, V}, {the, D}, {second, A}, {it, O},{turned, V}, {midnight, N}</td></tr>
<tr><td>Tweet Bigrams</td><td colspan="8">{My phone}, {phone crashed},{crashed the},{the second}, {second it},{it turned},{turned midnight}</td></tr>
<tr><td>Tweet Trigrams</td><td colspan="8">{My phone crashed},{phone crashed the},{crashed the second},{the second it},{second it turned},{it turned midnight}</td></tr>
<tr><td>{Word, POS} Pair Trigrams</td><td colspan="8">{{My, D}, {phone, N}, {crashed, V}}, {{phone, N}, {crashed, V}, {the, D}}, {{crashed, V}, {the, D}, {second, A}}, {{the, D}, {second, A},{it, O}}, {{second, A},{it, O},{turned, V}}, {{it, O},{turned, V}, {midnight, N}}</td></tr>
<tr><td>Tweet 5-grams</td><td colspan="8">{My phone crashed the second}, {phone crashed the second it}, {crashed the second it turned}, {the second it turned midnight}</td></tr>
</table>

N common noun, *O* pronoun, *V* verb, *A* adjective, *R* adverb, *D* determiner, *P* pre- or post-position

information about the use of language can be obtained. For instance, the verbs or adjectives co-occurring most frequently with a particular noun can be identified and used to infer the way an object is most frequently described. Such methods have been widely used for more advanced purposes such as language translation. Lastly, interpretation is applied to select semantically accurate (i.e. meaningful) phrases out of the anticipated large group of syntactically accurate information collected. Given this introduction to the methods to be used, the scale to be created is now introduced and preliminary results and analysis are presented.

12.3.2 The Need for a Technology Delights and Hassles Scale

The purpose of this scale development exercise is to develop a scale of process-based evaluations or appraisals of technology objects and technology usage behavior under the banner of a technology hassles and delights scale (THDS). This work follows in the tradition of early work on a computer hassles scale (Hudiburg 1989, 1992), a measure from the field of psychology that failed to gain traction within IS. The study of affective evaluations towards technology and technology use is an emerging area in the information systems (IS) discipline. It is a critical element of user behavior that is an important part of technology adoption research (Beaudry and Pinsonneault 2010). Social scientists have long held that emotions play a critical role in shaping human decisions as well as subsequent behavior. Despite this, the study of cognitions has dominated IS and only recently have efforts been undertaken to identify pertinent categories of affective concepts (e.g., Loiacono and Djamasbi 2010; Zhang 2013). Recent conceptual work such as the Affective Response Model (ARM) (Zhang 2013) provides the foundation for classifying existing affective evaluations towards technology objects and use, and shedding light on how they come about. For instance, experiences during use are a source of enduring positive or negative evaluations of particular technologies or technology in general (Propositions 6 and 7 in Zhang 2013). ARM identifies these factors as process-based factors (Category 5.1 and 6.1 in Zhang 2013), and thus provides an emerging theoretical space which can serve as the foundation for THDS.

When creating measurement scales in the affective domain, there needs to be clarity about the *target*, *intensity* and *direction* of the characteristic (McCoach et al. 2013). Target refers to the object, behavior, or idea the feeling is directed at, intensity refers to the degree or strength of the feeling and direction reflects whether the feeling is positive, neutral or negative. In this case, the target is the *user experience with hands on use of* specific technology objects (mobile phones, tablets and computers). The intensity refers to feelings strong enough to be expressed, and the direction includes both positive and negative expressions. The technology delights and hassles scale is being developed to test hypotheses related to at least two important IS research areas (1) usage continuance/switching intentions (Bhattacherjee 2001; Bhattacherjee et al. 2012) and (2) deep structure usage (Burton-Jones and Straub 2006), answering the call to open the black box of constructs (perceived ease of use and usefulness) which drive usage (Benbasat and Barki 2007). This paper takes the perspective that improving our understanding of process-based factors is a new approach that can help shed more light on these principal constructs and may eventually lend itself to theory creation (Goodhue 2007). Leaning on best practices for creating affective scales in the affective domain (McCoach et al. 2013), tweets and semi-structured interviews, will be used. The objective of the analysis is to identify major themes from a corpus of tweets in which people are speaking about three categories of technology: computers, mobile phones and tablets. In parallel, semi-structured interviews will also be used to generate items, and the resultant themes identified will be compared and condensed into a single THDS.

12.4 Analysis and Preliminary Results

The previously discussed methods are now employed towards identifying the main themes in a corpus of tweets using the steps illustrated in Fig. 12.1 below. Each step is explained in the sections that follow.

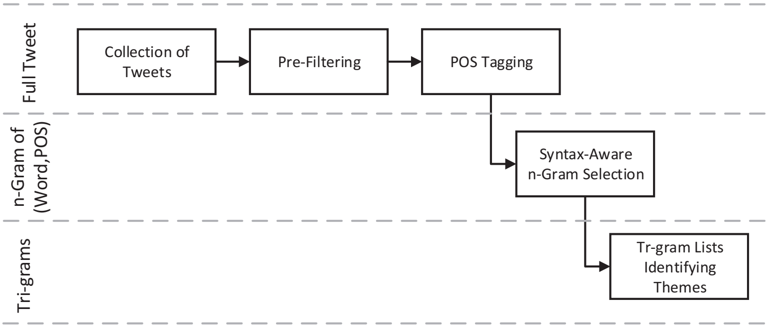


Fig. 12.1 Steps in NLP analysis

12.5 Analysis and Results

12.5.1 Collection of Tweets

The dataset used for this study consists of 146,315,059 tweets (N_f) from Jan 1, 2014 to March 31, 2015 (455 days). The keywords used to select tweets included: *computer, pc, laptop, desktop, phone, iphone, cellphone, smartphone, tablet, ipad*. This dataset represents an average of 321,571 tweets per day (range of 226,266–704,067). The maximum number of tweets was on Sept 9, 2014 the day of the iPhone 6 launch. Initial models and data cleansing approaches were developed on a smaller subset ($N_1 = 2$ million) due to the large processing requirements of repeated analysis using the full dataset and are reported alongside the final analysis with the full dataset.

12.5.2 Pre-filtering and POS Tagging

The first level of filtering involved excluding tweets that were retweets (i.e., duplicates of someone else's tweet), not written in English, or were automatic posts (e.g., by game apps). Tweets were filtered using the rich metadata that is part of each tweet (e.g. time of tweet, the application the tweet was sent from, the language encoding of the tweet, etc.). Unfortunately, these tags are not always accurate and may result in a substantial number of false negatives i.e. discarded tweets. The overall size of the dataset makes this a less severe issue. This initial filtering yielded 38,076,612 usable tweets (26% of N_f). TweetNLP (Owoputi et al. 2013) was used to tag the tweets for subsequent analysis. The 38 million usable tweets (26% of N_f) from the pre-filtering stage were tagged. To verify that TweetNLP appropriately tagged keywords of interest, tagger accuracy was verified on a random subset of two million tweets. The nine keywords appeared a total of 1,573,477 times, each being tagged as either a proper or common noun in at least 70% of appearances. Details are contained in Table 12.3 below.

Table 12.3 Verification of tagger accuracy

Category	Keyword	Proper noun (%)	Common noun (%)	Determiner (%)	Other POS (%)	Count
Computer	"computer"	2	74	11	14	108,963
	"desktop"	26	51	6	16	16,879
	"laptop"	2	76	10	12	78,310
	"pc"	45	26	7	22	52,441
Phone	"phone"	1	81	9	9	974,123
	"smartphone"	47	25	7	21	24,189
	"iphone"	69	7	6	18	202,797
Tablet	"tablet"	17	61	5	16	31,408
	"ipad"	66	8	10	16	83,332

12.5.3 Syntax-Aware n-Gram Selection

Following this, 5-grams which met the syntactic requirements were selected from the corpus of tweets. One grammatical syntax structure, the verb phrase, was used to identify themes. The verb phrase has an action word, a verb, at its core (e.g., walking, crash, froze), and also includes a complement or modifiers (e.g., *walking slowly back home*, *crash my computer*, *my phone froze)*. At its simplest, a verb phrase can capture the principal action in a sentence (e.g., "*My phone crashed*" is the verb phrase in the tweet {My phone crashed the second it turned midnight}). A more complete discussion of the different forms of verb phrases is beyond the scope of this paper. Since a single tweet may contain multiple verb phrases, adjacency of the verb phrase to one of the keywords was also a selection criteria. The sequence for selecting 5-grams from a sample tweet is shown in Table 12.3, and all selected 5-grams, all verb-phrases, and the full source tweets were saved separately. From the initial set of two million tweets, about half of the tweets (908,367) were found to meet this criteria and were further analyzed. Some additional filtering was done at this stage to exclude spam tweets not previously detected. After refining this algorithm on the smaller dataset, the analysis was run on the full data set, leading to 13,089,522 tweets (34% filtering rate).

12.5.4 Generating Themes from Tri-gram Lists

Broadly, the verb phrases identified through this filtering process were selected as primary themes for preliminary evaluation as either hassles or delights. Different lists of verb phrases were identified for each technology category, i.e., computers, phones, and tablets. Running the full sequence of steps in Table 12.4 on the full dataset of tweets (N_f) yielded 140,942 (phone), 58,649 (computer), and 20,977 (tablet) unique tri-grams that appeared more than once in the dataset. The top 200 of these tri-grams for each category were analyzed for themes. Where necessary, a

Table 12.4 Example Tweet showing steps and extracted verb

Step one	Select full Tweet	{My phone crashed the second it turned midnight}
Step two	Split into 5-grams and select if keyword is present (i.e., phone)	{My phone crashed the second}, **DNV**DA
		{phone crashed the second it}, NVDAO
		~~{crashed the second it turned}, VDAOV~~
		~~{the second it turned midnight}, DAOVN~~
Step three	Select 5-gram(s) containing verb phrase	{My phone crashed the second}, **DNV**DA
		~~{phone crashed the second it}, NVDAO~~
Step four	Select verb phrase	{My phone crashed}, **DNV**

random subset of tweets was retrieved to enable better interpretation. Due to space constraints, only themes from phone-related tri-grams are reported.

Tri-grams fell under meaningful themes such as operating the device, user clumsiness with the device, etc. as well as clear affective expressions (hate my phone…, love my phone…). In both the computer and tablet categories, meaningful themes were also identified. A cursory scan of the tweets under themes such as user rage and affective expression reveal that more detailed information about the reasons for these affective evaluations can be obtained and is planned as future research. The validation conducted using semi-structured interviews is presented in the following section.

12.5.5 *Cross-Validation of Themes from Twitter Data*

In order to cross-validate the themes derived from the NLP analysis, qualitative data from semi-structured interviews was collected independently. Participants were sought from the crowdsourcing platform, Amazon Mechanical Turk, an increasingly common source of data for social science research (Buhrmester et al. 2011; Steelman et al. 2014). A total of 45 participants completed the survey, responding to open ended questions asking them to list the most frequently occurring delights and hassles they had experienced using technology (21 for computers only, the rest for mobile phones and tablets). The sample was 56% female, 62% had a 4-year college degree or greater, and had an age range of 19–66 (mean = 36, S.D. = 14). As with the Twitter data, only the results for phones are presented due to space constraints. An example question provided is “During daily interaction with smartphones, some things happen that annoy and irritate you. Using short sentences, list some examples below.” The responses were coded to identify dominant themes. The themes

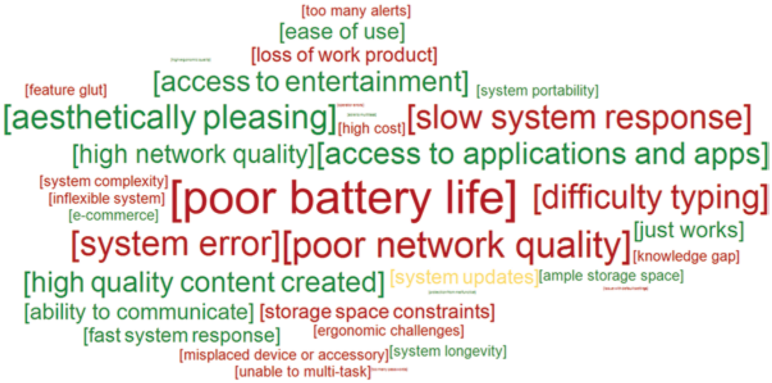


Fig. 12.2 Word cloud showing themes from semi-structured interviews for THDS scale. *Red*: hassle, *Green*: delight, *Yellow*: both hassle & delight; *Size* represents frequency

Themes from MTurk	% of Total
System Performance	**29.7%**
Ease of Use	16.2%
Battery Life	**15.3%**
Network Quality	12.6%
Access to Apps	11.7%
Screen quality	**7.2%**
Communication	**2.7%**
Storage Space	2.7%
Cost of Device	**0.9%**
Separation from Device	**0.9%**

Themes from Twitter	% of Total
(Not Relevant)	40.0%
Operating the device	12.3%
User clumsiness	10.6%
Desire to own/Purchase	9.4%
Communication	**8.1%**
Online Offers	4.9%
Battery Life	**4.7%**
Emotions towards Phone	3.7%
Separation from Device	**3.4%**
System Performance	**1.4%**
Questions about Device	1.1%
Cost of Device	**0.5%**
Screen quality	**0.3%**

Fig. 12.3 Comparison of themes for phone hassles & delights. % represents proportion of 1,711,038 tweets (in top 200 tri-grams) and proportion of 110 interview statements. *Shading* represents common themes across the two data sources

identified for phone use are shown in the word cloud in Fig. 12.2 below, with the size of the theme representing the frequency of occurrence and the color indicating the direction of the theme.

Finally, the themes from both methods were compared, with the semi-structured interviews yielding ten distinct themes and the Twitter NLP analysis yielding 12 themes as shown in Fig. 12.3. There was significant overlap between these themes, with six of ten themes from the semi-structured interviews present in the Twitter NLP themes. This overlap suggests that Twitter data may be a viable source for conducting content analysis with new scale development and in validating existing scales. The non-overlapping themes in the two data sets also provided insight. Unique themes from the Twitter data were Desire to Own/Purchase and Emotions towards Phone, which are more affective in content, as compared to Network Quality, Access to Apps, and Storage Space, which were the themes unique to the semi-structured interviews. While the themes presented below represent preliminary analysis of the two data sources, these findings seem appropriate given the inherent impulsive vs. reflective nature of these data sources. Twitter supports ambient awareness by enabling quick exchanges of tweets that describe real-time experiences of users with their phones. Thus, the Twitter data can be expected to reflect more impulsive, unfiltered expressions of affect, both positive and negative. In comparison, semi-structured interviews are initiated with an explanation of the context, and then participants are asked to express their beliefs or general experiences with their phones. In formulating responses, interviewees seem to provide recollections based on reflection, which produce more cognitive, functional aspects of their phone experiences.

Further examination of themes across and within the two data sources may yield additional insight. For example, the theme of Ease of Use from the semi-structured interview data and the theme of Operating the Device from the Twitter data, shared some common concepts. The affective themes from the Twitter data could be analyzed and decomposed into additional themes based on positive and negative evaluations, or based on some of the dimensions in the ARM framework. Other differences in themes could be examined based on the nature of the data source (i.e., size limitations), and the relative percentage of total themes across the data sources.

12.6 Discussion and Next Steps

The foregoing sections introduce a range of NLP techniques and discuss their application to a corpus of tweets for the purpose of extracting themes associated with technology use that can form a process-based experience scale—the THDS. By applying a syntax-aware filtering approach, lists of tri-grams that capture expressions related to specific technologies have been identified. The top 200 tri-grams related to mobile phones were analyzed and condensed into 12 themes. Following that, semi-structured interviews on the same topic were conducted independently and that data was also analyzed to arrive at a set of ten themes. Finally, themes from both sources were compared and a reasonable amount of overlap was discovered. Further, the potential for even more themes to be identified using further NLP analysis was noted. These findings are reported for a single category of technology, mobile phones.

12.7 Conclusion and Future Directions

This project started from a desire to put big data to the test in the context of scale creation, as well as a desire to create a much needed measure of process-based factors associated with technology use. Based on the progress reported thus far, the analysis of Twitter data using basic NLP techniques does yield themes which potentially represent delights and hassles using technology. These themes reasonably overlap with themes derived independently from semi-structured interviews, the traditional method of generating items for scale development. Deeper exploration of relevant tweets is needed, after which a more formal comparison of the themes identified in the twitter data and semi-structured interviews will be conducted. Upon completing this work, the extent to which twitter data and semi-structured interviews can be used in combination should be more evident. NLP methods may in fact become a meaningful, tool for scouring large datasets in support of scale and theory development.

Another critical aspect of this ongoing project is the rationale for how Twitter data may provide broader coverage than semi-structured interviews. One possible explanation, supported by preliminary analysis of themes, is that responses to questions asked in semi-structured interviews are fundamentally different from unprompted expressions posted to social media. Surveys rely on recollections and are subject to reflection and thought while tweets are more impulsive. When asked to recall feelings, survey instruments and interviews are inherently priming the participant, which may lead to inaccurate evaluations related to the prime (Russell 2003) {Citation} and even bias the information retrieval process from memory (Ratcliff and McKoon 1988). On the other hand, extracted tweets which occur unprompted are likely to represent feelings with high levels of activation at the time the tweet was created, and are therefore less likely to be subject to priming or memory biases. Given Twitter users' desire to create and sustain ambient awareness and virtually exhibit themselves, we expected our Twitter data set to result in a broad set of events with high levels of activation. The large number of tweets which contain explicit affective expressions are evidence of this. People are frequently tweeting about loving their phones, wishing to smash their phones, needing or missing their phones, etc., but such affective recollections do not emerge from the semi-structured interviews.

Finally, an exploration of how these methods may be applied beyond the THDS is necessary. Guidelines for affective scale development (McCoach et al. 2013) or the C-OAR-SE scale development method in marketing (Rossiter 2002) might be informative in determining the boundaries within which this approach might be suitable. Affective scales need to have a clear target, intensity and direction (Anderson and Bourke 2000; McCoach et al. 2013). The C-OAR-SE method requires the specification of the object, attribute and rater-entity associated with a scale. In the example above, the object was clearly defined (i.e., technology objects in three categories), the attributes were filtered (i.e. hassles or delights) and the rater-entity was individual users (who use Twitter). Future work on this subject will therefore explore how to expand this NLP-based approach to other areas of interest to IS researchers and other fields. This can potentially contribute significantly to broader utilization of big data sources in traditional research practices.

Practical applications of the THDS and associated concepts also abound. For instance, the insights gained from such analyses can inform design priorities when designing new features of software or hardware. Outside of this specific application, organizations can develop latent measures of consumer sentiment or consumer attitudes of relevance to business success and track these measures as part of customer relationship management or as a voice for customer initiatives. By automating the thematic discovery of technology-related hassles and delights from large unstructured data sources, more time and employee effort can be dedicated to interpretation and developing actions that are matched with the new insights gained.

Biographies

David Agogo (dagogo@som.umass.edu) is doctoral candidate of Operations and Information Management at the Isenberg School at UMass Amherst. His current research focuses on how affective evaluations influence different aspects of user performance on computing tasks. Other interests include privacy, consumer decision-making and data science and analytics.

Traci J. Hess (thess@isenberg.umass.edu) is the Douglas and Diana Berthiaume Endowed Professor in Information Systems in the Isenberg School of Management at UMass Amherst. She received her Ph.D. and M.A. degrees in IS at Virginia Tech, and a B.S. in Accounting from the University of Virginia. Her research interests include human-computer interaction, decision support systems, and user acceptance of IS. Her work has appeared in journals such as *MIS Quarterly, Journal of Management Information Systems, Journal of the Association for Information Systems, Decision Sciences, Decision Support Systems*, and *AIS Transactions on HCI*.

Acknowledgements The authors would like to thank anonymous reviewers from the SIG DSA 2015 Business Analytics Congress for comments that helped streamline and enhance this version of the paper. In addition, the authors would like to thank Brendan O'Connor, Computer Science Department, UMass Amherst, and co-creator of TweetNLP, for his guided support during the development of an earlier manuscript of this paper.

References

Agarwal R, Dhar V (2014) Editorial—big data, data science, and analytics: the opportunity and challenge for is research. Inf Syst Res 25(3):443–448

Anderson JC, Gerbing DW (1991) Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. J Appl Psychol 76(5):732–740

Anderson LW, Bourke SF (2000) Assessing affective characteristics in the schools. Routledge, New York

Asur S, Huberman B et al (2010) Predicting the future with social media. In: 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), vol 1. IEEE, pp 492–499

Beauchamp N (2013) Predicting and interpolating state-level polling using Twitter textual data. In: New directions in analyzing text as data workshop

Beaudry A, Pinsonneault A (2010) The other side of acceptance: studying the direct and indirect effects of emotions on information technology use. MIS Q 34(4):689–6A3

Benbasat I, Barki H (2007) Quo vadis, TAM? J Assoc Inf Syst 8(4):211–218

Bhattacherjee A (2001) Understanding information systems continuance: an expectation-confirmation model. MIS Q 25(3):351–370

Bhattacherjee A, Limayem M, Cheung CMK (2012) User switching of information technology: a theoretical synthesis and empirical test. Inf Manag 49(7):327–333

Boudreau M-C, Gefen D, Straub DW (2001) Validation in information systems research: a state-of-the-art assessment. MIS Q 25(1):1–16

Brill E (2000) Part-of-speech tagging. In: Handbook of natural language processing. CRC Press, Boca Raton, pp 403–414

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5

Burton-Jones A, Straub DW (2006) Reconceptualizing system usage: an approach and empirical test. Inf Syst Res 17(3):228–246

Churchill GA Jr (1979) A paradigm for developing better measures of marketing constructs. J Mark Res:64–73

Clark LA, Watson D (1995) Constructing validity: basic issues in objective scale development. Psychol Assess 7(3):309

De Choudhury M, Counts S, Horvitz E (2013) Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 3267–3276

Gayo-Avello D (2013) A meta-analysis of state-of-the-art electoral prediction from Twitter data. Soc Sci Comput Rev

Ghiselli EE, Campbell JP, Zedeck S (1981) Measurement theory for the behavioral sciences: origin & evolution. WH Freeman & Company

Goodhue DL (2007) Comment on Benbasat and Barki's 'Quo vadis TAM' article. J Assoc Inf Syst 8(4):15

Hinkin TR (1995) A review of scale development practices in the study of organizations. J Manag 21(5):967–988

Hinkin TR (1998) A brief tutorial on the development of measures for use in survey questionnaires. Organ Res Methods 1(1):104–121

Hirschberg J, Manning CD (2015) Advances in natural language processing. Science 349(6245):261–266

Hudiburg RA (1989) Psychology of computer use: Xvii the computer technology hassles scale: revision, reliability, and some correlates. Psychol Rep 65(3f):1387–1394

Hudiburg RA (1992) Factor analysis of the computer technology hassles scale. Psychol Rep 71(3):739–744

Kaplan AM, Haenlein M (2011) The early bird catches the news: nine things you should know about micro-blogging. Bus Horiz 54(2):105–113

Lampos V, Cristianini N (2010) Tracking the flu pandemic by monitoring the social web. In: 2010 2nd International workshop on cognitive information processing (CIP). IEEE, pp 411–416

Lampos V, Cristianini N (2012) Nowcasting events from the social web with statistical learning. ACM Trans Intell Syst Technol 3(4):72

Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Alstyne MV (2009) Computational social science. Science 323(5915):721–723

Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2015) Social media fingerprints of unemployment. PLoS One 10(5):e0128692

Loevinger J (1957) Objective tests as instruments of psychological theory: monograph supplement 9. Psychol Rep 3(3):635–694

Loiacono E, Djamasbi S (2010) Moods and their relevance to systems usage models within organizations: an extended framework. AIS Trans Hum-Comput Interaction 2(2):55–72

MacKenzie SB, Podsakoff PM, Podsakoff NP (2011) Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. MIS Q 35(2):293–334

Manning CD (2011) Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: Computational linguistics and intelligent text processing. Springer, pp 171–189

Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge

McCoach DB, Gable RK, Madura JP (2013) Instrument development in the affective domain. Springer

Nunnally J (1978) Psychometric methods. McGraw-Hill, New York, p 2013

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From Tweets to Polls: linking text sentiment to public opinion time series. ICWSM 11(122–129):1–2

Owoputi O, O'Connor B, Dyer C, Gimpel K, Schneider N, Smith NA (2013) Improved part-of-speech tagging for online conversational text with word clusters. In: HLT-NAACL, pp 380–390

Pennacchiotti M, Popescu A-M (2011) Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 430–438

Rao D, Yarowsky D, Shreevats A, Gupta M (2010) Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on search and mining user-generated contents. ACM, New York, pp 37–44

Ratcliff R, McKoon G (1988) A retrieval theory of priming in memory. Psychol Rev 95(3):385

Rossiter JR (2002) The C-OAR-SE procedure for scale development in marketing. Int J Res Mark 19(4):305–335

Russell JA (2003) Core affect and the psychological construction of emotion. Psychol Rev 110(1):145

Smith NA (2011) Linguistic structure prediction. Synth Lect Hum Lang Technol 4(2):1–274

Steelman ZR, Hammer BI, Limayem M (2014) Data collection in the digital age: innovative alternatives to student samples. MIS Q 38(2):355–378

Straub D, Boudreau M-C, Gefen D (2004) Validation guidelines for IS positivist research. Commun Assoc Inf Syst 13(1):63

Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Soc Networks 34(1):73–81

Zhang P (2013) The affective response model: a theoretical framework of affective concepts and their relationships in the ICT context. MIS Q 37(1):247–274

Chapter 13
Information Privacy on Online Social Networks: Illusion-in-Progress in the Age of Big Data?

Shwadhin Sharma and Babita Gupta

Abstract In the age of big data where vast amounts of data are collected, stored, and analyzed from all possible sources, the growth of social media and the culture of sharing personal information have created privacy and security related issues. Drawing on the prospect theory and rational apathy theory, we present a research model to investigate why people disclose personal information on Online Social Networks. This paper analyzes the impact of situational factors such as information control, ownership of personal information, and apathy towards privacy concern of users on Online Social Network. We describe the proposed research design for collecting our data and analysis using structural equation modeling to analyze the data. The findings and conclusions will be presented after the data is analyzed. This work contributes to the network analytics by developing new constructs using the Prospect Theory and the Rational Apathy theory from the fields of behavioral economics and social psychology respectively.

Keywords Online social networks • Big data • Information privacy • Information control • Ownership of personal information • Privacy apathy • Prospect theory • Information disclosure • Rational apathy

13.1 Introduction

The proliferation of social media and web 2.0 is enabling individuals and companies to engage with digital technologies at an unprecedented scale generating vast amounts of data, also referred to as "big data". Big data is characterized by higher volume, velocity, and variety (the three V's) of data that usually cannot be handled by traditional database management tools (Zikopoulos et al. 2012) and is often characterized as a massive volume of both structured and unstructured data that are

S. Sharma (✉) • B. Gupta
College of Business, California State University, Monterey Bay, Marina, CA, USA
e-mail: ssharma@csumb.edu; bgupta@csumb.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_13

generated at high velocity with veracity that adds value to the intended process (Demchenko et al. 2013; Kshetri 2014).

An Online Social Network (OSN) is an online platform that allows members to create public profiles within a bounded system, share texts, photos and videos, and other personal information, and thus, connect, develop, and maintain relationships (Boyd and Ellison 2007; Ellison et al. 2007). People may use OSNs for many different reasons including socialization, fun and enjoyment, usefulness in communicating and interacting with friends, bridging, bonding, and maintaining social capital. Use of OSNs have created humongous amount of structured and unstructured data. Indeed, the recent attention that big data is garnering can be largely credited to the rapid development of the Online Social Networks (OSNs). As OSNs have provided additional channels for interpersonal and business communication, huge volumes and variety of data are being generated for collection, storage, and aggregation from OSNs. These data can be used by the governments, business organizations, research agencies, marketing companies, etc. Manyika et al. (2011) estimated the value of big data for U.S. medical industry alone to be $300 billion. Companies in various industry sectors such as healthcare, retail, services market, supply chain and transportation, entertainment, and marketing and advertising have started to pay close attention to the big data phenomenon and thus, to OSNs as one of the primary sources of the big data (Tan et al. 2013). It is also important to note that despite several benefits of OSNs in the big data environment, the ability of organizations to collect, store, and analyze big data poses privacy and security related risks for the users.

The interaction of OSNs with users and the generation of big data on OSNs through these interactions are presented in Fig. 13.1 below. The interactions created in OSNs are accessed by many parties such as government, big organizations, third parties, and consumer and service firms. Figure 13.1 presents the simplistic view of how OSNs acts as a source of big data and thus, the source of privacy and security issues.

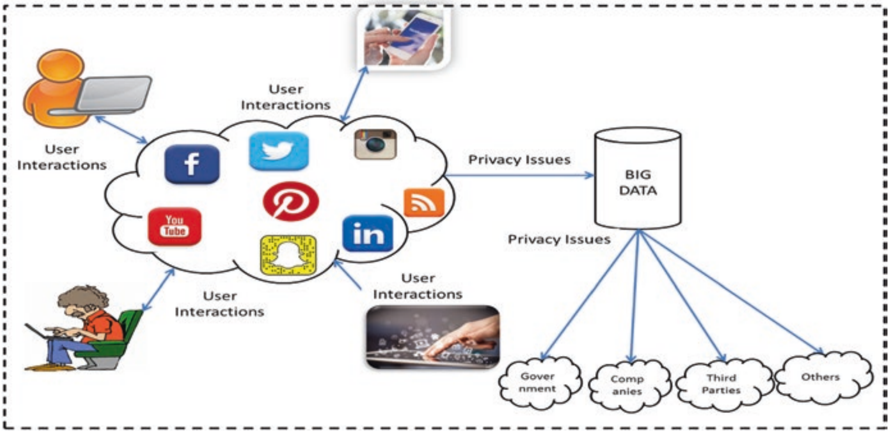


Fig. 13.1 OSN-big data-privacy

The growth of social media and the culture of sharing information have fueled the proliferation of OSNs such as Facebook, Instagram, Twitter, Google+, and Pinterest in individuals' daily life. These OSNs are becoming an important social platform for computer-mediated communication (Nadkarni and Hofmann 2012) at an exponential rate—be it for bonding, bridging, or maintaining social capital (Ellison et al. 2007), or using it as a medium of social interaction and exchanges (Boyd and Ellison 2007). With Facebook alone having more than one billion members (Sharma and Crossler 2014), it is no surprise to see that almost fourth-fifths of the Internet users use one or the other OSN (Conroy and Williams 2014). The exponential growth of OSNs has brought an intense focus on the privacy and security issues of its users. OSNs have been plagued with issues of privacy risks such as the surveillance, secondary use of Information, and collection of irrelevant information (Sharma and Crossler 2014). In the context of the big data, this already complex issue becomes further complicated.

An individual may feel a threat to their privacy when they lose control to their personal information. In an online environment, where users feel a certain amount of anonymity and the OSN providers have the freedom to aggregate and share the information easily, the issues of privacy and security may be more predominant. In a social network environment, information privacy may imply the level of identifiable information collected by the organization and the possible unauthorized uses of that information. These privacy concerns can range from information threats such as digital aggregation and improper access of personal data by third parties to dangers arising from the social environment such as online stalking, bullying, or leaking of private data to the world (Hogben 2007). The level of sophistication of technologies analyzing the big data generated by the OSNs has increased greatly over the last few years. In addition, cost-effective and innovative forms of collection and -processing of high volume, high velocity, and high-variety information assets has brought the privacy and security issues to the forefront (Kshetri 2014). As we start capturing life in digital reality in online social networks, it becomes easier for people and organizations with the right skills set to build an accurate portrait of our past, present, and future behavior, without our knowledge. Software such as Rapid Information Overlay Technology developed for the U.S. defense department (theguardian.com 2013) uses 'extreme-scale analytics' to gather information about individuals' online social network habits to predict their future behaviors. Internet giants like Google and Facebook (including Instagram) have been criticized for a long time for the lack of transparency on what's being done with the users' data they collect. An example of volume, velocity, and variety of data that Facebook stores and can retrieve is the Facebook Graph Search function that was launched in March 2013. This function can give answers to user's natural language queries by combining the big data acquired from its billions of members and external data into a search engine. These results can link Facebook activities such as pictures liked, relationship status, and comments made between a user's friends from the time they joined Facebook. Big data and it tools and techniques that are being used by many OSN companies are opaque, masked by the layers of technical, legal, physical design (Richards and King 2013), making the data being collected and used by these companies questionable. On top of it, there are several third-party applications on OSNs that also collect user information in real-time. As such, real-time structured and unstructured data provided and shared on OSNs such as Facebook, Instagram, Twitter, and Foursquare generally carry privacy risks.

However, even though social media is taking on the role of primary communication, people, especially the millennials, may be in a state of indifference when it comes to their privacy (Yoo et al. 2012). Some of these individuals using OSNs may not be aware of the risk associated with the release of personal information. Others may have experienced privacy invasion and thus, may not consider their information to be private anymore (Solove 2008) becoming apathetic towards their own privacy over time. Some people are comfortable giving up their privacy for patriotic reasons such as for national security while others believe that they have nothing to hide anyway as all of their information is already collected by big organizations like Google or the government (Goitein 2013).

Information Systems (IS) research has focused on privacy and its value. However, we do not yet fully understand why people, despite valuing privacy, still choose to freely share their personal information online. Thus, the concept of privacy on OSN is an interesting one to study as the value of privacy for each individual is situational in nature (Acquisti et al. 2013)—some users may modulate their privacy boundaries; for others, the definition of privacy in itself might vary from one situation/timeline to other. Using the prospect theory and rational apathy theory, this paper analyzes the impact of situational factors such as information control, ownership of personal information, benefits of information disclosure, and apathy towards user's privacy concern on OSN.

13.2 Literature Review

As OSNs are public platforms by design, any information shared on it carries a significant risk of being collected, stored and used without authorization as organizations and third parties such as advertising agents, employers, law enforcement agents, creditors, and tax authorities are increasingly seeking information shared and provided on OSNs users (Hogben 2007; Krasnova et al. 2012; Stieglitz et al. 2014). Privacy related issues can range from negative impact on personal and family lives, damages to reputation (Afroz et al. 2013), identity theft, and psychological pain such as embarrassment and addiction (Turel and Serenko 2012). With the increasing use of OSNs and big data, privacy concern construct has become one of the most widely used variables in IS research to predict the privacy-related behaviors (Dhami et al. 2013; Johnson et al. 2012; Xu et al. 2008). The findings of these privacy-related behavior studies have been often different from one another and sometimes, even contradictory. Some studies found that privacy concerns are more prevalent among the OSN users and negatively impact OSN usage behavior (O'Brien and Torres 2012; Xu et al. 2013). Other studies found that the OSNs users seem to be oblivious to privacy risks and thus, comfortable sharing their personal information on a social network (Hugl 2011; Rosenblum 2007). Despite the privacy risk, the users still use OSNs and share their personal information (Acquisti and Gross 2006; Tufekci 2008). This study explores how the introduction and rise of big data on OSNs would affect the perception of the users toward the privacy concerns and affect their OSN usage behavior.

It is important to study privacy in relation to big data as most of the hacking and privacy violations are now on bigger and broader terms. In 2010, Julian Assange used WikiLeaks to upload 90,000 documents related to Afghan War and started an unprecedented big data leak in the U.S. military history. Edward Snowden followed the trend by publishing 20 times as many documents. The data that was leaked provided a glimpse of how the U.S. government has been performing surveillance activities on its own citizens as well as leaders around the world such as Angela Merkel, Germany's chancellor. Recent big data breaches in Anthem Inc. and Ashley Madison are bringing a lot of attention to privacy violations as well. Big data has allowed people to extract implicit, previously unknown, and potentially personally identifiable information about the individuals.

13.3 Theoretical Framework and Hypotheses

13.3.1 Prospect Theory

Prospect theory states that while making decisions, individuals appraise a set of decision alternatives based on personal heuristics, and then select the alternative that brings the highest satisfaction and outcome (Keith et al. 2012). However, such personal decision heuristics may demonstrate bounded rationality (Simon 1982) as the theory is based on the assumption that utility comes from the returns and not the value of assets. Thus, an individual's reference point would strongly affect the choice of their heuristics (Kahneman and Tversky 1979). This phenomenon has fascinating implications for individuals' decision to share their personal information on OSNs as users compare the utility derived from information sharing to the loss of information through privacy risk.

13.3.2 Rational Apathy Theory

In many cases, individuals are rationally apathetic towards a cause. When a voter feels that his vote would not have any real influence on the conclusion of an election or change the political scenario, he could develop apathy towards the election. Similarly, a rational shareholder would not put an extra effort to go through the length and complexity of proxy statements unless he feels that his effort will make a difference (Karuitha et al. 2013). Apathy is basically defined as a state of indifference or reasoned assessment where an individual has an absence of interest or concern to certain aspects of emotional, social or physical life often caused by "learned helplessness" (Sarfaraz et al. 2012). Similarly, individuals using OSNs may show a non-pathological lack of interest towards their privacy as they may not consider it important (Solmitz 2000) or may believe that their privacy is already too diluted by the companies collecting data to care about it anymore (Yoo et al. 2012). In a privacy context, a person with a reference point of complete information control may quickly travel to the point of privacy apathy.

Drawing from the Prospect theory which is an extension of expected utility hypothesis and from the Rational Apathy theory, we visualize our research model for this study in the Fig. 13.2. This research model has two dimensions to it: one is the privacy calculus that examines risk and benefits of information disclosure, and the other is the reference point for their privacy control, ownership, and belief. In this paper, we investigate how reference points such as information control, perceived ownership, and existing offline benefits affect user's protection belief, risk belief, their state of apathy towards privacy, and perceived benefits from using the OSNs. We further study how all these constructs would affect the user's tendency to disclose information on OSNs. Thus, this paper seeks to answer the following research questions:

1. How do the perceived ownership, perceived information control, and existing offline benefits affect protection belief, risk belief, perceived benefits and state of user's privacy apathy on OSNs?
2. How do privacy apathy, protection belief, risk belief, and benefits affect information disclosure on OSNs in the age of big data?

To control for an explanation of results due to extraneous factors, prior research on OSN and information systems identified a number of factors that may impact the actual behavior of respondents. Analyzing the impact of control variables is essen-

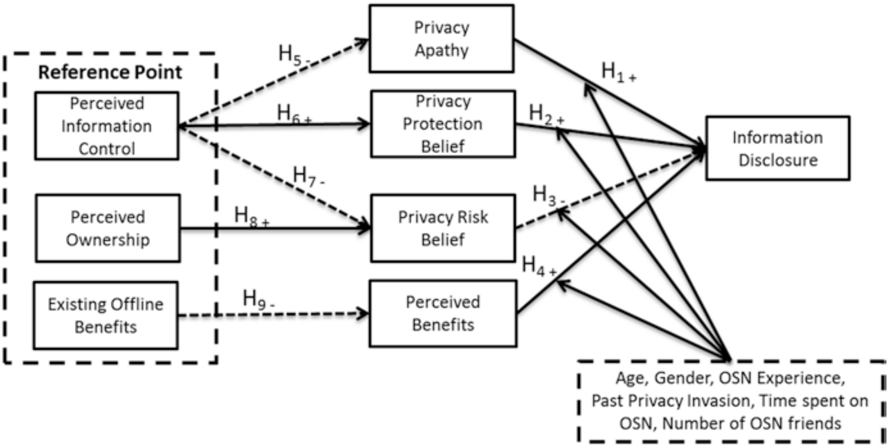


Fig. 13.2 Research model

tial for a research model as it removes any confounding variables (Ormond 2014). Thus, for this research model, age, gender, OSN experience, past privacy invasion, number of OSN friends, number of years of experience on Facebook, and time spent on OSN were included as the control variables to see if they impact the dependent variable.

13.4 Hypotheses Testing

While privacy apathy is a relatively newer concept in IS, the concept has been gaining momentum as a way to gauge the indifference of a user towards privacy concerns (Sharma and Crossler 2014). With big data being collected and stored by millions of web sites, applications, agencies, and third parties, individuals may believe that there is no such thing as privacy in the age of Web 2.0 technologies. As stated by Mark Zuckerberg, the co-founder of Facebook on January 2010, privacy is no more a "social norm". Similar sentiments were echoed by the United States Senate majority leader Harry Reid when he advised to "just calm down and understand that National Security Agency's (NSA) PRISM isn't anything that is brand new" (csmonitor.com 2013). Similarly, a recent survey showed that almost half of the Americans take NSA's PRISM program of data surveillance as "no big deal" as these people believe that "they're being tracked all over the Internet by companies like Google and Facebook" (csmonitor.com 2013). Thus, it is safe to hypothesize that users with privacy apathy put lower value and price to their personal information and thus, care less about information disclosure (Yoo et al. 2012).

H1: Privacy apathy would positively influence intention to disclose information on OSNs despite the threat of big data.

Privacy protection belief is the subjective possibility that consumers believe that their private information is protected as anticipated (Metzger 2004). In an online setting, users who exemplify higher protection beliefs are believed to have more control over their information and thus, are more in control over information disclosure and are more likely to disclose their personal information (Raschke et al. 2014). Thus, it is predicted that:

H2: Privacy protection belief would positively influence intention to disclose information on OSNs despite the threat of big data.

Privacy risk belief implies the probability of potential loss because of disclosure of personal information (Malhotra et al. 2004). It is deemed to be the cost of privacy as disclosing information is often considered risky. Such cost and risks associated to OSN can range from unintended third parties receiving users' personal information to hacking of personal account based on information shared on OSN (Hogben 2007). Several studies have verified the negative effect of perceived privacy risk on people's intention to disclose personal information on online transactions and activities (Li et al. 2010; Malhotra et al. 2004).

H3: Privacy risk belief would negatively influence intention to disclose information on OSNs despite the threat of big data.

Perceived benefits refer to a user's overall expectation of positive outcomes from an OSN without any significant privacy threats (Bulgurcu 2012). Individuals are likely to give up a degree of privacy in return for potential benefits related to OSNs. In an OSN environment, the user's fear in the form of losing control of personal information is compensated by the several benefits such as information, enjoyment, and convenience (Hogben 2007). Thus, the following is hypothesized:

H4: Perceived benefit would positively influence intention to disclose information on OSNs despite the threat of big data.

Privacy and control has often been linked together in prior work (Westin 1967). The ability of people to control their information has been emphasized as critical in any concept of privacy (Wolfe and Laufer 1974). Thus, it is no surprise to see that there has been an outcry regarding how users have lost control of their information on OSNs (Boyd 2008; Hoadley et al. 2010). When people tend to share information on OSN, it is often broadcasted to the network of friends. Sometimes, such information is accessed by the third party applications installed by the users. An individual believing lower information control on OSNs would believe that such information has been collected and stored by OSNs and third parties and thus, would have higher privacy apathy. Similarly, a sense of higher information control would lead to a positive privacy protection belief and a lower privacy risk belief. Thus, we posit:

H5: Perceived information control would negatively influence privacy apathy.
H6: Perceived information control would positively influence privacy protection belief.
H7: Perceived information control would negatively influence privacy risk belief.

Perceived ownership implies a sense of possession and entitlement (Furby 1978). In the case of information and OSNs, perceived ownership implies the sense of entitlement, possession, and attachment towards the information shared on OSNs (Feuchtl and Kamleitner 2009; Sharma and Crossler 2014). When individuals believe that the information shared on OSN is their information and contains some level of attachment with their identity and privacy, it positively influences their privacy risk belief (Sharma and Crossler 2014). Thus, it is hypothesized that:

H8: Perceived ownership would positively influence privacy risk belief.

Away from OSN, there are tremendous opportunities for people to maintain a relationship, enjoy life, consume information, and develop an offline real-life image. Users of OSN will perceive lower benefits from OSN use when they are enjoying many of similar benefits offline in their real life. Thus, existing benefits decrease the perceived benefits of future disclosure (Keith et al. 2012). Thus, we propose:

H9: Existing offline benefits would negatively influence perceived benefits of OSNs.

13.5 Hypotheses Testing

The proposed conceptual model will be evaluated using survey design. An online questionnaire survey has been developed to collect data and perform empirical tests of the relationships proposed in the research model. The survey design technique fits the research phenomenon being studied in this research as the objective of this research is to explore user's information disclosure behavior on OSN. Also, a survey design provides the benefit of generalizability to the study as data could be collected from a wider range of respondents.

All the items used in the survey instrument are adapted from previous studies. The items are reflective likert-scale and have been adapted to fit the context of this study. The items for this study along with their respective original source/s have been presented in Table 13.1 below:

Although constructs adopted from earlier studies have been rigorously tested for reliability and validity, additional content validation using a multi-stage iterative procedure is recommended (Churchill 1979). Podsakoff et al. (2003) also have suggested using an ex-ante approach such as expert panel review and a pilot test to control Common Method Bias (CMV). Thus, a preliminary investigation consisting of expert panel review, pretest, and pilot test will be conducted to ensure measurement validity of the instrument. The changes suggested by the expert panel review and pre- and pilot tests such as revisions to wordings to improve clarity and precision, dropping items to make the survey fatigue-free, revision of items to make them unambiguous, etc. will be incorporated. This will ensure content validity of our survey instruments and also reduce CMV. Similarly, to reduce CMV we will keep our survey anonymous, optional, and relatively short. We will also assess the extent of common method variance with two statistical tests. First, we will perform Harman's single factor test by loading all of the items in a principal component factor analysis (Podsakoff et al. 2003). If the results show that there is more than a single factor that accounts for a majority of covariance, it would suggest absence of CMV in our study. However, as Harman's single factor test is increasingly contested for its ability to detect common method bias, we will also use Lindell and Whitney's (2001) test that uses a theoretically unrelated construct (termed a marker variable) to assess CMV. We will use "Perceived effectiveness of credit card guarantees" as our marker variable construct for this study and will use it to adjust the correlations among the principal constructs (Pavlou et al. 2007). The absence of high correlations among any of the items of the study's principal constructs and perceived effectiveness of credit card guarantees would indicate that the study doesn't have serious concerns about common method bias as the construct perceived effectiveness of credit card guarantees is expected to be weakly related to the study's principal constructs.

Undergraduate students from different classes within a public university in California will be invited to complete the survey. The invitations will be sent through emails as well through classroom visits. As the age group of 18–25 years that are

Table 13.1 Pilot survey instrument

Survey instrument

Your Personal Information implies information that is related to you, can be used on its own or with other information to identify, contact, or locate you. Some of the examples of Personal Information can be address, location, race, relationship history, purchasing behavior, phone number, pictures (and tagging), etc.

OSN refers to Online Social Network that includes interactions on social media and web 2.0 technology platforms such as Facebook, Instagram, Snapchat, Vine, Youtube, Twitter, etc.

Construct	**Adapted item**	**Original source**
Perceived ownership (PO)	Information I share while on/to OSN is MY personal information.	Van Dyne and Pierce (2004)
	I sense that the information I provide on/to OSN is my own.	
	I feel a very high degree of personal ownership for the information I provide on/to OSN.	
	I sense that the information I provide on/to OSN is personal.	
	I believe that the information I disclose on/to OSN belongs to me.	
Privacy apathy (PA)	I have little interest in information privacy issues on information provided on/to OSN.	Yoo et al. (2012); Sharma and Crossler (2014)
	I care less about information privacy anymore on information provided on/to OSN.	
	I do not worry about privacy issues anymore on information provided on/to OSN.	
Privacy protection belief	I am confident that I know all the parties who would collect information that I share on/to OSN.	Li et al. (2011)
	I am aware of the exact nature of information that will be collected, stored, and used by OSN.	
	I believe there is an effective mechanism to address any violation of the information I provide on/to OSN	
	I am confident that I know all the parties who would collect information that I share on/to OSN.	
Privacy risk belief of information disclosure (PRB)	Sharing information on/to OSN would involve many unexpected problems.	Malhotra et al. (2004); Xu et al. (2009)
	It would be risky to disclose information on/to OSN.	
	There would be too much uncertainty with providing information on/to OSN	
	There would be high potential for loss in disclosing information on/to OSN.	

(continued)

Table 13.1 (continued)

Perceived information control	I believe I have control over the amount of your personal information collected on OSN.	Xu (2007)
	I believe I have control over who can get access to my personal information on OSN.	
	I believe I have control over my personal information that has been released on OSN.	
	I believe I have control over how my personal information is being used by OSN.	
	I believe I have control over my personal information that I provided on/to OSN.	
Perceived benefits	OSN is useful to exchange personal information with my friends.	Ellison et al. (2007); Krasnova et al. (2010)
	OSN is useful for me to monitor what others share about themselves.	
	Sharing personal information on OSN is fun.	
	By sharing personal information on OSN, I get more popular with my OSN-friends.	
	I share personal information via OSN because it's better than the alternatives.	
Existing offline benefits	I have more time to spend with my family and friends around me.	Self-developed
	Staying offline has several benefits than staying online.	
	I can build real relationships and stay happy and healthy when I am offline.	
	I have more time to pursue my hobbies and pursuits and form network with people I know.	
Behavioral intent to disclose information (BINT)	I am likely to provide my personal information on/to OSN.	Xu and Teo (2004)
	I plan to provide my personal information on/to OSN.	
	I intend to provide my personal information on/to OSN.	

educated and college students are the ones that use the OSNs the most (Lenhart et al. 2010), it is appropriate to have undergraduate students as the sample for this study.

A primary investigation consisting of reliability and validity testing, model fit test (i.e. goodness of fit), common method bias test, and t-test is conducted to ensure the validity of the structural model. We will use SmartPLS 2.0, SPSS along with AMOS for our instrument validation and testing of the structural model. SmartPLS uses a Partial Least Square (PLS) regression technique that employs a component-based approach for estimation and places minimal restrictions on sample size, measurement scales and residual distributions (Chin and Todd 1995) and it does not impose normality requirements on the data. We will also use AMOS which is a covariance based structured equation model that provides various overall goodness-of-fit indices for assessing model fit and method variance.

Before testing the hypothesized structural model, we evaluate the psychometric properties of the measures. All the constructs in this study are measured with multiple items. A PLS confirmatory analysis will be conducted to examine convergent validity, discriminant validity, and reliability using commonly accepted guidelines (Churchill 1979). Reliability for the constructs will be measured using composite reliability score and Cronbach's alpha. The Cronbach's alpha and composite reliability examine the internal consistency among the data. For all the constructs in our study, we will also perform descriptive statistics of all the constructs including means and standard deviations and the level of each item's contribution to the overall factor.

To further examine the validity of the measurement model, we will analyze how well the model fits the data with the help of model fit statistics available through AMOS (Anderson and Gerbing 1988). The goodness of fit index (GFI), comparative fit index (CFI), normed fit index (NFI) and incremental fit index (IFI) all assess the goodness of fit of the model with the data and should be above 0.90 to show model fit. Root mean square error of approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) which measures the "badness of fit" should both be below 0.05. Similarly, we will also assess if the relative chi-square (i.e. CMIN/df), which is also a "badness of fit", is below the threshold of 3 (Kline 1998) and is thus non-significant. Together, the result would indicate if our hypothesized measurement model is "fitting" the observed data.

The hypotheses and the relationships used for this study will be tested by examining the structural model of our study. A bootstrapping resampling procedure will be performed to assess the significance of the path coefficients within the structural model. The proposed hypotheses for this research model will be tested using t-statistics (p-value) for the standardized path coefficients. The t-statistics (p-value) provided by PLS structural model analysis would show us if the hypothesis is supported or not while the standardized path coefficients would determine the direction and strength of the relationship between exogenous and endogenous variables. The study will have a satisfactory and substantive model if the dependent factors have R-square (the variance explained by the independent variables) greater than 0.10 (Falk and Miller 1992). Thus, we will examine if the proposed paths were mostly significant and how well our model explained the variances in our endogenous variables. Also, we will analyze the effect of our control variables age, gender, education, experience in social networking, and experience on the Internet on the intention to disclose information on OSNs.

13.6 Conclusion

13.6.1 Study Summarization

The objective of this research is to explore the factors that may affect user's information disclosure behavior on online social networks. There has been limited research on why consumers choose to disclose personal information on OSN despite valuing in privacy. With big data analytics taking the center stage and OSNs becoming as the primary source of big data, privacy and security in social networks have become increasingly important. This research looks at factors that may explain individual's information disclosure behavior. We use prospect theory and rational apathy theory in our research model. As outlined in our research model, the information disclosure decision of the invidiual depends on variables such as the privacy risk, protection beliefs, and perceived benefits to disclose information on OSNs. The users information disclosure decisions would also be affected by their belief about who owns the information being provided and how the information that has already been collected and stored by social media companies is being used by these companies. Thus, this study may help us expand the concept of apathy, risk belief and privacy calculus in regard to the context of information disclosure behavior.

We will be using survey research as our research methodology as this helps in increasing the generalizability of this study. The survey instrument for this research will be hosted in Qualtrics. The main data will be collected from students as they represent the general demographic that use online social network the most. Prior to collecting the data, a preliminary investigation will include expert panel reviews, pre-test, and pilot studies to confirm the reliability and validity of the survey instrument. The loadings, cross-loadings, content and face validity, and reliability will also be examined during the pilot study. Then the structural model will be used to test the path coefficient and t-values for our hypotheses.

13.6.2 Key Findings

We will discuss the key findings based on our data analysis in the future publication.

13.6.3 Contribution and Implications

This paper has several theoretical and practical implications. First, this paper brings together the concept of big data and OSNs to analyze the privacy behavior of OSN users. Previous research on privacy and information disclosure has focused on internet transactions, eCommerce, and social networks (Dinev and Hart 2006; Keith et al. 2012) but the concept of big data and its impact on privacy behavior of OSN users has been studied by very few researchers. Second, this paper brings the concept of privacy apathy and perceived ownership into focus. Users of OSNs are believed to be worried about losing privacy in the age of big data. This paper seeks to study if some of the users would care less about privacy if they believe that they have already lost the ownership of their information on the internet. Third, we are also expanding the prospect theory and apathy theory and the concept of the reference point in guiding OSN users to decide about their privacy-related behavior.

IS research has regularly faced the criticism of lacking relevance to practice (Baskerville and Myers 2004; Benbasat and Zmud 1999). As such, this paper provides value to practitioners in many different ways. First, this study is helpful to OSN providers and third parties as these organizations would now understand how consumer's information disclosure behavior works. Second, this study helps us understand why people tend to disclose too much of their personal information on OSNs.

13.6.4 Limitations of this Study

McGrath (1995) stated that all research methods are inherently flawed, though each is flawed differently. Thus, the role of the researcher is always to minimize the flaws associated with the research by maximizing the three criteria of good research: generalizability, precision, and realism). This research is no exception to other research and thus, has its limitations. Some of the limitations of this study pertain to the generalizability of the study due to sample frame used for this study, theoretical constructs excluded from the study to achieve parsimonious research model, research method used for testing the proposed model, and use of self-reported scales. However, understanding these limitations also provides with the opportunities for future research. As this research is a research-in-progress, understanding these limitations will also provide us with the opportunities for strengthening the next steps in our research plan and our future research.

Biographies

Shwadhin Sharma is an Assistant Professor in the College of Business at California State University Monterey Bay. His research interests are in the areas of technical and behavioral aspects of big data analytics, privacy and security, electronic commerce and social commerce, role of dispositional factors in IT, and IT adoption and discontinuation. He has published his research in journals such as Journal of Computers Information Systems, Electronic Commerce Research and Applications and Computers & Security and academic conferences. He has served as reviewer for several reputed journals and conferences. He is currently serving on the editorial board of International Journal of the Internet of Things and Cyber-Assurance. He has also co-chaired as a SIGDSA mini-track on "Social Network Analytics in Big Data Environment" in AMCIS 2016.

Babita Gupta is Professor of Information Systems and the Director of AACSB Accreditation at the College of Business, California State University Monterey Bay. She teaches courses in information technology innovation strategies, business intelligence & analytics, database management, and information systems for decision making. Her research interests are in the areas of online security and privacy, business intelligence strategies, technology adoption, and the role of culture in IT. She has published in journals such as the *Communications of the Association for Information Systems (CAIS)*, *Journal of Electronic Commerce Research (JECR)*, the *Journal of Strategic Information Systems (JSIS)*, the *Communications of the ACM (CACM)*, the *Journal of Industrial Management and Data System*, the *Journal of Scientific and Industrial Research*, the *Journal of Information Technology Cases and Applications*, and the *Journal of Computing and Information Technology*.

She serves as an *Advisory Board Member* of the Teradata University Network, a non-profit division of Teradata.com. She was elected as the *Program-Chair-Elect Officer* in 2012 for the Special Interest Group on Decision Support and Analytics (SIGDSA) under the Association for information Systems (AIS) organization. She has also served as a Board Member of the California Coastal Rural Development Corporation for over a decade.

References

Acquisti A, Gross R (2006) Imagined communities: awareness, information sharing, and privacy on the Facebook, privacy enhancing technologies. Springer, Berlin, pp 36–58

Acquisti A, John LK, Loewenstein G (2013) What is privacy worth? J Leg Stud 42(2):249–274

Afroz S, Islam AC, Santell J, Chapin A, Greenstadt R (2013) How privacy flaws affect consumer perception. Socio-Technical Aspects in Security and Trust (STAST), 2013 Third Workshop on: IEEE, pp 10–17

Anderson JC, Gerbing DW (1988) Structural equation modeling in practice: a review and recommended two-step approach. Psychol Bull 103(3):411–423

Baskerville R, Myers MD (2004) Special issue on action research in information systems: making is research relevant to practice foreword. Manag Inf Syst Q 28(3):329–335

Benbasat I, Zmud RW (1999) Empirical research in information systems: the practice of relevance. Manag Inf Syst Q 23(1):3–16

Boyd D (2008) Facebook privacy trainwreck: exposure, invasion and social convergence. Convergence 14(1):13–20

Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13(1):210–230

Bulgurcu B (2012) Understanding the information privacy-related perceptions and behaviors of an online social network user. University of British Columbia, Vancouver

Chin WW, Todd PA (1995) On the use, usefulness, and ease of use of structural equation modeling in MIS research: a note of caution. Manag Inf Syst Q 19(2):237–246

Churchill GA (1979) A paradigm for developing better measures of marketing constructs. J Mark Res 16(1):64–73

Conroy S, Williams A (2014) Use of internet, social networking sites, and mobile technology for volunteerism. AARP Office of Volunteerism and Service

csmonitor.com (2013) The danger of American apathy on NSA surveillance. The Christian Science Monitor. http://www.csmonitor.com/Commentary/Opinion/2013/0731/The-danger-of-American-apathy-on-NSA-surveillance. Accessed 19 Jan 2014

Demchenko Y, Grosso P, De Laat C, Membrey P (2013) Addressing big data issues in scientific data infrastructure. In: Proceedings of international conference on collaboration technologies and systems (CTS), San Diego, CA, pp 48–55

Dhami A, Agarwal N, Chakraborty TK, Singh, BP, Minj J (2013) Impact of trust, security and privacy concerns in social networking: an exploratory study to understand the pattern of information revelation in Facebook. In: Proceedings of 3rd international advance computing conference (IACC), pp 465–469

Dinev T, Hart P (2006) An extended privacy calculus model for E-commerce transactions. Inf Syst Res 17(1):61–80

Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "friends:" social capital and college students' use of online social network sites. J Comput-Mediat Commun 12(4):1143–1168

Falk RF, Miller NB (1992) A primer for soft modeling. University of Akron Press, Akron

Feuchtl S, Kamleitner B (2009) Mental ownership as important imagery content. Adv Consum Res 36(2):995–996

Furby L (1978) Possession in humans: an exploratory study of its meaning and motivation. Soc Behav Personal 6(1):49–65

Goitein E (2013) The danger of American apathy on NSA surveillance. The Christian Science Monitor. Available on November 3 from http://www.csmonitor.com/Commentary/Opinion/2013/0731/The-danger-of-American-apathy-on-NSA-surveillance. Accessed 31 Jul 2013

Hoadley MC, Xu H, Lee J, Rosson MB (2010) Privacy as information access and illusory control: the case of the Facebook news feed privacy outcry. Electron Commer Res Appl 9(1):50–60

Hogben G (2007) Security issues and recommendations for online social networks. Enisa Position Paper 1:1–36

Hugl U (2011) Reviewing person's value of privacy of online social networking. Internet Res 21(4):384–407

Johnson M, Egelman S, Bellovin SM (2012) Facebook and privacy: it's complicated. In: Proceedings of the eighth symposium on usable privacy and security, New York, USA, pp 9–15

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–291

Karuitha JK, Onyuma SO, Mugo R (2013) Do stock splits affect ownership concentration of firms listed at the Nairobi securities exchange? Res J Finan Acc 4(15):105–117

Keith MJ, Thompson SC, Hale J, Greer C (2012) Examining the rationality of information disclosure through mobile devices. In: Proceedings of 33rd international conference on information systems, Orlando, USA, pp 1–17

Kline RB (1998) Principles and practice of structural equation modeling. Guilford Press, New York

Krasnova H, Spiekermann S, Koroleva K, Hildebrand T (2010) Online social networks: why we disclose. J Inf Technol 25(2):109–125

Krasnova H, Veltri NF, Günther O (2012) Self-disclosure and privacy calculus on social networking sites: the role of culture. Bus Inf Syst Eng 4(3):127–135

Kshetri N (2014) Big data's impact on privacy, security and consumer welfare. Telecommun Policy 38(11):1134–1145

Lenhart A, Purcell K, Smith A, Zickuhr K (2010) Social media and mobile internet use among teens and young adults. Millennial. Pew Internet and American Life Project, Washington

Li H, Sarathy R, Xu H (2010) Understanding situational online information disclosure as privacy calculus. J Comput Inf Syst 51(1):62–71

Li H, Sarathy R, Xu H (2011) The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. Decis Support Syst 51(3):434–445

Lindell MK, Whitney DJ (2001) Accounting for common method variance in cross-sectional research designs. J Appl Psychol 86(1):114–121

Malhotra NK, Kim SS, Agarwal J (2004) Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Inf Syst Res 15(4):336–355

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute

McGrath E (1995) Methodology matters: doing research in the behavioral and social sciences. In: Human-computer interaction. Morgan Kaufmann, San Francisco, pp 152–169

Metzger MJ (2004) Privacy, trust, and disclosure: exploring barriers to electronic commerce. J Comput-Mediat Commun 9(4)

Nadkarni A, Hofmann SG (2012) Why do people use Facebook? Personal Individ Differ 52(3):243–249

O'Brien D, Torres AM (2012) Social networking and online privacy: Facebook users' perceptions. Ir J Manag 31(2):63–97

Ormond DK (2014) The impact of affective flow on information security policy compliance. In: M.S. University (ed) pp 1–178

Pavlou PA, Liang H, Xue Y (2007) Understanding and mitigating uncertainty in online exchange relationships: a principal-agent perspective. Manag Inf Syst Q 31(1):105–136

Podsakoff PM, MacKenzie SB, Lee J-Y, Podsakoff NP (2003) Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 88(5):879–903

Raschke RL, Krishen AS, Kachroo P (2014) Understanding the components of information privacy threats for location-based services. J Inf Syst 28(1):227–242

Richards NM, King JH (2013) Three paradoxes of big data. Stanford Law Review Online 66(41):41–46

Rosenblum D (2007) What anyone can know: the privacy risks of social networking sites. IEEE Secur Priv 5(3):40–49

Sarfaraz A, Ahmed S, Khalid A, Ajmal MA (2012) Reasons for political interest and apathy among university students: a qualitative study. Pak J Soc Clin Psychol 10(1):61–67

Sharma S, Crossler RE (2014) Disclosing too much? Situational factors affecting information disclosure in social commerce environment. Electron Commer Res Appl 13(5):305–319

Simon HA (1982) Models of bounded rationality. MIT Press, Cambridge

Solmitz DO (2000) The roots of apathy and how schools can reduce apathy. Available from https://dwaynehoward.wordpress.com/2012/05/14/the-roots-of-apathy/

Solove DJ (2008) Understanding privacy. Harvard University Press, Cambridge

Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C (2014) Social media analytics. Bus Inf Syst Eng 6(2):89–96

Tan W, Blake MB, Saleh I, Dustdar S (2013) Social-network-sourced big data analytics. IEEE Internet Comput 5:62–69

theguardian.com (2013) Software that tracks people on social media created by defence firm. Available from http://www.theguardian.com/world/2013/feb/10/software-tracks-social-media-defence

Tufekci Z (2008) Can you see me now? Audience and disclosure regulation in online social network sites. Bull Sci Technol Soc 28(1):20–36

Turel O, Serenko A (2012) The benefits and dangers of enjoyment with social networking websites. Eur J Inf Syst 21(5):512–528

Van Dyne L, Pierce JL (2004) Psychological ownership and feelings of possession: three field studies predicting employee attitudes and organizational citizenship behavior. J Organ Behav 25(4):439–459

Westin AF (1967) Privacy and freedom. Atheneum, New York

Wolfe M, Laufer RS (1974) The concept of privacy in childhood and adolescence. In: Margulis ST (ed) Privacy as a behavioral phenomenon, symposium presented at the meeting of the environmental design research association, Milwaukee

Xu H (2007) The effects of self-construal and perceived control on privacy concerns. In: Proceedings of international conference on information systems, Montreal, Canada, pp 1–14

Xu H, Teo HH (2004) Alleviating consumer's privacy concern in location-based services: a psychological control perspective. In: Proceedings of the twenty-fifth international conference on information systems, Charlottesville, Virginia, pp 793–806

Xu H, Dinev T, Smith HJ, Hart P (2008) Examining the formation of individual's privacy concerns: toward an integrative view. In: Proceedings of 29th international conference on information, Paris, France, pp 1–16

Xu H, Teo HH, Tan BC, Agarwal R (2009) The role of push-pull technology in privacy calculus: the case of location-based services. J Manag Inf Syst 26(3):135–174

Xu F, Michael K, Chen X (2013) Factors affecting privacy disclosure on social network sites: an integrated model. Electron Commer Res 13(2):151–168

Yoo CW, Ahn HJ, Rao HR (2012) An exploration of the impact of information privacy invasion. In: Proceeding of thirty third international conference on information systems, Orlando, Florida, pp 1–18

Zikopoulos PC, Eaton C, Deroos D, Deutsch T, Lapis G (2012) Understanding big data: analytics for enterprise class and streaming data. McGraw-Hills books. eBook, http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF

Chapter 14
Online Information Processing of Scent-Related Words and Implications for Decision Making

Meng-Hsien (Jenny) Lin, Samantha N.N. Cross, William Jones, and Terry L. Childers

Abstract This paper takes a multi-method approach, combining neuroscience methods and behavioral experiments to investigate emotions triggered by olfactory-related information and related consumer decision-making outcomes. In the online context, olfactory information is limited to visual forms of triggering olfactory sensations. The effectiveness of using sensory congruent brand names in online ads to trigger emotions, and the influence on attitudes toward the ad, brand and purchase intentions are examined. Moreover, individual differences in olfactory sensitivity were considered, revealing moderating effects on cognitive and emotional processes. Findings provide managerial and organizational implications for online advertising, branding decisions and market segmentation decisions.

Keywords Emotions • Information processing • Neuroscience • Olfactory • Sensory

14.1 Introduction

With the rise of e-commerce and online-based shopping, the trend from retail e-commerce sales in the U.S. is growing from 225.5 billion U. S. dollars revenue in 2015 and is predicted to almost double to 434.2 billion in 2017 (Statista 2015). Online advertising and promotional strategy decisions become all the more

M.-H. Lin, Ph.D. (✉)
College of Business, California State University, Monterey Bay, Seaside, CA 93955, USA
e-mail: jelin@csumb.edu

S.N.N. Cross • T.L. Childers
Department of Marketing, Iowa State University, Ames, IA 50011, USA
e-mail: snncross@iastate.edu; tchilders@iastate.edu

W. Jones
College of Business, University of South Dakota, Vermillion, SD 57069, USA
e-mail: William.Jones@usd.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information Systems 21, DOI 10.1007/978-3-319-58097-5_14

influential in sales and purchase decisions of consumers. Product categories that rely on attributes that are processed by senses other than visual may be challenged to reinvent and accommodate the lack of sensory input. For example, consumers typically base their purchase decisions for products such as laundry detergent on a set of important attributes, including olfactory input. However, scented products have less leverage in online platforms. To compensate for the lack of odor and scent information provided online for decision-making processes, visual olfactory-related information is the main source of reliance for judging and influencing scent-based purchase decisions made online. In this paper, we examine how purchase decisions are made in the absence of actual scent in online shopping scenarios, focusing on visual information, such as branding and advertising strategy decisions.

Scent is strongly associated with emotions and memory (Goldkuhl and Styvén 2007; Morrin and Ratneshwar 2003) which are formed early in human development and are enhanced through life (Holland et al. 2005). Factors contributing to individual differences in olfaction include demographic factors such as gender (Ship and Weiffenbach 1993), age (Doty et al. 1984; Ship and Weiffenbach 1993) and culture (Herz 2007). However, other evidence indicates that sensitivity to scent exists across individuals in the population (Chebat et al. 2009; Cross et al. 2015). We argue that varying differences in olfactory ability across consumers can have an impact on the intensity of emotions perceived. We also contend that examining valence effects can assist us in understanding how valence is processed across the different olfactory groups.

Our research question is twofold. First, how do individuals process scent-related words emotionally, in the absence of actual scent, in an online environment? Further, how do individual differences in sensitivity to smell play a role in providing a nuanced understanding of purchase decisions (of products where scent is relevant) and emotional processing of scented-brand names (using scent associated words in brand names)? To address these questions, two studies were conducted to understand purchase behaviors online and the underlying emotional processes.

Two different methods were combined in this paper to explore the online purchase behaviors of individuals with varying olfactory orientations and investigate the underlying affective processes involved. In the first study, emotional reactions in response to reading scent-related words (vs. non-scent-related words) across individual olfactory ability were examined using electroencephalography (EEG). A passive task (bottom-up processing) of reading is compared to a task involving a more elaborate process incorporating olfactory imagery (top-down processing; cf. Hajcak and Nieuwenhuis 2006). Another dimension included in the investigation was the valence (pleasant vs. unpleasant) of odor-associated words. Valence differences are compared for scent-related words relative to non-scent related words in order to examine the emotional processes underlying passive reading and olfactory imagery. In a second study, the impact of scented brand names (vs. non scent related brand names) on online purchase decisions, attitude towards the product and product performance were investigated using online ads. Discoveries from the first study provide a deeper understanding of how regulation of emotions varies across individuals based on olfactory orientations and hence influences online decision making process.

Findings from this paper provide implications for branding and online advertising managerial decisions. These decisions involve very different considerations when compared to offline branding and advertising decisions, particularly for organizations in the scent product industry. This paper focuses on understanding the nuances in consumer online information processing and provides additional insight for supporting managerial and organizational decisions on market segmentation and targeting strategies.

14.2 Study 1: Individual Differences in Affective Responses to Scent-Related Words

14.2.1 Literature Review and Hypotheses

The relationship between odor and emotions are strongly connected (Herz et al. 2004) and consequently influences the perception and decisions of consumers (Chebat and Michon 2003; Bone and Ellen 1999). Previous studies have found that in the absence of actual scent, olfactory imagery can play a significant role in inducing sensations similar to that of processing actual odors, as evidenced by neuroscience data (Bensafi et al. 2003). Past research has focused on the effect of odors in the marketplace and its impact on purchase decisions and behavior. However, the "experience of odor" can be elicited without the scent being present, as in the form of imagined odors. Stevenson and Case (2005, p. 244), defined olfactory imagery as "being able to experience the sensation of smell when an appropriate stimulus is absent." They noted how this had resulted from cumulative evidence, mostly self-reported data, in three forms: (1) participants report such experiences; (2) descriptions of these experiences are similar to those of actual smelling; and (3) their reactions to certain forms of these experiences involve appropriate behavioral responses.

Odor valence, pleasant versus unpleasant, is weighted asymmetrically within individuals. In particular, unpleasant odors have a functional purpose—human survival. Thus we believe that the effect of odor valence (represented by pleasant or unpleasant odor-associated words in this study) will vary across the two olfactory groups: (1) individuals with a normal sense of smell and (2) individuals with a heightened sense of smell. However, regardless of individual sensitivity to smell, which can be categorized into: heightened, normal or decreased (Cross et al. 2015), unpleasant odor associations, represented by unpleasant odor-associated words, will elicit increased emotions compared to non-odor associations. This reflects the function of unpleasant odors as a warning against exposure to ingestion of hazardous or harmful substances. Also, there is a higher probability of activation of the amygdala for negative emotions, such as fear and disgust, relative to positive emo-

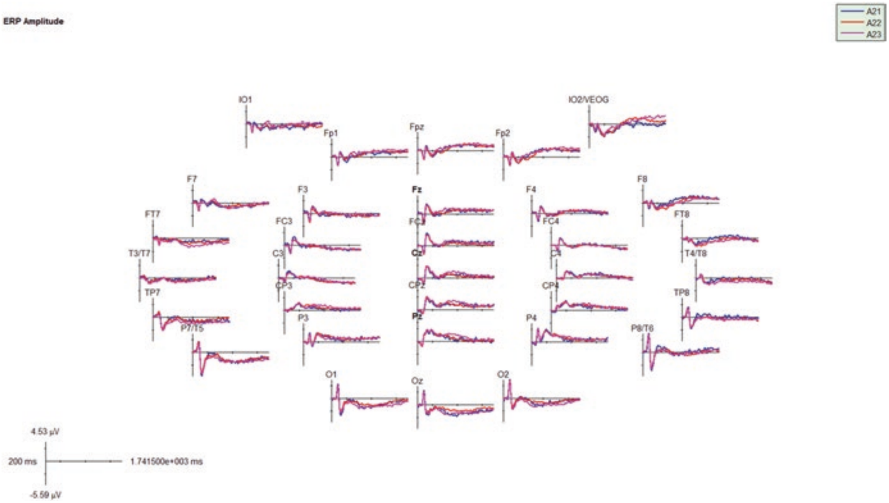


Fig. 14.1 Scalp distribution of Late positive potential (LPP). Displayed are grand average taken from read task. Non-olfactory words (*blue*); pleasant olfactory words (*red*); unpleasant olfactory words (*pink*)

tions such as happiness. This is particularly so for the gustatory-olfactory modalities (Costafreda et al. 2008). In individuals with a normal sense of smell, positive affect is expected to increase when reading positive olfactory words. On the other hand, attenuation of emotions is expected in sensitive individuals. Hence, we predict pleasant odor-associated words will not elicit enhanced LPP, resulting in comparable levels of LPP as the controlled condition. We contend that this reaction is likely due either to the keen olfactory sensitivity in these individuals, negative associations from past experiences through perceived intensity of the scent or suffering from ill effects of scent (Cross et al. 2015).

Event-related potential (ERP) studies utilize what are referred to in the literature as “components.” These components have both temporal and spatial characteristics. One such component is the late positive potential (LPP), which is commonly identified spatially in the brain as a midline centroparietal activation that temporally occurs after 300 ms post stimulus and may last up to 1500 ms (see Fig. 14.1 for scalp distribution of LPP). LPP evidences information processing operations associated with emotions and arousal, although it is important to point out that no specific event-related potential (ERP) component has been identified for a certain type of emotion (e.g., disgust). Researchers have used the LPP component to study emotion-relevant stimuli in comparison to neutral stimuli (Cunningham et al. 2005; Hajcak and Nieuwenhuis 2006; Lang et al. 1998; Schupp et al. 2006). Another important characteristic of LPP is its potential to reflect a negativity bias; whereby there is a stronger LPP among negative stimuli (Cacioppo and Berntson 1994; Ito et al. 1998). This negativity bias may be rooted in our evolution, as negative emotion-relevant stimuli may be highly motivationally salient (Weinberg et al. 2013).

However, as reviewed by Weinberg et al. (2013), LPP is also sensitive to motivationally salient pleasant stimuli when contextually appropriate, is strongly predictive of behavioral slowing to task-irrelevant emotional stimuli, and is stably modulated over (does not habituate to) repeated presentations of emotional stimuli.

H1: Odor-associated words will induce elevated emotions (reflected in LPP) in comparison to non-odor associated words. However, the effect of word valence (pleasant or unpleasant) will vary across the two olfactory groups.

H1a: For individuals with a normal sense of smell, emotions (LPP mean amplitude): non-odor associated words < pleasant odor-associated words < unpleasant odor-associated words

H1b: For individuals with a heightened sense of smell, emotions (LPP mean amplitude): non-odor associated words = pleasant odor-associated words < unpleasant odor-associated words

Brain regions involved in odor processing, such as orbitofrontal cortex, anterior insula and piriform cortex, are activated during mental imaging of odors, as evidenced in positron emission tomography (PET) methods (Djordjevic et al. 2005). Recently, researchers have provided evidence using functional magnetic resonance imaging (fMRI) showing hedonic patterns for differences in mentally imaging pleasant odors compared to unpleasant odors, which matches activity in the brain when exposed to real odorants (Bensafi et al. 2007). The ability to perform olfactory imagery has also been shown to vary across individuals. Stevenson and Case (2005) reported that olfactory experts reported more vivid olfactory images than did non-experts.

In Part 2 of our study, the role of olfactory imagery on emotions is investigated by explicitly instructing individuals to perform olfactory imagery, in contrast to passively reading olfactory related words. For individuals who are sensitive to smell (also known as hyperosmics in medical terms), we expect an automatic suppression mechanism to kick in. During passive processing, high olfactory imagery ability will allow hyperosmics to experience intense odor-related emotions because of strong associations from past experience and memory (Stevenson and Case 2005). For hyperosmics, the olfactory imagery task will in fact trigger an automatic protective mechanism to prevent the elicitation of added overwhelming odor-associated emotions compared to the passive read task. This will be reflected in some suppression of the emotions elicited, shown through reduced levels of LPP, especially for unpleasant odor-associated words.

However, for individuals with a normal sense of smell, we do not expect odor-associated word imagery to significantly increase or decrease the emotions elicited. In marketing, a multisensory study that investigated the effect of visual (pictorial) stimuli on the ease of forming olfactory imagery (Krishna et al. 2010) found that using actual scent can aid visual imagery which led to better verbal recall. However, on the contrary, visual stimuli did not enhance better olfactory imagery. In other words, having a picture in an ad does not assist in scent recall (Krishna 2010). This study did not account for the variability of individuals' sense of smell, and hence we can infer the findings from their sample would be more reflective of individuals with a normal sense of smell (or 70% of the population). For these individuals, the ability to smell is generally taken for granted and is relatively not as meaningful, compared to those who feel its absence or suffer its heightened presence (Cross et al. 2015). Odor-related experiences for normal individuals also should not be as strong or as emotionally charged as those of hyperosmics. Individuals with a normal sense of smell possess good, but less fluent olfactory imagery ability in comparison to individuals with a heightened sense of smell. Thus, we do not expect explicit odor-imagery instructions to further enhance (e.g., ceiling effect) odor-induced emotions, although there may be a slight increase stemming from the unpleasant odor-associated word imagery due to the negativity effect.

H2: The effect of olfactory imagery, elicited by mental imagery triggered by olfactory-related words, on emotions will vary across different olfactory groups.

H2a: For individuals with a normal sense of smell, olfactory imagery will not further enhance emotions (LPP mean amplitude) for odor-associated words in comparison to the passive reading task.

H2b: For individuals with a heightened sense of smell, olfactory imagery will suppress emotions (reflected in lower LPP mean amplitude) in odor-associated words in comparison to the passive reading task.

14.2.2 Methods and Procedures

The emotional processes occurring during the viewing of scent-related words (vs. non-olfactory related words) are explored using a neuroscience tool, electroencephalography (EEG), to understand the brain's responses during olfactory imagery.

A screener survey was distributed across campus to students and staff members in a large university in the Midwest to recruit participants from the two olfactory categories for the purpose of the study. A self-reported screener question, validated by Lin et al. (2017), asked individuals to select a category, out of the four, that described their sense of smell best: heightened sense of smell, normal sense of smell, decreased sensitivity to smell, and impaired with no sense of smell. This resulted in 24 individuals with normal sense of smell and 23 individuals with a heightened sense of smell. The other two smell categories were not further investigated in this current study due to our interest in understanding the impact of scent-related words for sensitive individuals, which make up approximately 20–25% of the population (Aron 1998).

The study was a three valence (non-olfactory words vs. pleasant olfactory words vs. unpleasant olfactory words) within subject × 2 olfactory ability (normal vs. sensitive) between subject mixed design. In the first task, participants were asked to silently read the words presented to them on the screen. They were shown a list of 72 words displayed on a computer screen one word at a time. The list of words is taken from Royet et al. (2003) and supplemented with words from González et al. (2006). The list consists of 12 non-olfactory related words (e.g., needle, button, saucer), 36 words with pleasant olfactory associations (e.g., rose, coffee, honey) and 24 words with unpleasant olfactory associations (e.g., dumpster, feces, trash). The 72 words are presented in 3 blocks of 24 words apiece, consisting of 4 non-olfactory related words, 12 pleasant olfactory words, and 8 unpleasant olfactory words. The procedure consists of 72 trials showing a fixation cross (+) for 1 s, then a word is displayed for 750 ms. Followed by a blank screen for the intertrial interval (ITI) of 3 s. This is repeated for the first block of 24 words. Then there is a short pause of 1 min, followed by the next block. This continues until the three blocks are completed. Block order is randomized across participants. In the second task, participants were instructed to silently read the word *and also* form a mental image of the corresponding smell represented by the word. For example, to form an image related to the smell of garlic for "garlic"). Practice trials were included to ensure participants understood the instructions.

14.2.3 Electrophysiological Recordings

The electroencephalogram (EEG) (bandpass 0.01–500 Hz, digitized at 2048 Hz, gain 1000, 16-bit A/D conversion) was recorded from an array of 33 sintered silver-chloride electrodes based on a modified 10–20 system using an Electrode Arrays cap (Electrode Arrays, El Paso, TX). These electrodes include the midline sites and occipital sites which are of particular interest for the purpose of our study. All electrodes were referenced to an electrode placed on the nose during recording (Sensorium Inc.), and then re-referenced to an average reference for data analysis. Electrode impedance was lower than 20 kΩ for all participants. Vertical eye movements were recorded from two additional electrodes placed below the right and left eyes. The ground electrode was located 10 mm anterior to the medial electrode (Fz). Processing and averaging of the EEG data was done using EMSE 5.3 (Source-Signal Imaging, San Diego). Ocular artifacts associated with blinks and saccades were corrected using the Ocular Artifact Correction filter in EMSE. Trials contaminated by other high amplitude artifacts (i.e., ±100 μV) were eliminated during averaging. ERPs were averaged for trials related to control, minor violation, and major violation scenarios from −200 to 1500 ms around onset of the decision cue. ERP waveforms presented here are plotted in MATLAB. To correct for violations of sphericity, the Greenhouse-Geisser correction was applied.

14.2.4 Results

We took measurements for the Late Positive Potential (LPP) ERP component, using the window of 600–900 ms recorded at the electrode site Pz (Cacioppo and Berntson 1994; Schupp et al. 2003). Under the passive read task, an ANOVA test of the individuals with a normal sense of smell revealed a strong olfactory words valence

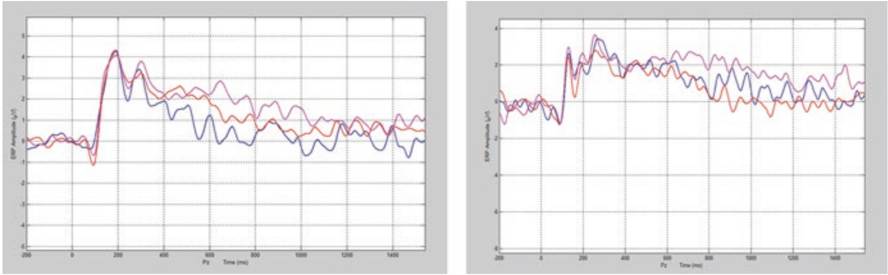
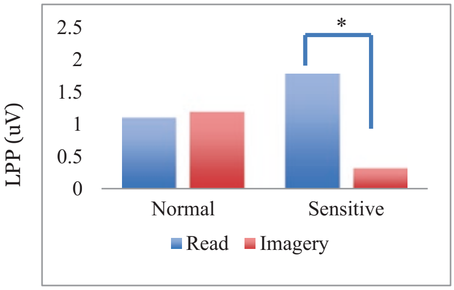


Fig. 14.2 ERP results for LPP at Pz for the reading task for individuals normal (*left*) vs. sensitive (*right*) to smell. Non-olfactory word (*blue*); pleasant olfactory words (*red*); unpleasant olfactory words (*pink*)

Fig. 14.3 Interaction effects of olfactory orientation (Normal vs. Sensitive individuals) and task (Read vs. Imagery of words) on affect (reflected in LPP levels)



effect, F(2, 44) = 13.00, p < 0.001. The LPP is significantly increased for pleasant olfactory words ($M_{pleasant}$ = 1.02 μV vs. $M_{non\text{-}olfactory}$ = 0.25 μV, $p < 0.05$) while significantly increased for unpleasant olfactory words ($M_{unpleasant}$ = 2.17 m μV vs. $M_{non\text{-}olfactory}$ = 0.25 μV, $p < 0.05$) compared to non-olfactory words (Fig. 14.2). These results confirm H1a, where emotions are elicited most under unpleasant odor- association words, followed by pleasant odor-associated words in comparison to non-olfactory words.

Olfactory word valence for the LPP is also significant with hyperosmics (F(2, 42) = 5.65, $p < 0.05$). As predicted, the LPP is not increased for pleasant olfactory words ($M_{pleasant}$ = 0.78 μV vs. $M_{non\text{-}olfactory}$ = 0.38 μV, $p > 0.1$) but is strongly increased for unpleasant olfactory words ($M_{unpleasant}$ = 1.71 μV vs. $M_{non\text{-}olfactory}$ = 0.38 μV, $p <$ 0.05) compared to non-olfactory words. Results are consistent with H1b.

For effects due to the olfactory imagery task, there are additional differences between the groups (Fig. 14.3). Individuals with a normal sense of smell were not affected by the imagery task instruction. Neither pleasant nor unpleasant olfactory word stimuli were differentially affected by the more passive read task versus the more resource-demanding imagery task. In contrast, hyperosmics were affected by the imagery instructions and resulted in suppressed LPP. Imagery did not affect the processing of pleasant words, however, under the imagery task, words related to unpleasant smells *reduced* the LPP magnitude in relation to the response to non-olfactory related words.

For individuals with a normal sense of smell, the effect of imagery is not significant, $M_{read} = 1.11\ \mu V$ vs. $M_{imagery} = 1.20\ \mu V$; $F(1, 22) = 0.012$, $p > 0.1$. This non-effect is consistent for both valence comparisons across tasks (p's > 0.1), which confirms H2a.

In hyperosmics, there is a significant difference in the effect of the task on the LPP amplitude ($M_{read} = 1.78\ \mu V$ vs. $M_{imagery} = 0.33\ \mu V$; $F(1, 21) = 5.1$, $p < 0.05$) confirming H2b. Individuals sensitive to smell appear to be automatically processing the affect information by just engaging in the reading task. When instructed to perform olfactory imagery, affect reflected by LPP is suppressed (Fig. 14.2). Further examination shows no difference under the pleasant olfactory word condition ($M_{read} = 1.2\ \mu V$ vs. $M_{imagery} = 0.48\ \mu V$; $F(1, 23) = 0.69$, $p > 0.1$). For the unpleasant olfactory condition, there is a significant task effect ($M_{read} = 2.72\ \mu V$ vs. $M_{imagery} = 0.92\ \mu V$; $F(1, 23) = 5.0$, $p < 0.05$). Confirming H2b, the olfactory imagery task results in a suppression effect in hyperosmics, reflected in reduced LPP, in the unpleasant condition.

14.2.5 Discussion

Summarizing results for the two valence categories of pleasant versus unpleasant olfactory words during the read task shows there is a clear negativity bias. Unpleasant olfactory words generated significantly higher LPP amplitudes for both olfactory groups. This is consistent with the role of smell in warning against unsafe conditions and substances.

However, emotional reactions during the reading of pleasant olfactory words were not different from non-olfactory words in individuals with a heightened sense of smell. To further understand the relationship of olfactory imagery fluency and olfactory orientation, the following analysis was conducted from the data gathered in the prescreener survey.

Fluency in performing scent related imagery through reporting of the level of vividness of their olfactory imagery varies across individuals. Olfactory imagery ability can be measured through the Vividness of Olfactory Imagery Questionnaire (VOIQ; Gilbert et al. 1998); an imagery scale modeled after the visual imagery scale by Marks (1973). We believe that the ability to perform olfactory imagery, reflected by the VOIQ scale, will be highly correlated with the level of olfactory sensitivity in individuals. In other words, hyperosmics have better olfactory imagery abilities compared to individuals with a normal sense of smell, and hyposmics have the lowest performance in olfactory imagery. To examine this aspect of individual differences, we surveyed undergraduates using the VOIQ scale (N = 518). Results showed that smell category strongly predicts VOIQ scores ($F(2, 514) = 17.62$, $p < 0.001$) while gender was not significant. Hyperosmics reported lower scores (reflecting higher vividness in olfactory imagery) ($M_{hyperosmic} = 33.56$) followed by individuals with a normal sense of smell ($M_{normal} = 39.82$). Those with a diminished sense of smell (hyposmics) reported the highest scores and the least vividness in olfactory imagery ($M_{hyposmic}$ = 46.98). The correlation and direction between VOIQ and the three smell categories confirmed that vividness likely plays a role in the effects of olfactory imagery on emotions.

With these findings in mind, the suppressed affect during imagery in sensitive individuals is even more likely an effect of automatic suppression. On the contrary, automatic affect response to reading the olfactory pleasant words is supported by the fact that individuals who are sensitive to smell implicitly (hence automatically) process affective associations of olfactory words.

14.3 Study 2: Evaluations and Behavioral Intentions to Scented Brand Names

14.3.1 Literature Review and Hypotheses

The congruency theory predicts that processing pieces of congruent information increases fluency of processing and enhances the likeability and evaluation of the product (Bosmans 2006). The advantages of congruent information across multiple sensory inputs have been demonstrated using the visual and auditory senses to increase learning (Kim et al. 2008). The congruency theory further suggests that congruent visual and olfactory information also assisted better recall of the product (Herz 1997). Similar findings were demonstrated when studied across tactile and scent-associated information (Krishna et al. 2010). The congruent (vs. incongruent) role of ambient scent has also been studied in the context of variety seeking behaviors in order to understand the nuances in the consumer decision making process. Holistic processing is enhanced when congruent information, such as the scent and target product, are presented. Time spent in store and browsing is increased (Mitchell et al. 1995). Product evaluation is also influenced by the level of congruency between ambient scent and the target product (Bosmans 2006). In this study, evidenced by what congruency theory predicts, visual information of odor-associated products (i.e., home fragrance, cookies) was presented in ads along with scent (vs. no scent) brand names. Based on previous findings in sensory congruency research, we expect that brand, ad, product evaluations are rated higher during the more congruent relationship of odor-associated product and scent-related brand names, in comparison to the pairing odor-associated product and non- scent-related brand names.

H3: Individuals will rate products with a scent-related brand name (vs. a non-scent-related brand name) higher in attitudes toward the ad (Aad), attitudes toward the brand (Abrand), attitude toward the product (AProduct), and beliefs of product functionality (Beliefs) and purchase intentions (PI).

To reiterate, individual differences in sensitivity to smell do play a role in purchase decisions and consumption behavior. While congruency theory suggests that attitudes toward the product are enhanced under the influence of congruent information presented in multisensory situations, we expect deviations to such expectations are fostered by the influence of sensitivity to smell. In a common setting, individuals with a normal sense of smell are likely to follow the congruency theory predictions and rate odor-associated products with scent-related brand names more favorably.

However, based on earlier ERP findings revealed in Study 1, the differentiated effects between normal (vs. heightened) individuals are expected. In support of the congruency effect, affect (reflected in LPP) is enhanced in individuals with a normal sense of smell while reading olfactory-related words. However, findings in Study 1 also suggest that processing of emotional reactions to olfactory-related words is automatically suppressed in sensitive individuals. Based on these findings, we expect individuals sensitive to smell will suppress affect responses towards olfactory related information. Hence, ratings of Aad, Abrand Aproduct and Beliefs of a product paired with a scent-related-brand name will not be rated higher than a product with a non-scent-related brand name. These affective outcomes are likely to result in lower product evaluations. Fishbein and Ajzen (1975) conventional theory of reasoned action suggest that attitudes are strong predictors of behavioral intentions. We contend that the environment (online) and individual difference (olfactory orientation) will have different consequences for online purchase behaviors. Without the presence of an actual scent, accompanied by the resultant suppression of emotional reactions to scent-related information, individuals with a heightened sense of smell are still able to make sound purchase decisions. These individuals are likely to rate online purchase intentions comparably to individuals with a normal sense of smell.

H4: Differential effects of affective vs. behavioral responses will vary across individual differences in sense of smell. In particular,

H4a: Individuals with a normal sense of smell will rate perception and affect ratings of ad (Aad) and brand (Abrand) higher for scent-related brand names (vs. non-scent-related brand names). Affect toward scent-related brand names (vs. non-scent-related brand names) will not be rated higher in sensitive individuals.

H4b: Individuals with a heightened sense of smell will rate behavioral purchase intentions (PI), target product (Product) and cognitive beliefs of product performance (Belief) higher for scent-related brand names (vs. non-scent-related brand names). Scent-related brand names (vs. non-scent-related brand names) will not increase behavioral purchase intentions of individuals with a normal sense of smell.

14.3.2 Method and Procedures

Two product categories often associated with a scent were selected for creating the online ads, including a home fragrance product and a food item (cookie). Ads were pretested for likeability and product association with scent.

Words used to construct brand names in this study were chosen from a database with normality ratings to ensure gender and mood neutral words. Scent-related words (e.g., lavender, orange blossom, caramel) and non-scent-related words (e.g., bingo, symphonies, compass) were also selected from normality ratings on olfaction association levels. The scent-related brand name version and non-scent-related brand name version of the ads were constructed so the only variation is the brand name between the two conditions (Fig. 14.4).

A total of 256 participants from a mid-size university in the United States were recruited and given class credit for their participation. A screener question to survey the smell orientation of the individuals was asked and used to group participation into four groups based on their sensitivity to smell: no sense of smell, decreased sense of smell, normal sense of smell and increased sensitivity to smell. The two olfactory groups of interest for our study, sensitive and normal, resulted in 66 sensitive individuals and 163 individuals with a normal sense of smell. A total number of 229 participants were included in the analyses.

Participants were randomly assigned into the two ad conditions: scent-related word brand ad and non- scent-related word brand ad. This resulted in 35 sensitive individuals in the scent-related word brand condition and 31 in the non-scent-related word brand. Eighty-four and 79 normal individuals were included in the two conditions respectively. The ads were presented on a computer screen simulating an online shopping scenario and participants were instructed to perform olfactory imagery, "Please take a minute or two to try to form a mental image in your of what

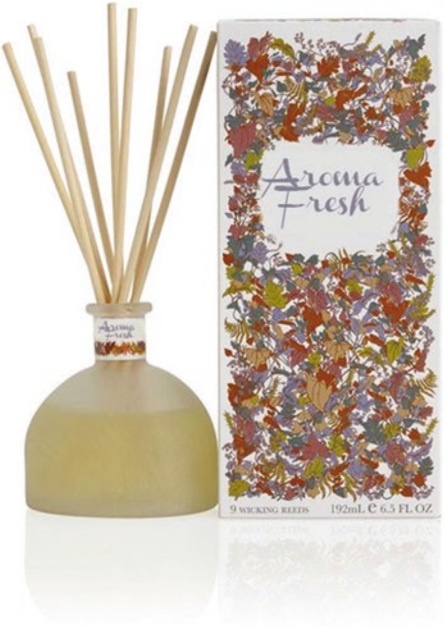


Fig. 14.4 Example of the home fragrance product with the scented brand "Aroma Fresh" used in the online ad study

the product must smell like." Questions related to vividness of imagery (e.g., "please identify the strength of the smell that came to mind when thinking about the product") on a 5-point scale were included in the survey and later used as a manipulation check. Ratings were included as covariate. Participants were then asked to rate their attitudes towards the ad (A_{ad}), brand (A_{brand}), product ($A_{product}$), and purchase intentions (PI). Beliefs related to the functionality/performance of the product was asked for each product (e.g., "judging from the ad and brand, I believe this home fragrance gets rid of odors effectively").

14.3.3 Results

Control variables included gender, imagery vividness level and product involvement. There were marginal significant effects on the outcome variables, hence these were not included in the following analyses.

In a between subject design, approximately half of the participants were randomly given the scent-related ad and the others were shown the non-scent-related ad. MANOVA test reveals significant main effects on the impact of scent-related brand (vs. non-scent-related brand) in home fragrance ads on the A_{brand} ($M_{scent} = 3.31$ vs. $M_{noscent} = 2.98$, $F(1, 225) = 3.63$, $p < 0.5$) and $A_{product}$ ($M_{scent} = 3.38$ vs. $M_{noscent} = 3.15$, $F(1, 225) = 2.93$, $p < 0.05$). There were marginal effects on purchase intentions ($M_{scent} = 3.75$ vs. $M_{noscent} = 3.50$, $F = 1.55$, $p < 0.08$) and on the Belief of how well the product performs ($M_{scent} = 2.73$ vs. $M_{noscent} = 2.42$, $F = 1.53$, $p < 0.08$). (In the case of home fragrance, how effectively did it get rid of odors?) No significant effect of the brand name on A_{ad} ($M_{scent} = 3.09$ vs. $M_{noscent} = 2.96$ $F(1, 225) = 0.41$, $p > 0.1$). This was due to the removal of confounding effects through pretesting the ads and brand names. Findings confirm H3.

Smell orientation (normal vs. sensitive) also had significant main effects on the A_{ad} ($M_{normal} = 3.16$ vs. $M_{sensitive} = 2.9$, $F(1. 225) = 4.03$, $p < 0.05$) and Belief ($M_{normal} = 2.65$ vs. $M_{sensitive} = 2.22$, $F(1, 225) = 5.59$, $p < 0.01$). There was a marginal significant effect of smell orientation on A_{brand} ($M_{normal} = 3.26$ vs. $M_{sensitive} = 3.05$, $F(1, 225) = 2.51$, $p < 0.08$). There were no main effects of smell orientation on $A_{product}$ ($M_{normal} = 3.35$ vs. $M_{sensitive} = 3.19$, $F(1, 225) = 1.88$, $p > 0.1$) and PI ($M_{normal} = 2.52$ vs. $M_{sensitive} = 2.36$, $F(1, 225) = 0.51$, $p > 0.1$). Results in general confirm H4.

Further, there are marginal interaction effects of smell orientation (normal vs. sensitive) × fragrance scent (scent vs. no scent) on A_{brand} ($F(1, 225) = 1.88$, $p > 0.1$), A_{ad} ($F(1, 225) = 1.84$, $p < 0.05$), $A_{product}$ ($F(1, 225) = 1.88$, $p > 0.1$) and Belief ($F(1, 225) = 1.91$, $p < 0.08$). Planned post-hoc tests in individuals with a normal sense of smell demonstrate that a scent-related brand name, in comparison with non-scent-related brand name, has a significant impact on A_{brand} ($M_{scent} = 3.47$ vs. $M_{noscent} = 3.08$, $t(162) = 2.64$, $p < 0.01$), A_{ad} ($M_{scent} = 3.29$ vs. $M_{noscent} = 3.02$, $t(162) = 1.92$, $p < 0.01$), $A_{product}$ ($M_{scent} = 3.49$ vs. $M_{noscent} = 3.24$, $t(162) = 1.85$, $p < 0.05$) and Belief ($M_{scent} = 2.98$ vs. $M_{noscent} = 2.66$, $t(159) = 1.83$, $p < 0.05$). Scent-related brand name

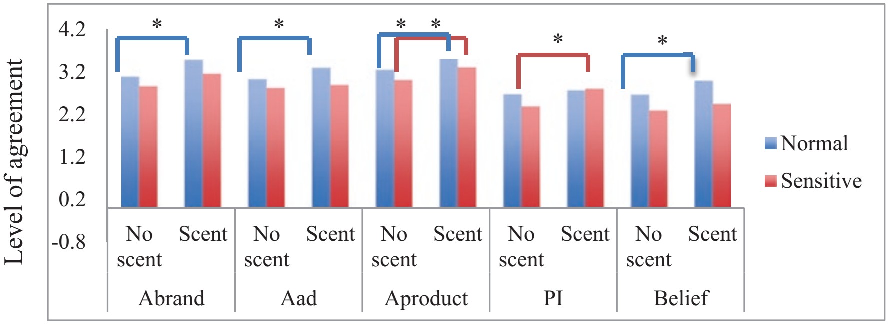


Fig. 14.5 Mean affect related ratings (Aad, Abrand, Aproduct and Belief) and behavioral intention ratings (Purchase intensions) across the two brand conditions (Non-scent-related brand vs. scent-related brand) for the home fragrance ad in the two olfactory orientation groups (normal vs. sensitive). $**p < 0.01$, $*p < 0.05$

did not have an impact on PI ($M_{scent} = 2.76$ vs. $M_{noscent} = 2.67$, $t(162) = 0.48$, $p > 0.1$). Findings confirm H4a (Fig. 14.3).

In contrast, the scent-related brand name had marginal influences on $A_{product}$ ($M_{scent} = 3.3$ vs. $M_{noscent} = 3.0$, $t(64) = 1.43$, $p < 0.05$) and PI ($M_{scent} = 2.80$ vs. $M_{noscent} = 2.38$, $t(64) = 1.74$, $p < 0.05$) for individuals with a heightened sense of smell. Scent-related brand name had no significant impact on affect driven evaluations, A_{brand}, A_{ad}, or Belief. Results support H4b (Fig. 14.5).

In a different product category using a cookie ad, similar effects were found in the evaluation of the ad. Overall, individuals with a normal sense of smell are more likely to be influenced by the scent-related brand resulting in higher attitude ratings. Individuals with a normal sense of smell rated A_{brand} ($M_{normal} = 3.24$ vs. $M_{sensitive} = 2.88$, $F(1, 225) = 4.95$, $p < 0.01$), A_{ad} ($M_{normal} = 3.35$ vs. $M_{sensitive} = 3.08$, $F(1, 225) = 3.14$, $p < 0.05$), PI($M_{normal} = 3.82$ vs. $M_{sensitive} = 3.43$, $F(1, 225) = 3.14$, $p < 0.05$) and Belief “believed the cookie would taste better” ($M_{normal} = 3.70$ vs. $M_{sensitive} = 3.41$, $F(1, 225) = 5.46$, $p < 0.01$) higher than sensitive individuals.

Overall, Scent-related brand names in the cookie ad were not rated significantly higher than non-scent-related brand ads on A_{brand} ($M_{scent} = 3.14$ vs. $M_{noscent} = 2.97$, $F(1, 225) = 1.09$, $p > 0.1$), A_{ad} ($M_{scent} = 3.29$ vs. $M_{noscent} = 3.15$, $F(1, 225) = 0.86$, $p > 0.1$), $A_{product}$ ($M_{scent} = 3.94$ vs. $M_{noscent} = 3.84$, $F(1, 225) = 0.63$, $p > 0.1$) and Belief ($M_{scent} = 3.56$ vs. $M_{noscent} = 3.54$, $F(1, 225) = 0.02$, $p > 0.1$). Purchase intentions ($M_{scent} = 3.75$ vs. $M_{noscent} = 3.50$, $F(1, 225) = 2.42$, $p < 0.08$) were marginally higher in the scent-related brand ad.

However, individuals with a normal sense of smell were significantly influenced by the scent-related brand name (vs. non-scent-related brand name) and rated A_{brand} ($M_{scent} = 3.43$ vs. $M_{noscent} = 3.05$, $t(162) = 2.41$, $p < 0.01$) and A_{ad} ($M_{scent} = 3.51$ vs. $M_{noscent} = 3.13$, $t(162) = 2.41$, $p < 0.01$) higher in the scent-related brand name condition. The effects were not significant on PI and belief.

Individuals with a heightened sense of smell are not influenced by the scent-related brand name; thus do not rate A_{brand} ($M_{scent} = 2.90$ vs. $M_{noscent} = 2.91$, $t(64) = -0.08$, $p > 0.1$), A_{ad} ($M_{scent} = 3.09$ vs. $M_{noscent} = 3.15$, $t(64) = -0.24$, $p > 0.1$), or Belief ($M_{scent} = 3.50$ vs. $M_{noscent} = 3.36$, $t(64) = 0.51$, $p > 0.1$) higher in the scent-related brand name condition (vs. non-scent-related brand name). However, PI are marginally higher in sensitive individuals ($M_{scent} = 3.66$ vs. $M_{noscent} = 3.25$, $t(64) = 1.43$, $p < 0.08$) when the scent-related brand name was presented.

14.3.4 Discussion

As congruency theory predicted, and in support of H3, findings overall show that scent-related brand names are better perceived and rated higher in positive attitudes towards A_{brand}, A_{ad}, Belief and PI in the home fragrances ad. The scent-related brand name did not strongly influence attitudes for the cookie and could be likely a result of the product category. Home fragrances are normally more frequently associated with a scent than cookies, where taste is the determinant attribute.

The main effects for olfactory orientation were significant in both ads for A_{brand}, A_{ad} and Beliefs where individuals with a normal sense of smell rated the products with scent-related brand names more favorably. The image of a product presented in the ad, which is automatically associated with a scent, triggered lower attitudes toward the brand, ad, and product in the sensitive individuals. Attitudes are not elevated by the brand name for these individuals, whereas the scent-related brand name is seen and rated higher in the normal individuals. Findings in Study 2 coincide and support the ERP results, revealing suppressed affective reactions and responses in processing scent-related information.

However, purchase intention was rated higher in sensitive individuals when scent-related brand name was presented, despite lower attitudes toward the brand, ad and belief. On the surface, this seems to contradict the suggestions offered by the theory of reasoned action. However, we argue that the attitudinal reaction, reflected in non-significant effects of scented brand names on ad and brand ratings, was masked by emotional suppression for the purpose of overall behavioral function in individuals sensitive to smell. Such affect regulation, which has been suggested to foster feeling “right” (vs. feeling “good”) based on the demands of the situation (Koole et al. 2008). Further, the online environment was able to mitigate the negative physiological responses that might otherwise yield in a different behavioral outcome.

Findings from this study argue against the perception that attitudes (revealed at the surface level) alone accurately predict beliefs and behaviors. Our results suggest individual difference factors should be considered in the predictive model. Higher ratings of attitudes, as observed in normal individuals, might not translate into purchase behaviors. On the other hand, lower ratings of positive attitudes can still result in increased purchase intentions and beliefs. Underlying explanation for this contrary to conventional belief lies in the automatic suppressing processes of affect discovered in study 1.

14.4 General Conclusion and Discussion

The results in this paper revealed differentiated underlying emotional processes during online purchase decisions. Product decisions that normally use scent as one of the main attributes in driving purchase decisions are constrained by the online environment, in the case of e-commerce or online ads. Our findings suggest individual differences in sensitivity to smell plays a crucial role in purchase decisions related to scent-related products. Further, strongly correlated with olfactory sensitivity is the effectiveness of performing olfactory imagery. These differential effects based on individual difference factors investigated in this paper have ramifications for managerial decisions and strategy planning. In particular, understanding consumer responses to online advertising and sensory related information has implications for organizational branding and online advertising decisions. Customer relationship management and marketing communication efforts should consider (and/or reconsider) these individual difference elements when communicating with their consumers.

Normal individuals (vs. sensitive) in general are attracted to scented products and are less concerned about the “side effects” scented products might have on individuals sensitive to smell. Further, they rate the ad, product and brand significantly higher when a scent-related brand (vs. non-scent-related brand) was used. This was replicated in both home fragrance product ad food items. In the case of product performance, normal individuals believed the effectiveness of the home fragrance was enhanced when a scent-related brand name (vs. no scent) was used. Evidenced from findings in the ERP study (Study 1), affective responses in the pleasant valence olfactory words resulted in an attenuated emotional response. These findings are supported by other studies investigating automatic emotional suppression responses from sensitive individuals as a reaction to reduce unpleasant affect (Lin et al. 2017). Gaining these nuanced understandings of decision making processes involved in consumer’s online shopping experiences and attitudes, can enhance quality decisions made at the organizational level.

Suppression of affect demonstrated in the ERP study (Study 1) is consistent with findings from online brand attitudes in Study 2. Our combined findings suggest an inhibited processing of emotions in individuals sensitive to smell when olfactory words were presented. This reaction can be viewed as a form of protective mechanism for individuals who have strong memory associations with scent from accumulating experiences in the past. By considering individual difference factors, implicit emotional reactions to scent were demonstrated. Future research should consider generalizing these findings in other areas of individual differences, including personality research and individual differences in other sensory perceptions. Additionally, the two studies suggest that the mind (emotional reactions) develops an automatic emotional suppression mechanism for regulating negative associations, so that the body (behavior) can normally perform and make cognitive driven decisions. The balancing mechanism between emotions and cognitions can occur implicitly and automatically. These findings open doors for future research on emotional regulation and other emotional intelligent related streams of research.

Results from our paper provide insight into understanding online purchase decisions and behaviors that are relevant to product purchases that are often associated with scent attributes. The paper also demonstrates the advantages of utilizing mixed methods. Fundamental mechanisms underlying the differential effects demonstrated through behavioral experiments and surveys were explained and supported through empirical data utilizing neuroscience methods. Through the combined use of methods described, the role of valence, automatic processes during passive view/read of words and images presented through online advertisements was investigated. Particularly, ERP findings provide implicit and almost real time data on emotional processes of scent related information, presented in visual format, which provides an explanation for the inconsistency between self-reported attitudes and behaviors observed in the behavioral study.

Other implications for understanding the interplay between the input of multiple sensory affect, attitudes and behavior are warranted. Clearly, online decision making processes and purchase behaviors may diverge from traditional decisions and behaviors taking place in block and mortar stores. Yet, as this paper shows, the role of scent and the influence of individual differences in sensitivity to scent in online purchase forums remain salient.

One of the limitations to the paper is we only included one end of the olfactory sensitivity spectrum in our investigation. The purpose for the study was to understand vulnerable consumers, sensitive individuals, and their cognitive and emotional processing of online information. Future studies should consider investigating individuals who fall on the other end of the spectrum, individuals who have decreased sense of smell. Previous studies have found that hyposmics reported lower levels of quality of life and lacked enjoyment of many daily consumption activities such as dining in restaurants friends (Miwa et al. 2001). Others have also investigated the full spectrum and discovered that sense of smell is often viewed as part of the consumer's identity and provide many implications for marketers and businesses (Cross et al. 2015). Individuals who find themselves deviated from the societal norms and expectations of following consumption rules in the marketplace have often been neglected and marginalized by the society (Lin et al. 2014). Further, the use of self-reported measure to screen and recruit individuals for the specific olfactory categories has its disadvantages. However, in a separate study, the validation of the scale and support effective use of the scale is demonstrated (Lin et al. 2017).

Biographies

Meng-Hsien (Jenny) Lin is an Assistant Professor of Marketing at California State University, Monterey Bay. Her research interests include studying various individual differences factors in the context of sensory marketing (influence of olfactory sensitivity on consumer behavior), advertising (gender differences and information processing in children), and focuses the mediating role of emotions on these relationships. She studies these topics using multi-methods, including behavioral experiments, survey research, neuroscience, and in-depth interviews. Her work also involves pedagogical research in marketing. Some of Dr. Lin's work has been published in *Neuroscience and Journal for Advancement of Marketing Education*. Her research has implications for theory, public policy, and consumer well-being issues. Dr. Lin received her Ph.D. in marketing and an M.B.A. from Iowa State University.

Terry L. Childers is Emeritus Professor of Marketing at Iowa State University. Prior to this, he was the Dean's Chair in Marketing at Iowa State University. Childers conducts neuromarketing research, or the level of consciousness consumers have when making purchasing decisions. He received the Distinguished Service Award for his exceptional dedication to the Journal of Consumer Research in 2013 at the Association for Consumer Research Conference.

Samantha N.N. Cross is an Associate Professor in Marketing in the College of Business at Iowa State University. Her research examines how diverse entities, identities, perspectives, beliefs, ways of sensing, and consuming co-exist in individuals, households, and society. Current research streams examine diverse cultural influences on decision-making, consumption, and innovation within the home; the impact of sensory influences on consumer identity and purchase behavior within the marketplace; and innovations in research methodology. She has received several awards for her research, including the Jane K. Fenyo Best Paper Award for Student Research, the ACR/Sheth Foundation Dissertation Award, and the Best Paper in Track Award at the American Marketing Association (AMA) Winter Conference. She has presented her work in several forums, both nationally and internationally. Her work has been accepted for publication in the *Journal of Marketing, the International Journal of Research in Marketing, Journal of Public Policy and Marketing, Journal of Business Research, Journal of Macromarketing, and Consumption, Markets and Culture*. Dr. Cross received her Ph.D. in marketing from the University of California, Irvine, her M.B.A. in international business from DePaul University, and a B.Sc. in management studies from the University of the West Indies.

William Jones is an Assistant Professor of Marketing at the Beacom School of Business at the University of South Dakota. Billy's work has explored consumers' use of numbers, behavioral pricing, individual differences in consumers' sensory processes, and issues in marketing education. Dr. Jones's work has been or will be published in *Biological Psychology, Journal for Advancement of Marketing Education,* and *Psychology & Marketing* among other journals and presentations at national and international conferences. Dr. Jones received his Ph.D. in marketing from the University of Kentucky, an M.B.A. from Georgia Southern University, and a bachelor's degree in psychology from the University of Scranton.

References

Aron EN (1998) The highly sensitive person: how to thrive when the world overwhelms you. Three Rivers Press, New York

Bensafi M, Porter J, Pouliot S, Mainland J, Johnson B, Zelano C, Young N et al (2003) Olfactomotor activity during imagery mimics that during perception. Nat Neurosci 6(11):1142–1144

Bensafi M, Sobel N, Khan RM (2007) Hedonic-specific activity in Piriform cortex during odor imagery mimics that during odor perception. J Neurophysiol 98(6):3254–3262

Bone PF, Ellen PS (1999) Scents in the marketplace: explaining a fraction of olfaction. J Retail 75(2):243–262

Bosmans A (2006) Scents and sensibility: when do (in) congruent ambient scents influence product evaluations? J Mark 70(3):32–43

Cacioppo JT, Berntson GG (1994) Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. Psychol Bull 115(3):401

Chebat J-C, Michon R (2003) Impact of ambient odors on mall shoppers' emotions, Cognition, and spending: a test of competitive causal theories. J Bus Res 56(7):529–539

Chebat J-C, Morrin R, Chebat D-R (2009) Does age attenuate the impact of pleasant ambient scent on consumer response? Environ Behav 41(2):258–267

Costafreda SG, Brammer MJ, David AS, Fu CH (2008) Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and Fmri studies. Brain Res Rev 58(1):57–70

Cross SNN, Lin M-H, Childers TL (2015) Sensory identity: the impact of olfaction on consumption. In: Belk R, Murray J, Thyroff A (eds) Research in consumer behavior series. Emerald, Bradford

Cunningham WA, Espinet SD, DeYoung CD, Zelazo PD (2005) Attitudes to the right-and left: frontal ERP asymmetries associated with stimulus valence and processing goals. NeuroImage 28(4):827–834

Djordjevic J, Zatorre RJ, Petrides M, Boyle JA, Jones-Gotman M (2005) Functional neuroimaging of odor imagery. NeuroImage 24(3):791–801

Doty RL, Shaman P, Dann M (1984) Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiol Behav 32(3):489–502

Fishbein M, Ajzen I (1975) Belief, attitude, intention and behavior: an introduction to theory and research. Addison-Wesley, Reading, MA

Gilbert AN, Crouch M, Kemp SE (1998) Olfactory and visual mental imagery. J Ment Imag 22:137–146

Goldkuhl L, Styvén M (2007) Sensing the scent of service success. Eur J Mark 41(11/12):1297–1305

González J, Barros-Loscertales A, Pulvermüller F, Meseguer V, Sanjuán A, Belloch V, Ávila C (2006) Reading cinnamon activates olfactory brain regions. NeuroImage 32(2):906–912

Hajcak G, Nieuwenhuis S (2006) Reappraisal modulates the electrocortical response to unpleasant pictures. Cogn Affect Behav Neurosci 6(4):291–297

Herz RS (1997) Emotion experienced during encoding enhances odor retrieval cue effectiveness. Am J Psychol 110:489–506

Herz R (2007) The scent of desire. William Morrow, New York

Herz RS, Schankler C, Beland S (2004) Olfaction, emotion and associative learning: effects on motivated behavior. Motiv Emot 28(4):363–383

Holland RW, Hendriks M, Aarts H (2005) Smells like clean spirit nonconscious effects of scent on cognition and behavior. Psychol Sci 16(9):689–693

Ito TA, Larsen JT, Smith NK, Cacioppo JT (1998) Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. J Pers Soc Psychol 75(4):887

Kim RS, Seitz AR, Shams L (2008) Benefits of stimulus congruency for multisensory facilitation of visual learning. PLoS One 3(1):1532

Koole SL, Kuhl J, Shah J, Gardner W (2008) Dealing with unwanted feelings. In: Handbook of motivation science. Guilford Press, New York, pp 295–307

Krishna A (2010) An integrative review of sensory marketing: engaging the senses to affect perception judgment and behavior. J Consum Psychol 22(3):332–351

Krishna A, Elder R, Caldara C (2010) Feminine to smell but masculine to touch? Multisensory congruence and its effect on the aesthetic experience. J Consum Psychol 20:410–418

Lang PJ, Bradley MM, Cuthbert BN (1998) Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. Biol Psychiatry 44(12):1248–1263

Lin MH, Cross SNN, Childers TL (2014) Two ends of the olfactory sensitivity continuum: too much and too little. In: Proceedings for the American marketing association winter conference

Lin MH, Cross SNN, Childers TL (2017) Sensitive to the servicescape: the impact of individual differences in sense of smell in response to ambient scent. Working paper

Marks DF (1973) Visual imagery differences in the recall of Pictures. Br J Psychol 64(1):17–24

Mitchell DJ, Kahn BE, Knasko SC (1995) There's something in the air: effects of congruent or incongruent ambient odor on consumer decision making. J Consum Res:229–238

Miwa T, Furukawa M, Tsukatani T, Costanzo RM, DiNardo LJ, Reiter ER (2001) Impact of olfactory impairment on quality of life and disability. Arch Otolaryngol Head Neck Surg 127(5):497–503

Morrin M, Ratneshwar S (2003) Does it make sense to use scents to enhance brand memory? J Mark Res 40(1):10–25

Royet J-P, Plailly J, Delon-Martin C, Kareken DA, Segebarth C (2003) fMRI of emotional responses to odors: influence of hedonic valence and judgment, handedness, and gender. NeuroImage 20(2):713–728

Schupp HT, Markus J, Weike AI, Hamm AO (2003) Emotional facilitation of sensory processing in the visual cortex. Psychol Sci 14(1):7–13

Schupp HT, Flaisch T, Stockburger J, Junghöfer J (2006) Emotion and attention: event-related brain potential studies. Prog Brain Res 156:31–51

Ship JA, Weiffenbach JM (1993) Age, gender, medical treatment, and medication effects on smell identification. J Gerontol 48(1):26–32

Statista (2015). http://www.statista.com/statistics/272391/us-retail-e-commerce-sales-forecast/

Stevenson RJ, Case TI (2005) Olfactory imagery: a review. Psychon Bull Rev 12(2):244–264

Weinberg A, Ferri J, Hajcak G (2013) Interactions between attention and emotion. In: Handbook of cognition and emotion. Guilford Press, New York, pp 35–54

Chapter 15
Say It Right: IS Prototype to Enable Evidence-Based Communication Using Big Data

Simon Alfano, Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann

Abstract Financial investors face an increasing information abundance when making their valuation decisions of financial assets. As Information Systems research demonstrated, valuation not only builds on quantitative facts, but also on qualitative information such as the language used in financial disclosures and the readability of the texts. As an originator of financial disclosures, e.g. a company, it is thus essential to thoughtfully steer the creation of new textual information. While regulators provide guidelines on what content to publish, corporate communication departments can flexibly steer how they communicate. We have developed an IS prototype that accounts for the importance of textual information and provides corporate communications with a decision-support tool to assure a high readability and a positive sentiment. Our IS prototype builds on a two-step process. First, we extract a dictionary with the most relevant words for investors from a large inventory of regulatory filings with Bayesian learning algorithms. Second, we use this dictionary as input for a Microsoft Word add-in that highlights positively or negatively connoted words and suggests alternative words with a more positive investor perception to corporate communications professionals.

Keywords Bayesian learning • Capital markets communication • Decision-support system • Evidence-based communication • IS prototype • Readability • Sentiment

S. Alfano (✉) • S. Feuerriegel • D. Neumann
Department of Information Systems Research, Faculty for Economics, Albert-Ludwigs-University Freiburg, Platz der Alten Synagoge, Freiburg 79098, Germany
e-mail: simon.alfano@is.uni-freiburg.de

N. Pröllochs
Department of Information Systems Research, Faculty for Economics, Albert-Ludwigs-University Freiburg, Platz der Alten Synagoge, Freiburg 79098, Germany

TonalityTech, Freiburg, Germany

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information Systems 21, DOI 10.1007/978-3-319-58097-5_15

15.1 Introduction

Stakeholders in electronic markets nowadays face an overwhelming volume of available data, from which the notion of the term *"data"* is frequently restricted to a bundle of quantitative facts. However, Information Systems (IS) research has proven that investors also account for the language used in communication, commonly referred to as *sentiment* (e.g. Antweiler and Frank 2004; Schumaker and Chen 2009; Tetlock 2007; Tetlock et al. 2008). Analyzing sentiment is positioned at the heart of IS research, although understanding the wording that drives positive or negative sentiment seems to be addressed only rarely (Pröllochs et al. 2015). By filling this gap, corporations can eliminate subjective experience and, in contrast, advance their communication *based on evidence* (Cornelissen 2014). Knowledge regarding evidence-based communication can thus be of immense value to corporates in order to facilitate or provide a decision-support framework for altering e.g. investor communication, brand image and employer perception.

15.2 IS Prototype Architecture

We propose an IS prototype for corporate communications practitioners that enables evidence-based communication. Our IS prototype allows for the improvement of written texts based on two key metrics of corporate communication: readability and sentiment (e.g. Rennekamp 2012; Tan et al. 2014). We interviewed more than 15 corporate communications practitioners (from mid-to large cap companies and PR agencies) during the process of requirement engineering in order to ensure high relevance for practitioners. To assure that the identified requirements meet the needs of practitioners, we have held regular meetings with a focus group among our interviewees where we presented new features and collected direct user feedback for further iterations. The architecture of our research prototype comprises of two building blocks (see Fig. 15.1) as follows:

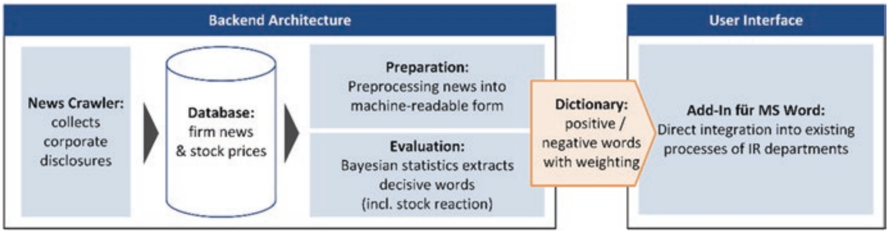

Backend Architecture
News Crawler: collects corporate disclosures
Database: firm news & stock prices
Preparation: Preprocessing news into machine-readable form
Evaluation: Bayesian statistics extracts decisive words (incl. stock reaction)
Dictionary: positive / negative words with weighting
User Interface
Add-In für MS Word: Direct integration into existing processes of IR departments

Fig. 15.1 End-to-end process of the IS prototype for evidence-based communication

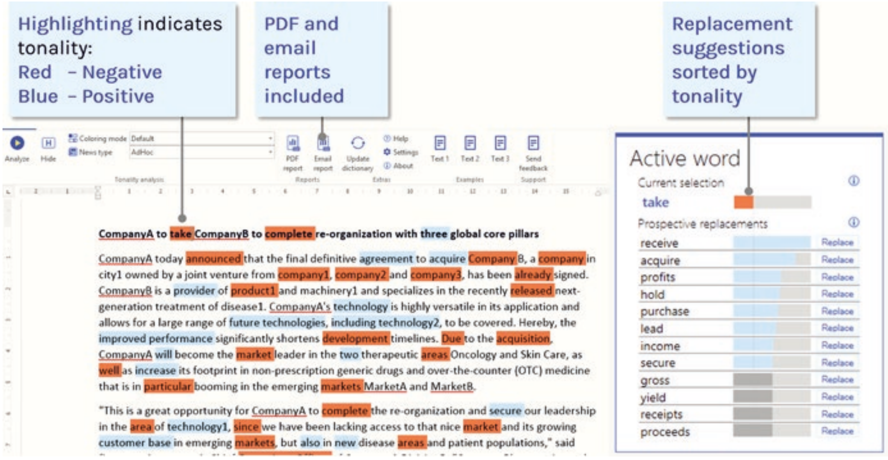
15.2.1 Building Block 1: Backend Architecture with Big Data Analytics

In our backend architecture, we collect and store a large set of more than 14,000 corporate disclosures from stock-listed companies. For each company, we match the firm news with the corresponding stock price movements and abnormal following the release of a new financial disclosure. We then preprocess the financial disclosures into a machine-readable format (Manning and Schütze 1999). Subsequently, we utilize Bayesian learning as a method from Big Data analytics (Hastie et al. 2013; Zou and Hastie 2005). Thereby, we extract decisive words that influence investors as measured by the stock market reaction (Pröllochs et al. 2015). Finally, we generate a dictionary containing all the decisive words extracted and assign each word a positive or negative sentiment score.

15.2.2 Building Block 2: User Interface

As requested by the corporate communications practitioners during the interviews during the requirements engineering phase, the IS prototype must seamlessly integrate into the writing process when using Microsoft Word.

The user interface (see Fig. 15.2) offers an add-in for Microsoft Word with several functions. In the ribbon bar, users can choose a suitable dictionary for their domain and for the purpose of the underlying communication (e.g. regulatory filing in the United States). Users can then analyze their corporate communication throughout the writing process in order to objectively assess both the sentiment and



CompanyA to take CompanyB to complete re-organization with three global core pillars

CompanyA today announced that the final definitive agreement to acquire Company B, a company in city1 owned by a joint venture from company1, company2 and company3, has been already signed. CompanyB is a provider of product1 and machinery1 and specializes in the recently released next-generation treatment of disease1. CompanyA's technology is highly versatile in its application and allows for a large range of future technologies, including technology2, to be covered. Hereby, the improved performance significantly shortens development timelines. Due to the acquisition, CompanyA will become the market leader in the two therapeutic areas Oncology and Skin Care, as well as increase its footprint in non-prescription generic drugs and over-the-counter (OTC) medicine that is in particular booming in the emerging markets MarketA and MarketB.

"This is a great opportunity for CompanyA to complete the re-organization and secure our leadership in the area of technology1, since we have been lacking access to that nice market and its growing customer base in emerging markets, but also in new disease areas and patient populations," said

Fig. 15.2 User interface for evidence-based communication as an add-in for Microsoft word

the readability. The analysis further highlights words in the text that investors perceive negatively (red) and positively (blue). Our IS prototype supports users by proposing alternative words as a replacement. In addition, a dashboard shows an aggregated review that reports the overall sentiment and readability. The dashboard also displays the sentiment and readability of each sentence to simplify the identification of areas for improvement.

15.3 Conclusion

In summary, our IS prototype puts forward three beneficial managerial implications for corporate communications practitioners. First, practitioners can steer the readability to facilitate easier-to-read financial disclosures and to reduce ambiguous interpretations by investors. Second, our IS prototype simplifies coordination in the drafting process of disclosures since evidence can complement subjectivity. Third, actively shaping sentiment and readability can improve corporate value due to the positive relationship between sentiment and stock returns.

Going forward, we will provide the interviewees of our requirements engineering phase with ‘demo’ versions such that we can collect ‘live’ feedback from users of the tool and utilize their feedback to reiterate the tool for more advanced application cases.

15.4 Biographies

Simon Alfano is a Ph.D. student in the Finance Research Group at the Chair of Information Systems Research at Freiburg University’s Department of Economics. In his research at the intersection of behavioral economics and data science, Simon studies how investors process the textual information. As a member of the Finance Research Group, Simon is also supporting the start-up TonalityTech to provide corporate communications access to the knowledge on how investors process qualitative aspects of financial disclosures. Prior to his Ph.D. studies, Simon worked as a management consultant.

https://www.is.uni-freiburg.de/mitarbeiter-en/team/simon-alfano?set_language=en

Nicolas Pröllochs is a Ph.D. student at the Chair of Information Systems of the University of Freiburg with a focus on text mining and sentiment analysis of financial news. He holds a Master of Science in Economics and Information Systems from the University of Freiburg. He has co-authored research publications for the Hawaii International Conference on System Sciences, the European Conference on Information Systems and the Conference on Information Systems and Technology.

https://www.is.uni-freiburg.de/mitarbeiter-en/team/nicolas-proellochs

Stefan Feuerriegel is a post-doctoral researcher at the Chair of Information Systems Research of the University of Freiburg with a focus on text mining and sentiment analysis of financial news. Previously, he obtained his Ph.D. degree from the same research institution. He also holds a Master of Science in Simulation Sciences from the RWTH Aachen University. Among others, he has co-authored research publications in the European Journal of Operational Research, Optimization Engineering and the Journal of Decision Systems.

https://www.is.uni-freiburg.de/mitarbeiter-en/team/stefan-feuerriegel

Dirk Neumann is Full Professor with the Chair of Information Systems of the University of Freiburg, Germany. His research topics include Business Analytics, Text Mining and Cloud Computing. He studied information systems in Giessen (Diploma), Economics in Milwaukee, WI, USA (Master) and received a Ph.D. from Karlsruhe Institute of Technology (KIT) in 2004. He has (co-)authored many research publications at European Journal of Operational Research, ACM Transactions on Internet Technology, Journal of Management of Management Information Systems or Decision Support Systems.

https://www.is.uni-freiburg.de/mitarbeiter-en/team/dirk-neumann

Acknowledgements We want to acknowledge the support of the start-up TonalityTech, our research collaboration partner on this IS prototype development project, for the provision of the above screenshot (Fig. 15.2).

References

Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of internet stock message boards. J Financ 59(3):1259–1294

Cornelissen J (2014) Corporate communication: a guide to theory & practice. Sage, London

Hastie TJ, Tibshirani RJ, Friedman JH (2013) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge

Pröllochs N, Feuerriegel S, Neumann D (2015) Generating domain-specific dictionaries using bayesian learning. In: 23rd European conference on information systems (ECIS 2015), Münster, Germany

Rennekamp K (2012) Processing fluency and investors' reactions to disclosure readability. J Account Res 50(5):1319–1354

Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news. ACM Trans Inf Syst 27(2):1–19

Tan HT, Ying Wang E, Zhuo BO (2014) When the use of positive language backfires: the joint effect of tone, readability, and investor sophistication on earnings judgments. J Account Res 52(1):273–302

Tetlock PC (2007) Giving content to investor sentiment: the role of media in the stock market. J Financ 62(3):1139–1168

Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: quantifying language to measure firms' fundamentals. J Financ 63(3):1437–1467

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Statist Soc B 67(2):301–320

Chapter 16
Introduction: Pedagogy in Analytics and Data Science

Nicholas Evangelopoulos, Joseph W. Clark, and Sule Balkan

Abstract Keeping with the "Exploring the Information Frontier" theme of the ICIS 2015 conference, the Pre-ICIS Business Analytics Congress workshop sought forward-thinking research in the areas of data science, business intelligence, analytics and decision support with a special focus on the state of business analytics from the perspectives of organizations, faculty, and students. The teaching track aimed to promote comprehensive research or research-in-progress in teaching and learning addressing topics including business analytics curriculum development, pedagogical innovation, organizational case studies, tutorial exercises, and the use of analytics software in the classroom. This work has been summarized in this chapter.

Keywords Data science • Analytics • Business intelligence • Decision support systems • Curriculum design • Pedagogy

16.1 Introduction

Emerging technologies in business intelligence and social media are fueling a need for innovative curricula in online, traditional and hybrid delivery formats that meet the industry needs. In keeping with the BA Congress theme of "Exploring the Analytics Frontier", we sought pedagogical research contributions, teaching materials, and pedagogical practices/cases that address acquisition, application, and continued development of the knowledge and skills required in the usage of business analytics in the classroom, with emphasis on business intelligence, social media

N. Evangelopoulos (✉)
University of North Texas, 365D Business Leadership Building,
1307 West Highland Street, Denton, TX 76201, USA
e-mail: nick.evangelopoulos@unt.edu

J.W. Clark
University of Maine, DP Corbett Business Building, Rm. 315, Orono, ME 04469, USA
e-mail: joe.clark@maine.edu

S. Balkan
Portland State University, Fourth Avenue Bldg, 1900,
1900 SW Harrison St., Portland, OR 97201, USA
e-mail: balkansule@gmail.com

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information Systems 21, DOI 10.1007/978-3-319-58097-5_16

analytics, big data analytics, high performance analytics, data science, visualization, and other emerging analytic technologies.

With the explosion of data, the demand for business intelligence and analytics is increasing at a faster rate now than ever before. The measurable value from data is created only after its interpretation and implementation in business processes. It is becoming a mainstream for large and small businesses alike to use data driven decision making. This created a high demand for data scientists and business analysts. According to a 2015 MIT Sloan Management Review study, 40% of the companies surveyed were struggling to find and retain the data analytics talent (Ransbotham et al. 2015). International Data Corporation (IDC) predicts a need by 2018 for 181,000 people with deep analytical skills, and a requirement five times that number for jobs with the need for data management and interpretation skills (Deloitte 2016).

In an effort to close the big talent gap, top business schools are adjusting their curricula to incorporate state of the art tools and techniques in the fields of business intelligence and analytics with the objective of training students to meet demand. The BA Congress Teaching track brought together a community of scholars who have developed cutting edge curriculum in the areas of pedagogical innovation, organizational case studies, tutorial exercises and software. BA Congress received several contributions, ranging from survey of fields and data types to analytics software tutorials and case studies. In this chapter a small sample of these are presented.

16.2 The Papers in the Teaching Track

In this section, we briefly introduce four papers from the teaching track of the BAC 2015 workshop.

In the first paper, titled *Tools for Academic Business Intelligence & Analytics Teaching—Results of an Evaluation*, Kollwitz et al. (2017) survey the field of tools for business intelligence and analytics, systematically evaluating them for classroom use. Dividing the field into five subdomains that correspond to popular skill profiles—Big Data, text, web, network, and mobile analytics—they identify and compare state-of-the-art tools in each major area. Taking into account the practical importance of free licenses for academic use, the availability of documentation and training materials for each piece of software, and compatibility with Windows, Mac OS, and Linux platforms, their disciplined research should provide excellent recommendations to instructors building a curriculum for analytics teaching.

Business analytics tools come in a wide range of algorithmic complexity, platform availability, ease of use, and application domains. As educators prepare to introduce their students to these tools, brief tutorials can come handy. The second paper, titled *Neural Net Tutorial* (Huguenard and Ballou 2017), provides such a tutorial for artificial neural networks. Its step-by-step directions guide students to build a working neural net and train it with a sample data set to make predictions about horse racing outcomes. Prefaced with a brief but accessible introduction to the concept of machine learning generally and neural networks specifically, the tutorial could stand on its own in an undergraduate or graduate analytics course.

The third paper, titled *An Examination of ERP Learning Outcomes: A Text Mining Approach* (Dunaway 2017), studies the effectiveness of student role-playing in the ERP simulation ERPSIM, which has become increasingly popular in teaching ERP in the classroom. Informed by situated learning theory, Dunaway hones in on the question of how well the learning gained from ERPSIM translates into real-world knowledge transfer. Using text mining techniques, she analyzes students' written reflections on the roles they played in the simulation. Her findings indicate that role play amplifies the learning gleaned from the simulation activity.

Last, but not least, the fourth paper, titled *Data Science for All: A University-Wide Course in Data Literacy* (Schuff 2017), addresses the need for analytical thinking and information literacy as part of the foundational education of all students. The author provides a blueprint for an undergraduate course targeted not at MIS or analytics students but at all majors, which has progressed from pilot testing to a multi-section course now offered as part of the "general education" curriculum at Temple University. Schuff's paper is a rich resource that other instructors may mine for ideas to use in their own courses targeted at non-majors, including four modules mapped to ten teaching objectives and a few examples of exercises and assignments. College administrators should take note, too, of this proof that universities should, and can, empower students of all majors with a certain amount of data savvy.

Biographies

Nicholas Evangelopoulos is a Professor in the Department of Information Technology and Decision Sciences at the University of North Texas. He received his Ph.D. in decision sciences from Washington State University and his M.S. in computer science from the University of Kansas. His current research interests include text Analytics, change-point analysis, probabilistic models, and applied statistics. His articles appear in MIS Quarterly, Decision Sciences, Decision Support Systems, Communications of the ACM, the European Journal of Information Systems, Information Systems Journal, Communications in Statistics, Computational Statistics & Data Analysis, and many others. He has taught a number of courses in business statistics, data mining, data warehousing, big data analytics, and programming. His consulting experience includes corporate projects in text analytics, predictive modeling, customer attrition modeling, and policy evaluation. In 2010 he received the UNT College of Business Outstanding Teaching Innovation Award, in 2013 the UNT College of Business Outstanding Senior Faculty Research Award, and in 2016 he was named the 2016–2017 Professional Development Institute Fellow for the UNT College of Business.

Joseph W. Clark was born and raised in Maine, then went away for higher education, earning his B.A. and Ph.D. from USC and his M.B.A. from Tulane. He was one of the first generation of web developers during the dot-com boom of 1997–2001. As an academic, he has held appointments at China Agricultural University, the

University of Nebraska Omaha, and Arizona State University. In 2016, he finally came to his senses and moved home, joining the University of Maine as a Lecturer in Management Information Systems. Dr. Clark's research touches on business intelligence and analytics, decision making, entrepreneurship and design thinking. His research and teaching cases have been presented at international conferences such as HICSS and ICIS, and he has authored an e-book on data engineering. In addition to university teaching, Dr. Clark offers consulting services in software development, databases, and big data analytics. He is looking forward to raising his four children in the north woods of Maine where he grew up.

Sule Balkan got her Ph.D. in Economics from the University of Arizona Eller College of Management in 1998. She recently moved to Portland Oregon after living in Taiwan for 4 years where she was the Director of Big Data Certificate Program and an Associate Professor at the Institute of Business and Management of National Chiao Tung University. Her research and teaching interests include business intelligence topics such as predictive modeling, advanced analytics, information driven campaign management and Big Data. Before moving to Taiwan, Sule worked as a Clinical Associate Professor of Information Systems in the W.P. Carey School of Business, Arizona State University for 4 years. She has more than 10 years of professional experience in information management, predictive modeling and analytics, and campaign execution fields. She worked as Director of Information Management at Ameriprise Financial prior to joining to academia. Currently she is working at Portland State University Department of Engineering and Technology Management.

References

Deloitte (2016) Analytics trends: the next evolution. Downloaded on October 31, 2016, from http://www.deloitte.com/us/AnalyticsTrends

Dunaway MM (2017) An examination of ERP learning outcomes: a text mining approach. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series (http://www.springer.com/series/7573), Vol. 21, 2016–2017

Huguenard BR, Ballou DJ (2017) Neural net tutorial. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series (http://www.springer.com/series/7573), Vol. 21, 2016–2017

Kollwitz C, Dinter B, Krawatzeck R (2017) Tools for academic business intelligence & analytics teaching—results of an evaluation. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series (http://www.springer.com/series/7573), Vol. 21, 2016–2017

Ransbotham S, Kiron D, Prentice P (2015) Minding the analytics gap. MIT Sloan Manag Rev 56(3):63–68

Schuff D (2017) Data science for all: a university-wide course in data literacy. In: Deokar A, Gupta A, Iyer L, Jones MC (eds) Analytics and data science: advances in research and pedagogy. Springer annals of information systems series (http://www.springer.com/series/7573), Vol. 21, 2016–2017

Chapter 17
Tools for Academic Business Intelligence and Analytics Teaching: Results of an Evaluation

Christoph Kollwitz, Barbara Dinter, and Robert Krawatzeck

Abstract The trend towards big data and business intelligence & analytics (BI&A) is still continuing. In the upcoming years, thousands of new jobs for data scientists will be established by the economy. Therefore, there is a need for well-educated graduates with deep analytical skills. In order to prepare students for their later profession and to teach them in analytics tools relevant for practice, related academic education is required. The BI&A sub-domains and tool categories (like text mining and web analytics) correspond to popular skill profiles. Since the market for BI&A tools is very large and hence hard to survey, this paper identifies and evaluates a number of tools for each BI&A sub-domain. The tools are evaluated with regard to university-specific requirements (such as expenses and available learning resources) and BI&A category-specific requirements (such as functionality). Based on the evaluation results recommendations for each tool category are given.

Keywords Business intelligence & analytics • BI&A • Big data • Higher education • Teaching • Tool evaluation

17.1 Introduction

According to Accenture (2014), 89 percent of the organizations expect big data to revolutionize the business in the same way as the internet had done it before. The ubiquitous trend towards big data has not only had an impact on organizations, it has also led to new challenges in academic education, such as the revision of existing or the design of new curricula. Considering the recent IT advances and the massive growth of data, there is already a huge demand for data scientists and managers with

C. Kollwitz • B. Dinter • R. Krawatzeck (✉)
Chemnitz University of Technology, Faculty of Economics and Business Administration, Professorship Business Information Systems I – Business Process and Information Management, 09107 Chemnitz, Germany
e-mail: christoph.kollwitz@wirtschaft.tu-chemnitz.de;
barbara.dinter@wirtschaft.tu-chemnitz.de; robert.krawatzeck@wirtschaft.tu-chemnitz.de

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_17

a deep analytical understanding and it might even increase over the next years (Jacobi et al. 2014; Manyika et al. 2011). Furthermore, the interdisciplinary and rapid development of big data technologies causes additional new challenges for education in this field (Jacobi et al. 2014). Al-Sakran (2015) has already indicated a lack of specific knowledge about business intelligence & analytics (BI&A) tools with respect to potential data scientists.

A very first approach to overcome this issue from an academic side is the design of more practical and tool-oriented curricula. However, universities face the problem that a wide range of potential vendors and tools (Al-Sakran 2015) could be considered within such curricula. Consequently, faculty should be supported when evaluating and selecting BI&A tools for teaching purposes. Therefore, the main research question of this contribution is: Which BI&A tools are most suitable for the practical use in academic education?

Many overviews and evaluations of BI&A tools have been published in the past (e.g., Combe et al. 2010; Davis and Woratschek 2015; for a full list c.f. Wang 2015). However, these contributions usually take a broad and rather generic look on BI&A and/or do not take specific aspects for academic education (such as the expenses) into account. Since the broad field of BI&A can be further differentiated into various subdomains (e.g., data analytics, streaming analytics, text analytics, etc.), as for example proposed by (Chen et al. 2012), there is a need to train students in various tool categories in order to prepare them adequately for practice. Various other requirements—besides the academic-specific ones—related to the characteristics of each BI&A sub-domain need to be identified and considered. Therefore, the paper at hand gives a short overview of tools which are suitable for different BI&A sub-domains and for academic education. To provide further guidance for faculty and academic staff, we additionally propose one specific tool for each BI&A sub-domain. These suggestions are the result of the sub-domain-specific tool evaluations.

The remainder of the paper is organized as follows. The following section "Theoretical Foundations" further motivates the need of practical, tool-oriented lessons within academic education and introduces the BI&A framework by (Chen et al. 2012) which is used to define the various BI&A sub-domains. Following, we describe the research methodology which will be used for the tool evaluations and deduce the university-specific requirements. Subsequently, we present for each identified BI&A sub-domain separately the specific tool requirements, an overview of possible tools, the evaluation report, and—based on the evaluation results—a tool recommendation. Concluding, we summarize our findings, discuss the limitations and provide an outlook to future research.

17.2 Theoretical Foundations

In this section we show the importance of practical lessons in academic education which make use of tools. In addition, we present the BI&A research framework by (Chen et al. 2012) which our work is based on.

17.2.1 The Value of Hands-on Lessons

The rise of big data results in an increasing demand for professionals who have in-depth analytical skills and knowledge (Manyika et al. 2011). However, the shortage of skilled data scientists has become a serious challenge for industry (Davenport and Patil 2012). Current curricula in the field of BI&A do not address all skills relevant in practice (Wang 2015). Al-Sakran (2015) showed that especially graduates lack of deep knowledge in technical tools. To overcome this issue, we suggest to include more hands-on exercises in academic teaching by using tools which are suitable for both, educational use and the potential later usage in industry (i.e., covering common tool features required in real-world settings). In general, hands-on exercises are an integral part of BI&A curricula (Wixom et al. 2014). Within literature, the relevance of hands-on lessons for education is highlighted by various authors. Teaching information systems students the necessary skills for using BI&A tools is regarded as one of the key results of academic education (Al-Sakran 2015; Wixom et al. 2014). Therefore, universities should cooperate with software vendors and provide "state-of-the-art" software in appropriate classes (Wixom et al. 2011). The data sets for practical hands-on analytical experience should come from industry (Schiller et al. 2015) or can be obtained from public sources (for a list of possible sources for data sets cf. Schiller et al. 2015 p. 820) However, since the BI&A tool market is very large and hard to survey, the final selection of specific "state-of-the-art" tools for teaching out of all possible solutions is challenging.

17.2.2 The BI&A Framework

The subsequent evaluations of software solutions are based on the BI&A research framework as introduced by Chen et al. (2012). Chen et al. (2012) differentiate between five technical sub-domains in analytics research and have identified several foundational technologies and emerging research trends for each sub-domain. Foundational technologies encompass a variety of underlying mature technologies. Emerging research describes trends and developments and shows an overview of the current focus in research. We use this framework as it constitutes a complete list of shapes within BI&A from which we deduce different tool categories for our evaluations. Table 17.1 shows an excerpt of the framework (for a complete version see Chen et al. 2012).

Since we will evaluate different analytics tools for each sub-domain in the remainder of the paper, in the following a brief overview for the sub-domains is presented.

First, Chen et al. (2012) identified *big data analytics* as a sub-domain. A commonly used definition for big data, proposed by Gartner research institute (Laney 2001), includes the "three Vs" (Volume, Velocity, and Variety). Within this sub-domain, there is a variety of research areas dealing with different kinds of (structured)

Table 17.1 BI&A framework (based on Chen et al. 2012)

	Foundational technologies	Emerging research
(Big) Data analytics	• Relational database management systems • Data warehousing • Extract, transform, and load processes • Online analytical processing • Business process management	• Statistical machine learning • Sequential/temporal mining • Spatial mining • Data mining for high-speed data streams and sensor data
Text analytics	• Information retrieval • Computational linguistics • Search engines	• Statistical natural language processing • Information extraction • Opinion mining • Sentiment analysis
Web analytics	• Information retrieval • Computational linguistics • Search engines • Web site ranking • Search log analysis	• Cloud services • Cloud computing • Social media analysis • Social marketing
Network analytics	• Bibliometric analysis • Citation network • Social network theory	• Link mining • Dynamic network modelling
Mobile analytics	• Web services • Smartphone platforms	• Mobile web services • Mobile social innovation • Gamification

data and with a broad range of advanced analytical techniques (Chen et al. 2012) requiring high computing power. Within the big data analytics sub-domain a wide range of tools addresses different characteristics of big data. Since the big data characteristic “Variety” is already covered by the other BI&A sub-domains (e.g., text analytics uses text data, web analytics uses web data, etc.) and the characteristic “Volume” has mainly an impact on the underlying big data architectures (e.g., using distributed systems for high volume data computing), we will focus on the remaining characteristic “Velocity”. Since the high data generation and processing speed (i.e., velocity) is best addressed by streaming analytics, we consequently focus our tool evaluation for the sub-domain “big data analytics” on streaming analytics tools.

In addition, Chen et al. (2012) considered *text analytics* as a sub-domain targeting the analysis of unstructured text data (such as e-mails, corporate documents, and web pages). Following Ghosh et al. (2012), we understand the term text analytics as a synonym for text mining and vice versa. The technique of text analytics is based on information retrieval and computational linguistics which are already widely applied in industry. For instance, text analytics can provide valuable information about customers from social networks through sentiment analysis. So far, seven application areas for text analytics have been identified: web mining, classification, clustering, natural language processing (NLP), concept extraction, information extraction (IE), and information retrieval (IR) (Ghosh et al. 2012).

The third sub-domain is referred to as *web analytics*. It has gained growing attention with the emergence of web applications as services since the early twenty-first century. In addition, this trend is reinforced by the emergence of cloud computing and cloud services (Chen et al. 2012). Web analytics is used mainly in e-marketing to monitor customer behavior and get deeper insights into their motivation (Nakatani and Chuang 2011).

Network analytics also has gained attention with the increasing importance of social media. Grounded in scientific bibliometrics, network analytics can be applied in many research areas such as sociology, computer science, mathematics, and physics (Combe et al. 2010). Major application areas of network analytics are internet communities and social networks. The main task of network analytics is to identify relationships and network characteristics between different entities (e.g., centrality, betweenness) (Chen et al. 2012).

Finally, *mobile analytics* is regarded as a BI&A sub-domain. It is based on the rise of mobile devices and smartphone platforms such as Android or iOS (Chen et al. 2012). It is usually divided into two different categories, depending on whether mobile websites or mobile apps are analyzed. This most recent sub-domain can be described as the measurement and analysis of various social, behavioral and economic data which are generated by mobile platforms and environments and the usage of these patterns in gamification, mobile advertising, and social marketing (Chen et al. 2012).

17.3 Methodology

In order to evaluate and subsequently recommend different BI&A tools, we adapt the methodology for the evaluation and selection of computer-aided software engineering (CASE) tools suggested by IEEE (2010). Although the methodology is specifically designed for CASE tools, it fits also very well to the evaluation of BI&A tools due to its generic character. Following IEEE (2010), the evaluation and selection of software tools is divided into four mayor process steps:

S1 Preparation process,
S2 Structuring process,
S3 Evaluation process, and
S4 Selection process.

In the preparation step (S1), the overall goals of the evaluation have to be defined. As mentioned within the introduction, we seek for tools which are suitable for usage in higher education (Goal G1) as well as prepare the students for their future tasks in practice (Goal G2). Therefore the tools should be state of the art and should offer a broad range of functionality. Summarizing our evaluation goals are:

G1 The tool is suitable for use in academic teaching, and
G2 The tool represents the state of the art and offers a broad range of functionality.

In the structuring step (S2), we will derive requirements from our overall goals. While the university-specific requirements (UR, see below), related to the first goal (G1), are the same for all tool categories, the tool-specific requirements (TR), related to the second goal (G2), essentially depend on each sub-domain. Therefore, we will identify relevant criteria for each BI&A sub-domain separately. The third step (S3) corresponds to our actual evaluations, in which we assess and compare the tools related to the set requirements (both URs and TRs). The evaluation results can be found in the Tables 17.3, 17.4, 17.5, 17.6, and 17.7. In the last step (S4), we will select one tool per sub-domain that covers the requirements from both perspectives best. Based on the evaluation reports we justify our recommendations.

For the information gathering, we will mainly use documentations, publications, and web pages provided by the tool vendors as well as the BI&A literature.

17.3.1 University-Specific Requirements

The education sector is still under intense cost pressure and often subject of budget cuts. Hence, universities usually have limited budget for software, hardware and its maintenance. One option to face this challenge, is the usage of free learning resources and tools offered by academic programs (such as Teradata University Network [TUN], SAS Global Academic Program, and IBM Academic Initiative; for a full list see Table 17.2).

Due to the support by academic programs and the option to use open source solutions, we consider free of charge usage as one of the crucial requirements in the evaluation process (UR1). We will use the following rating to evaluate the criterion UR1:

○ The tool is not free of charge for academic use.
◉ A downgraded version of the tool is free of charge for academic use.
● The tool is free of charge for academic use.

In addition, a comprehensive documentation constitutes another key requirement as most faculty and students will get familiar with the tool on their own and through "learning by doing" (UR2). Moreover, the tool maintenance will be considerably easier if a good documentation is available. In addition, many vendors offer a variety of further information sources (such as frequently asked questions sections [FAQs] and tutorials). Tool related information can be provided either officially by the vendors or unofficially by user communities as it is often the case for open source tools. Consequently, our rating for the criterion UR2 is as follows:

○ The vendor provides a basic documentation only.
◉ The vendor provides a comprehensive documentation and additional information sources (like FAQs, tutorials, etc.).
● The vendor provides a comprehensive documentation and additional information sources. In addition, the vendor provides the option to share information with other users (e.g., via user groups, internet forums, etc.).

Table 17.2 Overview about academic programs by vendor

Vendor	Program name	Website
Google	Analytics Academy	https://analyticsacademy.withgoogle.com/explorer
IBM	IBM Academic Initiative	https://developer.ibm.com/academic/
IBM	IBM SPSS Mining in Academia	http://www-01.ibm.com/software/analytics/spss/academic/programs/
RapidMiner	RapidMiner Academia	https://rapidminer.com/academia/
SAP	SAP University Alliances	https://www.sap.com/training-certification/university-alliances.html
SAS	SAS Global Academic Program	http://support.sas.com/learn/ap/index.html
Software AG	University Relations	http://www.softwareag.com/corporate/community/uni/default.asp
Teradata	Teradata University Network (TUN)	http://www.teradatauniversitynetwork.com/
TIBCO	StreamBase University	http://www.streambase.com/community/streambase-university/

According to Wixom et al. (2011) there is a need for a wide range of free learning resources for students and for faculty. Especially, since on the one hand it enables students to improve their practical skills by self-instruction and on the other hand it provides additional course content for lecturers. Hence, the availability of free learning resources and teaching materials is represented in our evaluation of the BI&A tools as a university-specific requirement (UR3). We rate the criterion as follows:

○ The vendor does not provide learning resources.
⦿ The vendor provides either learning resources for students or teaching material for faculty.
● The vendor provides both, learning resources for students and teaching material for faculty.

Finally, the tools should be platform independent or at least available for the major operating systems (i.e., Windows, Linux, and Mac OS). Some tools are web-based and therefore independent of an operating system. The degree of platform independency constitutes the forth requirement (UR4) which will be assessed as follows:

○ The tool runs only on one of the major operating systems.
⦿ The tool runs on two of the three major operating systems.
● The tool runs on all major operating systems, namely Windows, Linux, and Mac OS, or is in fact platform-independent.

In summary, we suggest four university-specific requirements for the evaluations:

UR1 Free of charge,
UR2 Comprehensive documentation,
UR3 Free learning resources and teaching materials, and
UR4 Platform independency.

17.4 Tool Evaluations and Recommendations

Based on the aforementioned considerations we will select and evaluate different tools for each BI&A sub-domain. In the following we present for each sub-domain separately the sub-domain-specific requirements (TRs), the selected evaluation candidates, the evaluation report covering the assessment of all criteria, and the final tool recommendation based on the evaluation results.

17.4.1 Sub-domain "(Big) Data Analytics"

As justified within the section "Theoretical Foundations", we will evaluate streaming analytics tools as representatives for this sub-domain. Streaming analytics deals with the analysis of data from various infinite streams (Andrade et al. 2014). Such data is generated for example in the financial sector, in manufacturing or in intelligent transportation systems (Andrade et al. 2014). A particular challenge is the real-time analysis of data originating from multiple sources, i.e. the data processing has to be performed faster than new data is provided via the stream(s). In addition, a data stream often can be analyzed only in one single pass, as the data is not stored and therefore cannot be analyzed again later on (Ellis 2014).

We have selected the software solutions from the top four vendors of streaming analytics tools according to Forrester Wave report (Gualtieri and Curran 2014), namely IBM InfoSphere Streams, SAP Sybase Event Stream Processor, Software AG Apama Streaming Analytics, and TIBCO StreamBase LiveView (cf. Table 17.3), so all tools are widely accepted in practice. All of them represent stand-alone solutions. Assumed that most faculty and students have no experience with streaming analytics tools, it is especially important that those tools are user-friendly and provide a comprehensible graphical interface. Predefined streaming operators can help to facilitate the tool usage (Gualtieri and Curran 2014). Consequently, we consider predefined streaming operators as the first tool-specific requirement (TR-BDA-1). We rate the criterion by the fact if such operators are provided or not.

Another important characteristic of streaming analytics is the processing of data from different sources with various structures. Tools should be able to use a variety of data sources and provide different interfaces to access them. Therefore, we consider the range of accepted data sources as the second tool-specific criterion (TR-BDA-2). We have assessed this criterion by means of the Forrester Wave report

Table 17.3 Evaluation report for streaming analytics tools

	IBM InfoSphere Streams	SAP Sybase Event Stream Processor	Software AG Apama Streaming Analytics	TIBCO StreamBase LiveView
Website	http://www-03.ibm.com/software/products/en/infosphere-streams	https://www.sap.com/products/complex-event-processing.html	http://www.softwareag.com/corporate/products/apama_webmethods/analytics/overview/default.asp	http://www.tibco.com/products/event-processing/complex-event-processing/streambase-liveview
Latest version	4.0	5.1 SP 10	5.1	7.5
License	Proprietary	Proprietary	Proprietary	Proprietary
Platforms	○(Linux)	●(Windows, Linux, Unix)	●(Windows, Linux, Mac OS)	◉(Windows, Linux)
Free of charge	●	○	●	●
Documentation	●	●	●	●
Learning resources	●	◉	●	●
Streaming operators	●	●	●	●
Accepted data sources (from 0 "weak support" to 5 "strong support")	5	4	5	3

(Gualtieri and Curran 2014), which evaluates the supported data sources of streaming analytics tools on a scale from 0 (weak support) to 5 (strong support). Table 17.3 presents the evaluation report for streaming analytics tools, covering all UR and the big data analytics-specific requirements which are:

TR-BDA-1 Predefined streaming operators and
TR-BDA-2 Accepted data sources.

All vendors offer special programs for academic relations (for a full list see Table 17.2). IBM, Software AG and TIBCO feature their streaming analytics tools by these programs and offer free-of-charge versions for academic use. All tools are documented in detail. Additional information sources such as FAQs, tutorials and user guides are provided as well as community platforms, which are not tool-specific, but nevertheless allow the exchange of ideas and experiences with other users. As part of the IBM Academic Initiative free learning resources are available for InfoSphere Streams. For lecturers workshops, a faculty guide, and various reports are provided. Students can use a range of video tutorials, white papers, and reports. TIBCO provides a very extensive online learning course, various reports, white papers, and data sheets at leisure. Other material is available via the StreamBase University program. Software AG offers a variety of benefits for both, faculty and students. The University Relations program provides an education package which is tailored to the Apama Streaming Analytic tool. For students it includes an online training course, case studies, and the opportunity for free certification. For faculty teaching material, tutorials, and corresponding information are available. For the Sybase Event Stream Processor few free learning resources could be found. Although the SAP University Alliances program offers a wide range of learning courses, case studies, and podcasts, no specific material for the Event Stream Processor tool is available. Some resources, such as a starter guide, can be accessed via the SAP InfoCenter.

All tools feature predesigned streaming operators that simplify their use. Regarding the data sources that can be used, the solutions from IBM and Software AG have received the highest score, followed by the tool from SAP and with the lowest score TIBCO. A minor disadvantage of InfoSphere Streams is the fact, that it runs only on Linux-based operation systems.

From our point of view, we recommend the tool Apama Streaming Analytics from Software AG for use in academic education. Software AG offers the highest added value with regard to learning resources since the material corresponds well with the tool. As an example, the education package for students includes case studies and a scenario lasting over a period of 12 weeks (90 min/week). In the other categories Apama Streaming Analytics also reaches the highest scores. The tool offers an intuitive graphical user interface and useful predesigned streaming operators. Moreover, it provides the option to develop own custom streaming operators or interfaces and provides a component for data visualization. In summary, we see the Software AG solution as the best overall package for teaching a streaming analytics tool in higher education.

17.4.2 Sub-domain "Text Analytics"

Adequate text analytics tools are considered as another crucial factor in competition (Zikopoulos et al. 2012). Therefore, their usage in teaching should be aligned with the needs of business (Chiang et al. 2012). Paying attention to this, we have selected three tools from leading vendors in the field of advanced business analytics (Herschel et al. 2014) for evaluation, namely IBM SPSS Modeler Premium, SAS Text Enterprise Miner, and RapidMiner Studio (cf. Table 17.4), which are significant in practice and widely used. All three solutions are not dedicated text analytics tools, but comprehensive analytics packages with specific extensions for text analytics. We are focusing on these extensions in our evaluation.

Text analytics refers to a variety of application areas. As mentioned above, Miner et al. (2012) define seven areas which exhibit some overlaps. As the first tool-specific requirement we have selected the range of functionality covered by the tools (TR-TA-1). We have rated this criterion by the degree to which the tools are covering the seven areas.

Data sources for text analytics can be very heterogeneous. They consist of various text formats (e.g., Word, PDF, CSV, XML), online sources (HTML, RSS, e-mails), and databases (e.g., MySQL, MongoDB). A modern tool should be able to cover as many of these data sources as possible to provide a wide range of application areas. Accordingly, the second tool-specific requirement addresses the range of supported data sources by the text analytics tools (TR-TA-2). We rate the criterion as basic, if a selection of common text formats is supported, and as advanced, if database connections and/or web content are supported, too.

The results of text analytics can be visualized in different ways, for example by clustering diagrams or decision trees. State-of-the-art tools in this sub-domain should offer a wide range of visualization techniques without the need for additional software. Therefore, the third requirement is the capability to visualize the text analytic results (TR-TA-3). Summarizing the following domain-specific requirements were assessed (cf. Table 17.4):

TR-TA-1 Range of functionality,
TR-TA-2 Data sources supported, and
TR-TA-3 Visualization techniques.

All vendors offer their text analytics solution free-of-charge for academic use. SAS offers a free on-demand solution of the SAS Enterprise Miner, which also includes the SAS Text Miner via the SAS Global Academic Program. IBM provides a free-of-charge version of SPSS Modeler Professional including IBM SPSS Text Analytics for mining unstructured data sources via IBM SPSS Mining in the Academia program. Another solution offered by IBM is the SPSS Text Analytics for Surveys tool, which can be used for qualitative text analysis in surveys. RapidMiner provides in general a free version of its RapidMiner Studio. In addition, a professional version is free-of-charge for members of the RapidMiner Academia program. In contrast to the general free version, the professional version provides a wider

Table 17.4 Evaluation report for text analytics tools

	RapidMiner Studio	SAS Enterprise Miner	IBM SPSS Modeler Text Analytics
Website	https://rapidminer.com/solutions	http://www.sas.com/en_us/software/analytics/text-miner.html	http://www-01.ibm.com/software/analytics/spss/products/modeler
Latest version	6.5	14.1	17
License	GNU Affero General Public License (AGPL)	Proprietary	Proprietary
Platforms	●(Windows, Linux, Unix)	●(Web-based)	⦿(Windows, Linux)
Free of charge	●	●(Cloud-based)	●
Documentation	●	●	●
Learning resources	●	●	●
Range of functionality	Web mining, classification, clustering, NLP, concept, extraction, IE, IR	Web mining, classification, clustering, NLP, concept, extraction, IE, IR	Web mining, classification, clustering, NLP, concept, extraction, IE, IR
Data sources	Advanced	Advanced	Advanced
Visualization techniques	●	●	●

range of extensions (e.g., for Hadoop), more supported data sources and higher computing capacity. The tools are extensively documented and offer various teaching material for students and faculty. SAS provides a comprehensive documentation including tutorials, user guides, and factsheets. Furthermore, there is an internet forum, which addresses in particular text and content analytics topics. As part of the Global Academic Program, SAS provides various learning resources. Teaching material for faculty is available, which additionally can be accessed online via the SAS Live Web Classroom.

There are special certification programs and tutorials for students as well as a variety of white papers and reports. Furthermore, SAS is member of the TUN, which is one of the largest providers of learning resources in the BI&A community (Wang 2015). Via TUN a variety of additional free learning resources can be accessed. Besides the general Academic Initiative, IBM has created a special program for data and text mining, the Mining in Academia program. The detailed tool documentation (including an internet forum) is complemented by a series of free learning resources. One the on hand, IBM provides for faculty teaching material especially for the tool and for text analytics in general, on the other hand, students will benefit from various resources ranging from white papers and reports to video courses, a special support forum for students, and diverse certification options. RapidMiner provides a manual, a starter guide, and tutorials for documentation. The open source tool has an active community, including an internet forum and a marketplace for extensions and plug-ins. Furthermore, it offers a couple of free learning resources for students, namely white papers, webinars, and reports. For faculty free certifications and a repository with sample data are available.

Among the evaluated tools no significant differences could be found with regard to the TRs. All solutions cover a broad range of text analytics application areas, including all areas provided by Miner et al. (2012). The accepted data formats are manifold for all tools and comprise common text formats, web data, and database interfaces. In order to analyze data, SAS Enterprise Miner and SPSS Modeler need to transform them into a proprietary format. In contrast, a preprocessing is not necessary using RapidMiner for data analysis. The graphical representation of analysis results is also supported by all tools.

Overall, we can see no clear winner in this sub-domain. The decision depends on the preferences of faculty. If an open source solution with an active community and a variety of extensions is preferred, RapidMiner seems to be an appropriate solution. For a cloud-based solution without installation efforts and a comprehensive online course offering, SAS Enterprise Miner seems to be the means of choice. The in-house solution from IBM provides the advantage of an additional text analytics tool for surveys and offers a special student forum and other useful free learning resources. Therefore, also the SPSS Modeler can be recommended. Summarizing, we consider all evaluated tools as suitable solutions for teaching.

17.4.3 Sub-domain "Web Analytics"

The market for web analytics software is large and hence hard to survey. For the evaluation we have chosen a combination of two market leading vendors in this field (Google Analytics and Quantcast Measure) and two open source solutions, namely PIWIK and Open Web Analytics (cf. Table 17.5). While the proprietary tools are web-based software as a service solutions (SaaS solutions), the open source tools represent in-house solutions, requiring a separate server to store the raw data.

There are different ways to track user activities on the web. In practice, a distinction between web log analysis and page tagging has been established (Nakatani and Chuang 2011). Page tagging, in most cases via JavaScript Cookies or PHP, allows a more detailed look on user actions that are not tracked by web log analysis. However, such a client-side data collection requires cookies, which must be approved by web site visitors. If the cookies are deleted or expire, the quality of data collection decreases. Since web log analysis is based on stored data, historical log data can be analyzed. Moreover, it stores data in consistent log files on in-house servers. Both methods are widely applied in practice. We believe students should be taught in both methods to get a broad view on web analytics methods. Therefore we consider the tracking methods as the first tool-specific requirement (TR-WA-1).

The way the tracking data is stored in a database is another distinguishing criterion for web analytics tools. On the one hand, most in-house solutions require a separate database (e.g., MySQL) causing additional costs and efforts for server deployment and maintenance. One the other hand, keeping the database in-house allows access to raw data and prevents legal problems about data privacy. By means of the second tool-specific requirement (TR-WA-2) we note if the database is located in-house or in the cloud and which data format is used for storage.

Finally, Nakatani and Chuang (2011) point out that web analytics tools can be distinguished with regard to their capability to provide data analysis in real-time. We choose this aspect as the third tool-specific requirement (TR-WA-3) and check whether the tools are able to collect and analyze data in near real-time or not. This leads to the following three requirements specific for web analytics tools (cf. Table 17.5):

TR-WA-1 Tracking methods,
TR-WA-2 Data storage, and
TR-WA-3 Real-time functionality.

As stated in Table 17.5 all tools are free-of-charge. Furthermore, PIWIK and Google Analytics offer an enterprise version as a fee-based SaaS solution and a premium version for extremely high data volumes, respectively. All tools provide a detailed documentation in different forms, like guides, FAQs, and/or wikis. In addition, all tools except for Quantcast Measure, offer community functionalities in the form of internet forums. Additionally, Open Web Analytics also offers a user chat via Internet Relay Chat (IRC). All tools come with some free learning resources such as videos and tutorials, whereas only Google Analytics provides a comprehensive selection. In addition to some tutorials and an extensive video library on

Table 17.5 Evaluation report for web analytics tools

	Google Analytics	Open Web Analytics	PIWIK	Quantcast Measure
Website	http://www.google.com/intl/en/analytics	http://www.openwebanalytics.com	http://piwik.org	https://www.quantcast.com
Latest version	N.A.	1.5.7	2.14.3	N.A.
License	Proprietary	GNU General Public License (GPL)	GNU General Public License (GPL)	Proprietary
Platforms	●(Web-based)	⦿(Linux, Windows)	●(Windows, Linux, Mac OS)	●(Web-based)
Free of charge	●	●	●	●
Documentation	●	●	●	⦿
Learning resources	●	⦿	⦿	⦿
Tracking methods	Page tagging (via JavaScript)	Page tagging (via JavaScript, PHP)	Log files, page tagging (via JavaScript, PHP)	Page tagging (via JavaScript)
Data storage	Cloud (proprietary)	In-house (MySQL)	In-house (MySQL)	Cloud (proprietary)
Real-time functionality	●	●	●	●

YouTube, Google offers a variety of online training courses, the so-called Analytics Academy, including practical exercises and a learning community.

With regard to the tracking method, the tools from Google and Quantcast rely on page tagging via JavaScript. Open Web Analytics provides additional page tagging via PHP, while PIWIK allows both page tagging (via JavaScript or PHP) and log data analysis. Google Analytics and Quantcast Measure rely on a cloud-based database with proprietary data formats, while the open source tools require a separate MySQL database. All tools exhibit the capability for real-time data collection and analysis.

From our point of view, we consider PIWIK as a suitable solution for use in academic education. The main advantage of PIWIK is its option to use both data collection methods, page tagging and log data analysis. Thus, it offers a broader range of opportunities for teaching than the other tools. Also, in contrast to the proprietary solutions, it uses the popular open source database MySQL for data storage. Case studies and exercises with sample data to convey special aspects in the field of web analytics might be conducted more convenient due to the popularity of MySQL. Compared to Open Web Analytics, PIWIK exhibits a wider distribution resulting in a larger community. PIWIK provides a forum where an active community gives assistance and suggests ideas for future development. In addition, various extensions can be downloaded in form of third-party plug-ins via the PIWIK marketplace.

17.4.4 Sub-domain "Network Analytics"

There is a large pool of potential network analysis tools for usage in the academic context. To make a preliminary selection we have aligned our evaluation to the criteria provided by Combe et al. (2010). Thus, only tools are taken into account, which are able to process networks with a minimum of a five-digit number of nodes and which have at least basic functionalities in network analysis. In addition, we have checked whether the tools have already been used in an academic context. In consideration of these criteria we have identified the network analytics tools Gephi, Pajek, and UCINET (cf. Table 17.6).

For tool-specific requirements we consider a large capacity in processing network data as relevant. Especially in the big data context it is necessary to analyze large network structures with hundred thousands of nodes. For this reason the maximum number of nodes that can be processed constitutes the first tool-specific requirement (TR-NA-1).

Furthermore, there is a great variance in the range of supported network analytics functions (TR-NA-2). For our evaluation we distinguish between basic functionality, which covers the fundamental metrics of network analysis, and advanced functionality, which includes additional algorithms and calculations.

Table 17.6 Evaluation report for network analytics tools

	Gephi	Pajek	UCINET
Website	http://gephi.github.io	http://mrvar.fdv.uni-lj.si/pajek/	https://sites.google.com/site/ucinetsoftware
Latest version	0.8.2	4.05	6.586
License	GNU General Public License (GPL)	Closed source	Closed source
Platforms	●(Windows, Linux, Mac OS)	⦿(Windows; Linux, Mac OS via Wine)	⦿(Windows, Linux, Mac OS via Wine)
Free of charge	●	●	○
Documentation	●	⦿	⦿
Learning resources	●	●	●
Maximum number of nodes	150.000	500.000	32.767
Range of functionality	Basic	Advanced	Advanced
Visualization techniques	●	●	●

The third tool-specific criterion indicates whether the software has integrated visualization capabilities or if additional software is required for this purpose (TR-NA-3). The following network analytics-specific TRs were assessed (cf. Table 17.6):

TR-NA-1 Maximum number of nodes,
TR-NA-2 Range of functionality, and
TR-NA-3 Visualization techniques.

For use in the academic education, we recommend a combination of two tools. The free-of-charge open source tool Gephi provides extensive options for visualization of networks in real time, using a 3D engine (Bastian et al. 2009). Once the network has been visualized, it can be interactively modelled and comprehensively analyzed. Hereby it does not matter if complex, dynamic or hierarchical graphs are subject to analysis (Bastian et al. 2009). The program features a user-friendly, graphical user interface and can process up to 150.000 data nodes (Combe et al. 2010). In addition to a detailed documentation (e.g., tutorials, FAQs, Wiki,), there is an internet forum for sharing experiences, use cases, and a user chat via IRC. As learning resources, Gephi offers various official and user-generated (video) tutorials and instructions, facilitating self-studying for students. For faculty the vendor provides various sample data which can used for designing exercises for teaching. Gephi is Java-based and available for Windows, Linux and Mac OS. A disadvantage of Gephi is its limited functionality compared to other network analytics solutions. Since Gephi is a fairly new tool and still in the beta phase (latest Version 0.82) advanced functionalities might be available in the future.

For advanced analysis, we recommend the free-of-charge software Pajek. The tool has been developed at the University of Ljubljana and has continuously been improved and supplemented by various functions since that time. Further advantages are the solid performance and the high number of nodes that can be processed (up to 500.000) (Combe et al. 2010). The program includes comprehensive options for visualization, is well documented by the developers, and offers an extensive wiki. Among others, the Pajek-Wiki provides a variety of free learning resources (e.g., course documents, sample data, and academic papers). Pajek is designed for Windows, but also works with the help of a virtual runtime environment (e.g., Wine) on Linux. Due to the outdated design, Pajek's user interfaces needs some time to getting used to (Lombardi 2011). UCINET is not for free-of-charge and it cannot handle larger amounts of data. In addition, it requires a separate visualization tool, resulting in more effort and maybe more costs.

Overall, we propose Gephi as an entry-level solution, in particular suitable for self-studying and to teach fundamentals of network analytics. A big advantage is the active community that has emerged around the tool. To get deeper insights into the topic and to perform advanced analysis, we recommend Pajek. The main advantage of Pajek results from its maturity. It has been continuously developed since 1998 and therefore it constitutes a very robust version with many extra resources. A complementary use of both solutions is also recommended, because the native Pajek file format can also be imported into Gephi. Thus, it allows the design of teaching materials crossing tool boundaries.

17.4.5 Sub-domain "Mobile Analytics"

The market for mobile applications is constantly growing as more and more users access the internet via mobile devices. The newest of the five BI&A sub-domains has great similarities with the web analytics sub-domain, since most of the web analytics tools provide options to analyze data via mobile websites (e.g., Google Analytics). However, mobile analytics does not relate only to the analysis of mobile websites but also to the analysis of smartphone and tablet apps. In addition, the mobile market provides specific requirements for analytics, which we present within the next paragraphs. For this reason, we follow Chen et al. (2012) and consider mobile analytics as a separate sub-domain. We evaluate four vendors that have specialized in mobile analytics, i.e. tools which provide mobile analytics as part of their generic web analytics solution are excluded. The assessed tools are Apsalar, Crittercism, Flurry Analytics, and Localytics (cf. Table 17.7).

First, it is important that the most popular mobile platforms are supported by the tool. Based on their market share we consider especially Apples iOS and Googles Android as important, followed by Windows Phone, BlackBerry, and Amazons Fire OS. Consequently, the first tool-specific requirement is the range of supported platforms (TR-MA-1).

As we mentioned in the web analytics subsection, real-time functionality is important for web analytics and same is true for the analyses of mobile websites and applications. Real-time information help to improve app performance and timely troubleshooting (Ask 2013). The ability to process data in real-time is therefore the second tool-specific criterion (TR-MA-2).

Several metrics are used in practice to analyze mobile apps. The Forrester research institute differentiates five potential core metrics (Ask 2013). Besides the analysis of user engagement, Forrester mentions financial, performance, benchmarking, and qualitative metrics. For the evaluation, we summarize the capability to measure indicators in these categories as the range of functionality (TR-MA-3). We distinguish basic functionality (the tool covers 1–3 categories according to Ask 2013) and advanced functionality (4–5 categories according to Ask 2013).

The last tool-specific requirement for mobile analytics tools is the capability of data imports and exports (TR-MA-4). This leads to the following assessment criteria for mobile analytics tools (cf. Table 17.7):

TR-MA-1 Supported mobile platforms,
TR-MA-2 Real-time analytics,
TR-MA-3 Range of functionality, and
TR-MA-4 Data interfaces.

All tools are web-based and available as free versions. However, Localytics and Crittercism are only free-of-charge up to a limit of 10,000 or 30,000 monthly active users, respectively. The functionality of Crittercism is limited in the free version. All tools offer a solid documentation including FAQs and white papers. Crittercism and Apsalar feature in addition community functionalities, the former one in form of an internet forum, the latter one in the form of a FAQ section on the website. Free learning resources are offered by all vendors, e.g. in the form of video tutorials

Table 17.7 Evaluation report for mobile analytics tools

	Apsalar	Crittercism	Flurry Analytics	Localytics
Website	https://apsalar.com	http://www.crittercism.com	http://www.flurry.com	http://www.localytics.com
Latest version	N.A.	N.A.	N.A.	N.A.
License	Proprietary	Proprietary	Proprietary	Proprietary
Platforms	●(Web-based)	●(Web-based)	●(Web-based)	●(Web-based)
Free of charge	●	⦿	●	⦿
Documentation	●	●	⦿	⦿
Learning resources	●	●	⦿	●
Supported platforms	iOS, Android, Unity	iOS, Android, Windows Phone, Unity, Appcelerator	iOS, Android, Blackberry, Windows Phone, Java ME,	iOS, Android, BlackBerry, Windows Phone, Windows 8
Real-time analytics	●	●	○	●
Range of functionality	Basic	Basic	Advanced	Advanced
Data interfaces	Import/export	Export	Import/export	Import/export

and how-tos. Apsalar offers in addition a range of white papers and a couple of cheat-sheets which are suitable for teaching. Localytics and Apsalar also provide ebooks, case studies, and webinars. The leading mobile platforms (iOS and Android) are supported by all vendors. All tools except Flurry Analytics have the ability to perform real-time analysis (Ask 2013). The functional range for Apsalar and Crittercism is rated as basic and for Flurry Analytics and Localytics as advanced. Attention should be paid to Localytics which covers the full range of functionality (Ask 2013). Concerning the data interfaces all tools have the capability for data import and export, except for Crittercism, which only supports data export. Access to raw data in Localytics is only available for the commercial enterprise version.

Overall we recommend Apsalar as the most suitable tool for usage in academic context. The tool offers extensive documentation including community functionalities and a variety of free learning resources. Besides the interface for data export and import, the tool provides full access to the data. Apsalar accesses user data only in an anonymous form which improves data privacy. A shortcoming of Apsalar in comparison to the other tools is the limited functionality. However, Absalar achieved the best results with regard to the other criteria.

17.5 Conclusion, Limitations, and Further Work

The market for BI&A tools is very large and hence hard to survey. Many vendors want to benefit from the ongoing trend towards big data. In addition, more and more open source solutions gain maturity and compete with the established vendors. Universities face the challenge to select appropriate tools from this large range of offerings. In order to support universities in this challenge, the paper at hand presented different tool evaluations covering all five BI&A sub-domains.

For the evaluation, the BI&A research framework from Chen et al. (2012) was used to deduce different tool categories, from which potential tools have been selected as evaluation candidates. The evaluation requirements have been derived from the academic context and from specific characteristics of each BI&A sub-domain. We were able to identify appropriate tools for university education in all five BI&A sub-domains and gave a broad overview about their license models, functionalities and their provision of documentation and free learning resources.

The evaluations cannot provide an exhaustive investigation of all available tools, but we aimed at making a contribution to the selection of suitable tools for education. Currently, our tool recommendations are based on the evaluation of the available tool documentation resources and the current BI&A literature only. The evidence could be strengthened by surveys of instructors actually using these tools within their teaching sessions. A further limitation is the strict focus on streaming analytics tools only within the sub-domain "big data analytics". In addition, further tool categories, such as data mining tools or in-memory databases could be evaluated and currently remain subject to further work.

The evaluation considers tools from different, rather independent sub-domains. Future research should examine the options for an integrated solution that covers the full range of applications in the BI&A domain. Thus, the business suite of the major vendors, such as SAP, SAS or IBM could cover multiple sub-domains of the BI&A research framework. Furthermore, the Hadoop ecosystem and its various extensions seem to be an appropriate opportunity for use in higher education. For example, Apache Storm in combination with Hadoop can be used for streaming analytics as part of the sub-domain (big) data analytics. Storm is a scalable solution which became an Apache top level project in 2014. In future, Storm could turn into a real challenger for the established vendors. On the one hand, the implementation and operation of a Hadoop cluster requires extensive domain knowledge and a certain degree of self-development effort, on the other hand, the options for self-development and adjustments also add value to the education of students and to the faculty experiences.

Biographies

Christoph Kollwitz (christoph.kollwitz@wirtschaft.tu-chemnitz.de) is a Ph.D. candidate at Chemnitz University of Technology. His research interests include online communities, open data and open innovation. He holds a Bachelor's degree in business and economic science from Friedrich Schiller University Jena and a Master's degree in customer relationship management from Chemnitz University of Technology. Christoph is currently working in the research project CODIFeY—Community-based Service Innovation for E-mobility, where he is concerned with data analytics in the fields of co-creation communities.https://www.tu-chemnitz.de/wirtschaft/wi1/team/christoph-kollwitz/

Barbara Dinter (barbara.dinter@wirtschaft.tu-chemnitz.de) is Professor and Chair of Business Information Systems I at Chemnitz University of Technology. She holds a Ph.D. from Technische Universität München, Germany, where she previously earned a Master's degree in computer science. In her role as an IT consultant, she has worked with a variety of organizations. Her research interests include BI and analytics, big data, data driven innovation, and information management. She has published in renowned journals such as *Decision Support Systems, Information Systems Management, Journal of Database Management*, and *Journal of Decision Systems*, and on conferences such as ICIS, ECIS, and HICSS.https://www.tu-chemnitz.de/wirtschaft/wi1/team/barbara-dinter/

Robert Krawatzeck (robert.krawatzeck@wirtschaft.tu-chemnitz.de) is a doctoral candidate at Chemnitz University of Technology. His research interests include agile business intelligence, data warehouse/business intelligence testing, design science research together with test-driven and model-driven development. He received a diploma in computer science from Chemnitz University of Technology. Robert is a member of the agile business intelligence task force from the German Chapter of TDWI. He has published in journals like Information Systems Management, and on conferences like AMCIS, DESRIST, ECIS, ER, HICSS, ICIS, and PACIS. https://www.tu-chemnitz.de/wirtschaft/wi2/wp/en/team/robert-krawatzeck/

References

Accenture (2014) Big success with big data. https://www.accenture.com/us-en/insight-big-data-research.aspx

Al-Sakran HO (2015) Development of business analytics curricula to close skills gap for job demand in big data. Inf Knowl Manag 5(3):36–46

Andrade HC, Gedik B, Turaga DS (2014) Fundamentals of stream processing: application design, systems, and analytics. Cambridge University Press, New York, NY

Ask JA (2013) Forrester's shopping guide to mobile analytics vendors. https://www.forrester.com/Forresters+Shopping+Guide+To+Mobile+Analytics+Vendors/fulltext/-/E-RES99541

Bastian M, Heymann S, Jacomy M Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the international AAAI conference on weblogs and social media (ICWSM'2009), San Jose, 2009, pp 361–362

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188

Chiang RHL, Goes P, Stohr EA (2012) Business intelligence and analytics education, and program development: a unique opportunity for the information systems discipline. ACM Trans Manag Inf Syst 3(3):12:1–12:13

Combe D, Largeron C, Egyed-Zsigmond E, Géry M (2010) A comparative study of social network analysis tools. In: Proceedings of the international workshop on web intelligence and virtual enterprises, vol 2, pp 1–12

Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. Harv Bus Rev:70–76

Davis GA, Woratschek CR (2015) Evaluating business intelligence/business analytics software for use in the information systems curriculum. Inf Syst Educ J 13(1):23–29

Ellis B (2014) Real-time analytics: techniques to analyze and visualize streaming data. Wiley, Indianapolis, IN

Ghosh S, Roy S, Bandyopadhyay SK (2012) A tutorial review on text mining algorithms. Int J Adv Res Comput Commun Eng 1(4):223–233

Gualtieri M, Curran R (2014) The forrester wave: big data streaming analytics platforms, Q3 2014. https://www.forrester.com/The+Forrester+Wave+Big+Data+Streaming+Analytics+Platforms+Q3+2014/fulltext/-/E-RES113442

Herschel G, Linden A, Kart L (2014) Magic quadrant for advanced analytics platforms. https://www.gartner.com/doc/2667527

IEEE (2010) Information technology – guideline for the evaluation and selection of CASE tools (IEEE Std 14102-2010). New York, NY, IEEE

Jacobi F, Jahn S, Krawatzeck R, Dinter B, Lorenz A Towards a design model for interdisciplinary information systems curriculum development, as exemplified by big data analytics education. In: Proceedings of the 22nd European Conference on Information Systems (ECIS'2014), Tel Aviv, Israel, 2014, pp 1–15

Laney D (2001) 3D Data management: controlling data volume, velocity, and variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lombardi T (2011) Introduction to Pajek. http://youtu.be/PRrKo0maZ8Y

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute

Miner G, Elder J, Hill T, Nisbet R, Delen D (2012) Practical text mining and statistical analysis for non-structured text data applications. Academic, Oxford

Nakatani K, Chuang T-T (2011) A web analytics tool selection method: an analytical hierarchy process approach. Internet Res 21(2):171–186

Schiller S, Goul M, Iyer LS, Sharda R, Schrader D, Asamoah D (2015) Build your dream (not just big) analytics program. Commun Assoc Inf Syst 37:1. Article 40

Wang Y Business intelligence and analytics education: hermeneutic literature review and future directions in IS education. In: Proceedings of the Americas conference on information systems (AMCIS'2015), Puerto Rico, 2015, pp 1–10

Wixom B, Ariyachandra T, Douglas D, Goul M, Gupta B, Iyer L, Kulkarni U, Mooney JG, Phillips-Wren G, Turetken O (2014) The current state of business intelligence in academia: the arrival of big data. Commun Assoc Inf Syst 34(1):1–13. Article 1

Wixom B, Ariyachandra T, Goul M, Gray P, Kulkarni U, Phillips-Wren G (2011) The current state of business intelligence in academia. Commun Assoc Inf Syst 29(1):299–312. Article 16

Zikopoulos PC, DeRoos D, Parasuraman K, Deutsch T, Corrigan D, Giles J (2012) Harness the power of big data: the IBM big data platform. New York, NY, Mc Graw Hill

Chapter 18
Neural Net Tutorial

Brian R. Huguenard and Deborah J. Ballou

Abstract When problems are complex and cannot be solved through conventional methods such as statistical or management science models, and when human expertise is not sufficient for efficiently finding high-quality solutions, we can consider the use of machine learning techniques. One such technique is the artificial neural network (neural net), which can be used for predictive modeling. This chapter provides a brief introduction to the topic of neural nets, along with a tutorial in which a working neural net is built and then used to make predictions.

Keywords Neural net • Machine learning • Predictive modeling

This chapter provides instructions for creating an artificial neural network that will be used to predict winners in horse races. The same instructions could be used to create neural networks in other domains. Through reading this chapter the reader should be able to acquire a basic understanding of how to create and use a neural network to aid in a predictive decision-making task. This chapter would also be an appropriate assignment in a business analytics course.

18.1 Introduction

When problems are complex and cannot be solved through conventional methods such as statistical or management science models, and when human expertise is not sufficient for efficiently finding high-quality solutions, we can consider the use of machine learning techniques. The basic idea behind machine learning is that the program analyzes historical problems and examples of past solutions, looking for patterns in the data that can be used to develop strategies and rules for solving future problems. These learned strategies are then incorporated into the future behavior of

B.R. Huguenard (✉) • D.J. Ballou
Department of Decision Sciences and Management, Tennessee Technological University,
Room 306, 1105 Peachtree St., Johnson Hall, Cookeville, TN, USA
e-mail: bhuguenard@tntech.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_18

the system, so that new problems can be solved more effectively. One such technique is the artificial neural network, or neural net (Gurney 1997; Haykin 1999; Lawrence 1994).

Artificial Neural Networks (ANN) are loosely based on a biological neural network. An ANN (or just neural net) is implemented with software simulations of the massively parallel processes that occur in the human brain. Neural net models are not intended to be accurate representations of real biological systems…they are more of an analogy to the human brain than an accurate model of it. Because of the interconnected nature of these networks, neural net models are also often referred to as connectionist models.

Neural nets have the ability to learn from experience, where the experience is gained from viewing historical data consisting of sets of inputs and corresponding solutions to those inputs. This process is called training. Some examples of problem domains where neural nets have been used successfully include:

- recognizing handwritten characters
- training a computer to pronounce English text
- prediction of fraud in business transactions
- diagnosis of complex medical conditions

The remainder of this chapter begins with an overview of the structure and workings of a neural net, followed by a tutorial involving the creation and use of a simple neural net.

18.2 Overview of Neural Nets

Before going into any more detail about neural nets, let's talk a bit about their biological counterpart. The human brain is composed of billions of special cells called neurons. Neurons are organized into groups called networks. Each network contains several thousand neurons that are highly interconnected. Signals are sent from one neuron to another, and each neuron has the capability to either leave the signal strength unchanged, decrease the signal strength, or increase the signal strength before sending it along to other cells. Information is stored in the gaps between the cells, called *synapses*.

18.2.1 Structure of a Neurode

A typical artificial neural net consists of a number of interconnected artificial neurons (sometimes called neurodes)…one such neurode is shown in Fig. 18.1. At the left edge of Fig. 18.1 are the inputs. For example, if this neurode were for a loan application problem, the inputs X1, X2, through Xn might represent the loan applicant's income level, age, and marital status. Note that each input in the neurode has

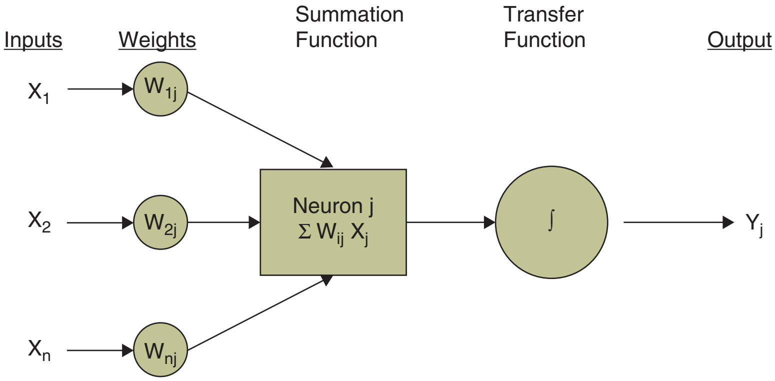


Fig. 18.1 Example neurode from a neural net

its own *weights*, the Wij's. These weights serve to represent the relative importance of the input. The neural net learns by adjusting these weights so that the desired output is created. The adjustment of these Wij's is analogous to the changing of synaptic strength in biological neurons. The *summation function* consolidates the various inputs along with their respective weights into a single weighted sum. This defines the internal activation level of the neurode. The *transfer function* processes this internal activation level and creates the output, usually transforming the internal activation into a value between 0 and 1. A low internal activation level may result in an output value of 0 or near 0, and high internal activation levels may result in an output near 1. The final output result, Yj, would then serve as an input to one or more other neurodes in the neural net, and the whole process would repeat itself for those other neurodes.

18.2.2 Layout of a Neural Net

In Fig. 18.2 we see the conceptual layout of a neural net. Remember that these components are all simulated in a computer program. The basic component of a neural net is a neurode. Each of the colored ovals in Fig. 18.2 represents a single neurode. A neural net is composed of a collection of neurodes grouped in layers. A typical structure is shown in Fig. 18.2, with three layers: an input layer, an intermediate layer called the hidden layer, and an output layer.

Each input to the input layer corresponds to an attribute of a problem, for example: if this were for a loan application problem, the inputs X1, X2, X3, and X4 might represent the loan applicant's income level, age, marital status, and gender. The inputs can be numeric or categorical in nature. A numeric representation of these attribute values serve as the input to the neural net. The input layer typically doesn't change the value of the input it receives…it just passes it along to the hidden layer.

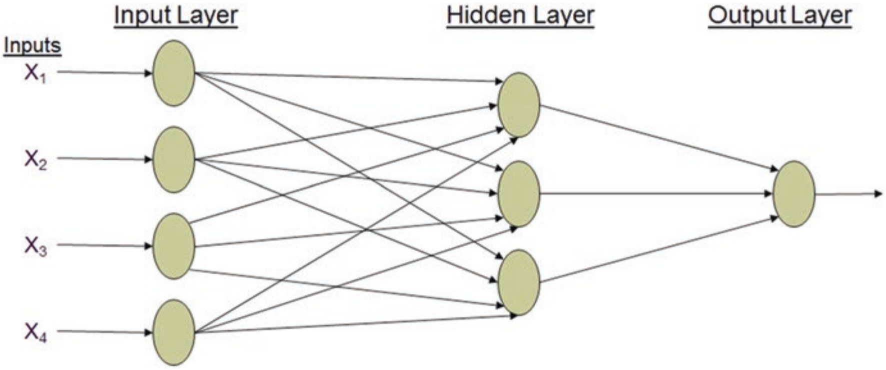


Fig. 18.2 Conceptual layout of a neural net

Once the input values arrive at the hidden layer, they are processed as shown in Fig. 18.1. The results of the hidden layer are then passed on to the output layer, where final processing occurs and the results of the neural net are computed. The output of the neural net might contain the answer to the problem, or it might provide input to another neural net. In the case of a loan application problem, the final answer coming from the net might be a yes or no.

18.2.3 Training a Neural Net

A neural net learns through a process called training. Training techniques fall into two categories: supervised and unsupervised. *Supervised learning* requires historical data giving cases of inputs and correct outputs. For example, we might have data where each case provides loan applicant characteristics, and the corresponding output is the correct classification of that application as being accepted or denied the loan. The neural net is provided the input, one case at a time, and when it produces a final answer for that case it compares its answer to the desired answer. If there is a difference between the two, then the weights of the net are adjusted in an attempt to correct the net's accuracy. *Unsupervised learning* involves using only input stimuli. No comparison to a desired output is performed, and the net develops its own set of categorizations for the input. Humans must then interpret the categorizations created by the net, and determine if they are useful or not. This approach can be useful for exploratory data analysis. Supervised learning is the most commonly used approach, and is the one we will discuss further.

The basic technique of supervised learning is one of repeatedly giving the neural net examples of solved problems, where the net has to first come up with a solution on its own, then is allowed to compare its answer to the correct one. If there is a difference between the two answers, that's where the learning occurs…the weights of the net are adjusted in an attempt to improve the performance of the net for future training

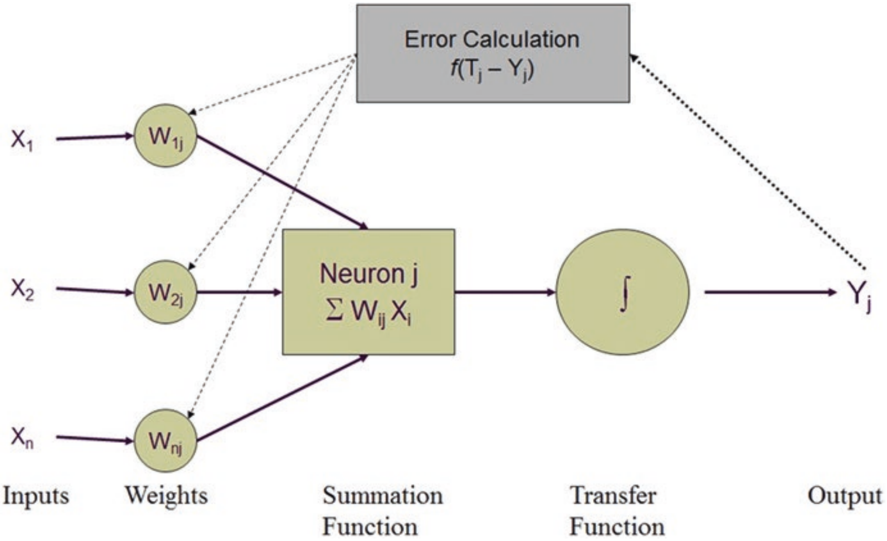


Fig. 18.3 Supervised learning through backpropagation

problems, and then another example solved problem is provided. This iterative process will continue until some threshold value of accuracy has been reached, or until some limit has been reached in the number of training problems to be processed.

The most commonly used supervised learning algorithm is called backpropagation. In Fig. 18.3 we once again have a representation of an individual artificial neuron, with representations of the inputs, the weights, the summation function, the transfer function, and the output.

If the output shown here, Yj, were a final output of the neural net for one training case, then during backpropagation the value Yj would be compared to the actual correct answer, or target answer, contained in the training case. We'll call that correct answer Tj. An error calculation is performed that is a function of the difference between the net's answer (Yj) and the target answer (Tj). If the net's answer is close enough to the target answer (based on some predetermined tolerance), then no further action is required for this particular training case, and the next training case would be started. However, if the error between the net's answer and the target answer is large enough (over the designated tolerance level), then the weights of this neuron will then be adjusted in proportion to the severity of the error. Now this was all for just one artificial neuron. The error from this neuron would then be propagated backwards through the hidden layers of the entire neural net, until adjustments have been made to weights as needed over the entire network.

The ultimate goal of backpropagation is to arrive at a set of weights that fits the training data so as to minimize the error between the network's answers and the target answers. Once a stable set of weights have been obtained, then training is over and the neural net is ready to accept new input, and its output can be used to help form a recommended decision.

18.2.4 Advantages and Disadvantages of Neural Nets

Neural nets can save time and effort in that manual design and creation of programming-language based models can be avoided. Since the neural nets train themselves, we don't have to begin with a deep understanding of how the inputs are related to the outputs… we don't have to have human experts in the problem domain in order to user neural nets. Neural nets are also naturally adaptable to changing input, rather than requiring a programmer to modify a more static application. Like humans, neural nets can process incomplete or inaccurate data and still produce useful output. In general, neural nets are very useful for pattern recognition (for example, interpreting handwritten messages), classification (example: classifying corporate bonds in terms of risk rating), and prediction problems (example: predicting future performance of the economy).

On the negative side, the inner workings of a neural net are a black box. There is no mechanism for providing an explanation of the decisions that it makes. Although training can be automated, it can still require a lot of time if there are many hidden layers. As a result, most neural nets have only one or two hidden layers. Finally, neural nets are not appropriate for number-crunching type problems…they are best at pattern recognition and classification.

18.3 Example Implementation of a Neural Net

18.3.1 Download the Neural Network Software

Before you work on this tutorial you need to download and install the software EasyNN-plus (the software runs on a PC only, not on a Mac). Go to the website http://www.easynn.com/dltrial.htm and click on the "Download as a Zip file" link. When asked whether to open or save the file, you should save it to your desktop. Once the file has finished downloading (it should be called something like "ennsetup.zip"), you need to unzip it (you should be able to right-click on it and choose the "Extract All" or "Unzip" command). After unzipping the file you should now have a folder on your desktop called "ennsetup". Go into the "ennsetup" folder and double click on the file "ennsetup.exe". When asked if you want to run the file, say yes, and that should start up the installer. Just give the default answers to questions it asks, and it should install the EasyNN-plus software on your machine. The free version of this software will run for 30 days and can accept a maximum of 100 rows of input data.

18.3.2 Download a Copy of the Data File

You can get download a copy of the Races98.txt file from [?URL?]. This file will serve as the input file for this tutorial. Once you have acquired a copy of Races98.txt, double-click on the file so you can have a look at its content (do not change anything in the file). You will see that the column titles are on the first line and the other lines start with the line number in square brackets. We will use the titles for column names and the numbers for row names. Some of the values in the race data are integers and some are Boolean (0 or 1).

	Runners	Distance	Handicap	Class	Stake>5k	Odds>2	Win
[1]	11	7	0	5	0	1	0
[2]	5	8	0	3	0	0	1
[3]	7	5	1	2	1	1	0
[4]	4	8	0	1	1	0	1
[5]	8	14	1	4	0	1	1
[6]	10	10	1	3	1	0	0
[7]	6	8	0	4	0	0	1
[8]	4	6	0	3	0	0	0
[9]	13	8	1	3	1	1	0
[10]	9	14	1	1	1	1	0
[11]	12	7	0	3	1	1	0
[12]	5	13	0	4	0	0	1
[13]	12	5	1	4	1	1	1
[14]	4	14	0	1	1	0	1
[15]	12	7	1	2	1	1	0
[16]	18	6	1	3	1	1	1
[17]	9	8	0	1	1	1	1
[18]	22	10	1	5	0	1	0
[19]	10	9	1	5	0	1	0
[20]	5	7	0	4	0	0	1
[21]	16	6	0	5	0	1	1
[22]	12	10	0	6	0	0	0
[23]	3	6	0	2	1	0	1
[24]	12	8	1	3	1	1	0
[25]	3	18	0	3	1	0	1
[26]	18	6	1	5	0	1	0
[27]	4	12	0	6	0	0	0
[28]	6	6	1	5	0	0	1
[29]	8	7	0	7	0	0	1
[30]	12	7	0	6	0	1	1
[31]	16	10	1	6	0	1	0
[32]	10	12	1	5	0	0	1
[33]	16	9	0	5	1	1	0
[34]	16	8	0	5	1	1	0
[35]	14	14	1	3	1	1	1
[36]	18	6	1	2	1	1	0
[37]	24	8	1	4	1	1	0
[38]	6	11	0	4	1	0	0
[39]	11	7	0	4	1	1	0
[40]	6	6	1	5	0	0	1

(continued)

	Runners	Distance	Handicap	Class	Stake>5k	Odds>2	Win
[41]	13	8	0	6	0	1	1
[42]	12	7	0	4	1	1	1
[43]	4	6	0	2	1	0	1
[44]	12	7	1	3	1	0	1
[45]	3	14	0	3	1	0	1
[46]	12	7	0	4	1	0	1
[47]	18	8	1	6	0	1	0
[48]	24	7	1	4	1	1	0
[49]	6	6	0	4	1	0	1
[50]	9	9	0	1	1	1	1
[51]	9	12	1	2	1	1	0
[52]	15	8	1	3	1	1	0
[53]	8	8	0	4	1	0	1
[54]	22	11	1	5	1	1	0
[55]	9	7	0	5	0	0	1
[56]	9	6	1	4	0	1	1
[57]	4	10	1	2	1	0	1
[58]	4	9	0	1	1	0	1
[59]	6	9	0	4	0	0	1
[60]	17	7	1	5	0	1	1
[61]	17	5	0	4	0	1	1
[62]	17	8	0	5	0	1	1
[63]	7	6	0	4	1	1	1
[64]	22	6	0	6	0	1	1
[65]	19	12	1	4	0	1	1
[66]	24	6	1	7	0	1	0
[67]	10	11	1	5	0	1	1
[68]	9	12	1	2	1	1	0
[69]	3	8	0	2	1	0	1
[70]	12	7	0	4	0	1	1
[71]	13	6	0	1	1	1	0
[72]	16	7	1	3	1	1	0
[73]	12	5	1	3	1	1	0
[74]	8	6	0	4	0	0	1
[75]	7	9	0	4	0	0	1
[76]	9	5	1	3	1	1	0
[77]	10	10	1	4	0	1	1
[78]	7	12	0	1	1	1	1
[79]	12	12	1	3	1	1	0
[80]	15	9	1	4	0	1	0
[81]	18	8	0	6	0	1	1
[82]	9	8	0	4	0	1	1
[83]	12	7	0	6	0	0	1
[84]	18	8	1	4	1	1	0
[85]	18	12	1	4	1	1	0

(continued)

	Runners	Distance	Handicap	Class	Stake>5k	Odds>2	Win
[86]	14	6	0	4	0	0	0
[87]	24	5	1	6	0	1	1
[88]	11	7	0	6	0	1	0
[89]	14	5	0	5	0	1	1
[90]	18	8	1	6	0	1	0
[91]	16	8	1	3	1	1	0
[92]	12	12	0	4	0	0	1
[93]	15	13	1	5	0	1	0
[94]	16	6	0	4	0	1	0
[95]	18	6	1	4	0	1	1
[96]	7	5	1	4	0	1	0
[97]	18	5	1	6	0	1	1
[98]	8	9	0	4	0	0	1

The Races98.txt file contains data from 98 horse races. The aim is to create a neural network that can be trained and validated using the horse race data. After the neural network has been trained and validated it can be used to help predict the winner of other races. When you are done inspecting the content of the Races98.txt file, close it.

18.3.3 Create the Neural Network

18.3.3.1 Start the Application

Start the EasyNN-plus application by double-clicking on the EasyNN-plus icon that should be on your desktop. On the first window you see, click on the “Continue without ordering” button. When the “Did you know” window comes up, click on “Close” (you can always go through the tutorials later if you wish by looking under the Help menu option). If a window comes up about checking the FAQ, click on the “Check the online FAQ later” button. You should now be in the EasyNN-plus application, with it waiting for you to specify an input file.

18.3.3.2 Define Input and Output

Press the **New** toolbar button or use the **File > New** menu command to produce a blank grid for a new neural network.

An empty **Grid** with a vertical line, a horizontal line and an underline marker will appear. Now select **File > Import...** and the file selection dialog window will appear. Navigate to your copy of the **Races98.txt** file, select it, and hit the **Open** button.

The first **Import** dialog will appear with the **Tab** delimiter already checked under the “Columns” section. Each line in the Races98.txt file is numbered, so those numbers can be used as the row names. Under the “Example row names” section, select “**Use first word(s) on each line for row names**”. Under “Example row types”, select “Training”. Now press **OK**.

The second **Import** dialog will appear. Press the **Set names** button.

The third import dialog called **Input/Output Column** will appear. The settings in this dialog are based on the values being imported for the first row of data from the input file, but they should be checked for possible errors. In particular the **mode** and **type** may not always be correct. There are seven columns of data to be imported, and all of them are of type input except the last column, which is of type output (the last column is the "Win" column, which we are trying to predict). Column 0 (the first column) is the first one you will setup. The column is called **Runners** and the first row of data for this column has a value of 11. The column settings will be correct so press **OK**.

The second column will now be shown (denoted as column 1). This column is **Distance** with a value of 7. The settings are correct so press **OK**.

Column 2 (the third column) is **Handicap** with a value of 0. The mode should be changed to **Bool** (short for Boolean, meaning this column has values of False (indicated by 0) or True (indicated by 1)). Press **OK**. Column 3 (the fourth column) is **Class** and is another integer. The mode will change back to **Integer**. Press **OK**.

Columns 4 and 5 (the fifth and sixth columns) are boolean so the mode should be set to **Bool**. Press **OK** for both columns.

The last column (denoted as column 6, but this is the seventh column) is **Win** and is also a boolean so the mode should be set to **Bool**. *The* ***type*** *is not correct and needs to be changed to* ***Output.*** Then press **OK**.

A "Save file as" dialog will now be shown. You will use this dialog to save the data file in a format used by EasyNN-plus. Save the file in the same directory as the original Races98.txt input file. Note that the file is saved with a "tvq" extension.

The data will be imported and the grid columns will be set to the correct mode and type. Those columns that were set with type of "input" are used to predict the value of the column(s) set with type "output". In the current exercise we will be using the first six columns to predict the "Win" column. *Note that if the amount of data you are importing has more than 100 rows then you will be warned that this trial version of EasyNN-plus can only process the first 100 rows. The "Races98.txt" input file you are using only has 98 rows, so that won't be a problem for this exercise.*

The **Grid** of training data is now complete.

18.3.3.3 Growing and Training the Network

To create the neural network press the "**Grow new network**" toolbar button or use the **Action > New Network** menu command. This will open the **New Network** dialog. There will be an input node for each of the input columns, and there will be an output node for each output node. In this dialog window you will specify how many hidden layers there will be in the network, and how many nodes will be allowed in each hidden layer. For the simple network used in this exercise we will use one hidden layer, and we will keep the default values for the min and max number of nodes allowed in that layer. Under the "Hidden layers" section, check **Grow layer number 1** and press **OK**. If you get a "Generating new network will reset learning." warning message, answer **Yes**. The neural network will be produced from the data you imported into the Grid.

In the window saying "Network created", note that you are told how many nodes are actually being used in the hidden layer(s) of the network. On the "Network created" window, click on the **Yes** button under the question "Do you want to set the controls?". This will open the **Controls** dialog.

On the Controls dialog box we will leave most of the settings alone. Look under the "Learning" section and check **Optimize** for both Learning Rate and Momentum. [*Learning Rate can be changed to any value from 0.1 to 10, representing the size of weight changes during learning. Very low values will result in slow learning and values above 1.5 will often result in erratic learning or oscillations. Checking "Optimize" allows EasyNN-plus to determine a reasonable value for Learning Rate by running a few learning cycles with different Learning Rate values. Momentum is used to prevent the neural network from getting "stuck" at a local minimum or maximum…leave momentum at its default value.*] We now need to indicate how many of the input rows will be kept aside for testing the trained network. We have 98 rows of total input in the "Races98.txt" file, so let's (arbitrarily) keep 30 of those rows aside to be used later for testing purposes...go to the "Validating" section and enter 30 where it says "Select __ examples at random from the...".

We next have to establish a stopping criteria so that the software knows when to stop training the neural network. There is no point in using more than 1000 cycles to train the simple network in this exercise, so look under the "Stops" section and put a check by "Stop on __ cycles" and enter a value of 1000.

Press OK. Answer **Yes** if you get the "Optimizing controls will reset learning." warning message. The controls will be set and the neural network will be ready to learn. *If you get a warning message in a "Validating Problems" window, click on the "Change All Out of Range Validating Rows to Training".*

A summary of the control settings and what is going to happen next will appear. Answer **Yes** to the question "Do you want *filename* to start learning?" (the *filename* will be whatever you named it near the end of step (3) earlier). Next the AutoSave dialog window will open...you don't need to change anything on that window so just press OK. This will start the training process for the neural network. [*Note: if for some reason you are not asked if you want to start the learning process, you can start it manually by using the **Action > Start Learning** menu option.*] Learning and validating will run for 1000 cycles.

18.3.3.4 Results of Training

Training for this small amount of data should complete almost instantly...if the status bar (the bar at the bottom of the main window) says "Fixed cycles stop: 1000 cycles", then training has completed. You can now use the **View > Information** menu command to see the results. Look under the "General" section for the line labeled "Validating results:"... this value tells you how accurately the neural net was able to predict the winner in the 30 test rows. If everything has worked as expected the neural network will have correctly predicted the results of at least 60% of the races.

18.3.3.5 Using the Neural Network to Make a Prediction

If the data grid is not in view, use the **View > Grid** menu command to bring the Grid to the front. Make sure that the first row of data is selected (just click on the row number [1]). Then use the **Insert > Querying Example Row** menu command to add a new row to the grid where you can enter values for a new horse. The "Example presets" window will come up...you should be sure the "set all values in row to 0" radio button is selected, then hit the "Set all row values" button. This should cause an empty row to appear at the top of the data grid, with 0 (or False for Boolean columns) for each value. You can use this query row to enter values for the input columns for a new horse, and then you can view what the prediction is in the output column. Use the following values in the query row: Runners = 10, Distance = 15, Handicap = False, Class = 3, Stake > 5 k = True, Odds > 2 = False. The resulting prediction should be that the horse will win.

18.4 Conclusion

Neural networks are flexible tools that can be used for a wide variety of decision tasks, such as classification, prediction, and pattern recognition. Neural networks do not require advanced knowledge of statistical techniques, are able to model complex nonlinear relationships, and can detect multiple interactions between predictor variables. This chapter provides an overview of an implementation of a simple neural network. The same instructions could be used to create neural networks in a variety of domains.

Biographies

Brian R. Huguenard is an Associate Professor in the Department of Decision Sciences and Management at Tennessee Technological University. He received his Ph.D. in Industrial Administration (Information Systems) from the Tepper School of Business, Carnegie Mellon University. He has also taught at the University of Notre Dame. He currently teaches programming, business analytics, and decision support courses. His research interests include human-computer interaction, real-time dynamic decision making, and the development of information systems to minimize end-user error. Prior to his academic career he worked as an Operations Research analyst for FedEx.

Deborah J. Ballou is an Associate Professor in the Department of Decision Sciences and Management at Tennessee Tech University. She received her Ph.D. in Industrial Administration (Information Systems) from the Tepper School of Business, Carnegie Mellon University. She has also taught at the University of Notre Dame and Concordia University. She currently teaches programming, advanced business analytics, and healthcare analytics courses. Her research interests include information technology professionals' training needs and career paths, healthcare informatics and analytics, and analytics-based decision support. She is co-leader of a new Data and Decision Sciences Collaborative sponsored by the Colleges of Business and Engineering at Tennessee Tech.

References

Gurney K (1997) An introduction to neural networks. Routledge, London

Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, Upper Saddle River

Lawrence J (1994) Introduction to neural networks. California Scientific Software Press, San Francisco

Chapter 19
An Examination of ERP Learning Outcomes: A Text Mining Approach

Mary M. Dunaway

Abstract Today's business colleges are attempting to meet the industry demand by developing marketable ERP (Enterprise Resource Planning) skills and delivering exposure to the realities of modern business into the curricula. Role adaptions in real-world settings such as ERP systems use can enhance students' ability to learn conceptual knowledge for practical application. The situated learning theory capitalizes on a specified context where the context extensively impacts learning. Education data text mining is emerging to produce new possibilities for gathering, analyzing, and presenting student learning outcomes. This chapter aims to reveal ERP learning patterns and themes as evidence of knowledge transfer in ERP role adaptions. The results demonstrate amplified learning through role play in a simulated ERP learning environment.

Keywords Text mining • Situated learning theory • Education data mining • Pedagogy

19.1 Introduction

The use of analytics in higher education is a relatively new area of practice and research. Corporate business practices have led the way in data mining across most all organizations to leverage competitive and profit-driven strategies (Chen et al. 2012; Holsapple et al. 2014). Data mining is gaining momentum in higher education, which is now using a variety of applications, most notably in enrollment, learning patterns, personalization, and threaded discussion analysis (Norris et al. 2008; Baepler and Murdoch 2010; Ravishanker 2011; Edgington 2011). By discovering hidden relationships, patterns, interdependencies, and correlating raw/unstructured

M.M. Dunaway, Ph.D. (✉)
22 N Lake Dr., #A1, Hamden, CT 06517, USA

School of Professional and Continuing Studies, University of Virginia,
Charlottesville, VA, USA
e-mail: marydunaway99@gmail.com

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_19

data, data mining is beginning to facilitate not only higher education institutional decision making but also student learning outcomes and performance.

Learning analytics does not have the same goals in the business streams as learning academic streams (e.g. Campbell et al. 2007; Baepler and Murdoch 2010; Van Barneveld et al. 2012; Ferguson 2012). The focus of learning analytics primarily targets two areas—learning effectiveness and learning operational excellence (Luan 2004; Zhang et al. 2010; Minami and Ohura 2013). The latter refers to the metrics that provide evidence of how the learning aligns with and meets the goals of the higher education institution. Learning analytics in the academic domain is focused on the learner, gathering data from course management and student course data to manage student success and performance outcomes (Van Barneveld et al. 2012). Different from other types of analytics by virtue of the fact that learning analytics is focused specifically on students and their learning behaviors (Baepler and Murdoch 2010).

Data mining techniques can enable institutions of higher education to rethink and improve students' learning experiences. Faculty can streamline their teaching and learning processes to extract and analyze students' learning outcomes and behaviors (e.g. Campbell et al. 2007; Baepler and Murdoch 2010; Edgington 2011). Data mining techniques can provide greater insights about student learning as their learning experiences unfold (Hung and Crooks 2009; García et al. 2011). Text mining, a type of data mining technique, provides a combination of explicit knowledge, analytical skill, and domain knowledge to uncover hidden trends and patterns from text. Text is a type of unstructured data where row and column attributes are not distinctive characteristics. Text mining involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, visualization, and predictive analytics.

AACSB (2015 August 1) states that the curricula facilitate and encourage active student engagement in learning. Most importantly, "in addition to time on task related to readings, course participation, knowledge development, projects, and assignments, students engage in experiential and active learning designed to improve skills and the application of knowledge in practice is expected" (AACSB 2015). Today's business colleges are attempting to meet the demands by developing marketable Enterprise Resource Planning (ERP) programs and delivering exposure to the realities of modern Information Systems (IS) into the curricula (Léger 2006; Seethamraju 2011; Cronan et al. 2011; Monk and Lycett 2014). Situation Learning theory posits the idea that much of what is learned is specific to the situation in which it is learned (e.g. Lave 1988; Lave and Wenger 1991). Particularly, important has been situated learning's emphasis on the mismatch between typical learning situations and "real world" situations such as the workplace in environments where complex IS are employed.

The aim of this research is to investigate the application of text mining to understand student learning outcomes gained in an ERP Fundamentals course. Specifically, this paper combines role play and simulations as forms of experiential learning. Students take on different business functional roles, interact, and participate in a diverse and complex learning settings where a simulated, real-work ERP system is engaged. Using the Situated Learning theory as a lens to examine student behaviors and applying a text mining approach can help to uncover and evaluate learner centric outcomes. These results can support faculty and student feedback for improved ERP teaching and student learning processes.

19.1.1 ERP Course Overview

ERP courses in the Information System (IS) curricula is at the forefront of experiential and active learning (Cronan and Douglas 2012; Cronan et al. 2011; Hepner and Dickson 2013). ERP courses at major universities provide the IS educational community with a goldmine of data about students' learning characteristics, learning outcomes, behaviors, attitudes, and use patterns (Léger 2006; Cronan and Douglas 2012; Cronan et al. 2011; Léger et al. 2011). Many IS courses utilize simulation games such as ERPSIM (Léger 2006) where students utilize a real-world SAP IS as the learning environment. Students learn real-time business decision making, dynamic functional processes, and Business Intelligence (BI) skills. The continuous availability of real-time SAP provides the students with large amounts of data where accelerated business decision experience can be gained.

19.2 Background and Theory

19.2.1 ERP Simulation and Learning

Role play and simulations are forms of experiential learning (Russell and Shepherd 2010). Learners take on different roles, assuming a profile of a character to interact and participate in diverse and complex learning settings. Computer simulation of business in higher education and experiential learning theory accelerate the use of simulation games in business education (Keys and Wolfe 1990). Good-quality learning design provides opportunities for situated and authentic learning.

The ERP Simulation (ERPSIM) game and the team interaction provide the environment for learning to transpire (Léger 2006). ERPSIM consists of several games (Distribution, Manufacturing, and Logistics) that are used for learning. Students learn cross-functional decision making to maximize firm profit and integrate business processes to achieve desire outcomes (Boudreau 2003; Kang and Santhanam 2003–2004). The ERPSIM game and the functional role play adaptions provide the specific learning context. According to Léger "as the learner's knowledge and skills increase, the role and status of the learner as a member of a community gradually evolves from that of novice or apprentice to expert" (Léger 2006, p. 39). Teams compete against each other during the simulation game. Moreover, students typically have had no experience with functional integrative ERP software.

19.2.2 Situational Learning Theory

An extension of constructivist theory (Land and Hannafin 2000), is the concept of situated learning. It is when all learning takes place in a specific context and the context significantly impacts learning (Alessi and Trollip 2001). When learning is removed from its context, the value of the knowledge and its relevance to the learner is weaken (Duffy and Cunningham 1996). Learning begins as the individual participates in environments and also engages with the communities formed within the environments (Sadler 2009). Hence, role playing or scenario-based learning activities are enriched by the situation or context of its environment (McLellan 1986).

Many scholars suggest greater emphasis on the relationship between what is learned in the classroom and what is needed outside the classroom (Lunce 2006; Léger et al. 2011; Karia et al. 2014). Situated learning involves a practice-based approach which integrates classroom learning and real-world situations where the environment is dynamic (Duffy and Cunningham 1996). A consequence of the learner is to recognize the practical utility of knowledge conveyed as well as the need to use it to interpret, analyze and to solve real-world problems (Chen and Hung 2002). Also, the collaborative process in which the student interacts with other members of a "community of practice" (Henning 1998). Duffy and Cunningham (1996) tends to be peer-based rather than the more formal teacher-student relationship of the classroom. The Situated Learning theory is important to the ERP role adaption offering support for the mechanisms that enable ERP learning. Situated learning reflects the thematic interactive activities, cognitive engagement, participation, and group social structure within the learning environment (Goel et al. 2010).

Benefits can be gained when combining situated learning and role adaptions. The benefits deliver content and interactions to facilitate learning. Different from a traditional learning environment, this learning design enables students to experience learning authentically.

Benefits gained from the role adaption:

- Decision-making and interpersonal communication skills
- Enable a professional environment decision-making
- Scenarios can be scaffolded, gradually increasing in complexity to ensure that students reach a sufficient level of competence
- Students' gain the ability to work under pressure and with others, including providing opportunities for cross-functional learning

Benefits gained from the situated learning context:

- Provides is a form of authentic context where exposure to active, experiential, reflective and contextual learning has a direct relevance to the educational experience and future profession
- Simulations enable instant feedback to students
- Effective means of providing content knowledge

19.2.3 Importance of ERP Learning

The increased importance of ERP and its pedagogical value to demonstrate business process integration, functional role play, and business decision making already have started to reengineer curricula (Magal and Word 2009). According to Chen et al. (2011), ERP system learning involves complex knowledge domains requiring a holistic curricula perspective to enhance student motivation and interest. Advances in pedagogical approaches that emphasize learning approaches such as active or learn-by-doing experience provide greater benefits to students than solely lecture-based (Chen et al. 2011). Prior research has shown learning limited to a lecture-based approach can make students passive learners (Bok 1986).

19.2.4 Role Adaptions in ERPSIM

The business process drives the functional role identity for the activities and tasks to be performed in the ERPSIM game. Typical business processes within an ERPSIM learning entail the planning, procurement, production, and sales processes. Each business process area is tied tasks similar to real-world job responsibilities within a company. Within each business process operational transactions, reporting, and BI inform decision making for the functional role. For example, within the Sales process, hands-on basic handling of market expense allocation, price list changes, and sales order reports are positioned for the Sales Analyst or Product Marketing role. Whereas in a Production process, finished goods forecast, materials requirement planning, and purchase supplier interactions are performed**.**

19.3 Research Methodology

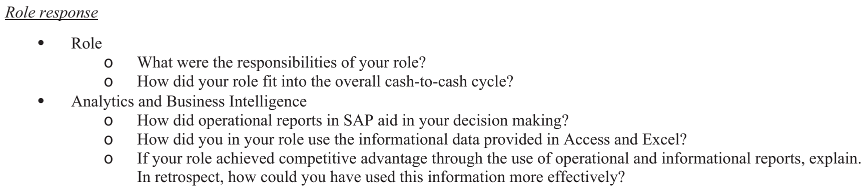
19.3.1 Background/Classroom Setting

Students were enrolled in an Introduction to ERP course where learning enterprise technology software SAP is the focus. The course consists of a traditional classroom setting with lectures, videos, team project, and hands-on lab assignments in addition to ERP simulation game. The ERPSIM game (Léger 2006) is developed by faculty at HEC Montreal and a type of simulation game where students operate a business and make decisions for a make-to-stock cereal product similar to what is done in today's large companies. The simulation game involves a cash-to-cash cycle consisting of the procurement, production, and sales processes. Students are organized into teams to manage their own fictitious manufacturing company to produce a cereal product. The ERPSIM is executed in simulated 30-day fast track time intervals representing a fiscal year quarter.

Teams of four students perform operations and tasks which requires them to interact with suppliers and customers by sending and receiving purchase orders, delivering products, and completing the entire cash-to-cash cycle. The simulation game, ERPSIM automates (1) the sales process where each firm receives in a large number of orders every minute, (2) the procurement process purchases raw materials, and (3) the production process utilize machine capacity, inventory, and warehouse functions. These operational functions are performed directly in the ERPSIM real-time. Many pre-defined SAP reports are available to help students evaluate their company's profit and operations of the business. Several key business decisions which student teams are required to make during the simulation are:

- Product formulation (raw materials and packaging)
- Target sales by region and market segment
- Product pricing adjusted throughout the business operations
- Sales forecasting to predict sales volumes by product for production planning
- Manufacturing resource planning and production
- Investment in production efficiencies to reduce cost/time delays
- Advertising to regional markets
- Debt management which consists of loan repayment

Once all of the ERP Simulation quarters have been completed, the teams must provide an overall reflection report of their simulation experience. The role response guidelines were written by the instructors and developed to emphasize the integration of the student role adaption and their use of the analytic and BI capabilities from the ERPSIM game for decision making. There is a section of the report where a role response is written by each team member. The role response includes the responsibilities of the role and how the role fits into the overall cash-to-cash cycle. Figure 19.1 shows the role response written guidelines given to students. Also, the Analytics and BI section of the report was to be written from a role response point of view, specifically how the operational reports in the ERP system were used to help with decision making, how the student's role utilized the information obtained through the metrics, and finally how the role contributed to the overall team's competitive advantage.



Role response

- Role
  - What were the responsibilities of your role?
  - How did your role fit into the overall cash-to-cash cycle?
- Analytics and Business Intelligence
  - How did operational reports in SAP aid in your decision making?
  - How did you in your role use the informational data provided in Access and Excel?
  - If your role achieved competitive advantage through the use of operational and informational reports, explain. In retrospect, how could you have used this information more effectively?

Fig. 19.1 Role response write-up guidelines

19.3.2 Situated Learning Adaption

Before the simulation game begins, each team makes a decision on what roles should be represented during the game and who will perform the activities and tasks accompanying the role. Teams usually follow a similar pattern for role division of labor across the simulation game. Common roles selected are procurement manager, production manager, business analyst, pricing specialist or analyst, operations manager, chief financial officer, financial analyst, and marketing director. Figure 19.2 describes the ERP role adaption examples from a previous class.

There are several objectives to be accomplished by functional role play adaption. Students will have an understanding of:

- How roles collectively work together is integrated in the business processes performed while using an ERP system.
- How the individual roles of which they represent contribute to business processes.
- The importance of selecting and using appropriate BI metrics based on their role to gain competitive advantage for the team.

19.3.3 ERP Role Play Strategy

There are three main areas of pedagogy of which roles can provide value infused by the role strategies. They are community learning, engagement encouragement, and motivation and interest improvement. Table 19.1 shows the steps of how the functional role play is imparted in the students' learning.

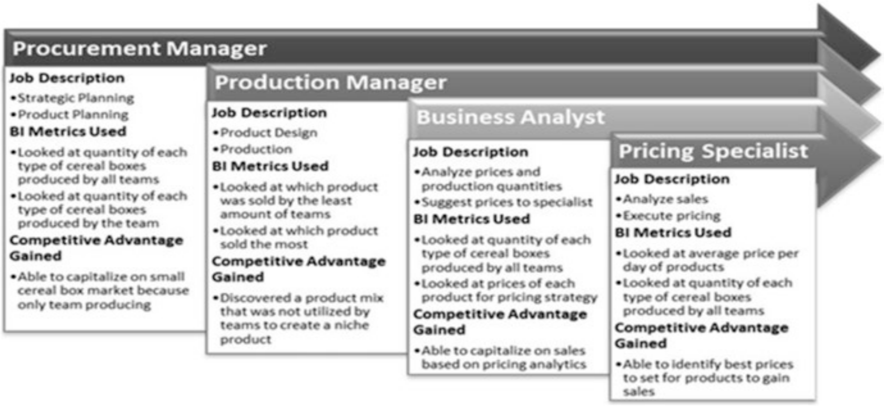

Procurement Manager
Job Description
•Strategic Planning
•Product Planning
BI Metrics Used
•Looked at quantity of each type of cereal boxes produced by all teams
•Looked at quantity of each type of cereal boxes produced by the team
Competitive Advantage Gained
•Able to capitalize on small cereal box market because only team producing
Production Manager
Job Description
•Product Design
•Production
BI Metrics Used
•Looked at which product was sold by the least amount of teams
•Looked at which product sold the most
Competitive Advantage Gained
•Discovered a product mix that was not utilized by teams to create a niche product
Business Analyst
Job Description
•Analyze prices and production quantities
•Suggest prices to specialist
BI Metrics Used
•Looked at quantity of each type of cereal boxes produced by all teams
•Looked at prices of each product for pricing strategy
Competitive Advantage Gained
•Able to capitalize on sales based on pricing analytics
Pricing Specialist
Job Description
•Analyze sales
•Execute pricing
BI Metrics Used
•Looked at average price per day of products
•Looked at quantity of each type of cereal boxes produced by all teams
Competitive Advantage Gained
•Able to identify best prices to set for products to gain sales

Fig. 19.2 ERP role adaption example from a previous class

Table 19.1 Pedagogical approach using role play strategy

Pedagogical approach	Steps	Infusion of role strategies
Community of learning	Team building	As students talk on their role orientation, team building occurs as each person will play their part and work together to complete the cash-to-cash cycle. Each person will depend on each other in their role to be successful
	Opportunities for interaction	As the quarters are played, the students communicate with each other to work through the business processes. The quarterly game BI metrics information will be conveyed within the team to learn from this data and to create competitive advantage
Engagement encouragement	Gain and sustain attention	The contribution by the collective roles, informed by the BI analytics and role performance provides the opportunity for the student to focus attention on the tasks at hand
Motivation and interest improvement	Establish relevance	When functional role orientation is introduced in the class, the importance of each student's role is emphasized. How the collective roles work together to accomplish the business processes and how they are applicable to the overall business strategy are infused in the discussion. Industry examples are used to relate class activity to the real world

19.4 Results

19.4.1 Qualitative Analysis of Student Role Responses

Teaching for transfer is one of the seldom-specified but most important goals in education. We want students to gain knowledge and skills that they can apply both in and outside of the university setting immediately and in the future. Transfer of learning is often done without conscious thought. The role responses written by the students demonstrate how contextualized learning in a real-world setting such as ERPSIM reinforces learning assurance, real industry roles, and business process knowledge. In Table 19.2 are excerpt responses of the role responses written by several students.

19.4.2 Quantitate Content Analysis of Student Role Responses

The data in this paper reflect student participation during the administration of the ERPSIM Manufacturing game. The sample (N = 62) was collected from students' final written projects spanning three semesters. The student teams were randomly selected and student roles were self-selected. Each team was required to write a

Table 19.2 Excerpt of student role responses

Student	Role response
Student 1	As Chief Operations Officer, my responsibilities dealt with production planning. I was tasked with forecasting, procuring, and producing the amount of product in a manner that was cost friendly to the company. My role played a vital part in the cash to cash cycle of our company because we had the potential to incur substantial storage cost for excess inventory and raw materials
Student 2	My duties as Chief Financial Officer primarily focused upon investment decisions. I managed decisions regarding gauging the cost vs. gain of investments in production capacity, increased efficiency through setup time reduction, budgeting marketing expenses, as deciding on whether to change the Bill of Materials on our products. My role was crucial to the cash to cash cycle in that every investing opportunity taken was completely focused on whether it would yield more of a profit than it cost the company. Through the simulation lesson were learned in regards to where to spend and not to spend, and when was the right time to stop because additional cost would not produce a worthwhile profit. IT was awesome for all of us to see how our roles were so intertwined with each other and how every step directly impacted another
Student 3	I had a very unique role throughout this whole process. I was Data Coordinator for this assignment, as well as I also was in charge of product design and BOM modification. My role was to use data from previous rounds to gauge how much our product should be modified to ensure that our product was differentiated without incurring too much cost. I created this metric to ensure that we were competitively differentiating our products and just as importantly, ourselves from the competition. Our team learned what the numbers were really telling us and that knowledge will translate to our future careers and will benefit us in the workplace

reflection paper as instructed in the ERP Simulation Guidelines. Within the guidelines, each team member was responsible for writing a role response. The response relates to each team member's role and application of the Analytics and BI.

KH Coder (Higuchi 2015) is an open source software for quantitative content analysis or text mining. Currently, there are almost 500 scholarly publications using this software (Wikipedia 2015). Several recent studies have used KH Coder to perform Text Mining in an education and learning context (Tsubakimoto 2011; Ishii et al. 2013; Minami and Ohura 2013). Also, KH Coder has been used for computational linguistics. Three words were removed from the sample that did not provide additional meaning for the analysis.

The words excluded: were, and, the, and of. A text mining analysis using the KH Coder (Higuchi 2015) software was performed across the students' role response data to uncover patterns and learning themes. Table 19.3 results show the word frequency distribution from the student data to describe their role during the ERP simulation experience. The words with the highest frequency used to describe the student role were Manager, Officer, and Chief. The business process areas related to the roles were Marketing, Pricing, Financial, and Production. The areas and roles align with the ERP system knowledge transfer where students learn Sales, Procurement, and Planning business processes. The results confirm known differentiated roles associated with real-world ERP system use.

Table 19.3 Results of word frequency in the student role response

Noun		Proper noun			
Word	Frequency	Word	Frequency	Word	Frequency
Pricing	8	Manager	12	Acquisitions	1
Production	7	Chief	9	Analytics	1
Coordinator	3	Marketing	9	Chain	1
Price	3	Officer	8	Executive	1
Sale	3	Financial	6	Improvements	1
Planner	2	Analyst	6	Materials	1
Strategist	2	Business	3	Optimizer	1
Activity	1	Controller	3	President	1
Cash	1	Director	3	Process	1
Cost	1	Product	3	Purchasing	1
Design	1	Inventory	2	Raw	1
Forecast	1	Material	2	Resource	1
Forecasting	1	Operating	2	Strategist	1
Market	1	Operations	2	Supply	1
Procurement	1	Procurement	2	Tracker	1
Specialist	1	Accounting	1		

Further analysis was performed using the Co-occurrence networks technique on the student role response data. This technique is generally used to provide a graphic visualization of potential relationships between people, organizations, concepts or other entities represented within text data. The generation and visualization of co-occurrence networks has become practical with electronically stored text amenable to text mining. Co-occurrence networks are the collective interconnection of terms based on their paired presence within a specified unit of text. Networks are generated by connecting pairs of terms using a set of criteria defining co-occurrence. For example, terms A and B may be said to "co-occur" if they both appear in a particular article. Another article may contain terms B and C. Linking A to B and B to C creates a co-occurrence network of these three terms.

Stronger co-occurrence are represented by thicker connector lines and weaker co-occurrences by less thick lines. Words close to each other do not always mean they have a strong co-occurrence. Instead, whether or not the words are connected with lines (edges) is significant. Although words may be plotted close to each other, if they are not connected with lines (edges), there is no strong co-occurrence. Additionally, larger nodes represent higher frequency words.

Figure 19.3 show the learning themes that emerged from the Co-occurrence network. The results show the emergence of key business processes and the interconnected role responses. Accounting Manager, Marketing Strategist, Production Manager, and Chief Officer were high frequency roles. The relationships connect pairs of role responses to define the co-occurrence. The role responses interconnectedness relate to the tasks engaged when using the ERP system to support the various business processes. In particular, the role of Chief is associated with operations and analytics tasks. Managers are associated with pricing, production,

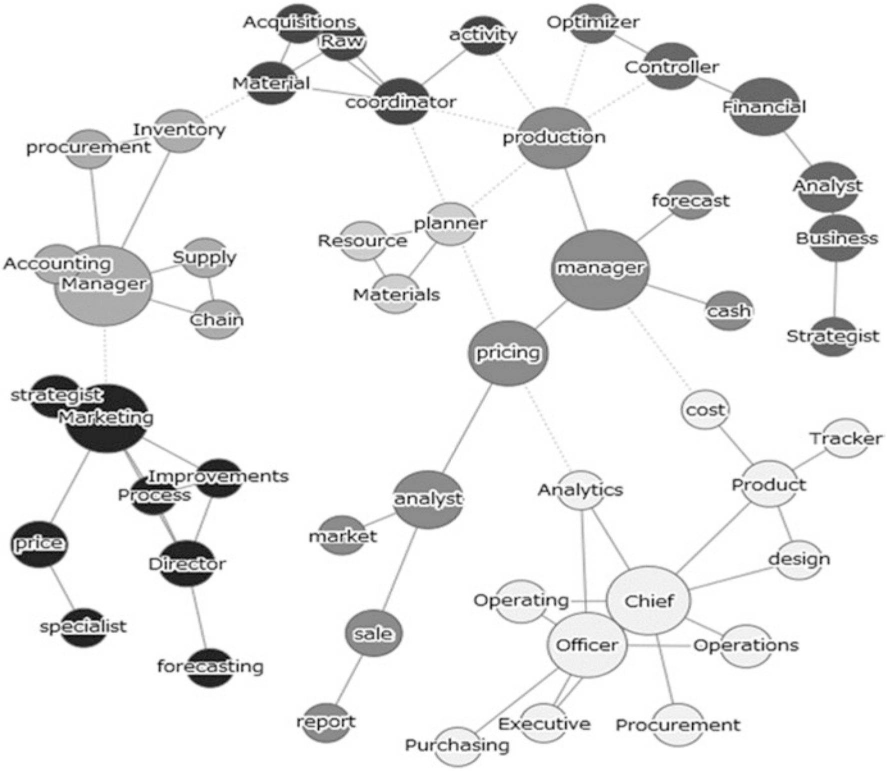


Fig. 19.3 ERP role adaption co-occurrence network

forecast, and financial related tasks. Managers are related to material planning and financial control. Marketing roles were interconnected with pricing, forecasting, and process improvements. Lastly, Accounting Manager roles were interconnected with Supply Chain, Procurement, and Inventory tasks. Stronger interactions exist mainly in Process Improvements, Procurement, and Accounting Manager roles.

Table 19.4 shows the overall ERP Role adaptions resulted from the student ERP knowledge transfer. The word cluster analysis results show the emerged ERP roles from the situated learning context using the ERP simulation and the related score. Forty-four word clusters emerged as role adaptions from the data. The highest three role adaptations were marketing managers, production managers, and pricing managers. Each word cluster is scored, and normally, highly scored clusters are reliable (Higuchi 2015).

The word cluster analysis uses a TermExtract transformation. The transformation is an automatic technical term (keyword) extraction system published by the Nakagawa Laboratory of the Digital Library Division, Information Technology Center of The University of Tokyo. The Term Extraction transformation can extract nouns only, noun phrases only, or both nouns and noun phases. Because the process in KH Coder is automatic, unintended word combinations may occur. Unintended words or combinations can be managed as nonexistent in the statistical analysis.

Table 19.4 Results of the role adaption word cluster analysis

Cluster	Score	Cluster	Score	Cluster	Score	Cluster	Score
Marketing manager	25.49	Director of marketing	3.57	Chief operations officer	2.85	Chief product design	2.33
Production manager	17.89	Marketing director	3.57	Pricing coordinator	2.71	Coordinator of raw material acquisitions	2.14
Pricing manager	15.49	Operations manager	3.50	Chief executive officer	2.67	Market analyst	2.11
Financial analyst	6.88	Chief financial officer	3.29	Cash manager	2.66	Chief analytics	2.11
Pricing analyst	6.16	Procurement manager	3.16	Accounting manager	2.66	Purchasing officer	2.06
Chief operating officer	6.11	Product cost manager	3.05	Forecast manager	2.66	Product tracker	2.00
Financial controller	5.09	Marketing strategist	3.00	Production planner	2.63	Forecasting director	1.86
Business analyst	5.03	Pricing officer	3.00	Chief procurement	2.51	Business strategist	1.86
Sales analyst	4.68	Inventory material manager	2.96	Supply chain manager	2.42	Materials resource planner	1.70
Financial manager	4.33	Inventory manager	2.94	Coordinator of production activities	2.40	Price specialist	1.57
Chief marketing officer	3.67	Price manager	2.94	Production optimizer	2.38	Process improvements	1.41

19.5 Conclusions and Limitations

The findings observed in this study closely align with the transfer of learning in ERP learning outcomes (Seethamraju 2011; Cronan and Douglas 2013; Monk and Lycett 2014). Role play in the ERP situated learning context enriches the experience for students to develop attributes and cognitive practices for real-world ERP roles. This type experience can help students transition into the workforce after graduation to anticipate the business acumen, value, and role responsibilities in an ERP profession. These results support prior literature where the aspects of the knowledge transfer enhancing activities lead to recall and application of what has been learned (Thayer and Teachout 1995). Teaching for transfer is one of the seldom-specified but most important goals in education. We want students to gain knowledge and skills that they can use both in school and outside of school, immediately and in the future. Transfer of learning is commonplace and often done without conscious thought.

The pedagogical approach as described provides business process learning that emphasizes the functional role play. The situated learning experience enhances students' knowledge capabilities for ERP, BI, and business process concepts beyond the lecture only. The use of functional role play and BI application are a strategies to improve learning and provide a deeper insight of business process knowledge. This approach is synergistic and reinforces the students' learning. This study supports (McLellan 1986) research findings to affirm that the hands-on experience of ERP systems indeed help students understand business process. Moreover, this examination demonstrated valuable student knowledge from the ERP course and that can impact practical application to business decisions in their future career roles (Cronan et al. 2011).

This study has a few limitations that need to be addressed in future research. The sample size is small (N = 62). Though a recommended sample size for text mining was not found in the literature, the ROI of text analytics increases exponentially with the size of the data. Thus, a larger sample may potentially reveal patterns and relationships not revealed. Another limitation is the teaching practice as described is specific to a southern university and its curriculum. While this paper asserts the benefits from the functional role play, other factors may play a direct or indirect role in the learning outcomes. Further research will utilize additional quantitative and qualitative methods to examine empirically the functional role play, attitudes, and team behaviors on learning outcomes.

Biography

Mary M. Dunaway, Ph.D. is the Director of Data Science Programs and Assistant Professor at the University of Virginia in the College of Continuing and Professional Studies. As a rising scholar, she has successfully published several journal articles, book chapters, and conference proceedings. Also, Dr. Dunaway is leading the effort to develop and launch an Applied Data Analytics graduate certificate program. She is a dynamic STEM academic who is a sought-out speaker and panelist for numerous conferences/workshops sharing her expertise in Information Systems and Data Science.

References

AACSB (2015) Business accreditation standards. http://www.aacsb.edu/accreditation/business/standards/aol/aol_approaches.asp

Alessi SM, Trollip SR (2001) Multimedia for learning: methods and development, 3rd edn. Allyn and Bacon, Boston, MA

Baepler P, Murdoch CJ (2010) Academic analytics and data mining in higher education. Int J Scholarship Teach Learn 4(2):1–9

Bok DC (1986) Higher learning. Harvard University Press, Cambridge, MA

Boudreau M (2003) Learning to use ERP technology: a causal model. In: Proceedings of the 36th annual Hawaii international conference on system sciences, vol 8, pp 235–244

Campbell JP, DeBlois PB, Oblinger DG (2007) Academic analytics: a new tool for a new era. Educ Rev 42(4):41–57

Chen D, Hung D (2002) Personalised knowledge representations: the missing half of online discussions. Br J Educ Technol 33(3):279–290

Chen K, Razi M, Rienzo T (2011) Intrinsic factors for continued ERP learning: a precursor to interdisciplinary ERP curriculum design. Decis Sci J Innov Educ 9(2):149–176

Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188

Cronan TP, Douglas DE (2012) A student ERP simulation game: a longitudinal study. J Comput Inf Syst 53(1):3–13

Cronan TP, Douglas DE, Alnuaimi O, Schmidt PJ (2011) Decision making in an integrated business process context: learning using an ERP simulation game. Decis Sci J Innov Educ 9(2):227–234

Cronan TP, Douglas DE (2013) Assessing ERP learning (management, business process, and skills) and attitudes," Journal of Organizational and End User Computing 25(2):59–74

Duffy T, Cunningham D (1996) Constructivism: implications for the design and delivery of instruction. In: Jonassen DH (ed) Handbook of research for educational communications and technology. Simon and Schuster, New York, pp 170–198

Edgington TM (2011) Introducing text analytics as a graduate business school course. J Inf Technol Educ 10:207–234

Ferguson R (2012) Learning analytics: drivers, developments and challenges. Int J Technol Enhanc Learn 4(5/6):304–317

García E, Romero C, Ventura S, de Castro C (2011) A collaborative educational association rule mining tool. Internet High Educ 14(2):77–88

Goel L, Johnson N, Junglas I, Ives B (2010) Situated learning: conceptualization and measurement. Decis Sci J Innov Educ 8(1):215–240

Henning PH (1998) Ways of learning: an ethnographic study of the work and situated learning of a group of refrigeration service technicians. J Contemp Ethnogr 27(1):85–136

Hepner M, Dickson W (2013) The value of ERP curriculum integration: perspectives from the research. J Inf Syst Educ 24(4):309–326

Higuchi K (2015) KH coder (version 2.0) [software]. http://khc.sourceforge.net/en/

Holsapple C, Lee-Post A, Pakath R (2014) A unified foundation for business analytics. Decis Support Syst 64:130–141

Hung JL, Crooks SM (2009) Examining online learning patterns with data mining techniques in peer-moderated and teacher-moderated courses. J Educ Comput Res 40(2):183–210

Ishii N, Suzuki Y, Fujii T, Fujiyoshi H (2013) Development and evaluation of question templates for text mining. Recent Prog Data Eng Internet Technol 156:469–474

Kang D, Santhanam R (2003–2004) A longitudinal field study of training practices in a collaborative application environment. J Inf Syst Educ 17(4):441–447

Karia M, Bathula H, Abbott M (2014) An experiential learning approach to teaching business planning: connecting students to the real world. In: Li M, Zhao Y (eds) Exploring learning and teaching in higher education. Springer, Berlin, pp 123–144

Keys B, Wolfe J (1990) The role of management games and simulation in education and research. J Manag 16(2):307–336

Land SM, Hannafin MJ (2000) Student-centered learning environments. In: Jonassen DH, Land SM (eds) Theoretical foundations of learning environments. Lawrence Erlbaum, Mahwah, NJ, pp 1–23

Lave J (1988) Cognition in practice: mind, mathematics and culture in everyday life. Cambridge University Press, New York, NY

Lave J, Wenger E (1991) Situated learning: legitimate peripheral participation. Cambridge University Press, New York, NY

Léger P (2006) Using a simulation game approach to teach enterprise resource planning concepts. J Inf Syst Educ 17(4):441–447

Léger P, Charland P, Feldstein HD, Robert J, Babin G, Lyle D (2011) Business simulation training in information technology education: guidelines for new approaches in IT training trends in higher education: from situational theory to simulation games. J Inf Technol Educ Res 10(1):39–53

Luan J (2004) Data mining applications in higher education, SPSS executive report. SPSS Inc. http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf

Lunce LM (2006) Simulations: bringing the benefits of situated learning to the traditional classroom. J Appl Educ Technol 3(1):37–45

Magal SR, Word J (2009) Essentials of business processes and information systems. Wiley, Hoboken, NJ

McLellan H (1986) Situated learning: multiple perspectives. In: McLellan H (ed) Situated learning perspectives. Educational Technology, Englewood Cliffs, NJ, pp 5–18

Minami T, Ohura Y (2015) How student's attitude influences on learning achievement? An analysis of attitude – representing words appearing, looking-back evaluation texts. Int J Database Theory App 8(2):129–144

Monk E, Lycett M (2014) Measuring business process learning with enterprise resource planning systems to improve the value of education. Educ Inf Technol 21:1–22

Norris D, Baer L, Leonard J, Pugliese L, Lefrere P (2008) Action analytics. Educ Rev 43(1):42–67

Ravishanker R (2011) Doing academic analytics right: intelligent answers to simple questions. Research bulletin 2, EDUCAUSE review. http://net.educause.edu/ir/library/pdf/ERB1102.pdf

Russell C, Shepherd J (2010) Online role-play environments for higher education. Br J Educ Technol 41(6):992–1002

Sadler TD (2009) Situated learning in science education: socio-scientific issues as contexts for practice. Stud Sci Educ 45(1):1–42

Seethamraju R (2011) Enhancing student learning of enterprise integration and business process orientation through an ERP business simulation game. J Inf Syst Educ 22(1):19–29

Thayer P, Teachout M (1995) A climate for transfer model. Report AL/HR-TP-1995-0035.Air Force Materiel Command, Brooks Air Force Base, TX. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA317057andLocation=U2anddoc=GetTRDoc.pdf

Tsubakimoto M (2011) Development and technical evaluation of an interactive environment for a term paper grading support system in higher education. In: World conference on e-learning in corporate, government, healthcare, and higher education, vol 1, pp 966–972

Van Barneveld A, Arnold K, Campbell J (2012) Analytics in higher education: establishing a common language. EDUCAUSE Learn Initiat 1(1):1–11

Wikipedia: The free encyclopedia. (2004, July 22). FL: Wikimedia Foundation, Inc. Retrieved August 1, 2015, from https://www.wikipedia.org

Zhang Y, Oussena S, Clark T, Kim H (2010) Use data mining to improve student retention in higher education—a case study. In: ICEIS 2010: proceedings of the 12th international conference on enterprise information systems, pp 190–197

Chapter 20
Data Science for All: A University-Wide Course in Data Literacy

David Schuff

Abstract Infusing data literacy into a curriculum is an unrealized opportunity for higher education to truly make an impact on the current generation as they prepare to move into the workforce. This chapter describes the design and structure of a new, unique undergraduate elective course introduced into the curriculum of a large, public University in the Northeastern United States. The design of the course is designed to inspire an “evidence-based” mindset, encouraging students to identify and use data relevant to them in their field of study and the larger world around them. The chapter includes the course goals mapped to specific learning objectives, examples of exercises and assignments, a reading list, and a course syllabus. Instructors and institutions interested in bringing data science concepts to a broad audience can use this course as a foundation to build their own curriculum in this area.

Keywords Data science • Data literacy • Curriculum design • Pedagogy

20.1 Introduction

Increasing attention has been paid to the demand for data scientists. In fact, Davenport and Patil (2012) declared the data scientist as the “sexiest job of the 21st century.” What exactly is meant by the term “data scientist,” however, is unclear. We often think of a data scientist as a highly quantitative, technically-trained professional with advanced knowledge of statistics and big data infrastructure technologies.

However, Davenport and Patil (2012) define a data scientist as “a high-ranking professional with the training and curiosity to make discoveries in the world of big data.” Press (2012) to defines the data scientist as “an engineer who employs the scientific method and applies data-discovery tools to find new insights in data.”

D. Schuff (✉)
Department of Management Information Systems, Fox School of Business,
Temple University, 210 Speakman Hall, 1810 North 13th Street, Philadelphia,
PA 19122-6083, USA
e-mail: schuff@temple.edu

A.V. Deokar et al. (eds.), *Analytics and Data Science*, Annals of Information
Systems 21, DOI 10.1007/978-3-319-58097-5_20

These definitions are broad and do not necessarily imply a data scientist is a statistician, a computer scientist, or even a business analyst. Davenport and Patil's definition specifically mentions "big data," while Press' definition does not.

What these definitions have in common is that they underscore the importance of data literacy (as opposed to statistical and technological proficiency) as a skill for discovery. Infusing data literacy into a curriculum is an unrealized opportunity for higher education to truly make an impact on the current generation as they prepare to move into the workforce. Universities are squarely focused on providing focused Master's Degrees in Business Analytics; there are now 117 such programs according to the website "Master's In Data Science" (www.mastersindatascience.org). However, a data-literate undergraduate population, through their sheer numbers, have a far greater potential impact on the way organizations operate.

This chapter describes the design and structure of a new, unique undergraduate elective course introduced last year into the curriculum of Temple University, a large, public University in the Northeastern United States. In its first year it has gone from a pilot to a regular, multi-section offering in the University's "General Education" curriculum by emphasizing practical data literacy through current events, readily available analysis tools and the methods of scientific inquiry.

20.2 The Environment

Temple University is a large, public, urban institution with over 37,000 students. Its primary mission is to educate the regional undergraduate population through 140 bachelor degree programs (the University also has 126 master's degree and 57 doctoral programs). There are 17 schools and colleges including liberal arts, business, education, law, media and communication, music and dance, and engineering.

Like many large Universities, there is an institution-wide core curriculum that covers several broad categories. To fulfill the "General Education" or "GenEd" requirements, students must select from a menu of courses in each category, which includes analytical reading and writing, humanities, quantitative literacy, arts, human behavior, race and diversity, science and technology, US society, and world society.

One of the stated goals of the University's GenEd program is that, in an environment where "the amount of information is available … and the speed with which we can access information … continues to expand," the University must teach students "how information is linked and how pieces of information are interrelated" (Temple University 2015). This is certainly a reasonable and important goal for undergraduate education regardless of a student's major field of study, with obvious ties to concepts of data science and data literacy.

Further, Information Systems is a field that is well-positioned to deliver this material to a broad audience. Key aspects of the IS2010 Model Curriculum includes "understanding and addressing information requirements" and "exploiting opportunities created by technology innovations" (Topi et al. 2010). Most importantly, Information Systems is one of the few fields with the orientation and skill set to teach data literacy to a non-technical audience. Its emphasis on training business professionals create an applied focus on the identification and collection of data for problem-solving and use of practical analysis tools.

With this in mind, we proposed, developed, and executed a new course for the University's GenEd curriculum that would employ this dual focus on data and technology, targeted at a non-technical audience. The design of the course set out to inspire an "evidence-based" mindset, encouraging students to identify and use data relevant to them in their field of study and the larger world around them.

20.3 Course Goals

The course was designed to address several of the University GenEd program's broad learning goals (Temple University 2015):

1. **Information literacy**, including the ability to recognize and articulate information needs; to locate, critically evaluate, and organize information for a specific purpose; and to recognize and reflect on the ethical use of information.
2. Development of **critical thinking** skills, including the evaluation of evidence, analysis and synthesis of multiple sources, and reflection on varied perspectives.
3. **Communications skills**, using spoken and written language to construct a message that demonstrates the communicator has established clear goals and has considered her or his audience.
4. **Retrieve, organize, and analyze data** associated with a scientific model.
5. Understand and communicate **how technology encourages the process of discovery**.
6. Recognize, use, and appreciate scientific or **technological thinking for solving problems that are part of everyday life**.

From these broad goals, we developed ten specific learning goals for the course that could be evaluated through assignments and exams. Because of the course's dual focus—literacy and skill-building—the course learning goals span Krathwohl's (2002) knowledge dimension, with goals focused on factual (e.g., knowing data science terminology), conceptual (e.g., applying data visualization principles to assess the effectiveness of a graphic), and procedural knowledge (e.g., how to clean a data set). The learning goals also span the entire range of Krathwohl's cognitive process dimension, requiring students to remember, understand, apply, analyze, evaluate, and create. This is in line with the purpose of the course, which is to impart terminology, teach basic skills, and have them apply those skills to produce original knowledge. Table 20.1 lists each learning goal, along with where the specific goal lies on each dimension. These learning goals can involve several components and therefore may span multiple levels in a dimension.

Table 20.1 Course learning goals mapped to Krathwohl's (2002) taxonomy

Learning goal	Knowledge dimension	Cognitive process dimension
1. **Describe how advances in technology enable the field of data science**—This includes topics such as the storage and retrieval of data, the difference between a relational database and a "flat-file" spreadsheet, and advances due to big data technologies.	Factual	Remember
2. **Locate sources of data relevant to their field of study**—A key goal of the course is to see the relevance of data and take a more data-driven approach within their own major. Students should be able to identify data sets relevant to a specific problem. When those data sets do not exist, they should be able to create them.	Procedural	Apply, analyze
3. **Identify and correct problems with data sets to facilitate analysis**—This includes both identifying "bad data" and determining appropriate remedies for correcting it, such as deleting bad data, replacing erroneous numeric values with the mean, and reconciling errors in text labels.	Procedural	Analyze, evaluate
4. **Combine data sets from different sources**—Students must be able to reconcile differences in numeric and textual representations across data sets so that they can be analyzed as a single data source.	Procedural	Evaluate, create
5. **Assess the quality of a data source**—Students assess the trustworthiness of a data set by judging its source and its content. This is a significant issue with freely available "open data."	Conceptual, procedural	Evaluate
6. **Convey meaningful insights from a data analysis through visualizations**—This includes learning the basic principles of effective visualizations in order to critique existing graphics and make effective choices when creating original ones.	Conceptual, procedural	Understand, apply, create
7. **Analyze a data set using pivot tables**—Students learn how to use pivot tables to aggregate and summarize data to reveal insights.	Conceptual, procedural	Apply, analyze
8. **Determine meaning in textual data using text mining**—Students learn the mechanics of sentiment analysis to understand its proper use and its limitations as a decision tool. They also use a simple sentiment analysis tool to analyze a body of text.	Conceptual, procedural	Apply, evaluate
9. **Identify when advanced analytics techniques are appropriate**—This includes differentiating between descriptive, prescriptive, and predictive analytics and what types of problems each can address.	Factual, conceptual	Remember, understand
10. **Predict events that will occur together using association mining**—Students perform simple association mining on a common problem (market basket analysis) to take analytics from the theoretical to the practical in a way that can be replicated using their own data.	Procedural	Apply, analyze

Table 20.2 Learning goal mapping by module

Learning goal	Module 1	Module 2	Module 3	Module 4
Information literacy	✔			✔
Critical thinking	✔	✔	✔	
Communications skills		✔		
Retrieve, organize, and analyze data			✔	✔
How technology encourages discovery	✔		✔	
Technological thinking for everyday problems	✔	✔	✔	✔

20.4 Course Structure

The course is divided into four multi-week modules:

- Module 1: Data in our Daily Lives
- Module 2: Telling Stories with Data
- Module 3: Working with Data in the Real World
- Module 4: Analyzing Data

Each class session has a discussion component (approximately one-third of the class time), supplemented with experiential learning both in and outside of class. In-class experiential activities (about two-thirds of class time) are ungraded and directly tied to the current discussion topic. Homework assignments build on concepts introduced in the lecture and in-class activities. They employ more in-depth research and hands-on, computer-based activities using software tools such as Tableau Desktop and Microsoft Excel. A final group project requires students to select an issue relevant to them, source a data set, perform an analysis, and communicate the insight from their analysis to a general audience.

We designed the course for medium-to-large class sizes (60 students and over) with multiple sections and instructors. The in-class activities encourage teamwork through small breakout groups of 4–5 students. The four modules of the course take students through the process of identifying and collecting, communicating, preparing, and analyzing data. The modules are deliberately ordered to first give students foundational skills they will use later in the course (identifying data and communicating results). Also, the course introduces students to concepts of data preparation before they do basic predictive analysis.

Table 20.2 summarizes how the six learning goals—information literacy, critical thinking, communications skills, the ability to retrieve, organize, and analyze data, understanding how technology encourages the process of discovery, and the use of technological thinking in solving problems that are part of everyday life—are covered in the course. The rest of this section explains how each module addresses these learning goals, including a specific example of an in-class activity that supports the course content. Appendix contains an abridged version of the course syllabus, including session-by-session topics, brief assignment descriptions, and a reading list.

20.4.1 Overview of Module 1: Data in Our Daily Lives

This module builds skills in support of the learning goals of "information literacy," "critical thinking," "how technology encourages discovery," and "technological thinking for everyday problems." The basics of scientific inquiry is discussed in this module, including the notion of theory and hypotheses formation. Students also learn to identify sources of relevant data. They will learn the role of data across many disciplines, with concrete examples from current events. For example, this module discussed the National Security Agency's collection and use of telephony metadata. This module will also cover how citizens and organizations can use government-published "open data" to understand the world around them.

Sample In-Class Exercise: Identifying Sources of Data
Objective: Identify uses for data from open data sites. (Open data is the blanket term for publicly available data sets that can be used freely and without restriction. Usually this data comes from government sources.)

Learning Outcomes:

- Navigate metadata repositories and explore the data sets
- Understand the types of data available through open data sources
- Formulate possible uses for new data sets

Step 1: Explore—individual (10 min)

1. Visit these open data sites: http://www.data.gov and http://www.opendataphilly.org
2. Browse each site to get a feel for what kind of data is available from each one. Navigating these sites can be a little cryptic; look for "Data" tabs and "View the Dataset" and "Download" buttons.
3. Identify two data sets from each site that are interesting to you. Make sure you write down the name of the data set so you can find it again!

Step 2: Prepare—group (15 min)

1. In groups of three or four, compare and discuss your lists.
2. Select two data sets from each site (Data.gov and OpenDataPhilly) that you want to share with the class.
3. Identify how that data could be used. Be creative! For example, imagine a new website or mobile app that would use the data.
4. Designate a member of your group to be the spokesperson.

Step 3: Class Discussion (20 min)
Each group will briefly report out with their best ideas.

20.4.2 Overview of Module 2: Telling Stories with Data

The module builds skills in support of the learning goals of "communications skills," "critical thinking," "retrieve, organize, and analyze data," and "technological thinking for everyday problems." This course has a unique take on the communication goal—beyond just oral and written skills, today's students should be able to communicate visually. They must be able to create clear, informative graphics to effectively communicate insights in a data set. Students learn key principles that differentiate good data visualizations from bad ones using guidelines from experts such as Edward Tufte. Students apply these principles by evaluating a series of examples pulled from current news sources. For example, Fig. 20.1 highlights an example from Fox News that highlights how differences in data points can be misrepresented by truncating the y-axis:

The sample exercise below requires students to locate examples of good and bad data visualizations and evaluate them using criteria based on a set of guidelines. The students then post their findings to the course site and present their examples to the rest of the class. This creates a library of positive and negative examples that they can refer to later.

Sample In-Class Exercise: Finding Good and Bad Data Visualizations
Objective: Understand the difference between effective and ineffective data visualizations.

Learning Outcomes:

- Identify the message a graphic is trying to convey
- Evaluate how successful the graphic is at conveying that message
- Explain why, according to the principles discussed in class, the graphic is (or is not) effective

Step 1: Explore—individual (15 min)

1. Visit one or more of these data visualization/infographic sites: http://www.flowingdata.com, http://www.coolinfographics.com, or http://dailyinfographic.com
2. Identify one example of a graphic that you think is very good, and one example of a graphic that you think isn't that great and could be improved. So you don't lose track of your graphics, copy and paste the URL where you found the graphic from your browser into a Word document. You can also use the document to make some notes.
3. Using the principles we discussed in class, come up with

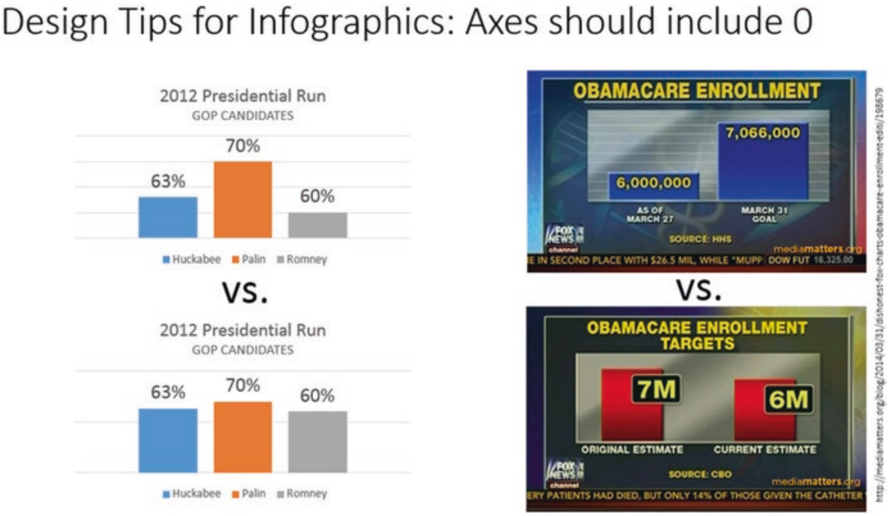
   (a) Reasons why the graphic you've chosen are good/bad.
   (b) Make recommendations for improvement.

Step 2: Prepare—group (15 min)

1. In groups of 3–5, compare and discuss your notes.
2. From your lists, choose one example of a good graphic and one example of a bad graphic.
3. For each group, post a comment to the entry on the course web site for this in-class exercise. Make sure the name of at least one member of your group is included in the post.

Step 3: Class Discussion (15 min)

Each group will report out on their choices. Make sure you enter the URLs correctly—I'll display your graphics as your talking about them!



Design Tips for Infographics: Axes should include 0

http://mediamatters.org/blog/2014/03/31/dishonest-fox-charts-obamacare-enrollment-edition/198679

Fig. 20.1 Visualization example inappropriately truncating the y-axis

20.4.3 Overview of Module 3: Working with Data in the Real World

This module builds skills in support of the "information literacy" GenEd learning goal, and the Science & Technology area goals to "retrieve, organize, and analyze data" and "technological thinking to solve everyday problems. Students learn how to fix problems in data sets. This builds on the course's first module where they learn how to identify data quality issues in data. Students learn how to address these problems through data cleansing and transformation to create a useable, reliable data set using Microsoft Excel. They will resolve inconsistencies within and across data sets, and determine when data is in the wrong form. The exercise below introduces students to the concept of a Key Performance Indicator. The exercise is intended to apply the SMART criteria (Specific, Measurable, Achievable, Relevant, and Time Phased) to evaluate candidate KPIs for a scenario. Another exercise requires the students to create several KPI scorecards using a more business-oriented scenario: on-time flight data for a set of airports.

Sample In-Class Exercise: Identifying Key Performance Indicators

Objective: Select Key Performance Indicators (KPIs) that facilitate evaluation for a given scenario.

Learning Outcomes:

- Identify “good” KPIs that adhere to the SMART criteria
- Select the best KPIs from a list of potential metrics
- Describe the limitations of using KPIs to make an evaluation

Scenario:

Working in groups is something students regularly do in their classes. However, everybody has worked in a group where the quality of individual contributions vary significantly among group members.

Your task is to come up with a set of KPIs to evaluate a group member. This can be used as a tool for groups to evaluate and give feedback to each other during group projects.

Step 1: Identify Key Performance Indicators (Individual, 5 min)

Working individually, come up with five KPIs that can be used to measure how a student is doing as a contributor to a group project. This should be done from the perspective of another student, not an instructor; however, an instructor might be interested in your list. Make sure each KPI is adheres to the SMART criteria discussed in class: Specific, Measurable, Achievable, Relevant, and Time Phased.

For example, “Student contributes quality work” is not a good KPI:

- It isn’t specific—What does “quality” mean? What does “contribute” mean?
- It isn’t measurable—It is unclear how quality would be measured?
- It isn’t time phased—A time period is never specified; is this weekly contributions, or over the course of the entire project?

However, “Number of ideas contributed each week” would be a much better KPI. It is certainly more specific and more measurable!

Step 2: Determine the Best KPIs (Group, 10 min)

In groups of 3–4, compare your lists. Choose the three best KPIs, taking ideas from your individual lists. Remember that you should choose items that best adhere to the SMART criteria, but you should also think about the items that are most important in determining who is a good member of a project group.

Step 3: Group Discussion (15 min)

Share your KPIs with the rest of the class. They should all meet the SMART criteria, so beyond that explain why the ones you've chosen are the best to use for evaluation—why are they the most helpful in differentiate good group members from poor ones.

20.4.4 Overview of Module 4: Analyzing Data

This module builds skills in support of the "critical thinking" GenEd learning goal and the Science & Technology area goals to "retrieve, organize, and analyze data," as "how technology encourages discovery," and "technological thinking for everyday problems." Students learn how data is stored and organized. Specifically, they will learn the differences between spreadsheets and databases and why each is used. They will also learn three analytics techniques to give them a sense of what can be done with data analytics, and also to give them hands-on experience with analytics tools such as Microsoft Excel and Tableau Desktop. For example, students learn how to use Pivot Tables to summarize large data sets (such as the crime activity assignment described below), and Association Analysis to discover which products are likely to be bought together at a store (such as graham crackers and marshmallows). Students also learn to interpret the output from these analyses and make inferences about underlying patterns in the data.

20.5 Final Project

The final project is a group project that requires students to bring together what they've learned throughout the course. Student teams source an "original" data set (i.e., not one already used in the course), develop a research question, and then answer that question by using one or more of the data analysis techniques and tools covered in the course. Students are encouraged to find a data set and question that is relevant and interesting to them.

The deliverable is a five-minute presentation with two minutes for questions from the instructor and the class. The short presentation is a deliberate choice because (1) it forces students to hone their presentation skills by being direct and to the point and (2) it allows the course to scale.

The suggested format for the final presentation is:

- Slide 1 should list the group members and the title of the presentation.
- Slide 2 will describe the scenario. What question will be answered and why is it important?

In-Class Exercise: Manually Determining the Sentiment of Textual Data
Objective: Differentiate between positive and negative sentiment in text.

Learning Outcomes:

- Perform a manual sentiment analysis of a Twitter stream
- Develop rules for classifying a message as positive or negative
- Explain the problems and issues with accurately describing sentiment within text

Part 1: Group (15 min)

1. In groups of two, visit two Twitter feeds for well-known brands. You can choose any brand you want for this, but if you need some ideas:

CocaCola	@CocaCola	http://twitter.com/cocacola
McDonalds	@McDonalds	http://twitter.com/mcdonalds
Honda Motors	@Honda	http://twitter.com/honda
Starbucks	@Starbucks	http://twitter.com/starbucks
Nike	@Nike	http://twitter.com/nike
H&M	@hm	http://twitter.com/hm

2. Within each feed, click on a few tweets and read the replies.
3. Find three examples of positive tweets, three examples of negative tweets, and three examples of neutral tweets (neither positive nor negative). Write them down in three lists.
4. Make a note of why you classified them as positive, negative, or neutral.

Part 2: Larger Group (5 min)

1. Find another group to form a group of four.
2. Share your lists of positive and negative tweets. See if you agree with each other's choices.
3. Come up with rules for determining whether a tweet is positive or negative. For example:
   (a) Are there certain words which increase your certainty of how to classify the tweet?
   (b) Are there certain tweets that sound positive but really are negative?
   (c) How do you detect sarcasm?
   (d) How would you explain to someone how to classify tweets?

Part 3: Class Discussion (20 min)
We'll compare notes. Specifically, we will discuss:

- What are some rules for determining positive versus negative sentiment?
- Were some tweets difficult to categorize? Why?
- In what ways would this be a good method of understanding how people felt about your brand? In what ways could it give you bad information?

- Slide 3 will describe the data. What are the key elements and how was it obtained?
- Slides 4 and 5 will describe the analysis and the results, making good use of data visualizations.
- Slide 6 will summarize the conclusions. What was learned? Students should support their conclusions using the results of the analysis, citing specific evidence.
- Slide 7 will list the references.

Some examples of final projects from past classes include:

- Exploring the question of whether the best soccer players are the highest paid.
- An analysis of the national origin of members of terrorist organizations.
- Investigating what time of day students are most likely to answer an online survey.
- Correlations between stock price and social media sentiment for a major aerospace firm.

20.6 Conclusions

Data analytics is no longer the prevue of data scientists. It is a fundamental skill for the twenty-first century workforce. This is both a challenge and an opportunity for higher education, one that the Information Systems discipline is uniquely positioned to meet. With courses such as the one described here, Information Systems can increase their reach beyond the business school and provide valuable, marketable skills to a broad audience across the University.

The key to seizing this opportunity is to recognize that "data literacy" is the true core skill for undergraduate students, not sophisticated analytics techniques. We must instill in our students an appreciation of evidence-based decision-making through an appreciation of what data can do and how even simple analysis can yield sophisticated insights.

Biography

David Schuff is Professor of Management Information Systems in the Fox School of Business at Temple University. David holds a B.A. in Economics from the University of Pittsburgh, an M.B.A. from Villanova University, an M.S. in Information Management from Arizona State University, and a Ph.D. in Business Administration from Arizona State University. His research interests include the application of information visualization to decision support systems, data warehousing, and the impact of user-generated content on organizations and society. David's work has appeared in numerous journals, including Management Information Systems Quarterly, Decision Sciences, Decision Support Systems, Information & Management, Communications of the ACM, IEEE Computer, AIS Transactions on Human-Computer Interaction, and Information Systems Journal.

Appendix: Abbreviated Course Syllabus for Data Science

Course Description

We are all drowning in data, and so is your future employer. Data pours in from sources as diverse as social media, customer loyalty programs, weather stations, smartphones, and credit card purchases. How can you make sense of it all? Those that can turn raw data into insight will be tomorrow's decision-makers; those that can solve problems and communicate using data will be tomorrow's leaders. This course will teach you how to harness the power of data by mastering the ways it is stored, organized, and analyzed to enable better decisions. You will get hands-on experience by solving problems using a variety of powerful, computer-based data tools virtually every organization uses. You will also learn to make more impactful and persuasive presentations by learning the key principles of presenting data visually.

Course Objectives

- Describe how advances in technology enable the field of data science
- Locate sources of data relevant to their field of study
- Identify and correct problems with data sets to facilitate analysis
- Combine data sets from different sources
- Assess the quality of a data source
- Convey meaningful insights from a data analysis through visualizations
- Analyze a data set using pivot tables
- Determine meaning in textual data using text mining
- Identify when advanced analytics techniques are appropriate
- Predict events that will occur together using association mining

Assignments

#	Assignment description
1	**Create a data analysis plan (individual)** Develop a plan for data analysis by forming hypothesis and finding data sets that will allow you to test those hypotheses. The scenario: Once students graduate, it's time for them to go get a job. But is staying in the area the best choice? Evaluate our city as a place to live, work, and play compared to the rest of the United States
2	**Analyze a data set using tableau (individual)** Use Tableau to analyze and reveal various relationships within a data set. Use the data set from the Environmental Protection Agency regarding fuel economy 2015 model year cars. Answer a series of questions by creating the most visually effective charts and graphs using the guidelines discussed in class

#	Assignment description
3	**Cleaning a data set (individual)** Correct the errors in a data set for the fictitious company "Vandelay Industries." The sales group is suspicious that there might be errors in the data for January. Work with a new data set of 3296 orders with 5192 line items from January 2014
4	**Group data analysis (group term project)** In groups, perform an original analysis on a data set of your choosing. The data set can come from any source as long as it is something you have not already worked on for this course. Possible sources of data include: open data from Data.gov, data sets from the Pew Research Center, sports statistics, a data set from your current employer, or an original survey conducted by your group Your analysis should clearly demonstrate the tools and techniques you've been exposed to in this course. This can take any form you'd like (i.e., comparison of averages across categories, mapping geographic data, sentiment analysis, developing and visualizing KPIs) Your group will present your work in class through a five-min presentation, with 2 min for questions

Schedule and Reading List (Current Configuration Is for Two 80-min Sessions per Week)

Week/ session	Topic/key questions	Readings
Module 1: Data in our daily lives		
1.1	Introduction • Course introduction/syllabus • What is the difference between data, information, and knowledge? • What makes "big data" big?	
1.2	Science and data science • What is data science? • What is the difference between a theory and a hypothesis? • What are the dangers of data analysis without a hypotheses?	Dhar, V. (2013). Data Science and Prediction. Communications of the ACM. Vol. 56, No. 12. pp. 64–73 Allain, R. (2013). Three Science Words We Should Stop Using. Wired.com. March 27
2.1	A brief introduction to data • What are the forms data can take? • Where does data come from? • What is metadata? A data dictionary?	Stein, G. (2013). I'm Beating the NSA to the Punch by Spying on Myself. Fastcolabs.com. June 12 Di Justto, P. (2013). What the N.S.A. Wants to Know about Your Phone Calls. The New Yorker. June 7
2.2	Identifying Sources of Data • What kinds of data are available in different disciplines (arts, sciences, medicine, business, government, etc.)? • What kinds of problems and issues can data insight address?	Silver, N. (2014). What the Fox Knows. FiveThirtyEight.com. March 17 Open Data. Wikipedia Silver, N. (2014). In Search of America's Best Burrito. FiveThirtyEight.com. June 5

Week/ session	Topic/key questions	Readings
3.1	Learning to (Mis)trust Data • How do you spot reliable sources of data? • How do you assess data quality? • What is the "Filter Bubble?"	Weisberg, J. (2011). Bubble Trouble: Is Web Personalization Turning Us Into Solipsistic Twits? Slate.com Crawford, K. (2013). The Hidden Biases in Big Data. Harvard Business Review Blog Network. April 1 Hayes, B. (2013). In Data We Trust. Business Over Broadway. November 4
3.2	Guest speaker	
Module 2: Telling stories with data		
4.1	Viewing data • What are different ways of viewing data? • When do you need to visualize data? • What are the basic techniques of data visualization?	Unwin, A. (2008). Chapter II.2: Good Graphics? Handbook of Data Visualization. Chen, Hardle, and Unwin (Eds.). pp. 57–78
4.2	Introduction to Tableau • What is Tableau? What can you do with it? • How is it different from Microsoft Excel?	Hoven, N. (n.d.). Stephen Few on Data Visualization: 8 Core Principles. Tableau Software Acohido, B. (2013). Watch Out, Terrorists: Big Data is on the Case. USAToday.com. July 29
5.1	Communicating using data • What are the principles of communicating data? • How do you communicate complex ideas using data? • How do you construct visualizations that complement a report? That stand on their own?	Davenport, T. (2013). Telling a Story with Data. Deloitte University Press Matlin, C. (2014). Visualizaing a Day in the Life of a New York City Cab. FiveThirtyEight.com. July 17
5.2	Storytelling with infographics • How are infographics different from other types of visualizations? • How do infographic tools differ from other data tools we've used so far?	Krum, R. (2014). Cool infographics: Effective Communication with Data Visualization. (Chapter 1: The Science of Infographics) Krum, R. (2014). Cool infographics: Effective Communication with Data Visualization. (Chapter 6: Designing Infographics)
6.1/6.2	Exam review/EXAM 1	
Module 3: Working with data in the real world		
7.1	Dirty Data • How does data get dirty? • What are the consequences (i.e., ethical, financial) of dirty data? • How do you clean it?	Redman, T. (2013). Data's Credibility Problem. Harvard Business Review. Vol. 91, No. 12. pp. 84–88 Gandel, S. (2013). Damn Excel! How the 'Most Important Software Application of All Time' Is Ruining the World. Fortune.com. April 17

Week/ session	Topic/key questions	Readings
7.2	Data cleansing • How do you identify data problems? • How do you correct data problems? • When is fixing the data not worth it?	Taber, D. (2010). Stupid Data Corruption Tricks: Take our CRM Quiz. CIO.com. November 2 Top Ten Ways to Clean Your Data. Microsoft
8.1	Choosing relevant data • How do you identify Key Performance Indicators (KPIs)? • How do you identify the right measure for the selected problem?	Performance Indicator. Wikipedia Schambra, W. (2013). The Tyranny of Success: Nonprofits and Metrics. NonprofitQuarterly.com. December 30
8.2	Evaluating key performance indicators • How do you categorize and visualize KPIs according to a threshold? • How do you use Tableau to evaluate KPIs? How would you use Excel?	Olson, P. (2014). Wearable Tech is Plugging into Health Insurance. Forbes.com. June 19 Bialik, C. (2014). Tracking Health One Step (and Clap, and Wave, and Fist Pump) at a Time. FiveThirtyEight.com. March 17
9.1	Connecting diverse data • How do you identify data sets that can be combined? • How do you combine data sets? • How do you resolve conflicts?	Strickland, J. (n.d.). How Data Integration Works. howstuffworks.com Gallagher, S. (2014). The GOP Arms Itself for the Next "War" in the Analytics Arms Race. arstechnica.com. February 7
9.2	Creating interactive dashboards • How does a dashboard differ from an Infographic? A chart? • How do dashboards facilitate decision-making?	Best Practices for Designing Views and Dashboards. Tableau Software Farmer, D. (2014). The One Skill You Really Need for Data Analysis
10.1/10.2	Exam Review/EXAM 2	
Module 4: Analyzing data		
11.1	Storing and retrieving data • What is a database? How are spreadsheets just a type of database? • How are technology advances changing how we think about storing data? • What are the core technologies of big data analytics?	Rosenblum, M. and Dorsey, P. (n.d.). Knowing Just Enough about Relational Databases. Dummies.com Bertolucci, J. (2013). How to Explain Hadoop to Non-Geeks. InformationWeek.com. November 19
11.2	Using Tableau to aggregate data • What can you learn from aggregation? • How does thinking of data dimensionally help solve problems?	Acampora, J. (2013). How to Structure Source Data for Excel Pivot Tables & Unpivot. July 18

Week/ session	Topic/key questions	Readings
12.1	Beyond numbers • What is the difference between structured and unstructured data? • What can you learn from text data that you can't from numeric data? • What are the tools for text analysis?	Hurwitz, J., Nugent, A., Halper, F., and Kaufman, M. (n.d.). Unstructured Data in a Big Data Environment. Dummies.com Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. Communications of the ACM. Vol. 56, No. 4. pp. 82–89
12.2	Twitter sentiment analysis using Excel and Google Drive • What are the steps in performing a sentiment analysis? • What are the challenges in deriving meaningful information from text?	Wohlsen, M. (2014). Don't Worry, Facebook Still Has No Clue How You Feel. Wired.com. July 2
13.1	Predicting the future • What is predictive analytics? What problems does it address? • What kinds of analysis can be done? • What kinds of data are needed for an analysis?	Paine, N. (2014). What Analytics Can Teach Us About the Beautiful Game. June 12 Bertolucci, J. (2013). Big Data Analytics: Descriptive vs. Predictive vs. Prescriptive. InformationWeek.com. December 31
13.2	Predictive analytics using Tableau • Perform a forecasting analysis • Perform a simple association analysis	Peck, D. (2013). They're Watching You at Work. TheAtlantic.com. November 20
13.1/13.2	Group presentations/FINAL EXAM review	

References

Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. Harv Bus Rev 90(10):70–76

Krathwohl D (2002) A revision of Bloom's taxonomy: an overview. Theory Pract 41(4):212–218

Press G (2012) Data scientists: the definition of sexy. Forbes. http://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/#526d00375187. Accessed 27 Sept 2012

Temple University (2015) General education program. http://gened.temple.edu. Accessed 29 Feb 2016

Topi H, Valacich J, Wright RT, Kaiser K, Nunamaker JF, Sipior JC, Jan de Vreede G (2010) IS 2010: curriculum guidelines for undergraduate degree programs in information systems. Commun Assoc Inf Syst 26:18