

Chapter 13

Social Impact and Social Media Analysis Relating to Big Data

Nirmala Dorasamy and Nataša Pomazalová

Abstract Social media is a component of a larger dynamic and complex media and information domain. As the connection with Big Data grows, its impact in the social media domain cannot be avoided. It is vital that while the positive impact needs to be recognized, the negative impact emerging from Big Data analysis as a social computational tool needs to be recognized and responded to by various agencies. There have been major investments in the development of more powerful digital infrastructure and tools to tackle new and more complex and interdisciplinary research challenges. While there is a need to size the opportunities offered by continuing advances in computational techniques for analyzing social media, the effective use of human expertise cannot be ignored. Using the right data, in the right way and for the right reasons, can change lives for the better, especially if Big Data is used discriminately and transparently. This chapter analyzes the impact of Big Data from social media platforms in the social, political, and economic spheres. Further, the discriminate use of Big Data analysis from social media platforms is explored, within the context of ethical conduct by potential users and proposes important imperatives to minimize, if not control, the negative impact of Big Data analysis from a social perspective.

Keywords Social media • Digital technology • Social networking sites • Database tools • Big Data • Big Data analysis • Social media services

N. Dorasamy (✉)

Department of Public management and Economics, Durban University of Technology,
Durban, South Africa

e-mail: nirmala@dut.ac.za

N. Pomazalová

Department of Regional Development, Mendel University, Brno, Czech Republic

Durban University of Technology, Durban, South Africa

e-mail: natta@atlas.cz

13.1 Introduction

Data, or segments of information, have been collected and used throughout history. However, the potential to collect, store, and analyze data has significantly increased with the advancement in digital technology. The emergence of Big Data has escalated its usefulness for decision-making at various levels of analysis, including individual, group, organizational, and national systems. Organizations have moved from using data stored in relational databases to using data from data mining in general ledger packages, weblogs, social media, e-mail, sensors, photographs, corporate enterprise resource planning (ERP) systems, custom relationship management (CRM) programs, and social networks. While the growth of Big Data has accelerated in the last few years, the ability to find useful information within the Big Data is of crucial importance and requires careful consideration. Managing such data has to be underpinned by quality, protecting privacy and ethical use.

The rapid growth of the web as a publishing tool and the recent explosion of social media and social networking sites have generated opportunities and challenges to social researchers. Currently, there are many types of social media services (SMS). The Personal SMS like Facebook allows users to create online profiles and connect with other users, focusing on social relationships and information sharing such as one's gender, age, interests, and job profile. The Status SMS like Twitter allows users to post short status updates to broadcast information quickly and publicly with other users. The Location SMS like Foursquare and Google Latitude, using GPS-based networks, broadcasts one's real-time location. The Content-sharing SMS like YouTube and Flickr is designed as platforms for sharing content, such as music, photographs, and videos [4]. The Shared-interest SMS like LinkedIn is more a network for a subset of professional users to share information interests like politics and education. These social media services provide datasets that have expanded in size and complexity to the extent that computer-based methods are now required to analyze mass volumes of information. Data, with datasets whose size can range from a few dozen terabytes (TB) to multiple petabytes (PB), is beyond the ability of typical database software tools to capture, process, and store, manage, and analyze. Big Data technologies are required to economically extract value and meaning from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and analysis [21].

Today, social media is a key source for Big Data analysis from platforms such as Twitter, Facebook, and Flickr, especially for businesses that depend on data-driven intelligence [3, 12]. IT companies such as Google, Amazon, Facebook, and IBM are fiercely competing in the Big Data analytics market. Social media systems provide valuable information in terms of its detail, personal nature, and accuracy. Since the data is not totally private, it is exposed to scrutiny within a user's network, which can increase the chances of accuracy when compared to data from other sources. Big Data integration and predictive analytics can help overcome the challenges of managing in an environment with increasing rates of change and innovation [12]. Studies have shown that Big Data analytics has resulted in improvements in

retail operating margins, reduction in national healthcare expenditures, and savings in operational efficiencies. It has great potential, in that it can generate significant value across sectors, such as healthcare, retail, manufacturing, and the public sector [12]. However, seeing that data is dynamic, there has to be continuous integration of existing data with “living data,” if companies want to reap the optimal benefits of data analytics.

Big Data volume is expanding due to the increase of social media, online data, and location data, often resulting from the accelerated usage of sensor-enabled devices. This has resulted in mobile cloud computing being made possible by focusing on Internet technologies that are built on web-based standards and protocols. The key drivers for cloud computing are bandwidth increase in networks, cost reduction in storage systems, and advances in database [12]. It has created a form of virtualization over the Internet, which involves data outsourcing that can be provisioned with minimal management effort or service provider interaction, with no up-front cost and provides just-in-time services. It is a model for enabling convenient and on-demand network access to a shared pool of computing resources and eliminates the cost for in house infrastructure [5]. An area of distinctive growth in the IT ecosystem is social media with smart devices, such as smartphones and tablets, providing a new communication platform with real-time access for customers. Various agencies have recognized such data as an important constituent of innovation and have developed means to use Big Data to develop solutions to business, technical, and social problems in innovative and collaborative ways. While the use of Big Data is invaluable, businesses are challenged by Big Data because it grows so large that they become awkward to work with using on-hand database management tools [12].

While the value of Big Data is clear for tackling complex technical and business problems, the question is on how well Big Data can solve complex social problems? While, business and science have shown the value of Big Data, the social sector needs to show how they can adopt this type of decision-making potential into their operations. The issues that are being addressed in the social sector are more complex than they are in business or science, making the use of Big Data more challenging. In addition, greater focus must be given to the rights, privacy, and dignity of different stakeholders. The large-scale collection, aggregation, analysis, and disclosure of detailed and triangulated information offer the possibility of powerful computational social science tools, but carries with it the potential for abuse by various entities, especially if datasets are not reliable and representative. In spite of these obstacles, progress continues.

13.2 Using Big Data and Social Media Innovatively

Big Data can encompass information such as transactions, social media, enterprise content, sensors, and mobile devices. Since Big Data refers to datasets that extend beyond single data repositories (databases or data warehouses), they are too large

and complex to be processed by traditional database management and processing tools. New Big Data technologies have helped to capture analyze and store data to solve problems. Organizations are employing computing power through hardware and software advances to manipulate huge amounts of data. Some new approaches include Hadoop, a software that takes a different approach to data management, and HANA, a hardware approach that handles data manipulation in raw memory to make real-time analytics of the Big Data a reality [7]. Such advances have enabled the cost reduction and volume capacity increase of digital storage mediums and unique ways of manipulating data from social media.

Big Data offers much potential for innovative use, thereby creating value. Fanning and Grant [7] identified the following ways in which Big Data is valuable:

- Big Data can reveal significant value by making information transparent and usable at a much higher frequency. More people looking at the data will bring different perspectives.
- As the data proliferates, there is more accurate and detailed performance information that may expose variability and issues that need attention.
- Controlled experiments using Big Data analysis can assist to make better management decisions regarding tailored products and services.
- Big Data can be used to improve the development of the next generation of products and services through proactive maintenance and preventive measures to minimize failures.

13.2.1 Social Media as Advertising and Marketing Medium

Social media platform providers provide numerous metrics for data analysis to profile users for advertisers by showing the right advertisements to the right people, using masses of hidden information to model which users are likely to respond to a particular advertisement. Facebook can optimize advertising revenue by targeting advertisements to achieve the greatest possible number of clicks. If used scrupulously, valuable information can reach the right users. An additional example is Amazon, which allows users to exercise some control over their data by allowing them to flag purchases not to be used to form recommendations [1]. This is useful both in cases of gift purchases and in cases where the product is of a sensitive nature. Further, by allowing users to instruct a website not to use a specific piece of personal data for a particular purpose represents a significant improvement on the current personal data free-for-all model used by both social networking companies and their corporate customers [17]. Such platforms can also play a beneficial role in exposing, predicting, and helping to eliminate disruptive behavior. Computational social science can serve a positive role in promoting the interests of the community in a social media platform, provided the ethical considerations are evaluated for various purposes of use.

The impact of Big Data can also be seen in marketing. There is a noticeable move away from large-scale mailings of catalogues and offers to various individuals, based primarily on purchased mailing lists or phone directories. This is due to businesses mining Big Data to target individuals based on knowing the preferences of individuals in the population. For example, Amazon can almost instantaneously offer additional purchase opportunities to individuals based on what others have also purchased given a similar purchase showing up in the cart. It can also target those who in the past purchased a product or those who are searching a particular topic [2]. Further, Google offers marketers the opportunity to provide relevant advertisements to individuals based on their search habits.

It is through the improvement in the tools to analyze and collect Big Data that larger markets can be targeted. According to Fanning and Grant [7], digital marketing uses a combination of push-and-pull Internet technologies to execute marketing campaigns. The use of software vendors such as Adobe allows customers to make each digital transaction layered, thus allowing the organization to see in real time how a particular advertising campaign is performing, in terms of what is being viewed, how often, how long, as well as other actions such as responses rates and purchases made [8]. This information provides key information for marketers to make real-time decisions. In addition, the proliferation of mobile phones, tablets, and other means of accessing the Internet has facilitated multichannel marketing by companies like Usablenet.

Applications such as Facebook, LinkedIn, and Google are driven by information sharing that can be used by business clients, government, other users within the social media platform, and the platform provider itself [7]. For example, business clients draw on this computational social science to target users based on constructs that range from age, gender, and geographical location, sexual preferences, education level, and employer.

13.2.2 Adding Value for Social Well-Being

The rapid use of mobile phones and Internet usage, especially in developing economies, gives people the opportunity to improve the quality of their lives. A mobile phone acts as an individual sensor, collecting pertinent information from its environment, which when aggregated and analyzed with information from thousands of other mobile phones can provide vital information, which can then be disseminated back to people on the ground via the same mobile phones. For example, Cell Life, a South African organization, created a mass messaging mobile service called Communicate, which reminds patients to take their medications, links patients to clinics, and offers peer-to-peer support services such as counseling and monitoring [6].

In addition, most modern mobile phones contain global positioning systems technology, which identifies the geographic location of the phone, and other information relating to social media postings. This is important, for example,

when researching migration patterns to understand the spread of infectious diseases like Ebola and to help stop the disease from spreading. Information on the patterns of human travel collected from mobile phone usage can be used to develop predictive models to combat diseases in specific regions [13].

A further consideration is census data collected in different countries, which is an important source of information for governments. In the 1800s, information from the national census was logged by hand, microfilmed, and sent to be stored in state archives, libraries, and universities. It took many years to properly tabulate census data after the initial collection. In recent years, countries have streamlined their data collection methods by adopting emerging technologies like geographic information systems, social media, videos, intelligent character recognition systems, and sophisticated data-processing software and processing tools to survey the populace [6].

Thorough data management during crisis communication can provide major benefits. The study by Proctera et al. [18] on the 2011 riots in England found that mainstream media lagged behind crowdsourced (“citizen journalism”) reports appearing in social media. Further, it was evidenced that collaborative efforts by large numbers of “producers” of data can provide competing and, at times, better coverage of events than mainstream media. While evidence cannot always be taken at face value, social media provides a platform for robust mechanisms for authentication, so that false rumors are identified more quickly. This is supported by Mendoza, Poblete, and Castillo [16] who noted that users deal with “true” and “false” rumors differently: the former are affirmed more than 90% of the time, whereas the latter are challenged (i.e., questioned or denied) 50% of the time.

There are several examples of social media being a valuable tool for information gathering, for keeping the public informed, and for providing advice during crisis situations. One such case was the August 2011 riots in England, which began as an isolated incident in Tottenham. A study by Proctera, Visb, and Voss [18] revealed that Twitter was used overwhelmingly for more positive means, especially for the organization of the riot cleanup. Even the police supported keeping social media sites open during the crisis. However, the study confirmed the conclusions of other studies that the police and government agencies in general still need to use social media platforms like Twitter effectively [18].

13.2.3 Gauging Business Performance

The business community has also been a heavy user of Big Data. For example, Netflix collects billions of hours of user data to analyze the titles, genres, time spent viewing, and video color schemes to gauge customer preferences and to give the customer the best possible experience. Following in the path of e-commerce, the rise of social media, Big Data, and cloud computing has impacted businesses in the following ways [11]:

- The ability to identify at the earliest opportunity those who are in danger of leaving organizations and to action the retention of best talent
- The ability to identify the high performers based on “new” live data highlighting performance and ratings and profitability rather than “old” assumptive data such as university attended and the grade of their degree
- The ability to measure the real drivers of performance within the business, thereby identifying “hidden gems” that make a real difference
- The ability to “fine-tune” businesses based on fact and evidence rather than fiction and emotion
- The ability to understand that datasets can be combined to form a more intricate and accurate picture from several data sources

13.2.4 Adding Value in Financial Markets

Financial markets have also benefitted from Big Data. Social media has played a significant role in growing financial markets with respect to the following [11]:

- Significant means by which crowdfunding helps the underserved small- and mid-sized entities (SMEs) and start-ups to access capital in a cost-effective manner.
- New technologies fund opportunities for growing segments of the economy not reached by traditional outlets and also create new jobs in a dynamic, technology-based business model.
- Algorithms and soft-/hardware technology related to “high-frequency trading” has exploded over the last 20 years, and the primary beneficiary has been the market for existing shares and other financial instruments (secondary market). With new cost-effective mechanisms to raise funds, underserved primary markets can now enjoy lower funding costs.
- Provides investors with protection from fraudsters through both proactive education and appropriate regulation by governments through the creation of a new and dynamic mechanism for allocating capital from traditional, institutional stakeholders (such as banks) to individual-driven operations using current and future technologies to touch millions of individuals looking for investment opportunities.

13.3 Discriminate Use of Social Media Analysis

There are multiple dimensions to Big Data, which include volume (considers the amount of data generated and collected), velocity (refers to the speed at which data are analyzed), variety (indicates the diversity of the types of data that are collected), viscosity (measures the resistance to flow of data), variability (measures the

unpredictable rate of flow and types), veracity (measures the biases, noise, abnormality, and reliability in datasets), and volatility (indicates how long data are valid and should be stored) [15]. The multiple dimensions make data searches and retrieval more complex, as organizations have to find economical ways of integrating heterogeneous datasets while allowing for newer sources of data (in origin and type) to be integrated within existing systems [6]. The proliferation of social networks and social media requires much of the data being collected to be thoroughly analyzed before decision-making, as the data can be easily manipulated. In view of the high cost of data collection and management, organizations need to analyze the trade-off between accuracy and the cost of inaccurate data. If the quality of data is poor and users cannot make sense of it, then the data has limited value and use.

13.3.1 Unstructured Data

Unstructured data from multimedia networks cannot be categorized or analyzed numerically, as it uses natural language. The explosive growth of social media implies that the variety and quantity of Big Data is growing. A great deal of the growth can be traced to unstructured data. For example, analyzing words and pictures and then collating everything into meaningful and accurately interpreted information requires diverse methods and can be time-consuming. The challenge is exacerbated when time-sensitive issues like monitoring civil riots require data to be aggregated and analyzed in the shortest time possible [19].

Some authors argue that there is no Big Data in the context of social problems, as data is highly unstructured and generally not limited to numbers. In the case of child pornography, the global industry lures thousands of children annually. Increasingly, the producers of child pornography make use of various Internet platforms like mobile phones, social media, and online classifieds. Although, many initiatives exist to curb the problem, few initiatives have attempted to use Big Data. Data from these technologies could be collected and used to identify, track, and prosecute offenders. The problem is that the illicit nature of child pornography makes it difficult to collect reliable primary data for some of the following reasons: there are no valid indicators to measure antipornographic success and information collected often meets organizational needs and not global needs [6]. In addition, because of data privacy and security issues, data held by various organizations are seldom shared in raw form, thereby limiting the creation of global datasets. This is accentuated by agencies combating social evils often competing with each other for scarce resources, therefore not being eager to share data.

Different service providers have their own network management systems. Organizations cannot connect their datasets across other organizations, if such data is immersed in their administrative systems for operational purposes. One such case is the US healthcare industry, which is characterized by large volumes of health plans offered by different service providers with their own network management. This

invariably results in data being stored in multiple formats in multiple places. If this data was more efficiently managed, then massive savings can accrue.

13.3.2 Gaps in Governance Standards

The lack of adequate data governance standards has failed to define how data is captured, stored, and used for accountability in the social arena. As a result some of the emergent challenges include integrating different datasets which lack good metadata (data that describe data), poor quality data, and difficult-to-manipulate forms such as PDFs or older file formats. As a result, large inconsistencies in the captured data exist, further complicated by the need to transform data for analysis, which is costly. The number of publicly available government datasets has accelerated, but only limited datasets where there is good metadata, ease of accessibility, and manipulability are ever used. Further, integrating information from multiple data sources requires skilled workers. In a report quoted by Fanning and Grant [7], by 2018 there would be a shortage of 140,000–190,000 people with deep analytical skills in the USA. This could be a serious problem for the analysis of Big Data to make effective decisions, if there are no initiatives to develop data analytics skills.

Without clarifying ethical issues on data storage, access, and use by different stakeholders, advancements in computational social science may put the public at increased risk. If society is to be protected, then there should be legal and ethical limitations on how social media as a computational social science tool can be used. Citing the sentiments of the CEO of Nasdaq, Mark Zuckerberg, to take risks is to “move fast and break things.” Oboler et al. [17] argue for external constraints to protect society from the cost of mistakes by social media innovators.

The risk posed by the capacity of computational social science tools and the explosion in the corpus of data, free of the ethical constraints placed on researchers, raises serious questions about the impact that those who control the data and the tools can have on society. Many social media companies are driven by online advertising revenue which places the individual’s interest in privacy in conflict with the interest of advertisers in extensive customer profiling. This is facilitated by Web 2.0 sites which have higher advertising rates for advertisers to target selected users. The increase in advertisers targeting specific users highlights unexpected consequences. For example, Target analyzed purchasing patterns to identify potential customers of baby paraphernalia. The analysis, based on purchasing history of unrelated items, highlighted potential pregnancies with a high degree of accuracy. Target sent advertising material to its target market, creating an angry response from a father whose teenage daughter received the advertising, but not knowing at the time that she was really pregnant. The daughter was forced to confirm her pregnancy because the retailer targeted her in the marketing as a result of the data analysis. This does raise ethical concerns regarding privacy of information [10].

Further, social media sites can scrape a great deal of data from users’ age and sexual preference to target advertisements for adult products, which can cause

distress or unauthorized disclosure of sensitive personal information. This poses both technical and ethical questions like: “Is any technically possible use of personal data ethically acceptable?” [17]. Social networking companies and advertisers need to consider such critical questions. By limiting data acquisition, sharing, and use and by raising public awareness of the implications of its availability through ethical considerations, the risk of abuse can be controlled.

13.3.3 Serving Self-Interest

Large volumes of data are not necessarily representative and reliable to solve problems relating to public interest. Big Data users can exploit Big Data with no regard for data quality, legality, data meaning, and process quality. For example, in 2011, the Rainforest Action Network in the USA discovered that the American Petroleum Institute and its oil lobby allies were able to manipulate social media opinion to show support for a pipeline project to carry oil from Canada to Texas by using fake Twitter accounts to send large numbers of tweets to show support for the project, which falsely represented public opinion. The Rainforest Action Network (RAN) discovered that 14 of 15 accounts were faked and the tweets were generated by an automated process [6].

Desouza and Smith [6] cite the case when public agencies and a newspaper in New York released information about gun owners after the Connecticut school mass shooting. Published information on the names and addresses of licensed gun owners living in the neighborhood can be used by the wrong people like criminals to target vulnerable homeowners who do not own guns or to target homeowners who have guns in order to steal them. According to Oboler, Welsh, and Cruz [17], methodology and ethics, drawn from the underlying fields of computational and social sciences, need to be considered. The authors argue that considerations apply not only to the research context but also to the worlds of government and commerce where philosophical concerns are less likely to counter immediate practical benefits. Most significantly, these concerns need to be considered in the context of social media platforms which have become computational social science tools that are easily accessible to businesses, governments, private citizens, and the platform operators themselves. Governments can exercise their power when they see social media acting against their interests. One such example was when the US government asked Twitter through a court order for data on WikiLeaks founder Julian Assange and those connected to him.

Social media can also promote the agenda of governments. Another example was when the Egyptian government cut off the Internet during the 2011 riots, after it realized that the US government provided training through the Internet to influence social change among Egyptian dissidents. Computational social science tools together with social media data can be used to reconstruct the movements of activists, to locate dissidents, and to map their networks. Governments and their security services have a strong interest in this activity [17].

13.3.4 *Moving Away from Comfort Metrics*

The analysis of general trends and the profiling of individuals can be investigated through social sciences. In this regard, Kettleborough [11] argues against “comfort metrics,” whereby data that is not relevant and focuses merely on the process is collected. There is a need to look at data differently and to be prepared to throw away many old beliefs like forced ranking, which looks at normal distribution to employee performance, as a good employee performance measurement tool. According to Kettleborough [11], there is evidence that these approaches have actually damaged organizations like Microsoft’s *lost decade*, which was as a direct result of misplaced or misunderstood data techniques.

Research studies have also shown that many businesses were preselecting and filtering candidates for employment based on social media. If job applicants do not protect their reputation online, then this can compromise their applications, if the interviewer has access to negative attributes posted online. Currently, business and government control large volumes of data used for computational social science analysis. The capacity to collect and analyze datasets on a vast scale provides the magnitude to disclose patterns of individual and group behavior. The potential damage from inappropriate disclosure of information is sometimes obvious. A lack of transparency in the way data is analyzed and aggregated, combined with a difficulty in predicting which pieces of information may later prove damaging, means that many individuals have little knowledge of potential adverse effects of the expansion in computational social science [17].

If data is not correctly understood, then massive mistakes can cause harm. Many employment recruiters are already “looking” at the social media lives of job applicants. Although it is seen as justifiable, especially using work-related sites such as LinkedIn, in cases where employers seek data from nonwork-related social media life, there could be some potentially negative consequences. According to the CIPD, “using social media in recruitment or as part of career progression carries the risk of a number of different claims if a candidate is not appointed as a result of information gleaned.” These include the following [11]:

- A breach of the Human Rights Act 1998 (incorporating Article 8 of the European Convention on Human Rights) to respect private and family life.
- A breach of the Data Protection Act 1998, which states that data controllers such as prospective employers should not hold excessive information and should process information in a fair way.
- It has been suggested that the over 50s age group will be more cautious with their social media presence than the under 30s, resulting in more potential for negative recruitment decisions for younger people.
- Information about marital status, number of children, and sexual orientation may incorrectly influence a selection decision.
- Information about physical or mental state, such as revealing depression to friends on social media, may be a disadvantage.

If life-changing judgments are to be made about people, then quality and accuracy must be beyond reproach. If employees are hired, promoted or dismissed based on Big Data discrimination, then there can be legal implications [11].

13.3.5 Using Power to Leverage Outcomes

Governments and powerful data-rich companies have the financial support and powerful resources to access data. Such organizations, by their nature, tend at times to assume that the risk of unjustified impacts on individuals is of little consequence when compared with the potential to avert perceived calamities [20]. It is easy to manipulate people, like using computational social science to guide political or product advertising, selling messages that people will favor or withhold information that may compromise support. Google, for example, can sway an election by predicting messages that would engage an individual voter (positively or negatively) and then disseminate content to influence that user's vote. The predictions could be highly accurate by making use of a user's e-mail in their Google-provided Gmail account, their search history, and social network connections. The dissemination of information could include "recommended" videos on YouTube to highlight where one political party agrees with the user's views – also articles in Google News could be given higher visibility to help sway voters into making *the right choice* [17]. Further, this can be complemented with negative messages to appear to create a balance, but in reality may have little or no impact. Such manipulation may not appear obvious, yet powerful to achieve the outcomes of the manipulator.

13.3.6 Risks Relating to Social Media Platforms

Social media platforms have added to their data either by acquiring other technology companies as Google did when acquiring YouTube or by moving into new fields as Facebook did when it created "Facebook Places" providing a geolocation service which generates high value information [14]. The value of information can be maximized by using a primary key that connects this data with existing information like a Facebook user ID or a Google account name, where a user is treated as a single user across all products of the company [14]. One account can connect to various types of online interactions, exposing greater breadth of a user's profile. In such a case all the data is immediately related and available to any query companies like Facebook and Google may have. This can be alarming as there is little privacy, since any information can be collected across platforms about users.

Accounts that are identity-verified, frequently updated, and used across multiple aspects of a person's life present the richest data and pose the greatest risk. For example, Facebook's Timeline feature allows users to mine social interactions that

had long been buried. Further, since Timeline is not an option in Facebook, masses of personal data can be held. Another challenge was the Beacon software which, developed by Facebook, connected people's purchases to their Facebook account. It indicated what users had purchased, where they got it, and whether they got a discount. It was eventually closed in view of legal, privacy, and ethical considerations. Further, the emergence of massive open online courses under MOOCs is now causing a stir in the world of Big Data with evidence that student details, including performance data, is being sold online. Recruiters looking for highly motivated candidates with wisdom can hunt potential candidates on this platform [11].

13.3.7 Research Methodology Challenges

The use of social media as a source of social research data can present various methodological challenges. There can be sampling bias which can distort findings, as any particular social medium is not representative of the population as a whole. Avoiding sampling bias in social media sources is a great challenge for researchers in the social sciences. If computational tools are to be appropriately used in social research, then it is important that users are aware of the strengths and weaknesses of such tools. Therefore, it is vital that the capacity of social researchers in developing skills relating to computational methods and tools is developed, so that they can decide when and how to apply them responsibly.

Anonymity among employees during surveys is also causing concern. While it is believed that employee responses will be more truthful if they remain anonymous, their identity can be traced from the demographic details in their social media profiles. If used incorrectly, this "honest data" could be turned against the employees [11]. Employees who are aware of this may not necessarily give the "correct picture."

In terms of reliability and validity, decisions cannot be made with incomplete or incomprehensive data. Rational and fair decisions have to be based on representation. For example, if 20 people are happy with the service at a state hospital, this does not exhibit behavior that is statistically significant for the whole population. As the old adage states, "one swallow does not a summer make." Data focusing on a few does not paint a correct picture. Further, Kettleborough [11] contends that correlation and causation should not be analyzed at face value, since if two items correlate, it does not mean that one causes the other.

Data mining or scrapping of social media sites can result in personal data being used against individuals, even if it has been cleaned to remove personal references. One such example is a study by researchers from the Université Catholique de Louvain in Belgium who identified "95% of the unique users by analyzing only four GPS time and location stamps per person." In addition, researchers at Carnegie Mellon University were able to create a system to uncover Social Security numbers

from birthday and hometown information listed on social networking sites like Facebook [11]. Large amounts of data can become the target of the unscrupulous.

13.3.8 Securing Big Data

Big Data, while sourcing data from multiple sources, relies on data that is available. Further, such data must be secured. The challenge is how the data is collected and stored. This raises security issues like internal employees adhering to confidentiality policies. Cases of storage abuses have occurred at Facebook sites. Data can also be lost due to hackers and employees. One example is the two Aviva employees who sold details of people who had accidents to claims companies. The fraud flag was raised when the claimants received calls from firms persuading them to take personal injury claims [11]. Information that is not secured can be used for blackmailing or espionage.

13.3.9 Limitations of Addressing Social Problems

In the social arena, a major gap exists between the potential of data-driven information and its actual use in addressing social problems. Certain social problems can be easily solved using Big Data, such as weather forecasting and areas with high disease rates. However, pandemic problems like drug trafficking and unemployment cannot be easily resolved in a sustainable way with Big Data. According to Desouza and Smith [6], these evil problems are more dynamic and complex than their technical counterparts, because of the diversity of stakeholders involved and the numerous feedback loops among the interrelated constituents. Government agencies and nonprofits are involved in tackling these problems but face the following challenges: limited cooperation and data sharing among them; inadequate information technology resources; their counterparts in the hard sciences work on technical problems or in business who have ready access to financial, product, and customer information; missing and incomplete data; and data stored in silos or in forms that are inaccessible to automated processing. In addition, there are regulatory constraints like policy relating to data sharing agreements, privacy and confidentiality of data, and collaboration protocols among various stakeholders tackling the same type of problem. While various agencies may invest in data technologies, the return on investment for solving social problems is yet to be convincing. This impact on the need is to be provided with information and advice via sources that they can trust in a more timely way.

13.4 Imperatives for Big Data Use from Social Media

In decision-making, context is the key; therefore knowledge of the domain such as social media is crucial. Analysis of Big Data is directly linked to decision-making, which has to be supported by very intricate techniques using wide and deep extensive data sources as shown in Table 13.1. Big Data is about massive amounts of different types of observational data, supporting different types of decisions and decision time frames. According to Goes [9], analytics moves from data to information to knowledge and finally to intelligence. The generation of knowledge and intelligence to support decision-making is critical as the Big Data world is moving toward real-time or close to real-time decision-making. Therefore, the need for context-dependent methodologies that strengthen prediction is pivotal for effective data analysis.

Big Data analytics from social media has to consider the tools, software, and the data to ensure quality results. While the technical ability may exist to gather data, the analytical capacity to draw meaning from such data needs to be developed. For example, visualization can be produced from real-time information as datasets emerge from user activity. However, such visualizations can only be considered powerful representations if visualization specialists are aware of which relationships benefit users. This requires an understanding of how meaning can be created through and across various datasets in social media platforms.

13.4.1 Responsibility of Analytic Role Players

Big Data has enormous potential to inform decision-making to help solve the world's toughest social problems. But for this to happen, issues relating to data collection, organization, and analysis must first be resolved. Much of this responsibility lies with the major analytic players, as shown in Table 13.2, who offer valuable services that help users cope with using Big Data effectively.

The aforementioned major analytic players have to ensure effective use of Big Data from social media platforms. This requires prudent use of analytic tools, incorporating the following guidelines [7]:

Table 13.1 Big data analytics

Decisions	Analytics	Techniques
Real time	Visualization	Statistics
Close to RT	Exploration	Econometrics
Hourly	Explanatory	Machine learning
Weekly	Predictive	Computational
Monthly		Linguistics
Yearly		Optimization
		Simulation

Adapted from [9]

Table 13.2 Major analytic players

Firm	Products	Website
SAP	HANA; Applied analytics	http://www.sap.com/solutions/technology/in-memory-computingplatform/hana/overview/index.epx ; http://www54.sap.com/solutions/analytics/applications/software/overview.html
Splunk	Splunk Hadoop Connect, Splunk enterprise	http://www.splunk.com/
Tibco	Spotfire	http://spotfire.tibco.com/
Google	Big query	https://cloud.google.com/products/big-query
IBM	Pure data system	http://www-01.ibm.com/software/data/puredata
SAS	High-performance analytics	http://www.sas.com/software/high-performance-analytics/index.html
Metamarkets	Data pipes; Druid	http://metamarkets.com/platform
Oracle	Big data appliance; Exadata; Exalytics	http://www.oracle.com/us/technologies/big-data/index.html
Tableau	Tableau	http://www.tableausoftware.com/
Cloudera	Hbase	http://www.cloudera.com/content/cloudera/en/products/clouderaenterprise-core.html
Qliktech	QlikView	http://www.qlikview.com/
GoodData	GoodData bashes	http://www.gooddata.com/what-is-gooddata/

Adapted from [7]

- Use of in-memory database technology that avoids resources swapping databases between the storage medium and the memory, but rather operating within memory with only limited accessing of alternative storage mediums.
- The sheer size and complexity of data cannot be handled by traditional technologies built on relational or multidimensional databases, as there is a need to have flexibility to have questions answered in real time.
- Use caves of data from unstructured data to improve service levels, reduce operations costs, mitigate security risks, and enable compliance.
- Use tools that break down traditional data silos and attain operational intelligence that benefits both IT and the business, which is valuable for capturing machine-generated data into a system that would provide operational real-time data.
- The need to use technology to efficiently store, manage, and analyze unlimited amounts of data that can process any type of data differently than relational databases.

Since information in the various social media platforms is not static, if it is not updated and cleaned, then “dirty data” will arise. Considering the garbage in, garbage out syndrome, poor quality data will produce poor results.

13.4.2 Evidence-Based Decision-Making

In addition, the following four recommendations have the potential to create datasets useful for evidence-based decision-making [6]:

- The global community needs to create large data banks on critical issues like homelessness and malnutrition, which must have the capacity to hold multiple different data types along with metadata that describes the datasets. This requires multi-sector alliances that promote and create data sharing on sectoral issues. At the 2012 G-8 Summit, leaders committed to the New Alliance for Food and Nutrition Security to help 50 million people out of poverty over the next 10 years through sustained agricultural growth. This is supported by a number of databases like Agrilinks.org, Feed the Future Initiative website, and Women's Empowerment in Agriculture Index.
- Citizens and professionals can help create and analyze these datasets. With the growth of data through open data platforms, citizens are creating new ideas and products through what has become known as "citizen science." A bike map and map of the London tube were created by citizens, using the raw data from the London Datastore which is managed by the Greater London Authority.
- Big Data cannot be left to the pure sciences and business, but needs analysts in the social sciences to be statistically equipped to collect data for large-scale datasets. Skills in data organization, preservation, visualization, search, and retrieval, identifying networked relationships among datasets, and how to uncover latent patterns in datasets need to be developed. These are valuable skills that go beyond simply searching the web for information.
- Virtual experimentation platforms which allow individuals to interact with different ideas and work collaboratively to find solutions to problems can create large datasets, develop innovative algorithms to analyze and visualize the data, and develop new knowledge for tackling social challenges. The use of open forums such as wikis and discussion groups can help the community share lessons learned, collaborate, and advance new solutions.

In addition, Oboler et al. [17] argue that social networking has provided a diverse range of datasets covering large sections of the population, granting researchers, governments, and businesses the powerful ability to identify trends in behavior among a large population and to find vast quantities of information on an individual user. As the industry develops, social media computational tools will increase the scope, accuracy, and usefulness of such datasets. In view of the ethical and privacy implications, regulatory barriers restricting the collection, retention, and use of personal information require consideration. While laws protect human rights, there is a need for greater protection of the customer.

13.4.3 Protection of Rights

The rights of users need to be protected, as social media platform providers and various agencies provide innovative services to targeted users. The debate is

whether consumers are protected from preventable harm only after proving damage or are rules set by law. In the first approach, advertisers have more freedom to mine data from various social media platforms, data over which the user has no control especially if it is outdated or hacked by third parties. The safeguarding of personal rights and freedoms is more favored through the setting of regulations and laws. This would place the burden on social media to restrict the storage, accessibility, and manipulation of data in ways that limit its usefulness. This can prevent unscrupulous use of data. However, this will require legislators to use multilateral legislating, since websites can freely choose the physical location of their hosting infrastructure where there is least regulation.

The ethical barriers for data use in the social sciences are much higher than pure science research as the data collection of personal information is higher in the social sciences. Oboler et al. [17] illustrate some suggestions to manage ethical use of social media data, as given below:

- In keeping with the code of ethics developed by professional bodies, example for engineers, these should be applied to social media as well. Such guidelines commit members to act in public interest, by not causing harm or violating the privacy of others. Social media platforms are a form of computational social science which requires recognition of the ethical concerns in the social sciences. This can reduce the opportunity for the abuse of a very powerful tool. Users of social media have an ethical responsibility to one another.
- A code of conduct for producers and consumers of online data which can highlight the issues to be considered when publishing information. For example, when a Twitter user uploads photographs, their action may reveal information about others in their network; the impact on those other people should be considered under a producer's code of ethics. A consumer code of ethics is also needed; such a code would cover users viewing information posted by others through a social media platform. A consumer code could raise questions of when it is appropriate to further share information, for example, by retweeting it.
- Guidelines for principles of engagement can help users determine what they are publishing and to create awareness of the potential impact of publishing information. The power of social media can be used to warn the owner when the content may pose a risk, especially when accesses open.
- A cultural mind shift is needed to become more forgiving of information exposed through social media, an acceptance that social media profiles are private and must be locked down with more intricate filters and used only in certain settings.

The aforementioned suggestions would change the nature of social media as a computational social science tool, by filtering what should be included out of the tools field of observation. As an instrument-based discipline, the way the field is understood can be changed either by changing the nature of the tool or by changing the way we allow it to be used.

13.4.4 Knowing the Context

Understanding and knowledge of the context is fundamental. For example, marketing depends, to a large extent, on information technology accruing from social networks. Researchers have to master the collection and analysis of web data and user-generated content, using advanced techniques. This is necessitated by the massive amounts of observational data, of different types, supporting different types of decisions and decision time frames [9]. In this regard, if researchers want to explain the growth in online shopping among teenagers in developing economies using social media networks, then it is imperative to use models like longitudinal models, latent models, and structured models to explain the causes within the context. Since the Big Data environment is targeting real-time decision-making, it is imperative that tools employed to analyze social media networks use context-dependent methodologies that enhance prediction in a valid and reliable way. The reason being that not only are the networks intricate but also require knowledge of the complex models. Consideration of these dynamics can produce valuable information from Big Data, allowing modeling of individuals at a very detailed level with a rich proliferation of the environment surrounding them [9].

Big Data analytics has the ability to yield deeper insights and predictions about the individuals. According to Waterman and Bruening [22], even though data may be processed accurately, the results may have profound effect on personal life choices. The authors argue that understanding the sources and limitations of data is critical to mitigate harm to individuals. This necessitates understanding and responding to the implications of choices about data and data analytic tools, integrity of analytic processes, and the consequences of applying the outcomes of analytic models to information about individuals [22].

13.5 Conclusion

The proliferation of Big Data has emerged as the new frontier in the wide arena of IT-enabled innovations and opportunities. Attention has focused on how to harness and analyze Big Data. Social media is one component of a larger dynamic and complex information domain, and their interrelationships need to be recognized. As the connection with Big Data grows, we cannot avoid its impact. Without being familiar with the data, the benefits of Big Data cannot be reaped. Large volumes of data cannot be analyzed using conventional media research methods and tools. The current Big Data analytics trend has seen the tools used to analyze and visualize data getting continuously better. There has been a major investment in the development of more powerful digital infrastructure and tools to tackle new and more complex and interdisciplinary research challenges.

Current programs have seen companies like Splunk, GoodData, and Tibco providing services to allow their users to benefit from Big Data. Users with the

ability to query and manipulate Big Data can achieve actionable information from Big Data to derive growth by making informed decisions. Access to data is critical. However, several issues require attention in order to benefit from the full potential of Big Data. Policies dealing with security, intellectual property, privacy, and even liability will need to be addressed in the Big Data environment. Organizations need to institutionalize the relevant talent, technology, structure workflows, and incentives to maximize the use of Big Data.

It is imperative that apart from the power users in marketing, financial, healthcare, science, and technical fields, those involved in daily decision-making must be empowered to use analytics. As more and more analytical power reaches decision-makers, enhanced and more accurate decision-making will emerge in the future. While there is a need to seize the opportunities offered by continuing advances in computational techniques for analyzing social media, the effective use of human expertise cannot be ignored. Using the right data in the right way and for the right reasons to innovate, compete, and capture value from deep and real-time Big Data information can change lives for the better. Big Data has to be used discriminately and transparently.

References

1. Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc* 15(5):662–679
2. Brown B, Chui M, Manyika J (2011) Are you ready for the era of ‘big data’? *McKinsey Q* 4:24–35
3. Bughin J, Chui M, Manyika J (2010) Clouds, big data, and smart assets: ten tech-enabled business trends to watch. *McKinsey Q* 56(1):75–86
4. Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from big data to big impact. *MIS Q* 36(4):1165–1188
5. Colgren D (2014) The rise of crowdfunding: social media, big data, cloud technologies. *Strategic Financ* 2014:55–57
6. Desouza KC, Smith KL (2014) Big data for social innovation. *Stanf Soc Innov Rev* 2014:39–43
7. Fanning K, Grant R (2013) Big data: implications for financial managers. *J Corporate Account Financ* 2013:23–30
8. Gandomi A, Haider M (2014) Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manag* 35:137–144
9. Goes PB (2014) Big data and IS research. *MIS Q* 38(3):iii–viii
10. Hill K (2012) How target figured out a teen girl was pregnant before her father did. *Forbes* (16 February). <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>. Accessed 9 Dec. 2014
11. Kettleborough J (2014) Big data. *Train J* 2014:14–19
12. Kim HJ, Pelaez A, Winston ER (2013) Experiencing big data analytics: analyzing social media data in financial sector as a case study. *Northeast Decision Sciences Institute Annual Meeting Proceedings*. Northeast Region Decision Sciences Institute (NEDSI), April 2013, 62–69
13. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2013) Big data, analytics and the path from insights to value. *MIT Sloan Manag Rev* 21:40–50

14. McCarthy J (2010) Blended learning environments: using social networking sites to enhance the first year experience. *Australas J Educ Technol* 26(6):729–740
15. McKelvey K, Rudnick A, Conover MD, Menczer F (2012) Visualizing communication on social media: making big data accessible. *arXiv* (3):10–14
16. Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we RT. In: 1st workshop on Social Media Analytics (SOMA '10). ACM Press, Washington, DC
17. Oboler A, Welsh K, Cruz L (2012) The danger of big data: social media as computational social science. *First Monday* 17(7):60–65
18. Proctera R, Visb F, Voss A (2013) Reading the riots on Twitter: methodological innovation for the analysis of big data. *Int J Soc Res Methodol* 16(3):197–214
19. Qualman E (2012) *Socialnomics: how social media transforms the way we live and do business*. Wiley, New York
20. Wigan MR, Clarke R (2013) Big data's big unintended consequences. *IEEE Comput Soc* 2013:46–53
21. Young SD (2014) Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends Microbiol* 22(11):601–602
22. Waterman KK, Bruening PJ (2014) Big data analytics: risks and responsibilities. *Int Data Privacy law* 4(2):89–95