Socio-Affective Computing 5

Erik Cambria
Dipankar Das
Sivaji Bandyopadhyay
Antonio Feraco *Editors*

A Practical Guide to Sentiment Analysis

Springer

Socio-Affective Computing

Volume 5

Series Editor
Amir Hussain, University of Stirling, Stirling, UK

Co-Editor
Erik Cambria, Nanyang Technological University, Singapore

This exciting Book Series aims to publish state-of-the-art research on socially intelligent, affective and multimodal human-machine interaction and systems. It will emphasize the role of affect in social interactions and the humanistic side of affective computing by promoting publications at the cross-roads between engineering and human sciences (including biological, social and cultural aspects of human life). Three broad domains of social and affective computing will be covered by the book series: (1) social computing, (2) affective computing, and (3) interplay of the first two domains (for example, augmenting social interaction through affective computing). Examples of the first domain will include but not limited to: all types of social interactions that contribute to the meaning, interest and richness of our daily life, for example, information produced by a group of people used to provide or enhance the functioning of a system. Examples of the second domain will include, but not limited to: computational and psychological models of emotions, bodily manifestations of affect (facial expressions, posture, behavior, physiology), and affective interfaces and applications (dialogue systems, games, learning etc.). This series will publish works of the highest quality that advance the understanding and practical application of social and affective computing techniques. Research monographs, introductory and advanced level textbooks, volume editions and proceedings will be considered.

More information about this series at http://www.springer.com/series/13199

Erik Cambria • Dipankar Das
Sivaji Bandyopadhyay • Antonio Feraco
Editors

A Practical Guide to Sentiment Analysis

Springer

Editors
Erik Cambria
School of Computer Science
and Engineering
Nanyang Technological University
Singapore, Singapore

Dipankar Das
Computer Science
and Engineering Department
Jadavpur University
Kolkata, India

Sivaji Bandyopadhyay
Computer Science
and Engineering Department
Jadavpur University
Kolkata, India

Antonio Feraco
Fraunhofer IDM@NTU
Nanyang Technological University
Singapore, Singapore

ISSN 2509-5706 ISSN 2509-5714 (electronic)
Socio-Affective Computing
ISBN 978-3-319-55392-4 ISBN 978-3-319-55394-8 (eBook)
DOI 10.1007/978-3-319-55394-8

Library of Congress Control Number: 2017938021

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

Preface

While sentiment analysis research has become very popular in the past ten years, most companies and researchers still approach it simply as a polarity detection problem. In reality, sentiment analysis is a “suitcase problem” that requires tackling many natural language processing (NLP) subtasks, including microtext analysis, sarcasm detection, anaphora resolution, subjectivity detection, and aspect extraction. In this book, we propose an overview of the main issues and challenges associated with current sentiment analysis research and provide some insights on practical tools and techniques that can be exploited to both advance the state of the art in all sentiment analysis subtasks and explore new areas in the same context.

In Chap. 1, we discuss the state of the art of affective computing and sentiment analysis research, including recent deep learning techniques and linguistic patterns for emotion and polarity detection from different modalities, e.g., text and video.

In Chap. 2, Bing Liu describes different aspects of sentiment analysis and different types of opinions. In particular, he uses product reviews as examples to introduce general key concepts and definitions that are applicable to all forms of formal and informal opinion text and all kinds of domains including social and political domains.

In Chap. 3, Jiwei Li and Eduard Hovy describe possible directions for deeper understanding about what opinions or sentiments are, why people hold them, and why and how their facets are chosen and expressed, helping bridge the gap between psychology/cognitive science and computational approaches.

In Chap. 4, Saif Mohammad discusses different sentiment analysis problems and the challenges that are to be faced in order to go beyond simply determining whether a piece of text is positive, negative, or neutral. In particular, the chapter aims to equip researchers and practitioners with pointers to the latest developments in sentiment analysis and encourage more work in the diverse landscape of problems, especially those areas that are relatively less explored.

In Chap. 5, Aditya Joshi, Pushpak Bhattacharyya, and Sagar Ahire contrast the process of lexicon creation for a new language or a resource-scarce language from a resource-rich one and, hence, show how the produced sentiment resources can be exploited to solve classic sentiment analysis problems.

In Chap. 6, Hongning Wang and ChengXiang Zhai show how generative models can be used to integrate opinionated text data and their companion numerical sentiment ratings, enabling deeper analysis of sentiment and opinions to obtain not only subtopic-level sentiment but also latent relative weights on different subtopics.

In Chap. 7, Vasudeva Varma, Litton Kurisinkel, and Priya Radhakrishnan present an overview of general approaches to automated text summarization with more emphasis on extractive summarization techniques. They also describe recent works on extractive summarization and the nature of scoring function for candidate summary.

In Chap. 8, Paolo Rosso and Leticia Cagnina describe the very challenging problems of deception detection and opinion spam detection, as lies and spam are becoming increasingly serious issues with the rise, both in size and importance, of social media and public opinion.

Finally, in Chap. 9 Federica Bisio et al. describe how to enhance the accuracy of any algorithm for emotion or polarity detection through the integration of commonsense reasoning resources, e.g., by embedding a concept-level knowledge base for sentiment analysis.

Singapore, Singapore — Erik Cambria
Kolkata, India — Dipankar Das
Kolkata, India — Sivaji Bandyopadhyay
Singapore, Singapore — Antonio Feraco

Contents

Chapter 1
Affective Computing and Sentiment Analysis

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco

Abstract Understanding emotions is one of the most important aspects of personal development and growth and, as such, it is a key tile for the emulation of human intelligence. Besides being a important for the advancement of AI, emotion processing is also important for the closely related task of polarity detection. The opportunity automatically to capture the sentiments of the general public about social events, political movements, marketing campaigns, and product preferences, in fact, has raised increasing interest both in the scientific community, for the exciting open challenges, and in the business world, for the remarkable fallouts in marketing and financial market prediction. This has led to the emerging fields of affective computing and sentiment analysis, which leverage on human-computer interaction, information retrieval, and multimodal signal processing for distilling people's sentiments from the ever-growing amount of online social data.

Keywords Affective computing • Sentiment analysis • Five eras of the Web • Jumping NLP curves • Hybrid approaches

1.1 Introduction

Emotions play an important role in successful and effective human-human relationships. In fact, in many situations, human 'emotional intelligence' is more important than IQ for successful interaction (Pantic et al. 2005). There is also significant evidence that rational learning in humans is dependent on emotions (Picard 1997).

E. Cambria (✉)
School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore, Singapore
e-mail: cambria@ntu.edu.sg

D. Das • S. Bandyopadhyay
Computer Science and Engineering Department, Jadavpur University, 700032, Kolkata, India
e-mail: das@cse.jdvu.ac.in; sbandyopadhyay@cse.jdvu.ac.in

A. Feraco
Fraunhofer IDM@NTU, Nanyang Technological University, Singapore, Singapore
e-mail: antonio.feraco@fraunhofer.sg

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_1

Affective computing and sentiment analysis, hence, are key for the advancement of AI (Minsky 2006) and all the research fields that stem from it. Moreover, they find applications in several different scenarios and there is a good number of companies, large and small, that include the analysis of emotions and sentiments as part of their mission. Sentiment mining techniques can be exploited for the creation and automated upkeep of review and opinion aggregation websites, in which opinionated text and videos are continuously gathered from the Web and not restricted to just product reviews, but also to wider topics such as political issues and brand perception.

Affective computing and sentiment analysis have also a great potential as a sub-component technology for other systems. They can enhance the capabilities of customer relationship management and recommendation systems allowing, for example, to find out which features customers are particularly happy about or to exclude from the recommendations items that have received very negative feedbacks. Similarly, they can be exploited for affective tutoring and affective entertainment or for troll filtering and spam detection in online social communication.

Business intelligence is also one of the main factors behind corporate interest in the fields of affective computing and sentiment analysis. Nowadays, companies invest an increasing amount of money in marketing strategies and they are constantly interested in both collecting and predicting the attitudes of the general public towards their products and brands. The design of automatic tools capable to mine sentiments over the Web in real-time and to create condensed versions of these represents one of the most active research and development areas. The development of such systems, moreover, is not only important for commercial purposes, but also for government intelligence applications able to monitor increases in hostile communications or to model cyber-issue diffusion.

Several commercial and academic tools, e.g., IBM,[1] SAS,[2] Oracle,[3] SenticNet[4] and Luminoso,[5] track public viewpoints on a large-scale by offering graphical summarizations of trends and opinions in the blogosphere. Nevertheless, most commercial off-the-shelf (COTS) tools are limited to a polarity evaluation or a mood classification according to a very limited set of emotions. In addition, such methods mainly rely on parts of text in which emotional states are explicitly expressed and, hence, they are unable to capture opinions and sentiments that are expressed implicitly. Because they are mainly based on statistical properties associated with words, in fact, many COTS tools are easily tricked by linguistic operators such as negation and disjunction.

The remainder of this chapter lists common tasks of affective computing and sentiment analysis and presents a general categorization for them, after which some concluding remarks are proposed.

[1] http://ibm.com/analytics

[2] http://sas.com/social

[3] http://oracle.com/social

[4] http://business.sentic.net

[5] http://luminoso.com

1.2 Common Tasks

The Web is evolving towards an era where communities will define future products and services.[6] In this context, big social data analysis (Cambria et al. 2014) is destined to attract increasing interest from both academia and business (Fig. 1.1).

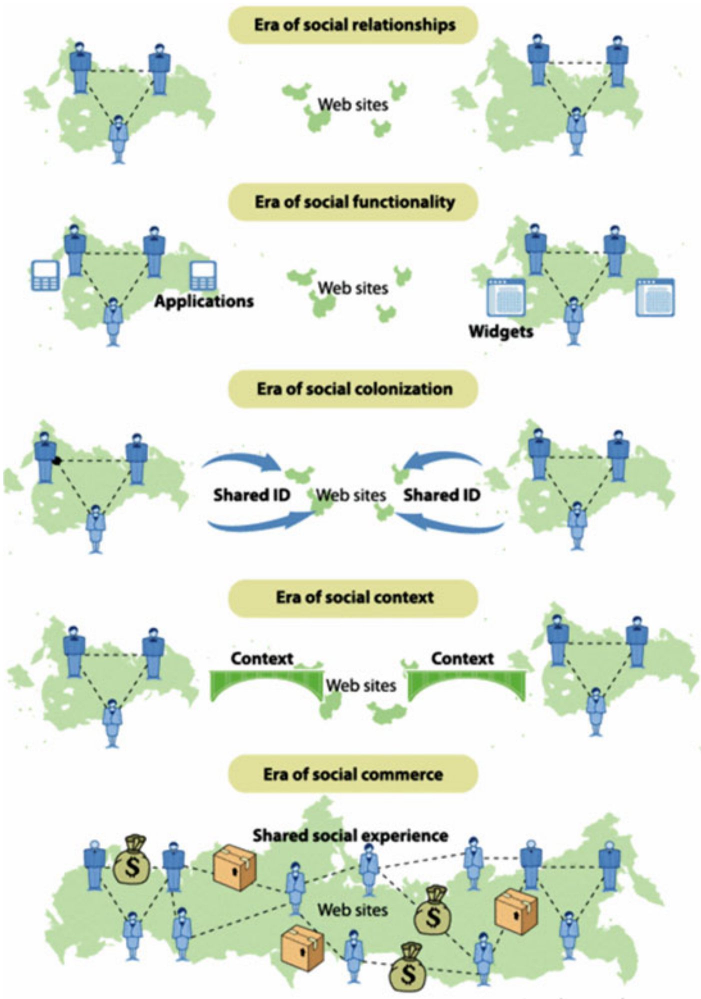


Fig. 1.1 Owyang's Five-Eras vision shows that mining sentiments from the general public is becoming increasingly important for the future of the Web

[6] http://web-strategist.com/blog/2009/04/27

The basic tasks of affective computing and sentiment analysis are emotion recognition (Picard 1997; Calvo and D'Mello 2010; Zeng et al. 2009; Schuller et al. 2011; Gunes and Schuller 2012) and polarity detection (Pang and Lee 2008; Liu 2012; Wilson et al. 2005; Cambria 2016). While the former focuses on extracting a set of emotion labels, the latter is usually a binary classification task with outputs such as 'positive' versus 'negative', 'thumbs up' versus 'thumbs down' or 'like' versus 'dislike'. These two tasks are highly inter-related and inter-dependent to the extent that some sentiment categorization models, e.g., the Hourglass of Emotions (Cambria et al. 2012), treat it as a unique task by inferring the polarity associated to a sentence directly from the emotions this conveys. In many cases, in fact, emotion recognition is considered a sub-task of polarity detection.

Polarity classification itself can also be viewed as a subtask of more advanced analyses. For example, it can be applied to identifying 'pro and con' expressions that can be used in individual reviews to evaluate the pros and cons that have influenced the judgements of a product and that make such judgements more trustworthy. Another instance of binary sentiment classification is agreement detection, that is, given a pair of affective inputs, deciding whether they should receive the same or differing sentiment-related labels.

Complementary to binary sentiment classification is the assignment of degrees of positivity to the detected polarity or valence to the inferred emotions. If we waive the assumption that the input under examination is opinionated and it is about one single issue or item, new challenging tasks arise, e.g., subjectivity detection, opinion target identification, and more (Cambria et al. 2015). The capability of distinguishing whether an input is subjective or objective, in particular, can be highly beneficial for a more effective sentiment classification. Moreover, a record can also have a polarity without necessarily containing an opinion, for example a news article can be classified into good or bad news without being subjective.

Typically, affective computing and sentiment analysis are performed over on-topic documents, e.g., on the result of a topic-based search engine. However, several studies suggested that managing these two task jointly can be beneficial for the overall performances. For example, off-topic passages of a document could contain irrelevant affective information and result misleading for the global sentiment polarity about the main topic. Also, a document can contain material on multiple topics that may be of interest to the user. In this case, it is therefore necessary to identify the topics and separate the opinions associated with each of them.

Similar to topic detection is aspect extraction, a subtask of sentiment analysis that consists in identifying opinion targets in opinionated text, i.e., in detecting the specific aspects of a product or service the opinion holder is either praising or complaining about. In a recent approach, Poria et al. (2016) used a 7-layer deep convolutional neural network to tag each word in opinionated sentences as either aspect or non-aspect word and developed a set of linguistic patterns for the same purpose in combination with the neural network.

Other sentiment analysis subtasks include aspect extraction (Poria et al. 2016), subjectivity detection (Chaturvedi et al. 2016), concept extraction (Rajagopal et al.

2013), named entity recognition (Ma et al. 2016), and sarcasm detection (Poria et al. 2016), but also complementary tasks such as personality recognition (Poria et al. 2013), user profiling (Mihalcea and Garimella 2016) and especially multimodal fusion (Poria et al. 2016). With increasing amounts of webcams installed in end-user devices such as smart phones, touchpads, or netbooks, there is an increasing amount of affective information posted to social online services in an audio or audiovisual format rather than on a pure textual basis. For a rough impression on the extent, consider that two days of video material are uploaded to YouTube on average per minute. Besides speech-to-text recognition, this allows for additional exploitation of acoustic information, facial expression and body movement analysis or even the "mood" of the background music or the color filters, etc.

Multimodal fusion is to integrate all single modalities into a combined single representation. There are basically two types of fusion techniques that have been used in most of the literature to improve reliability in emotion recognition from multimodal information: feature-level fusion and decision-level fusion (Konar and Chakraborty 2015). The authors in Raaijmakers et al. (2008) fuse acoustic and linguistic information. Yet, linguistic information is based on the transcript of the spoken content rather than on automatic speech recognition output. In Morency et al. (2011), acoustic, textual, and video features are combined for the assessment of opinion polarity in 47 YouTube videos. A significant improvement is demonstrated in a leave-one-video-out evaluation using Hidden-Markov-Models for classification. As relevant features the authors identify polarized words, smile, gaze, pauses, and voice pitch. Textual analysis is, however, also only based on the manual transcript of spoken words.

In Poria et al. (2016), finally, the authors propose a novel methodology for multimodal sentiment analysis, which consists in harvesting sentiments from Web videos by demonstrating a model that uses audio, visual and textual modalities as sources of information. They used both feature- and decision-level fusion methods to merge affective information extracted from multiple modalities, achieving an accuracy of nearly 80%.

1.3 General Categorization

Existing approaches to affective computing and sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches.

Knowledge-based techniques are very popular because of their accessibility and economy. Text is classified into affect categories based on the presence of fairly unambiguous affect words like 'happy', 'sad', 'afraid', and 'bored'. Popular sources of affect words or multi-word expressions are Ortony's Affective Lexicon (Ortony et al. 1988), Wiebe's linguistic annotation scheme (Wiebe et al. 2005), WordNet-

Affect (Strapparava and Valitutti 2004), SentiWordNet (Esuli and Sebastiani 2006), SenticNet (Cambria et al. 2016), and other probabilistic knowledge bases trained from linguistic corpora (Stevenson et al. 2007; Somasundaran et al. 2008; Rao and Ravichandran 2009). The major weakness of knowledge-based approaches is poor recognition of affect when linguistic rules are involved. For example, while a knowledge base can correctly classify the sentence "today was a happy day" as being happy, it is likely to fail on a sentence like "today wasn't a happy day at all". To this end, more sophisticated knowledge-based approaches exploit linguistics rules to distinguish how each specific knowledge base entry is used in text (Poria et al. 2015). The validity of knowledge-based approaches, moreover, heavily depends on the depth and breadth of the employed resources. Without a comprehensive knowledge base that encompasses human knowledge, in fact, it is not easy for a sentiment mining system to grasp the semantics associated with natural language or human behavior. Another limitation of knowledge-based approaches lies in the typicality of their knowledge representation, which is usually strictly defined and does not allow handling different concept nuances, as the inference of semantic and affective features associated with concepts is bounded by the fixed, flat representation.

Statistical methods, such as support vector machines and deep learning, have been popular for affect classification of texts and have been used by researchers on projects such as Pang's movie review classifier (Pang et al. 2002) and many others (Hu and Liu 2004; Glorot et al. 2011; Socher et al. 2013; Lau et al. 2014; Oneto et al. 2016). By feeding a machine learning algorithm a large training corpus of affectively annotated texts, it is possible for the system to not only learn the affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence of other arbitrary keywords (like lexical affinity) and word co-occurrence frequencies. However, statistical methods are generally semantically weak, i.e., lexical or co-occurrence elements in a statistical model have little predictive value individually. As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input. So, while these methods may be able to affectively classify user's text on the page- or paragraph-level, they do not work well on smaller text units such as sentences or clauses.

Hybrid approaches to affective computing and sentiment analysis, finally, exploit both knowledge-based techniques and statistical methods to perform tasks such as emotion recognition and polarity detection from text or multimodal data. Sentic computing (Cambria and Hussain 2015), for example, exploits an ensemble of knowledge-driven linguistic patterns and statistical methods to infer polarity from text. Xia et al. (2015) used SenticNet and a Bayesian model for contextual concept polarity disambiguation. Dragoni et al. (2014) proposed a fuzzy framework which merges WordNet, ConceptNet and SenticNet to extract key concepts from a sentence. iFeel (Araújo et al. 2014) is a system that allows users to create their own sentiment analysis framework by combing SenticNet, SentiWordNet and other sentiment analysis methods. Chenlo and Losada (2014) used SenticNet to extract bag of concepts and polarity features for subjectivity detection and other sentiment analysis tasks. Chung et al. (2014) used SenticNet concepts as seeds and proposed a method of random walk in ConceptNet to retrieve more concepts along with polarity scores. Other works propose the joint use of knowledge bases and machine learning for Twitter sentiment analysis (Bravo-Marquez et al. 2014), short text message classification (Gezici et al. 2013) and frame-based opinion mining (Recupero et al. 2014).

1.4 Conclusion

The passage from a read-only to a read-write Web made users more enthusiastic about sharing their emotion and opinions through social networks, online communities, blogs, wikis, and other online collaborative media. In recent years, this collective intelligence has spread to many different areas of the Web, with particular focus on fields related to our everyday life such as commerce, tourism, education, and health.

Despite significant progress, however, affective computing and sentiment analysis are still finding their own voice as new inter-disciplinary fields. Engineers and computer scientists use machine learning techniques for automatic affect classification from video, voice, text, and physiology. Psychologists use their long tradition of emotion research with their own discourse, models, and methods. Affective computing and sentiment analysis are research fields inextricably bound to the affective sciences that attempt to understand human emotions. Simply put, the development of affect-sensitive systems cannot be divorced from the century-long psychological research on emotion.

Hybrid approaches aim to better grasp the conceptual rules that govern sentiment and the clues that can convey these concepts from realization to verbalization in the human mind. In recent years, such approaches are gradually setting affective computing and sentiment analysis as interdisciplinary fields in between mere NLP and natural language understanding by gradually shifting from syntax-based techniques to more and more semantics-aware frameworks Cambria and White (2014), where both conceptual knowledge and sentence structure are taken into account (Fig. 1.2).

So far, sentiment mining approaches from text or speech have been mainly based on the bag-of-words model because, at first glance, the most basic unit of linguistic structure appears to be the word. Single-word expressions, however, are just a subset of concepts, multi-word expressions that carry specific semantics and sentics, that is, the denotative and connotative information commonly associated with objects, actions, events, and people. Sentics, in particular, specifies the affective information associated with real-world entities, which is key for emotion recognition and polarity detection, the basic tasks of affective computing and sentiment analysis.

The best way forward for these two fields, hence, is the ensemble application of semantic knowledge and machine learning, where different approaches can cover for each other's flaws. In particular, the combined application of linguistics and knowledge bases will allow sentiments to flow from concept to concept based on

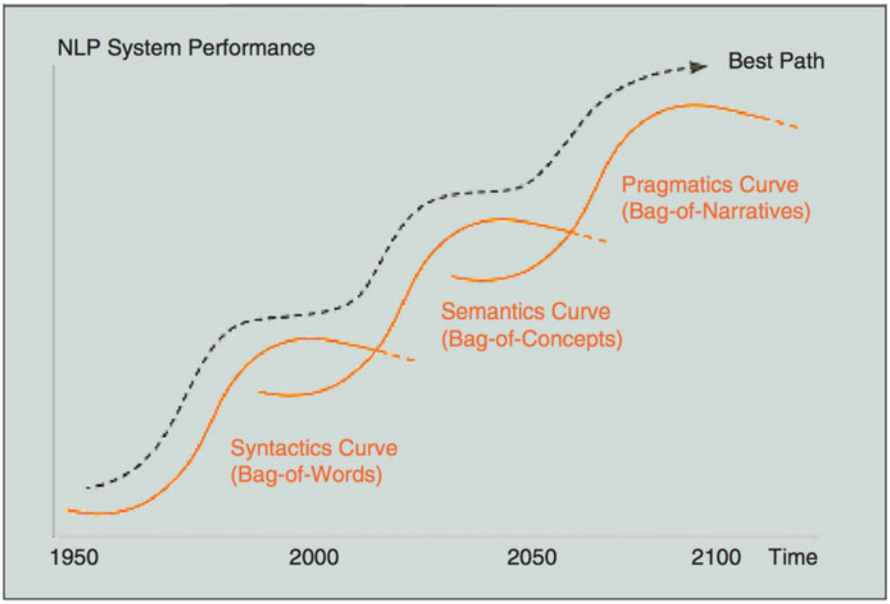


Fig. 1.2 Jumping NLP curves

the dependency relation of the input sentence, while machine learning will act as backup for missing concepts and unknown linguistic patterns.

Next-generation sentiment mining systems need broader and deeper common and commonsense knowledge bases, together with more brain-inspired and psychologically-motivated reasoning methods, in order to better understand natural language opinions and, hence, more efficiently bridge the gap between (unstructured) multimodal information and (structured) machine-processable data.

Looking ahead, blending scientific theories of emotion with the practical engineering goals of analyzing sentiments in natural language and human behavior will pave the way for development of more bio-inspired approaches to the design of intelligent sentiment mining systems capable of handling semantic knowledge, making analogies, learning new affective knowledge, and detecting, perceiving, and 'feeling' emotions.

References

Araújo, M., P. Gonçalves, M. Cha, and F. Benevenuto. 2014. iFeel: A system that compares and combines sentiment analysis methods. In *WWW*, 75–78.

Bravo-Marquez, F., M. Mendoza, and B. Poblete. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems* 69: 86–99.

Calvo, R., and S. D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1(1): 18–37.

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31(2): 102–107.

Cambria, E., and A. Hussain. 2015. *Sentic computing: A common-sense-based framework for concept-level sentiment analysis*. Cham: Springer.

Cambria, E., A. Livingstone, and A. Hussain. 2012. The hourglass of emotions. In *Cognitive behavioral systems*, ed. A. Esposito, A. Vinciarelli, and R. Hoffmann, V. Muller, Lecture notes in computer science, vol. 7403, 144–157. Berlin/Heidelberg: Springer.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*, 2666–2677.

Cambria, E., S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi. 2015. The CLSA model: A novel framework for concept-level sentiment analysis. In *Computational linguistics and intelligent text processing. CICLing 2015*, ed. A. Gelbukh, LNCS, vol. 9042, 3–22. Cham: Springer.

Cambria, E., H. Wang, and B. White. 2014. Guest editorial: Big social data analysis. *Knowledge-Based Systems* 69: 1–2.

Cambria, E., and B. White. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9(2): 48–57.

Chaturvedi, I., E. Cambria, and D. Vilares. 2016. Lyapunov filtering of objectivity for Spanish sentiment model. In: *IJCNN*, 4474–4481.

Chenlo, J.M., and D.E. Losada. 2014. An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences* 280: 275–288.

Chung, J.K.C., C.E. Wu, and R.T.H. Tsai. 2014. Improve polarity detection of online reviews with bag-of-sentimental-concepts. In *Proceedings of the 11th ESWC. Semantic Web Evaluation Challenge*. Crete: Springer.

Dragoni, M., A.G. Tettamanzi, and C. da Costa Pereira. 2014. A fuzzy system for concept-level sentiment analysis. In *Semantic web evaluation challenge*, 21–27. Cham: Springer.

Esuli, A., and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *LREC*.

Gezici, G., R. Dehkharghani, B. Yanikoglu, D. Tapucu, and Y. Saygin. 2013. Su-sentilab: A classification system for sentiment analysis in twitter. In *International Workshop on Semantic Evaluation*, 471–477.

Glorot, X., A. Bordes, and Y. Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, Bellevue.

Gunes, H., and B. Schuller. 2012. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31(2): 120–136.

Hu, M., and B. Liu. 2004. Mining and summarizing customer reviews. In *KDD*, Seattle.

Konar, A., and A. Chakraborty. 2015. *Emotion recognition: A pattern analysis approach*. Hoboken: Wiley & Sons.

Lau, R., Y. Xia, and Y. Ye. 2014. A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine* 9(1): 31–43

Liu, B. 2012. *Sentiment analysis and opinion mining*. San Rafael: Morgan and Claypool.

Ma, Y., E. Cambria, and S. Gao. 2016. Label embedding for zero-shot fine-grained named entity typing. In *COLING*, Osaka, 171–180.

Mihalcea, R., and A. Garimella. 2016. What men say, what women hear: Finding gender-specific meaning shades. *IEEE Intelligent Systems* 31(4): 62–67.

Minsky, M. 2006. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. New York: Simon & Schuster.

Morency, L.P., R. Mihalcea, and P. Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interfaces (ICMI)*, 169–176. New York: ACM.

Oneto, L., F. Bisio, E. Cambria, and D. Anguita. 2016. Statistical learning theory and ELM for big social data analysis. *IEEE Computational Intelligence Magazine* 11(3): 45–55.

Ortony, A., G. Clore, and A. Collins. 1988. *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Pang, B., and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2: 1–135.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, Philadelphia, 79–86.

Pantic, M., N. Sebe, J. Cohn, and T. Huang. 2005. Affective multimodal human-computer interaction. In *ACM International Conference on Multimedia*, New York, 669–676.

Picard, R. 1997. *Affective computing*. Boston: The MIT Press.

Poria, S., E. Cambria, and A. Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108: 42–49.

Poria, S., E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4): 26–36.

Poria, S., E. Cambria, D. Hazarika, and P. Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING*, 1601–1612.

Poria, S., E. Cambria, N. Howard, G.B. Huang, and A. Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174: 50–59.

Poria, S., I. Chaturvedi, E. Cambria, and A. Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *ICDM*, 439–448.

Poria, S., A. Gelbukh, B. Agarwal, E. Cambria, and N. Howard. 2013. Common sense knowledge based personality recognition from text. In *Advances in soft computing and its applications*, 484–496. Berlin/Heidelberg: Springer.

Raaijmakers, S., K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *EMNLP*, Edinburgh, 466–474.

Rajagopal, D., E. Cambria, D. Olsher, and K. Kwok. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *WWW*, Rio De Janeiro, 565–570.

Rao, D., and D. Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *EACL*, Athens, 675–682.

Recupero, D.R., V. Presutti, S. Consoli, A. Gangemi, and A. Nuzzolese. 2014. Sentilo: Frame-based sentiment analysis. *Cognitive Computation* 7(2): 211–225.

Schuller, B., A. Batliner, S. Steidl, and D. Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53(9/10): 1062–1087.

Socher, R., A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 1642–1654.

Somasundaran, S., J. Wiebe, and J. Ruppenhofer. 2008. Discourse level opinion interpretation. In *COLING*, Manchester, 801–808.

Stevenson, R., J. Mikels, and T. James. 2007. Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods* 39: 1020–1024.

Strapparava, C., and A. Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *LREC*, Lisbon, 1083–1086.

Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2): 165–210.

Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, Vancouver, 347–354.

Xia, Y., E. Cambria, A. Hussain, and H. Zhao. 2015. Word polarity disambiguation using bayesian model and opinion-level features. *Cognitive Computation* 7(3): 369–380.

Zeng, Z., M. Pantic, G. Roisman, and T. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1): 39–58.

Chapter 2
Many Facets of Sentiment Analysis

Bing Liu

Abstract Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, evaluations, attitudes, moods, and emotions. It is one of the most active research areas in natural language processing, data mining, information retrieval, and Web mining. In recent years, its research and applications have also spread to management sciences and social sciences due to its importance to business and society as a whole. This chapter defines the sentiment analysis problem and its related concepts such as sentiment, opinion, emotion, mood, and affect. The goal is to abstract a structure from the complex unstructured natural language text related to the problem and its pertinent concepts. The definitions not only enable us to see a rich set of inter-related sub-problems, but also a common framework that can unify existing research directions. They also help researchers design more robust solution techniques by exploiting the inter-relationships of the sub-problems.

Keywords Sentiment analysis • Opinion mining • Emotion • Mood • Affect • Subjectivity

Many people thought that sentiment analysis is just the problem of classifying whether a document or a sentence expresses a positive or negative sentiment or opinion. It is in fact a much more complex problem than that. It involves many facets and multiple sub-problems. In this chapter, I define an abstraction of the sentiment analysis problem. The definitions will enable us to see a rich set of inter-related sub-problems. It is often said that if we cannot structure a problem, we probably do not understand the problem. The objective of the definitions is to abstract a structure from the complex unstructured natural language text. The structure serves as a common framework to unify existing research directions and enable researchers to design more robust solution techniques by exploiting the inter-relationships of the sub-problems.

B. Liu (✉)
Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
e-mail: liub@cs.uic.edu

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_2

Unlike factual information, sentiment and opinion have an important characteristic, namely, being subjective. The subjectivity comes from many sources. First of all, different people may have different experiences and thus different opinions. Different people may also have different interests and/or different ideologies. Due to such different subjective experiences, views, interests and ideologies, it is important to examine a collection of opinions from many people rather than only one opinion from a single person because such an opinion represents only the subjective view of a single person, which is usually not sufficient for action. With a large number of opinions, some form of summary becomes necessary (Hu and Liu 2004). Thus, the problem definitions should also state what kind of summary may be desired. Along with the problem definitions, the chapter also discusses different types of opinions and the important concepts of affect, emotion and mood.

Throughout this chapter, I mainly use product reviews and sentences from such reviews as examples to introduce the key concepts, but the ideas and the resulting definitions are general and applicable to all forms of formal and informal opinion text such as news articles, tweets (Twitter posts), forum discussions, blogs, and Facebook posts, and all kinds of domains including social and political domains. The content of this chapter is mainly taken from my book "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions" (Liu 2015).

2.1 Definition of Opinion

Sentiment analysis mainly studies opinions that express or imply positive or negative sentiment. We define the problem in this context. We use the term *opinion* as a broad concept that covers sentiment, evaluation, appraisal, or attitude, and its associated information such as opinion target and the person who holds the opinion, and use the term *sentiment* to mean only the underlying positive or negative feeling implied by opinion. Due to the need to analyze a large volume of opinions, in defining opinion we consider two levels of abstraction: *a single opinion* and *a set of opinions*. In this section, we focus on defining a single opinion and describing the tasks involved in extracting an opinion. Section 2.2 focuses on a set of opinions, where we define *opinion summary*.

2.1.1 Opinion Definition

We use the following review (Review A) about a camera to introduce the problem (an id number is associated with each sentence for easy reference):

Review A

Posted by John Smith
Date: September 10, 2011

(1) *I bought a Canon G12 camera six months ago*. (2) *I simply love it*. (3) *The picture quality is amazing*. (4) *The battery life is also long*. (5) *However, my wife thinks it is too heavy for her*.

From this review, we can make the following important observation:

Opinion, sentiment and target: Review A has several opinions with positive or negative sentiment about the Canon G12 camera. Sentence (2) expresses a positive sentiment about the Canon camera as a whole. Sentence (3) expresses a positive sentiment about its picture equality. Sentence (4) expresses a positive sentiment about its battery life. Sentence (5) expresses a negative sentiment about the camera's weight.

From these opinions, we can make a crucial observation about sentiment analysis. That is, an opinion has two key components: a *target g* and a *sentiment s* on the target, i.e., (g, s), where g can be any entity or aspect of the entity on which an opinion has been expressed, and s can be a positive, negative, or neutral sentiment, or a numeric rating. *Positive*, *negative* and *neutral* are called *sentiment* or *opinion orientations*. For example, the target of the opinion in sentence (2) is the *Canon G12 camera*, the target of the opinion in sentence (3) is the *picture quality of Canon G12*, and the target of sentence (5) is the *weight of Canon G12* (*weight* is indicated by *heavy*). Target is also called *topic* by some researchers.

Opinion holder: Review A contains opinions from two persons, who are called *opinion sources* or *opinion holders* (Kim and Hovy 2004; Wiebe et al. 2005*). The* holder of the opinions in sentences (2), (3), and (4) is the author of the review ("John Smith"), but for sentence (5), it is the wife of the author.

Time of opinion: The date of the review was September 10, 2011. This date is useful because one often wants to know how opinions change over time or the opinion trend.

With this example, we can define opinion as a quadruple.

Definition 1 (Opinion) An *opinion* is a quadruple,

$$(g, s, h, t),$$

where g is the *sentiment target*, s is the *sentiment* of the opinion about the target g, h is the *opinion holder* (the person or organization who holds the opinion), and t is the *time* when the opinion is expressed.

The four components here are essential. It is generally problematic if any of them is missing. For example, the time component is important in practice because an opinion two years ago is not the same as an opinion today. Not having an opinion holder is also problematic. For example, an opinion from a very important person (e.g., the US President) is probably more important than that from the average Joe on the street.

One thing that we want to stress about the definition is that *opinion has target*. Recognizing this is important for two reasons: First, in a sentence with multiple targets, we need to identify the specific target for each positive or negative sentiment. For example, "*Apple is doing very well in this poor economy*" has a positive sentiment and a negative sentiment. The target for the positive sentiment is *Apple* and the target for the negative sentiment is *economy*. Second, words or phrases such as *good*, *amazing*, *bad* and *poor* that express sentiments (called *sentiment* or *opinion terms*) and opinion targets often have some syntactic relations (Hu and Liu 2004; Qiu et al. 2011; Zhuang et al. 2006), which allow us to design algorithms to extract both sentiment terms and opinion targets, which are two core tasks of sentiment analysis (see Sect. 2.1.6).

The opinion defined here is just one type of opinion, called a *regular opinion* (e.g., "*Coke taste great*"). Another type is *comparative opinion* (e.g., "*Coke tastes better than Pepsi*") which needs a different definition (Jindal and Liu 2006b; Liu 2006). Section 2.1.4 will further discuss different types of opinions. For the rest of this section, we focus on only regular opinions, which, for simplicity, we will just call opinions.

2.1.2 Sentiment Target

Definition 2 (Sentiment Target) The *sentiment target*, also known as the *opinion target*, of an opinion is the entity or a part or attribute of the entity that the sentiment has been expressed upon.

For example, in sentence (3) of Review A, the target is the *picture quality of Canon G12*, although the sentence mentioned only the *picture quality*. The target is not just the *picture quality* because without knowing that the picture quality belongs to the Canon G12 camera, the opinion in the sentence is of little use.

An entity can be decomposed and represented hierarchically (Liu 2006).

Definition 3 (Entity) An *entity e* is a product, service, topic, person, organization, issue or event. It is described with a pair, e: (T, W), where T is a hierarchy of *parts*, *sub-parts*, and so on, and W is a set of *attributes* of e. Each part or sub-part also has its own set of attributes.

For example, a particular camera model is an entity, e.g., Canon G12. It has a set of attributes, e.g., *picture quality*, *size*, and *weight*, and a set of parts, e.g., *lens*, *viewfinder*, and *battery*. *Battery* also has its own set of attributes, e.g., *battery life* and *battery weight*. A topic can be an entity too, e.g., *tax increase*, with its sub-topics or parts '*tax increase for the poor*,' '*tax increase for the middle class*' and '*tax increase for the rich*.'

This definition describes an entity hierarchy based on the *part-of* relation. The root node is the name of the entity, e.g., Canon G12 Review A. All the other nodes are parts and sub-parts, etc. An opinion can be expressed on any node and any attribute of the node. For instance, in Review A, sentence (2) expresses a positive opinion about the entity Canon G12 as a whole, and sentence (3) expresses a positive opinion about the picture quality attribute of the camera. Clearly, we can also express opinions about any part or component of the camera.

In the research literature, entities are also called *objects*, and attributes are also called *features* (as in product features) (Hu and Liu 2004; Liu 2010). The terms *object* and *feature* are not used in this Chapter because object can be confused with the term *object* used in grammar and feature can be confused with *feature* used in machine learning as data attribute. In recent years, the term *aspect* has become popular, which covers both *part* and *attribute* (see Sect. 2.1.4).

Entities may be called other names in specific application domains. For example, in politics, entities are usually *political candidates*, *issues*, and *events*. There is no term that is perfect for all application domains. The term *entity* is chosen because most current applications of sentiment analysis study opinions about various forms of named entities, e.g., products, services, brands, organizations, events, and people.

2.1.3 Sentiment and Its Intensity

Definition 4 (Sentiment) *Sentiment* is the underlying feeling, attitude, evaluation, or emotion associated with an opinion. It is represented as a triple,

$$(y, o, i),$$

where y is the *type* of the sentiment, o is the *orientation* of the sentiment, and i is the *intensity* of the sentiment.

Sentiment type: Sentiment can be classified into several types. There are linguistic-based, psychology-based, and consumer research-based classifications. Here I choose to use a consumer research-based classification as it is simple and easy to use in practice. Consumer research classifies sentiment broadly into two categories: *rational sentiment* and *emotional sentiment* (Chaudhuri 2006).

Definition 5 (Rational Sentiment) *Rational sentiments* are from rational reasoning, tangible beliefs, and utilitarian attitudes. They express no emotions.

We also call opinions expressing rational sentiment the *rational opinions*. The opinions in the following sentences imply rational sentiment: "*The voice of this phone is clear*," and "*This car is worth the price.*"

Definition 6 (Emotional Sentiment) *Emotional sentiments* are from non-tangible and emotional responses to entities which go deep into people's psychological state of mind.

We also call opinions expressing emotional sentiment the *emotional opinions*. The opinions in the following sentences imply emotional sentiment: “*I love iPhone*,” “*I am so angry with their service people*,” “*This is the best car ever*” and “*After our team won, I cried*.”

Emotional sentiment is stronger than rational sentiment, and is usually more important in practice. For example, in marketing, to guarantee the success of a new product in the market, the positive sentiment from a large population of consumers has to reach the emotional level. Rational positive may not be sufficient.

Each of these broad categories can be further divided into smaller categories. For example, there are many types of emotions, e.g., *anger*, *joy, fear*, and *sadness.* We will discuss some possible sub-divisions of rational sentiment in Sect. 2.4.2 and different emotions in Sect. 2.3. In applications, the user is also free to design their own sub-categories.

Sentiment orientation: It can be *positive*, *negative*, or *neutral*. Neutral usually means the absence of sentiment or no sentiment or opinion. Sentiment orientation is also called *polarity*, *semantic orientation*, or *valence* in the research literature.

Sentiment intensity: Sentiment can have different levels of strength or intensity. People often use two ways to express intensity of their feelings in text. The first is to choose sentiment terms (words or phrases) with suitable strengths. For example, *good* is weaker than *excellent*, and *dislike* is weaker than *detest*. *Sentiment words* are words in a language that are often used to express positive or negative sentiments. For example, *good*, *wonderful*, and *amazing* are positive sentiment words, and *bad*, *poor*, and *terrible* are negative sentiment words. The second is to use *intensifiers* and *diminishers*, which are terms that change the degree of the expressed sentiment. An intensifier increases the intensity of a positive/negative term, while a diminisher decreases the intensity of that term. Common English intensifiers include *very*, *so*, *extremely*, *dreadfully*, *really*, *awfully*, *terribly*, etc., and common English diminishers include *slightly*, *pretty*, *a little bit*, *a bit*, *somewhat*, *barely*, etc.

Sentiment rating: In applications, we commonly use some discrete ratings to express sentiment intensity. Five levels (e.g., 1–5 stars) are commonly employed, which can be interpreted as follows based on the two types of sentiment in Definitions 5 and 6:

- *emotional positive* (+2 or 5 stars)
- *rational positive* (+1 or 4 stars)
- *neutral* (0 or 3 stars)
- *rational negative* (−1 or 2 stars)
- *emotional negative* (−2 or 1 star)

Clearly, it is possible to have more rating levels, which, however, become difficult to differentiate based on the natural language text alone due to the highly subjective nature and the fact that people’s spoken or written expressions may not fully match with their psychological states of mind. For example, the sentence “*This is an excellent phone*” expresses a rational evaluation of the phone, while “*I love this* *phone*" expresses an emotional evaluation about the phone. However, whether they represent completely different psychology states of mind of the authors is hard to say. In practice, the above five levels are sufficient for most applications. If these five levels are not enough in some applications, I suggest dividing *emotional positive* (and, respectively, *emotional negative*) into two levels. Such applications are likely to involve sentiment about personal, social or political events or issues, for which people can be highly emotional.

2.1.4 Opinion Definition Simplified

Opinion as defined in Definition 1, although concise, may not be easy to use in practice especially in the domain of online reviews of products, services, and brands. Let us first look at the sentiment (or opinion) target. The central concept here is *entity*, which is represented as a hierarchy with an arbitrary number of levels. This can be too complex for practical applications because NLP is a very difficult task. Recognizing parts and attributes of an entity at different levels of details is extremely hard. Most applications also do not need such a complex analysis. Thus, we simplify the hierarchy to two levels and use the term *aspect* to denote both *part* and *attribute*. In the simplified tree, the root node is still the entity itself and the second level (also the leaf level) nodes are different aspects of the entity.

The definition of sentiment in Definition 4 can be simplified too. In many applications, positive (denoted by $+1$), negative (denoted by -1) and neutral (denoted by 0) orientations alone are already enough. In almost all applications, 5 levels of ratings are sufficient, e.g., 1–5 stars. In both cases, sentiment can be represented with a single value. The other two components in the triple can be folded into this value.

This simplified framework is what is typically used in practical sentiment analysis systems. We now redefine the concept of opinion (Hu and Liu 2004; Liu 2010).

Definition 7 (Opinion) An *opinion* is a quintuple,

$$(e, a, s, h, t),$$

where e is the target entity, a is the target aspect of entity e on which the opinion has been expressed, s is the sentiment of the opinion on aspect a of entity e, h is the opinion holder, and t is the opinion posting time. s can be *positive*, *negative*, or *neutral*, or a *rating* (e.g., 1–5 stars). When an opinion is only on the entity as a whole, the special aspect GENERAL is used to denote it. Here, e and a together represent the opinion target.

Sentiment analysis (or opinion mining) based on this definition is often called *aspect-based sentiment analysis,* or *feature-based sentiment analysis* as it was called earlier in (Hu and Liu 2004; Liu 2010).

We should note that due to the simplification, the quintuple representation of opinion may result in information loss. For example, *ink* is a part of *printer*. A printer review might say "*The ink of this printer is expensive*." This sentence does not say that the printer is expensive (*expensive* here indicates the aspect *price*). If one does not care about any attribute of the ink, this sentence just gives a negative opinion about the ink (which is an aspect of the printer entity). This results in information loss. However, if one also wants to study opinions about different aspects of the ink, then the ink needs to be treated as a separate entity. The quintuple representation still applies, but an extra mechanism will be required to record the part-of relationship between ink and printer. Of course, conceptually we can also extend the flat quintuple relation to a *nested relation* to make it more expressive. However, as we explained above, too complex a definition can make the problem extremely difficult to solve in practice. Despite this limitation, Definition 4 does cover the essential information of an opinion sufficiently for most applications.

In some applications, it may not be easy to distinguish entity and aspect or there is no need to distinguish them. Such cases often occur when people discuss political or social issues, e.g., "*I hate property tax increases*." We may deal with them in two ways. First, since the author regards '*property tax increase*' as a general issue and it thus does not belong to any specific entity. We can treat it as an entity with the aspect GENERAL. Second, we can regard '*property tax*' as an entity and '*property tax increases*' as one of its aspects to form a hierarchical relationship. Whether treating an issue/topic as an aspect or an entity can also depend on the specific context.

For example, in commenting about a local government, one says "*I hate the proposed property tax increase*." Since it is the local government that imposes and levies property taxes, the specific local government may be regarded as an entity and '*the proposed property tax increase*' as one of its aspects.

Not all applications need all five components of an opinion. In some applications, the user may not need the aspect information. For example, in brand management, the user typically is interested in only opinions about product brands (entities). This is sometimes called *entity-based sentiment analysis*. In some other applications, the user may not need to know the opinion holder or time of opinion. Then these components can be ignored.

2.1.5 Reason and Qualifier for Opinion

We can in fact perform an even finer-grained analysis of opinions. Let us use the sentence "*This car is too small for a tall person*" to explain. It expresses a negative sentiment about the *size* aspect of the car. However, only reporting the negative sentiment for size does not tell the whole story because it can mean *too small* or *too big*. In the above sentence, we call "*too small*" the *reason* for the negative sentiment about size. Furthermore, the sentence does not say that the car is too small for everyone, but only *for a tall person*. We call "*for a tall person*" the *qualifier* of the opinion. We now define these concepts.

Definition 8 (Reason for Opinion) A reason for an opinion is the cause of the opinion.

In practical applications, discovering the reasons for each positive or negative opinion can be very important because it may be these reasons that enable one to perform actions to remedy the situation. For example, the sentence "*I do not like the picture quality of this camera*" is not as useful as "*I do not like the picture quality of this camera because the pictures are quite dark*." The first sentence does not give the reason for the negative sentiment about the picture quality and it is thus difficult to know what to do to improve the picture quality. The second sentence is more informative because it gives the reason or cause for the negative sentiment. The camera manufacturer can make use of this piece of information to improve the picture quality of the camera. In most industrial applications, such reasons are called *problems* or *issues*. Knowing the issues allows businesses to find ways to address them.

Definition 9 (Qualifier of Opinion) A qualifier of an opinion limits or modifies the meaning of the opinion.

Knowing the qualifier is also important in practice because it tells what the opinion is good for. For example, "*This car is too small for a tall person*" does not say that the car is too small for everyone, but just for tall people. For a person who is not tall, this opinion does not apply.

However, as we have seen, not every opinion comes with an explicit reason and/or an explicit qualifier. "*The picture quality of this camera is not great*" does not have a reason or a qualifier. "*The picture quality of this camera is not good for night shots*" has a qualifier "*for night shots*," but does not give a specific reason for the negative sentiment. "*The picture quality of this camera is not good for night shots as the pictures are quite dark*" has a reason for the negative sentiment ('*the pictures are quite dark*') and also a qualifier ('*for night shots*'). Sometimes, the qualifier and the reason may not be in the same sentence and/or may be quite implicit, e.g., "*The picture quality of this camera is not great. Pictures of night shots are very dark*" and "*I am 6 feet 5 inches tall. This car is too small for me*." An expression can also serve multiple purposes. For example, '*too small*' in the above sentence indicates the *size* aspect of the car, a *negative sentiment* about the size, and also the *reason* for the negative sentiment/opinion.

2.1.6 Objective and Tasks of Sentiment Analysis

With the definitions in Sects. 2.1.1, 2.1.2, 2.1.3 and 2.1.4, we can now present the core objective and the key tasks of (aspect-based) sentiment analysis.

Objective of Sentiment Analysis Given an opinion document d, discover all opinion quintuples (e, a, s, h, t) in d. For more advanced analysis, discover the reason and qualifier for the sentiment in each opinion quintuple.

Key Tasks of Sentiment Analysis The key tasks of sentiment analysis can be derived from the five components of the quintuple (Definition 7). The first component is the entity and the first task is to extract entities. The task is similar to named entity recognition (NER) in information extraction (Hobbs and Riloff 2010; Sarawagi 2008). However, as defined in Definition 3, an entity can also be an event, issue, or topic, which is usually not a named entity. For example, in "*I hate tax increase*," the entity is '*tax increase*,' which is an issue or topic. In such cases, entity extraction is basically the same as aspect extraction and the difference between entity and aspect becomes blurry. In some applications, there may not be a need to distinguish them.

After extraction, we need to categorize the extracted entities as people often write the same entity in different ways. For example, Motorola may be written as Mot, Moto, and Motorola. We need to recognize that they all refer to the same entity (see (Liu 2015) for details).

Definition 10 (Entity Category and Entity Expression) An *entity category* represents a unique entity, while an *entity expression* or *mention* is an actual word or phrase that indicates an entity category in the text.

Each entity or entity category should have a unique name in a particular application. The process of grouping or clustering entity expressions into entity categories is called *entity resolution* or *grouping*.

For aspects of entities, the problem is basically the same as for entities. For example, *picture*, *image*, and *photo* refer to the same aspect for cameras. We thus need to extract aspect expressions and resolve them.

Definition 11 (Aspect Category and Aspect Expression) An *aspect category* of an entity represents a unique aspect of the entity, while an *aspect expression* or *mention* is an actual word or phrase that indicates an aspect category in the text.

Each aspect or aspect category should also have a unique name in a particular application. The process of grouping aspect expressions into aspect categories (aspects) is called *aspect resolution* or *grouping*.

Aspect expressions are usually nouns and noun phrases but can also be verbs, verb phrases, adjectives, and adverbs. They can also be explicit or implicit (Hu and Liu 2004).

Definition 12 (Explicit Aspect Expression) Aspect expressions that appear in an opinion text as nouns and noun phrases are called *explicit aspect expressions*.

For example, '*picture quality*' in "*The picture quality of this camera is great*" is an explicit aspect expression.

Definition 13 (Implicit Aspect Expression) Aspect expressions that are not nouns or noun phrases but indicate some aspects are called *implicit aspect expressions*.

For example, *expensive* is an implicit aspect expression in "*This camera is expensive*." It implies the aspect *price*. Many implicit aspect expressions are adjectives and adverbs used to describe or qualify some specific aspects, e.g., *expensive* (price), and *reliably* (reliability). They can also be verb and verb phrases, e.g., "*I can install the software easily*." *Install* indicates the aspect *installation*.

Implicit aspect expressions are not just individual adjectives, adverbs, verbs and verb phrases; they can be very complex. For example, in "*This camera will not easily fit in my pocket*," '*fit in my pocket*' indicates the aspect *size* (and/or *shape*). In the sentence "*This restaurant closes too early*," '*closes too early*' indicates the aspect of *closing time* of the restaurant. In both cases, some commonsense knowledge may be needed to recognize them.

Aspect extraction is a very challenging problem, especially when it involves verbs and verb phrases. In some cases, it is even very hard for human beings to recognize and to annotate. For example, in a vacuum cleaner review, one wrote "*The vacuum cleaner does not get the crumbs out of thick carpets*," which seems to describe only one very *specific* aspect, '*get the crumbs out of thick carpets*.' But in practice, it may be more useful to decompose it into three different aspects indicated by (1) '*get something out of*,' (2) *crumbs*, and (3) '*thick carpets*.' (1) represents the suction power of the vacuum cleaner in general, (2) represents suction related to *crumbs*, and (3) represents *suction* related to '*thick carpets*.' All three are important and useful because the user may be interested in knowing whether the vacuum can suck crumbs, and whether it works well with thick carpets.

The third component in the opinion definition is the sentiment. For this, we need to perform sentiment classification or regression to determine the sentiment orientation or score on the involved aspect and/or entity. The fourth component and fifth components are opinion holder and opinion posting time respectively. They also have expressions and categories as entities and aspects. I will not repeat their definitions. Note that opinion holders (Bethard et al. 2004; Choi et al. 2005; Kim and Hovy 2004) are also called *opinion sources* in (Wiebe et al. 2005).

Based on the above discussions, we can now define a model of entity and a model of opinion document (Liu 2006) and summarize the main sentiment analysis tasks.

Model of Entity An entity e is represented by itself as a whole and a finite set of its aspects $A = \{a1, a2, \ldots, an\}$. e can be expressed in text with any one of a finite set of its entity expressions $\{ee1, ee2, \ldots, ees\}$. Each aspect $a \in A$ of entity e can be expressed with any one of its finite set of aspect expressions $\{ae1, ae2, \ldots, aem\}$.

Model of Opinion Document An opinion document d contains opinions about a set of entities $\{e1, e2, \ldots, er\}$ and a subset of aspects of each entity. The opinions are from a set of opinion holders $\{h1, h2, \ldots, hp\}$ and are given at a particular time point t.

Given a set of opinion documents D, sentiment analysis performs the following eight (8) main tasks:

Task 1 (entity extraction and resolution): Extract all entity expressions in D, and group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster refers to a unique entity e.

Task 2 (aspect extraction and resolution): Extract all aspect expressions of the entities, and group these aspect expressions into clusters. Each aspect expression cluster of entity e represents a unique aspect a.

Task 3 (opinion holder extraction and resolution): Extract the holder expression of each opinion from the text or structured data and group them. The task is analogous to tasks 1 and 2.

Task 4 (time extraction and standardization): Extract the posting time of each opinion and standardize different time formats.

Task 5 (aspect sentiment classification or regression): Determine whether an opinion about an aspect *a* (or entity *e*) is positive, negative or neutral (classification), or assign a numeric sentiment rating score to the aspect (or entity) (regression).

Task 6 (opinion quintuple generation): Produce all opinion quintuples (*e*, *a*, *s*, *h*, *t*) expressed in *D* based on the results from tasks 1–5. This task is seemingly very simple but it is in fact quite difficult in many cases as Review B below shows.

For more advanced analysis, we also need to perform the following two additional tasks, which are analogous to task 2:

Task 7 (opinion reason extraction and resolution): Extract reason expressions for each opinion, and group all reason expressions for each aspect or entity and each sentiment orientation into clusters. Each cluster for an aspect (or entity) and a sentiment orientation represents a unique reason for the aspect (or entity) and the orientation.

Task 8 (opinion qualifier extraction and resolution): Extract qualifier expressions for each opinion, and group all qualifier expressions for each aspect (or entity) and each sentiment orientation into clusters. Each cluster for an aspect (or entity) and a sentiment orientation represents a unique qualifier for the aspect (or entity) and the orientation.

Although reasons for and qualifiers of opinions are useful, their extraction and categories are very challenging. Little research has been done about them so far.

We use an example review to illustrate the tasks (a sentence id is again associated with each sentence) and the mining results.

Review B

Posted by: bigJohn
Date: Sept. 15, 2011

> (1) *I bought a Samsung camera and my friend brought a Canon camera yesterday*. (2) *In the past week, we both used the cameras a lot*. (3) *The photos from my Samy are not clear for night shots, and the battery life is short too*. (4) *My friend was very happy with his camera and loves its picture quality*. (5) *I want a camera that can take good photos*. (6) *I am going to return it tomorrow*.

Task 1 should extract the entity expressions, *Samsung*, *Samy*, and *Canon*, and group *Samsung* and *Samy* together because they represent the same entity. Task 2 should extract aspect expressions *picture*, *photo*, and *battery life*, and group *picture* and *photo* together as they are synonyms for cameras. Task 3 should find that the holder of the opinions in sentence (3) is bigJohn (the blog author) and that the holder of the opinions in sentence (4) is bigJohn's friend. Task 4 should find that the time when the blog was posted is Sept-15-2011. Task 5 should find that sentence (3) gives a negative opinion to the *picture quality* of the Samsung camera and a negative opinion also to its *battery life*. Sentence (4) gives a positive opinion to the *Canon camera* as a whole and also to its *picture quality*. Sentence (5) seemingly expresses a positive opinion, but it does not. To generate opinion quintuples for sentence (4) we need to know what '*his camera*' and *its* refer to. Task 6 should finally generate the following opinion quintuples:

1. (Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
2. (Samsung, battery_life, negative, bigJohn, Sept-15-2011)
3. (Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
4. (Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

With more advanced mining and analysis, we also find the reasons and qualifiers of opinions. *None* below means unspecified.

1. (Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
   Reason for opinion: picture not clear
   Qualifier of opinion: night shots
2. (Samsung, battery_life, negative, bigJohn, Sept-15-2011)
   Reason for opinion: short battery life
   Qualifier of opinion: none
3. (Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
   Reason for opinion: none
   Qualifier of opinion: none
4. (Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)
   Reason for opinion: none
   Qualifier of opinion: none

2.2 Definition of Opinion Summary

Unlike facts, opinions are subjective (although they may not be all expressed in subjective sentences). An opinion from a single opinion holder is usually not sufficient for action. In almost all applications, the user needs to analyze opinions from a large number of opinion holders. This tells us that some form of summary of opinions is necessary. The question is what an opinion summary should be. On the surface, an opinion summary is just like a multi-document summary because we need to summarize multiple opinion documents, e.g., reviews. It is, however, very different from traditional multi-document summary. Although there are informal descriptions about what a traditional multi-document summary should be, it is never formally defined. A traditional multi-document summary is often just "defined" operationally based on each specific algorithm that produces the summary. Thus different algorithms produce different kinds of summaries. The resulting summaries are also hard to evaluate. An opinion summary in its core form, on the other hand, can be defined precisely based on the quintuple definition of opinion and easily evaluated. That is, all opinion summarization algorithms should aim to produce the same summary. Although they may still produce different final summaries, that is due to their different accuracies. This core form of opinion summary is called the *aspect-based opinion summary* (or *feature-based opinion summary*) (Hu and Liu 2004; Liu et al. 2005)

Definition 11 (Aspect-Based Opinion Summary) The *aspect-based opinion summary* about an entity e is of the following form:

GENERAL: number of opinion holders who are positive about entity e number of opinion holders who are negative about entity e

Aspect 1: number of opinion holders who are positive about aspect 1 of entity e number of opinion holders who are negative about aspect 1 of entity e

...

Aspect n: number of opinion holders who are positive about aspect n of entity e number of opinion holders who are negative about aspect n of entity e

where GENERAL represents the entity e itself and n is the total number of aspects of e.

The key features of this opinion summary definition are that it is based on positive and negative opinions about each entity and its aspects and that it is quantitative. The quantitative perspective is reflected by the numbers of positive and negative opinions. In an application, the number counts can also be replaced by percentages. The quantitative perspective is especially important in practice. For example, 20% of the people positive about a product is very different from 80% of the people positive about the product.

To illustrate this form of summary, we summarize a set of reviews of a digital camera, called *digital camera* 1, in Figure 2.1. This is called a *structured summary* in contrast to a traditional text summary of a short document generated from one or multiple long documents. In the figure, 105 reviews expressed positive opinions about the camera itself denoted by GENERAL and 12 expressed negative opinions. *Picture quality* and *battery life* are two camera aspects. 75 reviews expressed positive opinions about the picture quality, and 42 expressed negative opinions.

Digital Camera 1:
Aspect: **GENERAL**
Positive: 105 <Individual review sentences>
Negative: 12 <Individual review sentences>
Aspect: **Picture quality**
Positive: 75 <Individual review sentences>
Negative: 42 <Individual review sentences>
Aspect: **Battery life**
Positive: 50 <Individual review sentences>
Negative: 9 <Individual review sentences>
…

Fig. 2.1 An aspect-based opinion summary

We also added <*Individual review sentences*>, which can be a link pointing to the sentences and/or the whole reviews that contain the opinions (Hu and Liu 2004; Liu et al. 2005). With this summary, one can easily see how existing customers feel about the camera. If one is interested in a particular aspect and additional details, one can drill down by following the <*Individual review sentences*> link to see the actual opinion sentences or reviews.

In a more advanced analysis, we can also summarize opinion reasons and qualifiers in a similar way. Based on my experience, qualifiers for opinion statements are rare, but reasons for opinions are quite common. To perform the task, we need another level of summary. For example, in the example of Figure 2.1, we may want to summarize the reasons for the poor picture quality based on the sentences in <*Individual review sentences*>. We may find that 35 people say the pictures are not bright enough and 7 people say that the pictures are blurry.

Based on the idea of aspect-based summary, researchers have proposed many opinion summarization algorithms, and also extended this form of summary to some other more specialized forms (Liu 2015).

2.3 Affect, Emotion, and Mood

A*ffect*, *emotion*, and *mood* have been studied extensively in several fields, e.g., psychology, philosophy, and sociology. However, investigations in these fields are seldom concerned with the language expressions used to express such feelings. Their main concerns are people's psychological states of mind, theorizing what affect, emotion and mood are, what constitute basic emotions, what physiological reactions happen (e.g., heart rate changes, blood pressure, sweating and so on), what facial expressions, gestures and postures are, and measuring and investigating the impact of such mental states. These mental states have also been exploited extensively in application areas such as marketing, economics, and education.

However, even with such extensive research, understanding these concepts is still slippery and confusing because different theorists often have somewhat different definitions for them and even do not completely agree with each other about what emotion, mood, and affect are. For example, about emotion, diverse theorists have proposed that there are from two to twenty basic human emotions and some even do not believe there is such a thing called basic emotions (Ortony and Turner 1990). In most cases, emotion and affect are regarded as synonymous, and indeed, all three terms are sometimes used interchangeably. Affect is also used as an encompassing term covering all topics related to emotion, feeling, and mood. To make matters worse, in applications, researchers and practitioners use these concepts loosely in whatever way they feel like to without following any established definitions. Thus one is often left puzzled by just what an author means when the word emotion, mood, or affect is used. In most cases, the definition of each term also uses one or more of the other terms resulting in circular definitions, which causes further confusion. The good news for natural language processing researchers and practitioners is that in practical applications of sentiment analysis, we needn't be too concerned with such an unsettled state of affair because in practice we can pick up and use whatever emotion or mood states that are suitable for the applications at hand.

This section first tries to create a reasonable understanding of these concepts and their relationships for our tasks of natural language processing in general and sentiment analysis in particular. It then puts these three concepts in the context of sentiment analysis and discusses how they can be handled in sentiment analysis.

2.3.1 Affect, Emotion, and Mood in Psychology

We start the discussion with the dictionary definitions of affect, emotion, and mood[1]. The concept of *feeling* is also included as all three concepts are about human feelings. From the definitions, we can see how difficult it is to explain or to articulate these concepts:

- **Affect**: Feeling or emotion, especially as manifested by facial expression or body language.
- **Emotion**: A mental state that arises spontaneously rather than through conscious effort and is often accompanied by physiological changes.
- **Mood**: A state of mind or emotion.
- **Feeling**: An affective state of consciousness, such as that resulting from emotions, sentiments, or desires.

These definitions are confusing from a scientific point of view because we do not see a clear demarcation for each concept. We turn to the field of psychology to look for a better definition for each of them. The convergence of views and ideas among theorists in the past twenty years gives us a workable classification scheme.

An *affect* is commonly defined as an neurophysiological state consciously accessible as the simplest raw (nonreflective) feeling evident in moods and emotions (Russell 2003). The key point here is that such a feeling is primitive and not directed at an object. For example, you are watching a scary movie. If you are affected, it moves you and you experience a feeling of being scared. Your mind further processes this feeling and expresses it to yourself and the world around you. The feeling is then displayed as an *emotion*, such as crying, shock, and scream.

Emotion is thus the indicator of affect. Due to cognitive processing, emotion is a compound (rather than primitive) feeling concerned with a specific object, such as a person, an event, a thing, or a topic. It tends to be intense and focused and lasts a short period of time. *Mood*, like emotion, is a feeling or affective state but it typically lasts longer than emotion and tends to be more unfocused and diffused. Mood is also less intense than emotion. For example, you may wake up feeling happy and stay that way for most of the day.

[1] http://www.thefreedictionary.com/subjective

In short, emotions are quick and tense, while moods are more diffused and prolonged feelings. For example, we can get very angry very quickly, but it is difficult to stay very angry for a long time. The anger emotion may subside into an irritable mood that can last quite a long time. An emotion is usually very specific, triggered by noticeable events, which means that an emotion has a specific target. In this sense, emotion is like a rational opinion. On the other hand, a mood can be caused by multiple events, and sometimes it may not have any specific targets or causes. Mood typically also has a dimension of future expectation. It can involve a structured set of beliefs about general expectations of a future experience of pleasure or pain, or of positive or negative affect in the future (Batson et al. 1992).

Since sentiment analysis is not so much concerned with affect as defined above, below we focus only on *emotion* and *mood* in the psychological context. Let us start with emotion. Emotion has been frequently mentioned in sentiment analysis. Since it has a target or an involved entity, it fits the sentiment analysis context naturally. Almost all applications are interested in opinions and emotions about some target entities.

Theorists in psychology have grouped emotions into categories. However, as we mentioned earlier, there is still not a set of agreed basic (or primary) emotions among theorists. In (Ortony and Turner 1990), the basic emotions proposed by several theorists were compiled to show there is a great deal of disagreement. We reproduce them in Table 2.1.

In (Parrott 2001), apart from the basic emotions, secondary and tertiary emotions were also proposed (see Table 2.2). These secondary and tertiary are useful in some

Table 2.1 Basic emotions from different theorists

Source	Basic emotions
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman et al. (1982)	Anger, disgust, fear, joy, sadness, surprise
Gray (1982)	Anxiety, joy, rage, terror
Izard (1971)	Anger, contempt, disgust, distress, fear guilt, interest, joy, shame, surprise
James (1884)	Fear, grief, love, rage
McDougall (1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mowrer (1960)	Pain, pleasure
Oatley and Jobnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness
Panksepp (1982)	Expectancy, fear, rage, panic
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson (1930)	Fear, love, rage
Weiner and Graham (1984)	Happiness, sadness
Parrott (2001)	Anger, fear, joy, love, sadness, surprise

Table 2.2 Primary, Secondary and Tertiary emotions from Parrott (2001)

Primary emotion	Secondary emotion	Tertiary emotion
Anger	Disgust	Contempt, loathing, revulsion
	Envy	Jealousy
	Exasperation	Frustration
	Irritability	Aggravation, agitation, annoyance, crosspatch, grouchy, grumpy
	Rage	Anger, bitter, dislike, ferocity, fury, hatred, hostility, outrage, resentment, scorn, spite, vengefulness, wrath
	Torment	Torment
Fear	Horror	Alarm, fear, fright, horror, hysteria, mortification, panic, shock, terror
	Nervousness	Anxiety, apprehension (fear), distress, dread, suspense, uneasiness, worry
	Cheerfulness	Amusement, bliss, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
Joy	Contentment	Pleasure
	Enthrallment	Enthrallment, rapture
	Optimism	Eagerness, hope
	Pride	Triumph
	Relief	Relief
	Zest	Enthusiasm, excitement, exhilaration, thrill, zeal
Love	Affection	Adoration, attractiveness, caring, compassion, fondness, liking, sentimentality, tenderness
	Longing	Longing
	Lust/sexual desire	Desire, infatuation, passion
	Disappointment	Dismay, displeasure
	Neglect	Alienation, defeatism, dejection, embarrassment, homesickness, humiliation, insecurity, insult, isolation, loneliness, rejection
Sadness	Sadness	Depression, despair, gloom, glumness, grief, melancholy, misery, sorrow, unhappy, woe
	Shame	Guilt, regret, remorse
	Suffering	Agony, anguish, hurt
	Sympathy	Pity, sympathy
Surprise	Surprise	Amazement, astonishment

sentiment analysis applications because the set of basic emotions may not be fine-grained enough. For example, in one of the applications that I worked on, the client was interested in detecting *optimism* in the financial market. Optimism is not a basic emotion in the list of any theorist, but it is a secondary emotion for *joy* in Table 2.2. Note that although the words in Table 2.2 describe different emotions or states of mind, they can also be used as part of an emotion lexicon in sentiment analysis to spot different kinds of emotions. Of course, they need to be significantly expanded to include those synonymous words and phrases to form a reasonably complete emotion lexicon. In fact, there are some emotion lexicons that have been compiled by researchers, see (Liu 2015). Note also that for sentiment analysis, we do not need to be concerned with the disagreement of theorists. For a particular application, we can choose the types of emotion that are useful to the application. We also do not need to worry about whether they are primary, second or tertiary.

The *emotion annotation and representation language* (EARL) proposed by the Human-Machine Interaction Network on Emotion (HUMAINE) (HUMAINE 2006) has classified 48 emotions into different kinds of positive and negative orientations or valences (Table 2.3). This is useful to us because sentiment analysis is mainly interested in expressions with positive or negative orientations or polarities (also called *valences*). However, we should take note that some emotions do not have positive or negative orientations, e.g., *surprise* and *interest*. Some psychologists felt that these should not be regarded as emotions (Ortony and Turner 1990) simply because they do not have positive or negative orientations or valences. For the same reason, they are not commonly used in sentiment analysis.

Table 2.3 HUMAINE polarity annotations of emotions

Negative and forceful	**Negative and passive**	**Quiet positive**
Anger	Boredom	Calm
Annoyance	Despair	Content
Contempt	Disappointment	Relaxed
Disgust	Hurt	Relieved
Irritation	Sadness	Serene
Negative and not in control	**Positive and lively**	**Caring**
Anxiety	Amusement	Affection
Embarrassment	Delight	Empathy
Fear	Elation	Friendliness
Helplessness	Excitement	Love
Powerlessness	Happiness	
Worry	Joy	
	Pleasure	
Negative thoughts	**Positive thoughts**	**Reactive**
Doubt	Courage	Interest
Envy	Hope	Politeness
Frustration	Pride	Surprised
Guilt	Satisfaction	
Shame	Trust	
Agitation		
Stress		
Shock		
Tension		

We now turn to mood. The types of mood are similar to those of emotion except that the types of emotion that last only momentarily will not usually be moods, e.g., *surprise* and *shock*. Thus, the words or phrases used to express moods are similar to those for emotions too. However, since mood is a feeling that lasts a relatively long time, is diffused, and may not have a clear cause or target object, it is hard to recognize unless a person explicitly says it, e.g., *I feel sad today*. We can also monitor one's writings over a period of time to assess his/her prevailing mood in the period, which can help discover people with prolonged mental or other medical conditions (e.g., chronicle depression) and even the tendency to commit suicides or crimes.

It is also interesting to discover the mood of the general population, e.g., public mood, and the general atmosphere between organizations or countries, e.g., the mood of US and Russian relations, by monitoring the traditional news media and/or social media over a period of time.

2.3.2 Affect, Emotion, and Mood in Sentiment Analysis

The above discussions are only about people's states of mind, which are the subjects of study of psychologists. However, for sentiment analysis, we need to know how such feelings are expressed in natural language and how they can be recognized. This leads us to the linguistics of affect, emotion and mood. Affect as defined by psychologists as a primitive response or feeling with no target is not much of interest to us as almost everything written in text or displayed in the form of facial expressions and other visible signs have already gone through some cognitive processing to become emotion or mood. However, we note that the term affect is still commonly used in linguistics and many other fields to mean emotion and mood.

Wikipedia has a good page describing the linguistic aspect of emotion and mood. There are two main ways that human beings express themselves, speech and writing. In addition to choices of grammatical and lexical expressions, which are common to both speech and writing (see below), speaker emotion can also be conveyed through paralinguistic mechanisms such as intonations, facial expressions, body movements, biophysical signals or changes, gestures, and postures. In writing, special punctuations (e.g., repeated exclamation marks, !!!!), capitalization of all letters of a word, emoticons, and lengthening of words (e.g., *slooooow*) are frequently used, especially in social media.

Regarding choices of grammatical and lexical expressions, there are several common ways that people often employ to express emotions or moods:

1. use emotion or mood words or phrases such as love, disgusting, angry, and upset.
2. describe emotion-related behaviors, e.g., "He cried after he saw his mother" and "After received the news, he jumped up and down for a few minutes like a small boy."
3. use intensifiers. As we discussed in Sect. 2.1.3, common English intensifiers include very, so, extremely, dreadfully, really, awfully (e.g., *awfully bad*), *terribly*

(e.g., *terribly good*), *never* (e.g., "I will never buy any product from them again"), *the sheer number of, on earth* (e.g., "What on earth do you think you are doing?"), *the hell* (e.g., "What the hell are you doing?"), *a hell of a*, etc. To emphasize further, intensifiers may be repeated, e.g., "This car is very very good."

4. use superlatives. Arguably, many superlative expressions also express emotions, e.g., "This car is simply the best."
5. use pejorative (e.g., "*He is a fascist*."), laudatory (e.g., "He is a saint."), and sarcastic expressions (e.g., "What a great car, it broke the second day").
6. use swearing, cursing, insulting, blaming, accusing, and threatening expressions.

My experience is that using these clues is sufficient to recognize emotion and mood in text, although in linguistics, adversative forms, honorific and deferential language, interrogatives, tag questions, and the like may also be employed to express emotional feelings, but their uses are rare and are also hard to recognize computationally.

To design emotion detection algorithms, in addition to considering the above clues, we should be aware that there is a cognitive gap between people's true psychological states of mind and the language that they use to express such states. There are many reasons (e.g., being polite, and do not want people to know one's true feeling) that they may not fully match. Thus, language does not always represent psychological reality. For example, when one says "*I am happy with this car*," one may not have any emotional reaction towards the car although the emotion word *happy* is used. Furthermore, emotion and mood are very difficult to distinguish in written text (Alm 2008). We normally do not distinguish them. When we say emotion, we mean emotion or mood.

Since emotions have targets and most of them also imply positive or negative sentiment, they can be represented and handled in very much the same way as rational opinions. Although a rational opinion emphasizes a person's evaluation about an entity and an emotion emphasizes a person's feeling caused by an entity, emotion can essentially be regarded as sentiment with a stronger intensity (see Sect. 2.1.3). It is often the case that when the sentiment of a person becomes so strong, he/she becomes emotional. For example, "*The hotel manager is not professional*" expresses a rational opinion, while "*I almost cried when the hotel manager talked to me in a hostile manner*" indicates that the author's sentiment reached the emotional level of *sadness* and/or *anger*. The sentiment orientation of an emotion naturally inherits the polarity of the emotion, e.g., *sad*, *anger*, *disgust*, and *fear* are negative, and *love* and *joy* are positive. At the emotional level, sentiment becomes more fine-grained. Additional mechanisms are needed to recognize different types of emotions in writing.

Due to the similarity of emotion and rational opinion, we can still use the quadruple or quintuple representation of opinion (Definitions 1 and 7) to represent emotion. However, if we want to be more precise, we can give it a separate definition based on the quadruple (Definition 1) or quintuple (Definition 7) definitions as the meanings of some components in the tuple are not the exactly same as they were in the opinion definition because emotions focus on personal feelings, while rational opinions focus on evaluations.

Definition 14 (Emotion) An *emotion* is a quintuple,

$$(e, a, m, f, t),$$

where e is the target entity, a is the target aspect of e that is responsible for the emotion, m is the emotion type or a pair representing an emotion type and an intensity level, f is the feeler of the emotion, and t is the time when the emotion is expressed.

For example, for the emotion expressed in the sentence "*I am so upset with the manager of the hotel*," the entity is '*the hotel*,' the aspect is '*the manager*' of the hotel, the emotion type is *anger*, and the emotion feeler is *I* (the author). If we know the time when the emotion was expressed we can add it to the quintuple representation. As another example, in "*After hearing his brother's death, he burst into tears.*" the target entity is '*his brother's death*,' which is an event, and there is no aspect. The emotion type is *sadness* and the emotion feeler is *he*.

In practical applications, we should integrate the analysis of rational opinions and emotions, we may also want to add the sentiment orientation or polarity of an emotion, i.e., whether it is positive (desirable) or negative (undesirable) for the feeler. If that is required, a sentiment component can be included to Definition 14 to make it a sextuple.

Cause of Emotion In Sect. 2.1.5, we discussed the reasons for opinions. In a similar way, emotions have causes as emotions are usually caused by some internal or external events. Here we use the word *cause* instead of *reason* because an emotion is an effect produced by a cause (usually an event), rather than a justification or explanation in support of an opinion. In the above sentence, '*his brother's death*' is the cause for his *sadness* emotion. Actually, '*his brother's death*' is both the target entity and the cause. In many cases, the target and the cause of an emotion are different. For example, in "*I am so mad with the hotel manager because he refused to refund my booking fee,*" the target entity is the *hotel*, the target aspect is the *manager* of the hotel, and the cause of the *anger* emotion is '*he refused to refund my booking fee.*' There is a subtle difference between '*his brother's death*' and '*he refused to refund my booking fee.*' The latter states an action performed by *he* (the hotel manager) that causes the sadness emotion (negative). *He* is the agent of the undesirable action. The sentiment on the hotel manager is negative. The sentence also explicitly stated the *anger* is toward the hotel manager, In the case of '*his brother's death,*' '*his brother*' or *death* alone is not the target of the emotion. It is the whole event that is the target and the cause of the sadness emotion.

Unlike rational opinions, in many emotion and mood sentences, the authors may not explicitly state the entities (e.g., named entities, topics, issues, actions and events) that are responsible for the emotions or moods, e.g., "*I felt a bit sad this morning*" and "*There is sadness in her eyes*." The reason is that a rational opinion sentence focuses on both the opinion target and the sentiment on the target but the opinion holder is often omitted (e.g., "*The pictures from this camera are great*") while an emotion sentence focuses on the feeling of the feeler (e.g., "*There* *is sadness in her eyes*." This means that a rational opinion sentence contains both sentiments and their targets explicitly, but may or may not give the opinion holder. An emotion sentence always has feelers and emotion expressions, but may or may not state the emotion target or the cause (e.g., "*I love this car*" and "*I felt sad this morning*"). This does not mean that some emotions do not have targets or causes. They do, but the targets or the causes may be expressed in previous sentences or implied by the context, which makes extracting targets or causes very difficult. In the case of mood, the causes may be implicit or even unknown and are thus not stated in the text.

2.4 Different Types of Opinions

Opinions can actually be classified along many dimensions. We discuss some main classifications in this section.

2.4.1 Regular and Comparative Opinions

The type of opinion that we have defined is called the *regular opinion* (Liu 2006). Another type is *comparative opinion* (Jindal and Liu 2006b).

Regular Opinion A *regular opinion* is often referred to simply as an *opinion* in the literature. It has two main sub-types (Liu 2006):

Direct opinion: A *direct opinion* is an opinion that is expressed directly on an entity or an entity aspect, e.g., "*The picture quality is great*."

Indirect opinion: An *indirect opinion* is an opinion that is expressed indirectly on an entity or aspect of an entity based on some positive or negative effects on some other entities. This sub-type often occurs in the medical domain. For example, the sentence "*After injection of the drug, my joints felt worse*" describes an undesirable effect of the drug on '*my joints*,' which indirectly gives a negative opinion or sentiment to the drug. In this case, the entity is *the drug* and the aspect is the *effect on joints*. Indirect opinions also occur in other domains, although less frequently. In these cases, they are typically expressed *benefits* (positive) or *issues* (negative) of entities, e.g., "*With this machine, I can finish my work in one hour, which used to take me 5 hours*" and "*After switching to this laptop, my eyes felt much better*." In marketing, benefits of a product or service are regarded as the major selling points. Thus, extracting such benefits is of practical interest.

Comparative Opinion A *comparative opinion* expresses a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities (Jindal and Liu 2006a, b). For example, the sentences "*Coke tastes better than Pepsi*" and "*Coke tastes the best*" express two comparative opinions. A comparative opinion is usually expressed using the *comparative* or *superlative* form of an adjective or adverb, although not always (e.g., *prefer*). The definitions in Sects. 2.1 and 2.2 do not cover comparative opinion. Comparative opinions have many types. See (Liu 2015) for their definitions.

2.4.2 Subjective and Fact-Implied Opinions

Opinions and sentiments are by nature subjective because they are about people's subjective views, appraisals, evaluations, and feelings. But when they are expressed in actual text, they do not have to appear as subjective sentences. People can use objective or factual sentences to express their happiness and displeasure because facts can be desirable or undesirable. Conversely, not all subjective sentences express positive or negative sentiments, e.g., "*I think he went home*," which is a belief and has no positive or negative orientation. Based on subjectivity, we can classify opinions into two types, *subjective opinions* and *fact-implied opinions*. We define them below.

Subjective Opinion An *subjective opinion* is a regular or comparative opinion given in a subjective statement, e.g.,

> "*Coke tastes great*."
> "*I think Google's profit will go up next month*."
> "*This camera is a masterpiece*."
> "*We are seriously concerned about this new policy*."
> "*Coke tastes better than Pepsi*."

We can broadly classified subjective opinions into two categories: *rational opinions* and *emotional opinions* (Sect. 2.1.3).

Fact-Implied Opinion A *fact implied opinion* is a regular or comparative opinion implied in an objective or factual statement. Such an objective statement expresses a desirable or undesirable fact or action. This type of opinion can be further divided into two subtypes:

1. **Personal fact-implied opinion**: Such an opinion is implied by a factual statement about someone's personal experience, e.g.,

> "*I bought the mattress a week ago, and a valley has formed in the middle*."
> "*I bought the toy yesterday and I have already thrown it into the trash can*."
> "*My dad bought the car yesterday and it broke today*."
> "*The battery of this phone lasts longer than that of my previous Samsung phone.*"

Although factual, these sentences tell us whether the opinion holder is positive or negative about the product or his preference among different products. Thus, the opinions implied by these factual sentences are no different from subjective opinions.

2. **Non-personal fact-implied opinion**: This type is entirely different as it does not imply any personal opinion. It often comes from fact reporting and the reported fact does not give any opinion from anyone, e.g.,

   "*Google's revenue went up by 30%.*"
   "*The unemployment rate came down last week.*"
   "*Google made more money than Yahoo last month.*"

Unlike personal facts, these sentences do not express any experience or evaluation from any person. For instance, the first sentence above does not have the same meaning as a sentiment resulted from a person who has used a Google product and expresses a desirable or undesirable fact about the Google product. Since these sentences do not give any personal opinion, they do not have opinion holders although they do have the sources of information. For example, the source of the information in the first sentence above is likely to be Google itself, but it is a fact, not a Google's subjective opinion.

However, we can still treat them as a type of opinion sentences due to the following two reasons:

1. Each of the sentences above does indicate a desirable and/or undesirable state for the involved entities or topics (i.e., *Google*, *Yahoo* and *unemployment rate*) based on our commonsense knowledge.
2. The persons who post the above sentences might be expressing positive or negative opinions implicitly about the involved entities. For example, the person who posted the first sentence on Twitter is likely to have a positive sentiment about Google; otherwise, he/she would probably not post the fact. This kind of posts occur very frequently on Twitter, where Twitter users pick up some news headlines from the traditional media and post them on Twitter. Many people may also re-tweet them.

As we can see, it is important to distinguish personal facts and non-personal facts as opinions induced from non-personal facts represent a very different type of opinions and need a special treatment. How to deal with such facts depends on applications. My recommendation is to assign it the positive or negative orientation based on our commonsense knowledge whether the sentence is about a fact desirable or undesirable to the involved entity, e.g., Google. Users of the sentiment analysis system should be made aware of the convention so that they can make use the opinion appropriately based on their applications.

Sometimes the author who posts such a fact may also give an explicit opinion, e.g.,

"*I am so upset that Google's share price went up today.*"

The clause '*Google's share price went up today*' in the example gives a non-personal fact-implied positive opinion about Google, but the author is negative about it. This is called a *meta-opinion*, an opinion about an opinion.

Subjective opinions are usually easier to deal with because the number of words and phrases that can be used to explicitly express subjective feelings is limited, but this is not the case for fact-implied opinions. There seem to be an infinite number of desirable and undesirable facts and every domain is different. Much of the existing research has focused on subjective opinions. Limited work has been done about fact-implied opinions (Zhang and Liu 2011).

2.4.3 First-Person and Non-First-Person Opinions

In some applications, it is important to distinguish those statements expressing one's own opinions from those statements expressing beliefs about someone else's opinions. For example, in a political election, one votes based on one's belief of each candidate's stances on issues, rather than the true stances of the candidate, which may or may not be the same.

1. **First-person opinion:** Such an opinion states one's own attitude towards an entity. It can be from a person, a representative of a group, or an organization. Here are some example sentences expressing first-person opinions.

   "*Tax increase is bad for the economy*."
   "*I think Google's profit will go up next month*."
   "*We are seriously concerned about this new policy*."
   "*Coke tastes better than Pepsi*."

   Notice that not every sentence needs to explicitly use the first person pronoun "*I*" or "*we,*" or to mention an organization name.

2. **Non-first-person opinion:** Such an opinion is expressed by a person stating someone else's opinion. That is, it is a belief of someone else's opinion about some entities or topics, e.g.,

   "*I think John likes Lenovo PCs*."
   "*Jim loves his iPhone*."
   "*President Obama supports tax increase*."
   "*I believe Obama does not like wars*."

2.4.4 Meta-opinions

Meta-opinions are opinions about opinions. That is, a meta-opinion's target is also an opinion which is usually contained in a subordinate clause. The opinion in the subordinate clause can express either a fact with an implied opinion or a subjective opinion. Let us see some examples:

"*I am so upset that Google's profit went up*"
"*I am very happy that my daughter loves her new Ford car*"
"*I am so sad that Germany lost the game*."

These sentences look quite different from opinion sentences before. But they still follow the same opinion definition in Definition 7. It is just that the target of the meta-opinion in the main clause is now an opinion itself in the subordinate clause. For example, in the first sentence, the author is negative about '*Google's profit went up*,' which is the target of the meta-opinion in the main clause. So the meta-opinion is negative, but its target is a regular positive opinion about '*Google's profit*.' In practice, these two types of opinions should be treated differently. Since meta-opinions are rare, there is little research or practical work about them.

2.5 Author and Reader Standpoint

We can look at an opinion from two perspectives, that of the author (opinion holder) who posts the opinion, and that of the reader who reads the opinion. Since opinions are subjective, naturally the author and the reader may not see the same thing in the same way. Let us use the following two example sentences to illustrate the point:

"*This car is too small for me*."
"*Google's profits went up by 30%*."

Since the author or the opinion holder of the first sentence felt the car is too small, a sentiment analysis system should output a negative opinion about the size of the car. However, this does not mean that the car is too small for everyone. A reader may actually like the small size, and feel positive about it. This causes a problem because if the system outputs only a negative opinion about size, the reader will not know whether it is too small or too large and then he/she would not see this positive aspect for him/her. Fortunately, this problem can be dealt with by mining and summarizing opinion reasons (see Sect. 2.1.2). Here '*too small*' not only indicates a negative opinion about the size but also the reason for the negative opinion. With the reason, the reader can see a more complete picture of the opinion.

The second sentence represents a non-personal fact-implied opinion. As discussed in Sect. 2.4.1, the person who posts the fact is likely to be positive about Google. However, the readers may have different feelings. Those who have financial interests in Google should feel happy, but Google's competitors will not be thrilled. In Sect. 2.4.2, we choose to assign positive sentiment to the opinion because our commonsense knowledge says that the fact is desirable for Google. Users can decide how to use the opinion based on their application needs.

2.6 Summary

This chapter described many facets of sentiment analysis. It started with the definitions of the concepts of opinion, sentiment, and opinion summary. The definitions abstracted a structure from the unstructured natural language text, and also showed that sentiment analysis is a multi-faceted problem with many interrelated sub-problems. Researchers can exploit the inter-relationships to design more robust and accurate solution techniques. This chapter also classified and discussed different types of opinions. Along with these definitions and discussions, the important concepts of affect, emotion and mood were introduced and defined too. They are closely related to, but are also different from conventional rational opinions. Opinions emphasize evaluation or appraisal of some target objects, events or topics (which are collectively called entities in this chapter), while emotions emphasize people's feelings caused by such entities.

After reading this chapter, I am sure that you would agree with me that on the one hand, sentiment analysis is a challenging area of research involving many different tasks and perspectives, and on the other, it is also highly subjective in nature. Thus, I do not expect that you completely agree with me on everything in the chapter. I also do not claim that this chapter covered all important aspects of sentiment and opinion. My goal is to present a reasonably precise definition of sentiment analysis (or opinion mining) and its related concepts, issues, and tasks. I hope I have succeeded to some extent.

References

Alm, Ebba Cecilia Ovesdotter. 2008. *Affect in text and speech*: ProQuest.

Arnold, Magda B. 1960. *Emotion and personality*. New York: Columbia University Press.

Batson, C. Daniel, Laura L. Shaw, and Kathryn C. Oleson. 1992. Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions. *Emotion Review of Personality and Social Psychology* 13: 294–326.

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text*.

Chaudhuri, Arjun. 2006. *Emotion and reason in consumer behavior*. Oxford: Elsevier Butterworth-Heinemann.

Choi, Yejin, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the human language technology conference and the conference on empirical methods in natural language processing (HLT/EMNLP-2005)*.

Ekman, P., W.V. Friesen, and P. Ellsworth. 1982. What emotion categories or dimensions can observers judge from facial behavior? In *Emotion in the human face*, ed. P. Ekman, 98–110. Cambridge: Cambridge University Press.

Gray, Jeffrey A. 1982. *The neuropsychology of anxiety*. Oxford: Oxford University Press.

Hobbs, Jerry R., and Ellen Riloff. 2010. Information extraction. In *In handbook of natural language processing*, ed. N. Indurkhya and F.J. Damerau, 2nd ed. London: Chapman & Hall/CRC Press.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-2004)*.

HUMAINE. 2006. *Emotion annotation and representation language*. Available from: http://emotion-research.net/projects/humaine/earl

Izard, Carroll Ellis. 1971. *The face of emotion*. New York: Appleton-Century-Crofts.

James, William. 1884. What is an emotion? *Mind* 9: 188–205.

Jindal, Nitin and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of ACM SIGIR conference on research and development in information retrieval (SIGIR-2006)*.

———. 2006b. Mining comparative sentences and relations. In *Proceedings of national conference on artificial intelligence (AAAI-2006)*.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of interntional conference on computational linguistics (COLING-2004)*.

Liu, Bing. 2006. *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin: Springer.

———. 2010. Sentiment analysis and subjectivity, in *Handbook of natural language processing, Second Edition*, N. Indurkhya and F.J. Damerau, Editors.

———. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. *Proceedings of international conference on world wide web (WWW-2005)*.

McDougall, William. 1926. *An introduction to social psychology*. Boston: Luce.

Mowrer, Orval Hobart. 1960. *Learning theory and behavior*. New York: Wiley.

Oatley, K., and P.N. Jobnson-Laird. 1987. Towards a cognitive theory of emotions. *Cognition and Emotion* 1: 29–50.

Ortony, Andrew, and Terence J. Turner. 1990. What's basic about basic emotions? *Psychological Review* 97 (3): 315–331.

Panksepp, Jaak. 1982. Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences* 5 (3): 407–422.

Parrott, W. Gerrod. 2001. *Emotions in social psychology: Essential readings*. Philadelphia: Psychology Press.

Plutchik, Robert. 1980. A general psychoevolutionary theory of emotion. In *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, ed. R. Plutchik and H. Kellerman, 3–33. New York: Academic Press.

Qiu, Guang, Bing Liu, Bu Jiajun, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics* 37 (1): 9–27.

Russell, James A. 2003. Core affect and the psychological construction of emotion. *Psychological Review* 10 (1): 145–172.

Sarawagi, Sunita. 2008. Information extraction. *Foundations and Trends in Databases* 1 (3): 261–377.

Tomkins, Silvan. 1984. Affect theory. In *Approaches to emotion*, ed. K.R. Scherer and P. Ekman, 163–195. Hillsdale: Eribaum.

Watson, John B. 1930. *Behaviorism*. Chicago: Chicago University Press.

Weiner, B., and S. Graham. 1984. An attributional approach to emotional development. In *Emotion, cognition and behavior*, ed. C.E. Izard, J. Kagan, and R.B. Zajonc, 167–191. New York: Cambridge University Press.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39 (2): 165–210.

Zhang, Lei and Bing Liu. 2011. Identifying noun product features that imply opinions. In *Proceedings of the annual meeting of the Association for Computational Linguistics (short paper) (ACL-2011)*.

Zhuang, Li, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of ACM international conference on information and knowledge management (CIKM-2006)*.

Chapter 3
Reflections on Sentiment/Opinion Analysis

Jiwei Li and Eduard Hovy

Abstract The detection of expressions of sentiment in online text has become a popular Natural Language Processing application. The task is commonly defined as identifying the words or phrases in a given fragment of text in which the reader understands that the author expresses some person's positive, negative, or perhaps neutral attitude toward a topic. These four elements—expression words, attitude holder, topic, and attitude value—have evolved with hardly any discussion in the literature about their foundation or nature. Specifically, the use of two (or three) attitude values is far more simplistic than many examples of real language show. In this paper we ask: where do sentiments come from? We focus on two basic sources of human attitude—the holder's non-logical/emotional preferences and the fulfillment of the holder's goals. After exploring each source we provide a notional algorithm sketch and examples of how sentiment systems could provide richer and more realistic accounts of sentiment in text.

Keywords Sentiment analysis • Opinion mining • Natural language processing • Aspect extraction • Psychology of emotions

3.1 Introduction

Sentiment analysis is an application of natural language processing that focuses on identifying expressions that reflect authors' opinion-based attitude (i.e., good or bad, like or dislike) toward entities (e.g., products, topics, issues) or facets of them (e.g., price, quality).

Since the early 2000s, a large number of models and frameworks have been introduced to address this application, with emphasis on various aspects like opinion related entity exaction, review mining, topic mining, sentiment summarization, rec-

J. Li (✉)
Computer Science Department, Stanford University, Stanford, 94305, CA, USA
e-mail: jiweil@stanford.edu

E. Hovy
Language Technology Institute, Carnegie Mellon University, Pittsburgh, 15213, PA, USA
e-mail: hovy@cmu.edu

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_3

ommendation, and these extracted from significantly diverse text sources including product reviews, news articles, social media (blogs, Twitter, forum discussions), and so on.

However, despite this activity, disappointingly little has been published about what exactly a *sentiment* or *opinion* actually is. It is generally simply assumed that two (or perhaps three) polar values *positive, negative, neutral*) are enough, and that they are clear, and that anyone would agree on how to assign such labels to arbitrary texts. Further, existing methods, despite employing increasingly sophisticated (and of course more powerful) models (e.g., neural nets), still essentially boil down to considering individual or local combinations of words and matching them against predefined lists of words with fixed sentiment values, and thus hardly transcend what was described in the early work by Pang et al. (2002).

There is nothing against simple methods when they work, but they do not always work, and without some discussion of why not, and where to go next, this application remains rather technically uninteresting. The goal of this paper is to identify gaps in the current sentiment analysis literature and to outline practical computational ways to address these issues.

Goals, Expectations and Sentiments. We begin with the fundamental question "What make people hold positive attitudes towards some entities and negative attitudes toward others?". The answer to this question is a psychological state that relates to the opinion holder's satisfaction and dissatisfaction with some aspect of the topic in question. One of only two principal factors determines the answer: either (1) the holder's deep emotionally-driven, non-logical **native preferences**, or (2) whether (and how well) one of the holder's **goals** is fulfilled, and how (in what ways) the goal is fulfilled.

Examples of the former are reflected in sentences like "I just like red" or "seeing that makes me happy". They are typified by adverbs like "just" and "simply" that suggest that no further conscious psychological reflection or motivation obtains. Of this class of factor we can say nothing computationally, and do not address it in the rest of this chapter.

Fortunately, a large proportion of the attitudes people write about reflect the other factor, which one can summarize as goal-driven utility. This relates primarily to Consequentialism: both to Utilitarianism, in which pleasure, economic well-being and the lack of suffering are considered desirable, but also to the general case that morally justifiable actions (and the objects that enable them) are desirable. That is, the ultimate basis for any judgment about the rightness or wrongness of one's actions, and hence of the objects that support/enable them, is a consideration of their outcome, or consequence.

In everyday life, people establish and maintain goals or expectations, both long-term or short-term, urgent or not-urgent, ones. Achieving these goals would fill one with satisfaction, otherwise dissatisfaction: a man walks into a restaurant to achieve the goal of getting full, he cannot be satisfied if all food was sold out (the main goal not being achieved). A voter would not be satisfied if his candidate or party fails to win an election, since the longer-term consequences would generally work against his own preferences. The generation of sentiment-related texts is guided by such sorts of mental satisfaction and dissatisfaction induced by goals being achieved or needs being fulfilled.

We next provide some examples to illustrate why identifying these aspects is essential and fundamental for adequate sentiment/opinion analysis. Following the most popular motivation for computational sentiment analysis, suppose we wish to analyze customers' opinions towards a product or an offering. It is not sufficient to simply determine that someone likes or dislikes something; to make that knowledge useful and actionable, one also wants to know *why* that is the case. Especially when one would like to change the opinion, it is important to determine what it is about the topic that needs to be changed.

Case (1)

- **Question**: *Why did the customer like detergent X?*
- **Customer's review**: *The detergent removes stubborn stains.*

No general sentiment indicator is found in the above review. But the review directly provides the reason, and assuming his/her goal of clean clothing is achieved, it is evident that the opinion holder holds a positive opinion towards the detergent.

Case (2)

- **Question**: *Why did the traveller dislike flight Y?*
- **Customer's review**: *The food was good. The crew was helpful and took care of everything. The service was efficient. However the flight was supposed to take 1.5 h but was 3 h late, and I missed my next connecting flight.*

The major goal of taking a flight is to get to your destination, which is more important than goals like enjoying one's food and receiving pampering service. While multiple simultaneous goals induce competing opinion decisions, the presence of an importance ranking among them determines the overall sentiment.

Case (3)

- **Question**: *Why did the customer visit restaurant Z?*
- **Review1**: *The food is bad.*
- **Review2**: *The waiter was kind but the food was bad.*
- **Review3**: *The food was good but the waiter was rude.*

Although the primary goal of being sated may be achieved, secondary goals such as enjoying the food and receiving respectful service can be violated in various combinations. Often, these goals pertain to the method by which the primary goal was achieved; in other words, to the question "how?" rather than "why?".

A sentiment determination algorithm that can provide more than just a simple opinion label thus has to pay attention both to the primary reason behind the holder's involvement with the topic ("why?") and to the secondary reasons (both "why?" and "how?"), and has to be able to determine their relative importance and relationship to the primary goal.

Goals and Expectations are Personal. As different people (opinion holders) are from different backgrounds, have different personalities, and are in different situations, they have different goals, needs, and the expectations of life. This diversity generally leads to completely diverse opinions towards the same entity, the same action, and the same situation: a billionaire wouldn't be the least bit concerned with the price in a bread shop but would consider the quality, while a beggar might care only about the price. This rather banal observation is explained best by Maslow's famous hierarchy of needs (Maslow 1943), in which the beggar's attention focuses on Maslow's Physiological needs while the billionaire's focuses on Self-Actualization; more on this in Sect. 3.3.1.

Life Requires Trade-offs. Most situations in real life address many personal needs simultaneously. People thus face trade-offs between their goals, which entails sacrificing the achievement of one goal for the satisfaction of another. Given the variability among people, the rankings and decision procedures will also from individual to individual. However, Maslow's hierarchy describes the general behavioral trends of people in most societies and situations.

Complex Sentiment Expressions. As far as we see, current opinion analysis frameworks mostly fail to address the kinds of issues mentioned above, and thereby impair a deeper understanding about opinion or sentiment. As a result, they find it impossible to provide even rudimentary approaches to cases such as the following (from Hovy 2015):

1. *Peter thinks the pants are great and I cannot agree more.*
2. *Peter thinks the pants are great but I don't agree*.
3. *Sometime I like it but sometimes I hate it*.
4. *He was half excited, half terrified*.
5. *The movie is indeed wonderful, but for some reason, I just don't like it*.
6. *Why I won't buy this game even though I like it.*

In this paper, we explore the feasibility of addressing these issues in a practical way using machine learning techniques currently available.

3.2 A Review of Current Sentiment Analysis

Here we give a brief overview of tasks in current sentiment analysis literature. More details can be found in Liu (2010, 2012).

The key points involved at the algorithm level in the sentiment analysis literature follow the basic approaches of statistical machine learning, in which a gold-standard labeling of training data is obtained through manual annotation or other data harvesting approaches (e.g., semi-supervised or weakly supervised), and this is then used to train a variety of association-learning techniques who are then tested on new material. Usually, some text unit has to be identified and then associated with a sentiment label (e.g., positive, neutral, negative). Based on the annotated dataset, the techniques learn that vocabulary items like "bad", "awful", and "disgusting" are negative sentiment indicators while "good", "fantastic" and "awesome" are positive ones. The main complexity lies in learning which words carry some opinion and, especially, what to decide in cases where different words with opposite labels appear in the same clause.

Basic sentiment analysis identifies the simple polarity of a text unit (e.g., a token, a phrase, a sentence, or a document) and is framed as a binary or multi-class classification task; see for example Pang et al's work (2002) that uses a unigram/bigram feature-based SVM classifier. Over the past 15 years, techniques have evolved from simple rule-based word matching to more sophisticated feature and signal (e.g., local word composition, facets of topics, opinion holder) identification and combination, from the level of single tokens to entire documents, and from 'flat' word strings without any syntactic structure at all to incorporation of complex linguistic structures (e.g., discourse or mixed-affect sentences); see (Pang and Lee 2004; Hu and Liu 2004; Wiebe et al. 2005; Nakagawa et al. 2010; Maas et al. 2011; Tang et al. 2014a,b; Qiu et al. 2011; Wang and Manning 2012; Yang and Cardie 2014a; Snyder and Barzilay 2007). Recent progress in neural models provides new techniques for local composition of both opinion and structure (e.g., subordination, conjunction) using distributed representations of text units (e.g., Socher et al. 2013; Irsoy and Cardie 2014a,b; Tang 2015; Tang et al. 2014c).

A supporting line of research extends the basic sentiment classification to include related aspects and facets, such as identifying opinion holders, the topics of opinions, topics not explicitly mentioned in the text, etc.; see (Choi et al. 2006; Kim and Hovy 2006, 2004; Li and Hovy 2014; Jin et al. 2009; Breck et al. 2007; Johansson and Moschitti 2010; Yang and Cardie 2012, 2013, 2014b). These approaches usually employ sequence labeling models (e.g., CRF (Lafferty et al. 2001), HMM (Liu et al. 2004)) to identify whether the current token corresponds to a specific sentiment-related aspect or facet.

An important part of such supportive work is the identification of the relevant aspects or facets of the topic (e.g., the ambience of a restaurant vs. its food or staff or cleanliness) and the correspondent sentiment; see (Brody and Elhadad 2010; Lu et al. 2011; Titov and McDonald 2008; Jo and Oh 2011; Xueke et al. 2013; Kim et al. 2013; García-Moya et al. 2013; Wang et al. 2011; Moghaddam and Ester 2012). Online reviews (about products or offerings) in crowdsourcing and traditional sites (e.g., yelp, Amazon, Consumer Reports) include some sort of aspect-oriented star rating systems where more stars indicate higher level of satisfaction. Consumers rely on these user-generated online reviews when making purchase decisions. To tackle this issue, researchers invent aspect identification or target extraction approaches as one subfield of sentiment analysis. These approaches first identify 'aspects/facets of the principal Topic and then discover authors' corresponding opinions for each one; e.g., (Brody and Elhadad 2010; Titov and McDonald 2008). Aspects are usually identified either manually or automatically using word clustering models (e.g., LDA (Blei et al. 2003) or pLSA). However, real life is usually a lot more complex and much harder to break into a series of facets (e.g., quality of living, marriage, career).

Other related work includes opinion summarization, aiming to summary sentiment key points given long texts (e.g., Hu and Liu 2004; Liu et al. 2005; Zhuang et al. 2006; Ku et al. 2006), opinion spam detection aiming at identifying fictitious reviews generated to deceive readers (e.g., Ott et al. 2011; Li et al. 2014, 2013; Jindal and Liu 2008; Lim et al. 2010), sentiment text generation (e.g., Mohammad 2011; Blair-Goldensohn et al. 2008), and large-scale sentiment/mood analysis on social media for trend detecion (e.g., O'Connor et al. 2010; Bollen et al. 2011; Conover et al. 2011; Paul and Dredze 2011).

3.3 The Needs and Goals Behind Sentiments

As outlined in Sect. 3.1, this chapter argues that an adequate and complete account of utilitarian-based sentiment is possible only with reference to the goals of the opinion holder. In this section we discuss a classic model of human needs and associated goals and then outline a method for determining such goals from text.

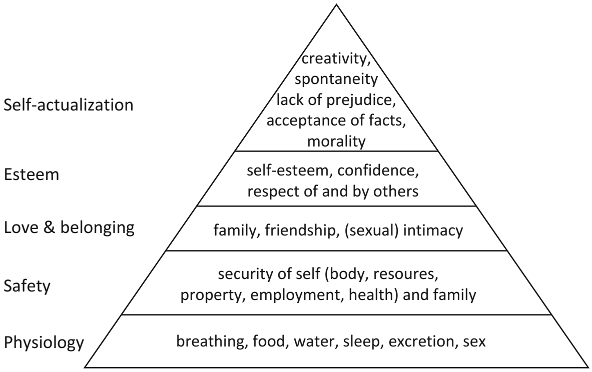
3.3.1 Maslow's Hierarchy of Needs

Abraham Maslow (Maslow 1943, 1967, 1971; Maslow et al. 1970) developed a theory of the basic human needs as being organized in a hierarchy of importance, visualized using a pyramid (shown in Fig. 3.1), where needs at the bottom are the most pressing, basic, and fundamental to human life (that is, the human will tend to choose to satisfy them first before progressing to needs higher up).

According to Maslow's theory, the most basic two levels of human needs are[1]:

- Physiological needs: breathing, food, water, sleep, sex, excretion, etc.
- Safety Needs: security of body, employment, property, heath, etc.

Fig. 3.1 Maslow's hierarchy of needs



[1]References from
https://en.wikipedia.org/wiki/Abraham_Maslow;
https://en.wikipedia.org/wiki/Maslow's_hierarchy_of_needs;
http://www.edpsycinteractive.org/topics/conation/maslow.html

which are essential for the physical survival of a person. Once these needs are satisfied, people tend to accomplish more and move to higher levels:

- Love and Belonging: psychological needs like friendship, family, sexual intimacy.
- Esteem: the need to be competent and recognized such as through status and level of success like achievement, respect by others, etc.

These four types of needs are also referred to as DEFICIT NEEDS (or D-NEEDS), meaning that for any human, if he or she doesn't have enough of any of them, he or she will experience the desire to obtain them. Less pressing than the D-needs are the so-called GROWTH NEEDS, including Cognitive, Aesthetic (need for harmony, order and beauty), and Self-actualization (described by Maslow as "the desire to accomplish everything that one can, to become the most that one can be"). Growth needs are more generalized, obscure, and computationally challenging. We focus in this chapter on deficit needs. For further reading, refer to Maslow's original papers (1943, 1967) or relevant Wikipedia pages.

We note that real life offers many situations in which an action does not easily align with a need listed in the hierarchy (for example, the goal of British troops to arrest an Irish Republican Army leader or of US troops to attack Iraq). Additionally, a single action (e.g., going to college, looking for a job) can simultaneously address multiple needs. Putting aside such complex situations in this chapter, we focus on more tractable situations to illustrate the key points.[2]

3.3.2 Finding Appropriate Goals for Actions and Entities

Typically, each deficit need gives rise to one or more goals that impel the agent (the opinion holder) to appropriate action. Following standard AI and Cognitive Science practice, we assume that the agent instantiates one or more plans to achieve his or her goals, where a plan is a sequence of actions intended to alter the state of the world from some situation (typically, the agent's initial state) to a situation in which the goal has been achieved and the need satisfied. In each plan, its actions, their preconditions, and the entities used in performing them (the plan's so-called *props*) constitute the material upon which sentiment analysis operates. For example, the goal to *sate one's hunger* may be achieved by plans such as *visit-restaurant, cook-and-eat-meal-at-home, buy-or-steal-ready-made-food, cadge-meal-invitation*, etc. In all these plans, *food* is one of the props. For the restaurant and buying-food plans, an affordable *price* is an important precondition.

[2]However, putting them aside them doesn't mean that we don't need to explore and explain these complex situations. On the contrary, these situations are essential and fundamental to the understanding of opinion and sentiment, but requires deeper and more systematic exploration in psychology, cognitive science, and AI.

A sentiment detection system that seeks to understand why the holder holds a specific opinion valence has to determine the specific actions, preconditions, and props that are relevant to the holder's goal, and to what degree they suffice. In principle, a complete account requires the system to infer from the given text:

1. what need is active,
2. which goal(s) have been activated to address the need,
3. which plan(s) is/are being followed to achieve the goal(s),
4. which actions, preconditions, and props appear in these plan(s),
5. which of these is/are being talked about in the text,
6. how well it/they actually have furthered the agent's plan(s),

from which the sentiment valence can be automatically deduced. When the valence is given in the text, one can work 'backwards' to infer step 6, and possibly even earlier steps.

Determining all this is a tall order for computational systems. Fortunately, it is possible to circumvent much of this reasoning in practice. For most common situations, a relatively small set of goals and plans obtains, and the relevant actions, preconditions, and props are usually quite standard. (In fact, they are precisely what is typically called 'facets' in the sentiment analysis literature, for which, as described in Sect. 3.2, various techniques have been investigated, albeit without a clear understanding of the reason these facets are important.)

Given this, the principal unaddressed computational problem today is the determination from the text of the original need or goal being experienced by the holder, since that is what ties together all the other (and currently investigated) aspects. How can one, for a given topic, determine the goals an agent would typically have for it, suggest likely plans, and potentially pinpoint specific actions, preconditions, and props?

One approach is to perform automated goal and plan harvesting, using typical text mining / pattern-matching approaches from Information Extraction. This is a relatively mature application of NLP (Hearst 1992; Riloff and Shepherd 1997; Riloff and Jones 1999; Snow et al. 2004; Davidov and Rappoport 2006; Etzioni et al. 2005; Banko 2009; Mitchell et al. 2009; Ritter et al. 2009; Kozareva and Hovy 2013), and the harvesting power and behavior of various styles of patterns has been investigated for over two decades. (In practice, the Double-Anchored Pattern (DAP) method (Kozareva and Hovy 2013) works better than most others.) Stated simply, one creates or automatically induces text patterns anchored on the topic (e.g., a camera) such as

"I want a camera because *"
"If I had a camera I could *"
"the main reason to get a camera is *"
"wanted to *, so he bought a camera" etc.

and then extracts from large amounts of text the matched VPs and NPs as being relevant to the topic. Appropriately rephrased and categorized, one obtains the information harvested by these patterns would provide typical goals (reasons) for buying and using cameras.

3.4 Toward a Practical Computational Approach

We are now ready to describe the overall approach necessary for a more complete sentiment analysis system. For illustrative purposes we focus on simple binary (positive/negative) valence identification. However, the framework applies to finer granularity (e.g., multi-class classification, regression) with minor adjustments. We first provide an overall algorithm sketch, provide a series of examples, and then suggest models for determining the still unexplored aspects required for deeper sentiment analysis.

First, we assume that standard techniques are employed to find the following from some given text:

1. Opinion Holder: Individual or organization holding the opinion.
2. Entity/Aspect/Theme/Facet: topic or aspect about which the opinion is held.
3. Sentiment Indicator: Sentiment-related text (tokens, phrases, sentences, etc.) that indicate the polarity of the holder.
4. Valence: *like, neutral*, or *dislike*.

These have been defined (or at least used with implicit definition) throughout the sentiment literature, and are defined for example in Hovy (2015). Of these, item 1 is usually achieved by simple matching. Item 2 can be partially addressed by recent topic/facet mining models, and item 3 can be addressed by existing sentiment related algorithms at the word-, sentence-, or text-level. Item 4 at its simplest is a matter of keyword matching, but the composition witin a sentence of contrasting valences has generated some interesting researech. Annotated corpora (or other semi-supervised data harvesting techniques) might be needed for goal and need identification, as discussed above.

Given this, the following sketch algorithm implements deeper sentiment analysis:

1. In the text, identify the key goal underlying the Theme.
2. Is there is no apparent goal?
   - If yes, the opinion is probably non-utilitarian, so find and return a valence if any, but return no reason for it.
   - If no, go to step 3.
3. Determine whether the goal is satisfied:
   - If yes, go to step 4,
   - If no, return a negative valence.
4. Identify the subgoals involved in achieving the major goal.
5. Identify how well the subgoals are satisfied.
6. Determine the final utilitarian sentiment based on the trade-off between different subgoals, and return it together with the trade-off analysis as the reasoning.

This procedure requires the determination of the Goals or Subgoals and the Condition/Situation under which the opinion holder holds that opinion. The former is discussed above; the latter can usually bet determined from the context of the given text.

3.4.1 Examples and Illustration

As a running example we use simple restaurant reviews, sentences in italics indicating original text from the reviews[3]:

Case 1

1. *My friends and I went to restaurant X.*
2. *So many people were waiting there and we left without eating.*

Following the algorithm sketch, the question **"was the major goal of going to a restaurant fulfilled?"** is answered **no**. The reviewer is predicted to hold a negative sentiment. Similar reasoning applies to Case 2 in Sect. 3.1.

Case 2

1. *My friends and I went to restaurant X.*
2. *The waiter was friendly and knowledgeable.*
3. *We ordered curry chicken, potato chips and italian sausage. The Italian sausage was delicious.*
4. *Overall the food was appetizing,*
5. *but I just didn't enjoy the experience.*

To the question "**was the major goal of being full fulfilled?**" the answer is **yes**, as the food was ordered and eaten. Next the algorithms addresses the *how* (manner of achievement) question described in steps 4–6, which involves the functional elements of goals/needs embedded in each sentence:

1. *My friends and I went to restaurant X.*
   Opinion Holder: I
   Entity/Aspect/Theme: restaurant X
   Need: sate hunger
   Goal: visit restaurant
   Sentiment Indicator: none
   Valence: neutral Condition: in restaurant X
2. *The waiter was friendly and knowledgeable.*
   Opinion Holder: I

[3]These reviews were originally from yelp reviews and revised by the authors for illustration purposes.

Entity/Aspect/Theme: waiter
Need: gather respect/friendship
Subgoal: order food
Sentiment Indicator: friendly, knowledgeable
Valence: positive
Condition: in restaurant X

3. *We ordered curry chicken, potato chips and italian sausage. Italian sausage was delicious.*
Opinion Holder: I
Entity/Aspect/Theme: Italian sausage
Need: sate hunger
Subgoal: eat food
Sentiment Indicator: delicious
Valence: positive
Condition: in restaurant X
4. *Overall the food was appetizing,*
Opinion Holder: I
Entity/Aspect/Theme: food
Need: sate hunger
Subgoal: eat enough to remove hunger
Sentiment Indicator: appetizing
Valence: positive
Condition: in restaurant X
5. *but I just didn't enjoy the experience.*
Opinion Holder: I
Entity/Aspect/Theme: restaurant visit experience
Need: none — this is not utilitarian
Goal: none
Sentiment Indicator: didn't enjoy
Sentiment Label: negative
Condition: in restaurant X

The analysis of the needs/goals and their respective positive and negative valences allows one to justify the various sentiment statements, and (in the case of tie final negative decision) also indicate that it is not based on utilitarian considerations.

3.4.2 A Computational Model of Each Part

Current computational models can be used to address each of the aspects involved in the sketch algorithm. We provide only a high-level description of each.

Deciding Functional Elements. Case 2 above involves three of the needs described in Maslow's hierarchy: food, respect/friendship, and emotion. The first two are stated to have been achieved. The third is a pure emotion, expressed without a reason, why the holder "just didn't enjoy the experience". Pure emotions usually have no overt utilitarian value but only relate to the holder's high-level goal of being happy. In this example, we have to conclude that since all overt goals were met, either some unstated utilitarian Maslow-type need was not met, or the holder's opinion stems from a deeper psychological/emotional bias, of the kind mentioned in Sect. 3.1, that goes beyond utilitarian value.

Whether the Major Goal is Achieved. To make a decision about goal achievement, one must: (1) identify the goal/subgoal of an action (e.g., buying the detergent, going to a restaurant); (2) identify whether that goal/subgoal is achieved. The two steps can be computed either separately or jointly using current machine learning models and techniques, including:

- **Joint Model**: Annotate corpora for satisfaction or not for all goals and subgoals together, and train a single machine learning algorithm.
- **Separate Model**:
    1. Determine the goal and its plans and subgoals either through annotation or as described in Sect. 3.3.2.
    2. Associate the actions or entities of the Theme (e.g., going to a restaurant; buying a car) with their respective (sub)goals.
    3. Align each subgoal with indicator sentence(s) in the document (e.g., "I got a small portion"; "the car was all it was supposed to be").
    4. Decide whether the subgoal is satisfied based on indicator sentence(s).

Learning Weights for Different Goals/Needs. One can clearly infer that the customer in case 2 assigns more weight to the emotional aspect, that being his or her final conclusion, and less to the food or respect/friendship (which comes last in this scenario). More formally, for a given text D, we discover L needs/(sub)goals, with indices $1, 2, \ldots, L$. Each type of need/(sub)goal $i \in [1, L]$ is associated with a weight that contributes to the final sentiment valence decision v_i. In document D, each type of need i is associated with achievement value a_i that indicates how the need or goal is satisfied. The sentiment score S_D for given document D is then given by:

$$S_D = \sum_{i \in [1,L]} v_i \cdot a_i$$

This simple approach is comparable to a regression model that assigns weights to relevant aspects, where gold standard examples can be the overall ratings of the labeled restaurant reviews. One can view such a weight decision procedure as a supervised regression model by assigning a weight value to each discovered need. Such a procedure is similar to latent aspect rating introduced in Wang et al. (2011); Zhao et al. (2010) by learning aspect weight (i.e., value, room, location, or service) for hotel review ratings. A simple illustrative example might be collaborative filtering in recommendation systems, e.g., Breese et al. (1998); Sarwar et al. (2001), optimizing need weight regarding each respective individual (which could be sampled from a uniform prior for humans' generally accepted weights).

Since individual expectations can differ, it would be advantageous to maintain opinion holder profiles (for example, both yelp and Amazon keep individual profiles for each customer) that record one's long-term activity. This would support individual analysis of background, personality, or social identity, and enable learning of specific goal weights for different individuals.

When these issues have been addressed, one can start asking deeper questions like:

- *Q: Why does John like his current job though his salary is low?*
  A: He weighs employment more highly than family.
- *Q: How wealthy is a particular opinion holder?*
  A: He might be rich as he places little concern (weight) on money.

or make user-oriented recommendations like:

- *Q: Should the system recommend an expensive–but-luxurious hotel or a cheap-but-poor hotel?*

3.4.3 Prior/Default Knowledge About Opinion Holders

Sentiment/opinion analysis can be considerably assisted by the existence of a knowledge base that provides information about the typical preferences of the holder.

Individuals' goals vary across backgrounds, ages, nationalities, genders, etc. An engineer would have different life goals from a businessman, or a doctor, a citizen living in South America would have different weighing systems from those in Europe or the United States, people in wartime would have different life expectations from when in peacetime. Two general methods exist today for practically collecting such standardized knowledge to construct a relevant knowledge base:

(1) Rule-based Approaches. Hierarchies of personality profiles have been proposed, and changes to them have long been explored in the social and developmental psychology literature, usually based on polls or surveys. For example, (1981) found that children have higher physical needs than other age groups, love needs emerging in the transitional period from childhood to adulthood; esteem needs are the highest among adolescents; the highest self-actualization levels are found with adults; and the highest levels of security are found at older ages. As another example, researchers (Tang and Ibrahim 1998; Tang et al. 2002; Tang and West 1997) have found that survival (i.e., physiological and safety) needs dominate during wartime while psychological needs (i.e., love, self-esteem, and self-actualization) surface during peacetime, which is in line with our expectations. For computational implementation, however, these sorts of studies provide very limited evidence, since only a few aspects are typically explored.

(2) **Computational Inference Approaches.** Despite the lack of information about individuals, reasonable preferences can be inferred from other resources such as online social media. A vast section of the Social Network Analysis research focuses on this problem, as well as much of the research of the large web search engine companies. Networking websites like Facebook, LinkedIn, and Google Plus provide rich repositories of personal information about individual attributes such as education, employment, nationality, religion, likes and dislikes, etc. Additionally, online posts usually offer direct evidence for such attributes. Some examples include age (Rao et al. 2010; Rao and Yarowsky 2010), gender (Ciot et al. 2013), living location (Sadilek et al. 2012), and education (Mislove et al. 2010).

3.5 Conclusion and Discussion

The past 15 years has witnessed significant performance improvements in training machine learning algorithms for the sentiment/opinion identification application. But little progress has been made toward a deeper understanding about what opinions or sentiments are, why people hold them, and why and how their facets are chosen and expressed. No-one can deny the unprecedented contributions of statistical learning algorithms in modern-day (post-1990s) NLP, for this application as for others. However, ignoring cognitive and psychological perspectives in favor of engineering alone inevitably hampers progress once the algorithms asymptote to their optimal performance, since understanding *how* to do something doesn't necessarily lead to better insight about *what* needs to be done, or how it is best represented. For example, when inter-annotator agreement on sentiment labels peaks at 0.79 even for the rather crude 3-way sentiment granularity of positive/neutral/negative (Ogneva 2010), is that the theoretical best that could be achieved? How could one ever know, without understanding what other aspects of sentiment/opinion are pertinent and investigating whether they could constrain the annotation task and help boost annotation agreement?

In this paper, we described possible directions for deeper understanding, helping bridge the gap between psychology / cognitive science and computational approaches. We focus on the opinion holder's underlying needs and their resultant goals, which, in a utilitarian model of sentiment, provides the basis for explaining the reason a sentiment valence is held. (The complementary non-utilitarian, purely intuitive preference-based basis for some sentiment decisions is a topic requiring altogether different treatment.) While these thoughts are still immature, scattered, unstructured, and even imaginary, we believe that these perspectives might suggest fruitful avenues for various kinds of future work.

References

Banko, Michelle. 2009. Ph.D. Dissertation, University of Washington.

Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, vol. 14.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1): 1–8.

Breck, Eric, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI*.

Breese, John S, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 43–52. Morgan Kaufmann Publishers Inc.

Brody, Samuel, and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 804–812. Association for Computational Linguistics.

Choi, Yejin, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *EMNLP*.

Ciot, Morgane, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-English contexts. In *EMNLP*, 1136–1145.

Conover, Michael, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *ICWSM*.

Davidov, A., and D. Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics COLING and the 44th Annual Meeting of the ACL*, 297–304.

Etzioni, O., M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, and S. Soderland et al. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence* 165(1): 91–134.

García-Moya, Lisette, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. 2013. Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems* 28(3): 0019–27.

Goebel, Barbara L, and Delores R Brown. 1981. Age differences in motivation related to Maslow's need hierarchy. *Developmental Psychology* 17(6): 809.

Hearst, Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, 539–545.

Hovy, Eduard H. 2015. What are sentiment, affect, and emotion? Applying the methodology of Michael zock to sentiment analysis. In *Language production, cognition, and the Lexicon*, 13–24. Cham: Springer.

Hu, Minqing, and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, vol. 4, 755–760.

Irsoy, Ozan, and Claire Cardie. 2014a. Deep recursive neural networks for compositionality in language. In *Advances in neural information processing systems*, 2096–2104. Cham: Springer.

Irsoy, Ozan, and Claire Cardie. 2014b. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 720–728.

Jin, Wei, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized HMM-based learning framework for web opinion mining. In *ICML*.

Jindal, Nitin, and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 219–230. ACM.

Jo, Yohan, and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 815–824. ACM.

Johansson, Richard, and Alessandro Moschitti. 2010. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*.

Kim, Soo-Min, and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367. Association for Computational Linguistics.

Kim, Soo-Min, and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*.

Kim, Suin, Jianwen Zhang, Zheng Chen, Alice H Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *AAAI*.

Kozareva, Z., and E.H Hovy. 2013. Tailoring the automated construction of large-scale taxonomies using the web. *Journal of Language Resources and Evaluation* 47: 859–890.

Ku, Lun-Wei, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. 100107.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*. Vol. 1.

Li, Jiwei, and Eduard H. Hovy. 2014. Sentiment analysis on the people's daily. In *EMNLP*, 467–476.

Li, Jiwei, Myle Ott, and Claire Cardie. 2013. Identifying manipulated offerings on review portals. In *EMNLP*, 1933–1942.

Li, Jiwei, Myle Ott, Claire Cardie, and Eduard H. Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *ACL* (1), 1566–1576.

Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 939–948. ACM.

Liu, Bing. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing* 2: 627–666.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1): 1–167.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, 342–351. ACM.

Liu, Yun-zhong, Ya-ping Lin, and Zhi-ping Chen. 2004. Text information extraction based on hidden Markov model [J]. *Acta Simulata Systematica Sinica* 3: 038.

Lu, Bin, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, 81–88. IEEE.

Maas, Andrew L, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 142–150. Association for Computational Linguistics.

Maslow, Abraham Harold, Robert Frager, James Fadiman, Cynthia McReynolds, and Ruth Cox. 1970. *Motivation and personality*, vol. 2. New York: Harper & Row.

Maslow, Abraham Harold. 1943. A theory of human motivation. *Psychological Review* 50(4): 370.

Maslow, Abraham H. 1967. A theory of metamotivation: The biological rooting of the value-life. *Journal of Humanistic Psychology* 7(2): 93–127.

Maslow, Abraham H. 1971. *The farther reaches of human nature*.

Mislove, Alan, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 251–260. ACM.

Mitchell, T.M., J. Betteridge, A. Carlson, E. Hruschka, and R. Wang. 2009. Populating the semantic web by macro-reading internet text. In *Proceedings of the 8th International Semantic Web Conference (ISWC)*.

Moghaddam, Samaneh, and Martin Ester. 2012. On the design of LDA models for aspect-based opinion mining. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 803–812. ACM.

Mohammad, Saif. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105–114. Association for Computational Linguistics.

Nakagawa, Tetsuji, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 786–794. Association for Computational Linguistics.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11: 122–129.

Ogneva, Maria. 2010. How companies can use sentiment analysis to improve their business. Mashable. http://mashable.com/2010/04/19/sentiment-analysis/

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 309–319. Association for Computational Linguistics.

Pang, Bo, and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, 79–86. Association for Computational Linguistics.

Paul, Michael J, and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 265–272.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics* 37(1): 9–27.

Rao, Delip, and David Yarowsky. 2010. Detecting latent user properties in social media. In *Proceedings of the NIPS MLSN Workshop*.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, 37–44. ACM.

Riloff, J., and E. Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 117–124.

Riloff, R., and E. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI)*, 474–479.

Ritter, A., S. Soderland, and O. Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read*.

Sadilek, Adam, Henry Kautz, and Jeffrey P Bigham. 2012. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 723–732. ACM.

Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, 285–295. ACM.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, vol. 17, 1297–1304.

Snyder, Benjamin, and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, 300–307.

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1631, 1642, Citeseer.

Tang, Thomas Li-Ping, and Abdul H Safwat Ibrahim. 1998. Importance of human needs during retrospective peacetime and the persian gulf war: Mideastern employees. *International Journal of Stress Management* 5(1): 25–37.

Tang, Thomas Li-Ping, and W. Beryl West. 1997. The importance of human needs during peacetime, retrospective peacetime, and the persian gulf war. *International Journal of Stress Management* 4(1): 47–62.

Tang, T.L.P, A.H.S Ibrahim, and W.B. West. 2002. Effects of war-related stress on the satisfaction of human needs: The united states and the middle east. *International Journal of Management Theory and Practices* 3(1): 35–53.

Tang, Duyu, Furu Wei, Bing Qin, Li Dong, Ting Liu, and Ming Zhou. 2014a. A joint segmentation and classification framework for sentiment analysis. In *EMNLP*, 477–487.

Tang, Duyu, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014b. Building large-scale Twitter-specific sentiment Lexicon: A representation learning approach. In *COLING*, 172–182.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014c. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1555–1565.

Tang, Duyu. 2015. Sentiment-specific representation learning for document-level sentiment analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 447–452. ACM.

Titov, Ivan, and Ryan T McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, vol. 8, 308–316. Citeseer.

Wang, Sida, and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 90–94. Association for Computational Linguistics.

Wang, Hongning, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 618–626. ACM.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2–3): 165–210.

Xueke, Xu, Cheng Xueqi, Tan Songbo, Liu Yue, and Shen Huawei. 2013. Aspect-level opinion mining of online customer reviews. *Communications, China* 10(3): 25–41.

Yang, Bishan, and Claire Cardie. 2012. Extracting opinion expressions with semi-Markov conditional random fields. In *EMNLP*.

Yang, Bishan, and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *ACL (1)*, 1640–1649.

Yang, Bishan, and Claire Cardie. 2014a. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of ACL*.

Yang, Bishan, and Claire Cardie. 2014b. Joint modeling of opinion expression extraction and attribute classification. *Transactions of the Association for Computational Linguistics* 2: 505–516.

Zhao, Wayne Xin, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 56–65. Association for Computational Linguistics.

Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 43–50. ACM.

Chapter 4
Challenges in Sentiment Analysis

Saif M. Mohammad

Abstract A vast majority of the work in Sentiment Analysis has been on developing more accurate sentiment classifiers, usually involving supervised machine learning algorithms and a battery of features. Surveys by Pang and Lee (Found Trends Inf Retr 2(1–2):1–135, 2008), Liu and Zhang (A survey of opinion mining and sentiment analysis. In: Aggarwal CC, Zhai C (eds) In: Mining text data. Springer, New York, pp 415–463, 2012), and Mohammad (Mohammad Sentiment analysis: detecting valence, emotions, and other effectual states from text. In: Meiselman H (ed) Emotion measurement. Elsevier, Amsterdam, 2016b) give summaries of the many automatic classifiers, features, and datasets used to detect sentiment. In this chapter, we flesh out some of the challenges that still remain, questions that have not been explored sufficiently, and new issues emerging from taking on new sentiment analysis problems. We also discuss proposals to deal with these challenges. The goal of this chapter is to equip researchers and practitioners with pointers to the latest developments in sentiment analysis and encourage more work in the diverse landscape of problems, especially those areas that are relatively less explored.

Keywords Sentiment analysis tasks • Sentiment of the writer, reader, and other entities • Sentiment towards aspects of an entity • Stance detection • Sentiment lexicons • Sentiment annotation • Multilingual sentiment analysis

4.1 Introduction

There has been a large volume of work in sentiment analysis over the past decade and it continues to rapidly develop in new directions. However, much of it is on developing more accurate sentiment classifiers. In this chapter, we flesh out some of the challenges that still remain. We start by discussing different sentiment analysis

S.M. Mohammad (✉)
National Research Council Canada, 1200 Montreal Rd., Ottawa, ON, Canada
e-mail: Saif.Mohammad@nrc-cnrc.gc.ca

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_4

problems and how one of the challenges is to explore new sentiment analysis problems that go beyond simply determining whether a piece of text is positive, negative, or neutral (Sect. 4.2). Some of the more ambitious problems that need more work include detecting sentiment at various levels of text granularities (terms, sentences, paragraphs, etc); detecting sentiment of the reader or sentiment of entities mentioned in the text; detecting sentiment towards aspects of products; detecting stance towards pre-specified targets that may not be explicitly mentioned in the text and that may not be the targets of opinion in the text; and detecting semantic roles of sentiment. Since many sentiment analysis systems rely on sentiment lexicons, we discuss capabilities and limitations of existing manually and automatically created sentiment lexicons in Sect. 4.3. In Sect. 4.4, we discuss the difficult problem of sentiment composition—how to predict the sentiment of a combination of terms. More specifically, we discuss the determination of sentiment of phrases (that may include negators, degree adverbs, and intensifiers) and sentiment of sentences and tweets. In Sect. 4.5, we discuss challenges in annotation of data for sentiment. We provide categories of sentences that are particularly challenging for sentiment annotation. Section 4.6 presents challenges in multilingual sentiment analysis. This is followed by a discussion on the challenges of applying sentiment analysis to downstream applications, and finally, some concluding remarks (Sect. 4.7).

4.2 The Array of Sentiment Analysis Tasks

Sentiment analysis is a generic name for a large number of opinion and affect related tasks, each of which present their own unique challenges. The sub-sections below provide an overview.

4.2.1 Sentiment at Different Text Granularities

Sentiment can be determined at various levels: from sentiment associations of words and phrases; to sentiment of sentences, SMS messages, chat messages, and tweets; to sentiment in product reviews, blog posts, and whole documents. A word–sentiment (or valence) association lexicon may have entries such as:

delighted – positive
killed – negative
shout – negative
desk – neutral

These lexicons can be created either by manual annotation or through automatic means. Manually created lexicons tend to be in the order of a few thousand entries, but automatically generated lexicons can capture sentiment associations for hundreds of thousands unigrams (single word strings) and even for larger expressions such as bigrams (two-word sequences) and trigrams (three-word sequences). Entries in an automatically generated lexicon often also include a real-valued score indicating the strength of association between the word and the valence category. These numbers are prior estimates of the sentiment of terms in an average usage of the term. While sentiment lexicons are often useful in sentence-level sentiment analysis,[1] the same terms may convey different sentiments in different contexts. The SemEval 2013 and 2014 *Sentiment Analysis in Twitter* shared tasks had a separate sub-task aimed at identifying sentiment of terms in context. Automatic systems have largely performed well in this task, obtaining F-scores close to 0.9. We discuss manually and automatically created sentiment lexicons in more detail in Sect. 4.3.

Sentence-level valence classification systems assign labels such as positive, negative, or neutral to whole sentences. It should be noted that the valence of a sentence is not simply the sum of the polarities of its constituent words. Automatic systems learn a model from labeled training data (instances that are already marked as positive, negative, or neutral) using a large number of features such as word and character ngrams, valence association lexicons, negation lists, word clusters, and even embeddings-based features. In recent years, there have been a number of shared task competitions on valence classification such as the 2013, 2014, and 2015 SemEval shared tasks titled *Sentiment Analysis in Twitter*, the 2014 and 2015 SemEval shared tasks on *Aspect Based Sentiment Analysis*, the 2015 SemEval shared task *Sentiment Analysis of Figurative Language in Twitter*, and the 2015 Kaggle competition *Sentiment Analysis on Movie Reviews*.[2] The NRC-Canada system (Mohammad et al. 2013a; Kiritchenko et al. 2014b), a supervised machine learning system, came first in the 2013 and 2014 competitions. Other sentiment analysis systems developed specifically for tweets include those by Pak and Paroubek (2010), Agarwal et al. (2011), Thelwall et al. (2011), Brody and Diakopoulos (2011), Aisopos et al. (2012), and Bakliwal et al. (2012). However, even the best systems currently obtain an F-score of only about 0.7.

Sentiment analysis involving many sentences is often broken down into the sentiment analysis of the component sentences. However, there is interesting work in sentiment analysis of documents to generate text summaries (Ku et al. 2006; Liu et al. 2007; Somprasertsri and Lalitrojwong 2010; Stoyanov and Cardie 2006; Lloret et al. 2009), as well as detecting the patterns of sentiment and detecting sentiment networks in novels and fairy tales (Nalisnick and Baird 2013a,b; Mohammad and Yang 2011).

[1]The top systems in the SemEval-2013 and 2014 *Sentiment Analysis in Twitter* tasks used large sentiment lexicons (Wilson et al. 2013; Rosenthal et al. 2014a).

[2]http://alt.qcri.org/semeval2015/task10/
http://alt.qcri.org/semeval2015/task12/
http://alt.qcri.org/semeval2015/task11/
http://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

4.2.2 Detecting Sentiment of the Writer, Reader, and Other Entities

On the surface, sentiment may seem unambiguous, but looking closer, it is easy to see how sentiment can be associated with any of the following: 1. the speaker or writer, 2. the listener or reader, or 3. one or more entities mentioned in the utterance. A large majority of research in sentiment analysis has focused on detecting the sentiment of the speaker, and this is often done by analyzing only the utterance. However, there are several instances where it is unclear whether the sentiment in the utterance is the same as the sentiment of the speaker. For example, consider:

James: *The pop star suffered a fatal overdose of heroine.*

The sentence describes a negative event (death of a person), but it is unclear whether to conclude that James (the speaker) is personally saddened by the event. It is possible that James is a news reader and merely communicating information about the event. Developers of sentiment systems have to decide before hand whether they wish to assign a negative or neutral sentiment to the speaker in such cases. More generally, they have to decide whether the speaker's sentiment will be chosen to be neutral in absence of clear signifiers of the speaker's own sentiment, or whether the speaker's sentiment will be chosen to be the same as the sentiment of events and topics mentioned in the utterance.

On the other hand, people can react differently to the same utterance, for example, people on opposite sides of a debate or rival sports fans. Thus modeling listener sentiment requires modeling listener profiles. This is an area of research not explored much by the community. Similarly, there is no work on modeling sentiment of entities mentioned in the text, for example, given:

Drew: *Jackson could not stop talking about the new Game of Thrones episode.*

It will be useful to develop automatic systems that can deduce that Jackson (not Drew) liked the new episode of *Game of Thrones* (a TV show).

4.2.3 Sentiment Towards Aspects of an Entity

A review of a product or service can express sentiment towards various aspects. For example, a restaurant review can speak positively about the service, but express a negative attitude towards the food. There is now a growing amount of work in detecting aspects of products in text and also in determining sentiment towards these aspects. In 2014, a shared task was organized for detecting aspect sentiment in restaurant and laptop reviews (Pontiki et al. 2014a). The best performing systems had a strong sentence-level sentiment analysis system to which they added localization features so that more weight was given to sentiment features close to the mention of the aspect. This task was repeated in 2015. It will be useful to develop aspect-based sentiment systems for other domains such as blogs and news articles as well. (See proceeding of SemEval-2014 and 2015 for details about participating aspect sentiment systems.)

4.2.4 Stance Detection

Stance detection is the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target. For example, given the following target and text pair:

Target of interest: *women have the right to abortion*
Text: *A foetus has rights too!*

Humans can deduce from the text that the speaker is against the proposition. However, this is a challenging task for computers. To successfully detect stance, automatic systems often have to identify relevant bits of information that may not be present in the focus text. The systems also have to first identify the target of opinion in the text and then determine its implication on the target of interest. Note that the target of opinion need not be the same as the target of interest. For example, that if one is actively supporting foetus rights (target of opinion), then he or she is likely against the right to abortion (target of interest). Automatic systems can obtain such information from large amounts of domain text.

Automatically detecting stance has widespread applications in information retrieval, text summarization, and textual entailment. In fact, one can argue that stance detection can bring complementary information to sentiment analysis, because we often care about the authors evaluative outlook towards *specific targets* and propositions rather than simply about whether the speaker was angry or happy.

Mohammad et al. (2016b) created the first dataset of tweets labeled for both stance and sentiment. More than 4000 tweets are annotated for whether one can deduce favorable or unfavorable stance towards one of five targets 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion'. Each of these tweets is also annotated for whether the target of opinion expressed in the tweet is the same as the given target of interest. Finally, each tweet is annotated for whether it conveys positive, negative, or neutral sentiment. Partitions of this stance-annotated data were used as training and test sets in the SemEval-2016 shared task competition, Task #6: Detecting Stance from Tweets Mohammad et al. (2016a). Participants were provided with 2,914 training instances labeled for stance for the five targets. The test data included 1,249 instances. All of the stance data is made freely available through the shared task website. The task received submissions from 19 teams. The best performing system obtained an overall average F-score of 67.8 in a three-way classification task: favour, against, or neither. They employed two recurrent neural network (RNN) classifiers: the first was trained to predict task-relevant hashtags on a large unlabeled Twitter corpus. This network was used to initialize a second RNN classifier, which was trained with the provided training data (Zarrella and Marsh 2016). Mohammad et al. (2016b) developed a SVM system that only uses features drawn from word and character ngrams and word embeddings to obtain an even better F-score of 70.3 on the shared task's test set. Yet, performance of systems is substantially lower on tweets where the target of opinion is an entity other than the target of interest.

Most of the earlier work focused on two-sided debates, for example on congressional debates (Thomas et al. 2006) or debates in online forums (Somasundaran and Wiebe 2009; Murakami and Raymond 2010; Anand et al. 2011; Walker et al. 2012; Hasan and Ng 2013; Sridhar et al. 2014). New research in domains such as social media texts, and approaches that combine traditional sentiment analysis with relation extraction can make a significant impact in improving the state-of-the-art in automatic stance detection.

4.2.5 Detecting Semantic Roles of Feeling

Past work in sentiment analysis has focused extensively on detecting polarity, and to a smaller extent on detecting the target of the sentiment (the stimulus) (Popescu and Etzioni 2005; Su et al. 2006; Xu et al. 2013; Qadir 2009; Zhang et al. 2010; Zhang and Liu 2011; Kessler and Nicolov 2009). However, there exist other aspects relevant to sentiment. Tables 4.1 and 4.2 show FrameNet (Baker et al. 1998) frames for 'feelings' and 'emotions', respectively. Observe that in addition to Evaluation, State, and Stimulus, several other roles such as Reason, Degree, Topic, and Circumstance are also of significance and beneficial to down-stream applications such as information retrieval, summarization, and textual entailment. Detecting these various roles is essentially a semantic role-labeling problem (Gildea and Jurafsky 2002; Màrquez et al. 2008; Palmer et al. 2010), and it is possible that they can be modeled jointly to improve detection accuracy. Li and Xu (2014) proposed a rule-based system to extract the event that was the cause of an emotional Weibo (Chinese microblogging service) message. Mohammad et al. (2015a) created a corpus of tweets from the run up to the 2012 US presidential elections, with annotations for sentiment, emotion, stimulus, and experiencer. The data also includes annotations for whether the tweet is sarcastic,

Table 4.1 The FrameNet frame for feeling

Role	Description
Core	
Emotion	The feeling that the experiencer experiences
State	The state the experiencer is in
Evaluation	A negative or positive assessment of the experiencer regarding his/her state
Experiencer	One who experiences the emotion and is in the state
Non-Core	
Explanation	The thing that leads to the experiencer feeling the emotion or state

Table 4.2 The FrameNet frame for emotions

Role	Description
Core	
Experiencer	The person that experiences or feels the emotion
State	The abstract noun that describes the experience
Stimulus	The person or event that evokes the emotional response in the experiencer.
Topic	The general area in which the emotion occurs
Non-Core	
Circumstances	The condition in which stimulus evokes response
Degree	The extent to which the experiencer's emotion deviates from the norm for the emotion
Empathy_target	The Empathy_target is the individual or individuals with which the experiencer identifies emotionally
Manner	Any description of the way in which the experiencer experiences the stimulus which is not covered by more specific frame elements
Reason	The explanation for why the stimulus evokes a certain emotional response

ironic, or hyperbolic. Diman Ghazi and Szpakowicz (2015) compiled FrameNet sentences that were tagged with the stimulus of certain emotions.

4.2.6 *Detecting Affect and Emotions*

Sentiment analysis is most commonly used to refer to the goal of determining the valence or polarity of a piece of text. However, it can refer more generally to determining one's attitude towards a particular target or topic. Here, attitude can even mean emotional or affectual attitude such as frustration, joy, anger, sadness, excitement, and so on. Russell (1980) developed a circumplex model of affect and showed that it can be characterized by two primary dimensions: valence (positive and negative dimension) and arousal (degree of reactivity to stimulus). Thus, it is not surprising that large amounts of work in sentiment analysis is focused on determining valence. However, there is barely any work on automatically detecting arousal and a relatively small amount of work on detecting emotions such as anger, frustration, sadness, and optimism (Strapparava and Mihalcea 2007; Aman and Szpakowicz 2007; Tokuhisa et al. 2008; Neviarouskaya et al. 2009; Bellegarda 2010; Mohammad 2012; Boucouvalas 2002; Zhe and Boucouvalas 2002; Holzman and Pottenger 2003; Ma et al. 2005; Mohammad 2012; John et al. 2006; Mihalcea and Liu 2006; Genereux and Evans 2006). Detecting these more subtle aspects of sentiment has wide-ranging applications, for example in developing customer relation models, public health, military intelligence, and the video games industry, where it is necessary to make distinctions between anger and sadness (both of which are negative), calm and excited (both of which are positive), and so on.

4.3 Sentiment of Words

Term–sentiment associations have been captured by manually created sentiment lexicons as well as automatically generated ones.

4.3.1 Manually Generated Term-Sentiment Association Lexicons

The General Inquirer (GI) has sentiment labels for about 3,600 terms (Stone et al. 1966). Hu and Liu (2004) manually labeled about 6,800 words and used them for detecting sentiment of customer reviews. The MPQA Subjectivity Lexicon, which draws from the General Inquirer and other sources, has sentiment labels for about 8,000 words (Wilson et al. 2005). The NRC Emotion Lexicon has sentiment and emotion labels for about 14,000 words (Mohammad and Turney 2010; Mohammad and Yang 2011). These labels were compiled through Mechanical Turk annotations.[3]

For people, assigning a score indicating the degree of sentiment is not natural. Different people may assign different scores to the same target item, and it is hard for even the same annotator to remain consistent when annotating a large number of items. In contrast, it is easier for annotators to determine whether one word is more positive (or more negative) than the other. However, the latter requires a much larger number of annotations than the former (in the order of N^2, where N is the number of items to be annotated).

An annotation scheme that retains the comparative aspect of annotation while still requiring only a small number of annotations comes from survey analysis techniques and is called MaxDiff (Louviere 1991). The annotator is presented with four terms and asked which word is the most positive and which is the least positive. By answering just these two questions five out of the six inequalities are known. If the respondent says that A is most positive and D is least positive, then:

$$A > B, A > C, A > D, B > D, C > D$$

Each of these MaxDiff questions can be presented to multiple annotators. The responses to the MaxDiff questions can then be easily translated into a ranking of all the terms and also a real-valued score for all the terms (Orme 2009). If two words have very different degrees of association (for example, $A \gg D$), then A will be chosen as most positive much more often than D and D will be chosen as least positive much more often than A. This will eventually lead to a ranked list such that A and D are significantly farther apart, and their real-valued association scores are also significantly different. On the other hand, if two words have similar degrees

[3] https://www.mturk.com/mturk/welcome

of association with positive sentiment (for example, *A* and *B*), then it is possible that for MaxDiff questions having both *A* and *B*, some annotators will choose *A* as most positive, and some will choose *B* as most positive. Further, both *A* and *B* will be chosen as most positive (or most negative) a similar number of times. This will result in a list such that *A* and *B* are ranked close to each other and their real-valued association scores will also be close in value.

MaxDiff was used for obtaining annotations of relation similarity of pairs of items in a SemEval-2012 shared task (Jurgens et al. 2012). Kiritchenko and Mohammad (2016a) applied Best–Worst Scaling to obtain real-valued sentiment association scores for words and phrases in three different domains: general English, English Twitter, and Arabic Twitter. They showed that on all three domains the ranking of words by sentiment remains remarkably consistent even when the annotation process is repeated with a different set of annotators. They also determine the minimum difference in sentiment association that is perceptible to native speakers of a language.

4.3.2 Automatically Generated Term-Sentiment Association Lexicons

Semi-supervised and automatic methods have also been proposed to detect the polarity of words. Hatzivassiloglou and McKeown (1997) proposed an algorithm to determine the polarity of adjectives. SentiWordNet was created using supervised classifiers as well as manual annotation (Esuli and Sebastiani 2006). Turney and Littman (2003) proposed a minimally supervised algorithm to calculate the polarity of a word by determining if its tendency to co-occur with a small set of positive seed words is greater than its tendency to co-occur with a small set of negative seed words. Mohammad et al. (2013b) employed the Turney method to generate a lexicon (Hashtag Sentiment Lexicon) from tweets with certain sentiment-bearing seed-word hashtags such as (*#excellent, #good, #terrible,* and so on) and another lexicon (Hashtag Sentiment Lexicon) from tweets with emoticons.[4] Since the lexicons themselves are generated from tweets, they even have entries for the creatively spelled words (e.g. *happpeee*), slang (e.g. *bling*), abbreviations (e.g. *lol*), and even hashtags and conjoined words (e.g. *#loveumom*). Cambria et al. (2016) created SenticNet that has sentiment entries for 30,000 words and multi-word expressions using information propagation to connect various parts of common-sense knowledge representations. Kiritchenko et al. (2014b) proposed a method to create separate lexicons for words found in negated context and those found in affirmative context; the idea being that the same word contributes to sentiment differently depending on whether it is negated or not. These lexicons contain sentiment associations for hundreds

[4] http://www.purl.com/net/lexicons

of thousands of unigrams and bigrams. However, they do not explicitly handle combinations of terms with modals, degree adverbs, and intensifiers.

4.4 Sentiment of Phrases, Sentences, and Tweets: Sentiment Composition

Semantic composition, which aims at determining a representation of the meaning of two words through manipulations of their individual representations, has gained substantial attention in recent years with work from Mitchell and Loapata (2010), Baroni and Zamparelli (2010), Rudolph and Giesbrecht (2010), Yessenalina and Cardie (2011), Grefenstette et al. (2013), Grefenstette and Sadrzadeh (2011), and Turney (2014). Socher et al. (2012) and Mikolov et al. (2013) introduced deep learning models and distributed word representations in vector space (word embeddings) to obtain substantial improvements over the state-of-the-art in semantic composition. Mikolov's word2vec tool for generating word embeddings is available publicly.[5]

Sentiment of a phrase or a sentence is often not simply the sum of the sentiments of its constituents. Sentiment composition is the determining of sentiment of a multi-word linguistic unit, such as a phrase or a sentence, based on its constituents. Lexicons that include sentiment associations for phrases as well as their constituent words are referred to as *sentiment composition lexicons (SCLs)*. Kiritchenko and Mohammad created sentiment composition lexicons for English and Arabic that included: (1) negated expressions Kiritchenko and Mohammad (2016a,b), (2) phrases with adverbs, modals, and intensifies Kiritchenko and Mohammad (2016a,b), and (3) opposing polarity phrases (where at least one word in the phrase is positive and at least one word is negative, for example, *happy accident* and *dark chocolate*) (Kiritchenko and Mohammad 2016c). Socher et al. (2013) took a dataset of movie review sentences that were annotated for sentiment and further annotated ever word and phrasal constituent within those sentences for sentiment. Such datasets where sentences, phrases, and their constituent words are annotated for sentiment are helping foster further research on how sentiment is composed. We discuss specific types of sentiment composition, and challenges for automatic methods that address them, in the sub-sections below.

4.4.1 Negated Expressions

Morante and Sporleder (2012) define negation to be "a grammatical category that allows the changing of the truth value of a proposition". Negation is often expressed

[5] https://code.google.com/p/word2vec

through the use of negative signals or negator words such as *not* and *never*, and it can significantly affect the sentiment of its scope. Understanding the impact of negation on sentiment improves automatic analysis of sentiment. Earlier works on negation handling employ simple heuristics such as flipping the polarity of the words in a negator's scope (Kennedy and Inkpen 2005; Choi and Cardie 2008) or changing the degree of sentiment of the modified word by a fixed constant (Taboada et al. 2011). Zhu et al. (2014) show that these simple heuristics fail to capture the true impact of negators on the words in their scope. They show that negators tend to often make positive words negative (albeit with lower intensity) and make negative words less negative (not positive). Zhu et al. also propose certain embeddings-based recursive neural network models to capture the impact of negators more precisely. As mentioned earlier, Kiritchenko et al. (2014b) capture the impact of negation by creating separate sentiment lexicons for words seen in affirmative context and those seen in negated contexts. They use a hand-chosen list of negators and determine scope to be starting from the negator and ending at the first punctuation (or end of sentence).

Several aspects about negation are still not understood though: for example, can negators be ranked in terms of their average impact on the sentiment of their scopes (which negators impact sentiment more and which impact sentiment less); in what contexts does the same negator impact the sentiment of its scope more and in what contexts is the impact less; how do people in different communities and cultures use negations differently; and how negations of sentiment expressions should be dealt with by paraphrase and textual entailment systems.

4.4.2 Phrases with Degree Adverbs, Intensifiers, and Modals

Degree adverbs such as *barely, moderately*, and *slightly* quantify the extent or amount of the predicate. Intensifiers such as *too* and *very* are modifiers that do not change the propositional content (or truth value) of the predicate they modify, but they add to the emotionality. However, even linguists are hard pressed to give out comprehensive lists of degree adverbs and intensifiers. Additionally, the boundaries between degree adverbs and intensifiers can sometimes be blurred, and so it is not surprising that the terms are occasionally used interchangeably. Impacting propositional content or not, both degree adverbs and intensifiers impact the sentiment of the predicate, and there is some work in exploring this interaction (Zhang et al. 2008; Wang and Wang 2012; Xu et al. 2008; Lu and Tsou 2010; Taboada et al. 2008). Most of this work focuses on identifying sentiment words by bootstrapping over patterns involving degree adverbs and intensifiers. Thus several areas remain unexplored, such as identifying patterns and regularities in how different kinds of degree adverbs and intensifiers impact sentiment, ranking degree adverbs and intensifiers in terms of how they impact sentiment, and determining when (in what contexts) the same modifier will impact sentiment differently than its usual behavior. (See Kiritchenko and Mohammad (2016b) for some recent work exploring these questions in manually annotated sentiment composition lexicons.)

Modals (a kind of auxiliary verb) are used to convey the degree of confidence, permission, or obligation to the predicate. Thus, if the predicate is sentiment bearing, then the sentiment of the combination of the modal and the predicate can be different from the sentiment of the predicate alone. For example, *cannot work* seems less positive than *work* or *will work* (*cannot* and *will* are modals). There is little work on automatically determining the impact of modals on sentiment.

4.4.3 Sentiment of Sentences, Tweets, and SMS messages

Bag-of-word models such as the NRC-Canada system (Mohammad et al. 2013a; Kiritchenko et al. 2014a,b) and Unitn Severyn and Moschitti (2015) have been very successful in recent shared task competitions on determining sentiment of whole tweets, SMS messages, and sentences. However, approaches that apply systematic sentiment composition of smaller units to determine sentiment of sentences are growing in popularity. Socher et al. (2013) proposed a word-embeddings based model that learns the *sentiment* of term compositions. They obtain state-of-the-art results in determining both the overall sentiment and sentiment of constituent phrases in movie review sentences. This has inspired tremendous interest in more embeddings-based work for sentiment composition (Dong et al. 2014; Kalchbrenner et al. 2014). These recursive models do not require any hand-crafted features or semantic knowledge, such as a list of negation words or sentiment lexicons. However, they are computationally intensive and need substantial additional annotations (word and phrase-level sentiment labeling). Nonetheless, use of word-embeddings in sentiment composition is still in its infancy, and we will likely see much more work using these techniques in the future.

4.4.4 Sentiment in Figurative Expressions

Figurative expressions in text, by definition, are not compositional. That is, their meaning cannot fully be derived from the meaning of their components in isolation. There is growing interest in detecting figurative language, especially irony and sarcasm (Carvalho et al. 2009; Reyes et al. 2013; Veale and Hao 2010; Filatova 2012; González-Ibánez et al. 2011). In 2015, a SemEval shared task was organized on detecting sentiment in tweets rich in metaphor and irony (Task 11).[6] Participants were asked to determine the degree of sentiment for each tweet where the score is a real number in the range from -5 (most negative) to $+5$ (most positive). One of

[6]The proceedings will be released later in 2015.

the characteristics of the data is that a large majority is negative; thereby suggesting that ironic tweets are largely negative. The SemEval 2014 shared task Sentiment Analysis in Twitter Rosenthal et al. (2014a) had a separate test set involving sarcastic tweets. Participants were asked *not* to train their system on sarcastic tweets, but rather apply their regular sentiment system on this new test set; the goal was to determine performance of regular sentiment systems on sarcastic tweets. It was observed that the performances dropped by about 25% to 70%, thereby showing that systems must be adjusted if they are to be applied to sarcastic tweets. We found little to no work exploring automatic sentiment detection in hyperbole, understatement, rhetorical questions, and other creative uses of language.

4.5 Challenges in Annotating for Sentiment

Clear and simple instructions are crucial for obtaining high-quality annotations. This is true even for seemingly simple annotation tasks, such as sentiment annotation, where one is to label instances as positive, negative, or neutral. For word annotations, researchers have often framed the task as 'is this word positive, negative, or neutral?' Hu and Liu (2004), 'does this word have associations with positive, negative, or neutral sentiment?' Mohammad and Turney (2013), or 'which word is more positive?'/'which word has a greater association with positive sentiment' (Kiritchenko et al. 2016; Kiritchenko and Mohammad 2016c). Similar instructions are also widely used for sentence-level sentiment annotations—'is this sentence positive, negative, or neutral?' (Rosenthal et al. 2015, 2014b; Mohammad et al. 2016a, 2015b). We will refer to such annotation schemes as *the simple sentiment questionnaires*. On the one hand, this characterization of the task is simple, terse, and reliant on the intuitions of native speakers of a language (rather than biasing the annotators by providing definitions of what it means to be positive, negative, and neutral). On the other hand, the lack of specification leaves the annotator in doubt over how to label certain kinds of instances—for example, sentences where one side wins against another, sarcastic sentences, or retweets.

A different approach to sentiment annotation is to ask respondents to identify the target of opinion, and the sentiment towards this target of opinion (Pontiki et al. 2014b; Mohammad et al. 2015b; Deng and Wiebe 2014). We will refer to such annotation schemes as *the semantic-role based sentiment questionnaires*. This approach of sentiment annotation is more specific, and more involved, than the simple sentiment questionnaire approach; however, it too is insufficient for handling several scenarios. Most notably, the emotional state of the speaker is not under the purview of this scheme. Many applications require that statements expressing positive or negative emotional state of the speaker should be marked as 'positive' or 'negative', respectively. Similarly, many applications require statements that describe positive or negative events or situations to be marked as 'positive' or 'negative', respectively. Instructions for annotating opinion towards targets do not specify how such instances are to be annotated, and worse still, possibly imply that such instances are to be labeled as neutral.

Some sentence types that are especially challenging for sentiment annotation (using either the simple sentiment questionnaire or the semantic-role based sentiment questionnaire) are listed below:

- *Speaker's emotional state:* The speaker's emotional state may or may not have the same polarity as the opinion expressed by the speaker. For example, a politician's tweet can imply both a negative opinion about a rival's past indiscretion, and a joyous mental state as the news will impact the rival adversely.
- *Success or failure of one side w.r.t. another*: Often sentences describe the success or failure of one side w.r.t. another side—for example, *'Yay! France beat Germany 3–1', 'Supreme court judges in favor of gay marriage'*, and *'the coalition captured the rebels'*. If one supports France, gay marriage, and the coalition, then these events are positive, but if one supports Germany, marriage as a union only between man and woman, and the rebels, then these events can be seen as negative.

  Also note that the framing of an event as the success of one party (or as the failure of another party) does not automatically imply that the speaker is expressing positive (or negative) opinion towards the mentioned party. For example, when Finland beat Russia in ice hockey in the 2014 Sochi Winter Olympics, the event was tweeted around the world predominantly as "Russia lost to Finland" as opposed to "Finland beat Russia". This is not because the speakers were expressing negative opinion towards the Russian team, but rather simply because Russia, being the host nation, was the focus of attention and traditionally Russian hockey teams have been strong.
- *Neutral reporting of valenced information*: If the speaker does not give any indication of her own emotional state but describes valenced events or situations, then it is unclear whether to consider these statements as neutral unemotional reporting of developments or whether to assume that the speaker is in a negative emotional state (sad, angry, etc.). Example:

  The war has created millions of refugees.

- *Sarcasm and ridicule*: Sarcasm and ridicule are tricky from the perspective of assigning a single label of sentiment because they can often indicate positive emotional state of the speaker (pleasure from mocking someone or something) even though they have a negative attitude towards someone or something.
- *Different sentiment towards different targets of opinion*: The speaker may express opinion about multiple targets, and sentiment towards the different targets might be different. The targets may be different people or objects (for example, an iPhone vs. an android phone), or they may be different aspects of the same entity (for example, quality of service vs. quality of food at a restaurant).
- *Precisely determining the target of opinion*: Sometimes it is difficult to precisely identify the target of opinion. For example, consider:

  Glad to see Hillary's lies being exposed.

It is unclear whether the target of opinion is ‘Hillary’, ‘Hillary’s lies’, or ‘Hillary’s lies being exposed’. One reasonable interpretation is that positive sentiment is expressed about ‘Hillary’s lies being exposed’. However, one can also infer that the speaker has a negative attitude towards ‘Hillary’s lies’ and probably ‘Hillary’ in general. It is unclear whether annotators should be asked to provide all three opinion–target pairs or only one (in which case, which one?).

- *Supplications and requests*: Many tweets convey positive supplications to God or positive requests to people in the context of a (usually) negative situation. Examples include:

  May god help those displaced by war.
  Let us all come together and say no to fear mongering and divisive politics.

- *Rhetorical questions*: Rhetorical questions can be treated simply as queries (and thus neutral) or as utterances that give away the emotional state of the speaker. For example, consider:

  Why do we have to quibble every time?

  On the one hand, this tweet can be treated as a neutral question, but on the other hand, it can be seen as negative because the utterance betrays a sense of frustration on the part of the speaker.
- *Quoting somebody else or re-tweeting*: Quotes and retweets are difficult to annotate for sentiment because it is often unclear and not explicitly evident whether the one who quotes (or retweets) holds the same opinions as that expressed by the quotee.

The challenges listed above can be addressed to varying degrees by providing instructions to the annotators on how such instances are to be labeled. However, detailed and complicated instructions can be counter-productive as the annotators may not understand or may not have the inclination to understand the subtleties involved. See Mohammad (2016a) for annotation schemes that address some of these challenges.

4.6 Challenges in Multilingual Sentiment Analysis

Work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages. Mihalcea et al. (2007) use English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon is then used to classify Romanian text. Wan (2008) translated Chinese customer reviews to English using a machine translation system. The translated reviews are then annotated using rule-based system that uses English lexicons. A higher accuracy is achieved when using ensemble methods and combining knowledge from Chinese and English resources. Balahur and Turchi (2014) conducted a study to assess the performance of statistical sentiment analysis techniques on machine-translated texts. Opinion-bearing phrases from the New York Times Text (2002–2005) corpus were automatically translated using publicly available machine-translation engines (Google, Bing, and Moses). Then, the accuracy of a sentiment analysis system trained on original English texts was compared to the accuracy of the system trained on automatic translations to German, Spanish, and French. The authors conclude that the quality of machine translation is acceptable for sentiment analysis to be performed on automatically translated texts. Salameh et al. (2015) conducted experiments to determine loss in sentiment predictability when they translate Arabic social media posts into English, manually and automatically. As benchmarks, they use manually and automatically determined sentiment labels of the Arabic texts. They show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. They also claim that even though translation significantly reduces human ability to recover sentiment, automatic sentiment systems are affected relatively less by this.

Some of the areas less explored in the realm of multilingual sentiment analysis include: how to translate text so as to preserve the degree of sentiment in the source text; how sentiment modifiers such as negators and modals differ in function across languages; understanding how automatic translations differ from manual translations in terms of sentiment; and how to translate figurative language without losing its affectual gist.

4.7 Challenges in Applying Sentiment Analysis

Applications of sentiment analysis benefit from the fact that even though systems are not extremely accurate at determining sentiment of individual sentences, they can accurately capture significant changes in the proportion of instances that are positive (or negative). It is also worth noting that such sentiment tracking systems are more effective when incorporating carefully chosen baselines. For example, knowing the percentage of tweets that are negative towards Russian President, Vladimir Putin, is less useful than, for instance, knowing: the percentage of tweets that are negative towards Putin before vs. after the invasion of Crimea; or, the percentage of tweets that are negative towards Putin in Russia vs. the rest of the world; or, the percentage of tweets negative towards Putin vs. Barack Obama (US president).

Sentiment analysis is commonly applied in several areas including tracking sentiment towards products, movies, politicians, and companies (O'Connor et al. 2010; Pang and Lee 2008), improving customer relation models (Bougie et al. 2003), detecting happiness and well-being (Schwartz et al. 2013), tracking the stock market (Bollen et al. 2011), and improving automatic dialogue systems (Velásquez 1997; Ravaja et al. 2006). The sheer volume of work in this area precludes detailed summarization here. Nonetheless, it should be noted that often the desired application can help direct certain design choices in the sentiment analysis system. For example, the threshold between neutral and positive sentiment and the threshold between neutral and negative sentiment can be determined empirically by what is most suitable for the target application. Similarly, as suggested earlier, some applications may require only the identification of strongly positive and strongly negative instances.

Abundant availability of product reviews and their ratings has powered a lot of the initial research in sentiment analysis, however, as we look forward, one can be optimistic that the future holds more diverse and more compelling applications of sentiment analysis. Some recent examples include predicting heart attack rates through sentiment word usage in tweets (Eichstaedt et al. 2015), corpus-based poetry generation (Colton et al. 2012), generating music that captures the sentiment in novels (Davis and Mohammad 2014), confirming theories in literary analysis (Hassan et al. 2012), and automatically detecting Cyber-bullying (Nahar et al. 2012).

References

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of Language in Social Media*, 30–38. Portland.

Aisopos, F., G. Papadakis, K. Tserpes, and T. Varvarigou. 2012. Textual and contextual patterns for sentiment analysis over microblogs. In *Proceedings of the 21st WWW Companion*, New York, 453–454.

Aman, S., and S. Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, *Lecture notes in computer science*, vol. 4629, 196–205.

Anand, Pranav, et al. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the ACL workshop on computational approaches to subjectivity and sentiment analysis*, Portland.

Baker, C.F., C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley framenet project. In *Proceedings of ACL*, Stroudsburg, 86–90.

Bakliwal, A., P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma. 2012. Mining sentiments from tweets. In *Proceedings of WASSA'12*, 11–18, Jeju.

Balahur, A., and M. Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language* 28(1): 56–75.

Baroni, M., and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193.

Bellegarda, J. 2010. Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles.

Bollen, J., H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1): 1–8.

Boucouvalas, A.C. 2002. Real time text-to-emotion engine for expressive internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication* 5: 305–318.

Bougie, J.R.G., R. Pieters, and M. Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from Tilburg university, Tilburg University.

Brody, S., and N. Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, 562–570.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: *COLING*, 2666–2677.

Carvalho, P., L. Sarmento, M.J. Silva, and E. De Oliveira, 2009. Clues for detecting irony in user-generated contents: oh…!! it's so easy;-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, 53–56. ACM.

Choi, Y., and C. Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, Honolulu, 793–801.

Colton, S., J. Goodwin, and T. Veale. 2012. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, 95–102.

Davis, H., and S. Mohammad. 2014. Generating music from literature. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, Gothenburg, 1–10.

Deng, L., and J. Wiebe. 2014. Sentiment propagation via implicature constraints. In *EACL*, 377–385.

Diman Ghazi, D.I., and S. Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In: *Proceedings of the 2015 Conference on Intelligent Text Processing and Computational Linguistics*.

Dong, L., F. Wei, M. Zhou, and K. Xu. 2014. Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In *Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*.

Eichstaedt, J.C., H.A. Schwartz, M.L. Kern, G. Park, D.R. Labarthe, R.M. Merchant, S. Jha, M. Agrawal, L.A. Dziurzynski, and M. Sap et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science* 2: 159–169.

Esuli, A., and F. Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC'06, 417–422.

Filatova, E. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, 392–398.

Genereux, M., and R.P. Evans. 2006. Distinguishing affective states in weblogs. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, 27–29.

Gildea, D., and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3): 245–288.

González-Ibánez, R., S. Muresan, and N. Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the ACL*, 581–586.

Grefenstette, E., G. Dinu, Y.-Z. Zhang, M. Sadrzadeh, and M. Baroni. 2013. Multi-step regression learning for compositional distributional semantics. arXiv preprint arXiv:1301.6939.

Grefenstette, E., and M. Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1394–1404.

Hasan, Kazi Saidul, and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *The 6th international joint conference on natural language processing*, Nagoya.

Hassan, A., A. Abu-Jbara, and D. Radev. 2012. Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7 on Graph-Based Methods for Natural Language Processing*, 6–14.

Hatzivassiloglou, V., and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference of European Chapter of the Association for Computational Linguistics*, Madrid, 174–181.

Holzman, L.E., and W.M. Pottenger. 2003. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

Hu, M., and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'04, New York, 168–177.

John, D., A.C. Boucouvalas, and Z. Xu. 2006. Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications*, 183–188. Anaheim: ACTA Press.

Jurgens, D., S.M. Mohammad, P. Turney, and K. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval'12, Montréal, 356–364.
Kalchbrenner, N., E. Grefenstette, and P. Blunsom. 2014. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.
Kennedy, A., and D. Inkpen. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange During Negotiations*, Ottawa
Kessler, J.S., and N. Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009).*
Kiritchenko, S., and S.M. Mohammad. 2016a. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego.
Kiritchenko, S., and S.M. Mohammad. 2016b. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA).*
Kiritchenko, S., and S.M. Mohammad. 2016c. Sentiment composition of words with opposing polarities. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego.
Kiritchenko, S., S.M. Mohammad, and M. Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of English and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval-2016, San Diego.
Kiritchenko, S., X. Zhu, C. Cherry, and S. Mohammad. 2014a. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, 437–442.
Kiritchenko, S., X. Zhu, and S.M. Mohammad. 2014b. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50: 723–762.
Ku, L.-W., Y.-T. Liang, and H.-H. Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. 100107.
Li, W., and H. Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications* 41(4, Part 2): 1742–1749.
Liu, B., and L. Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, ed. C.C. Aggarwal and C. Zhai, 415–463. New York: Springer
Liu, J., Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. 2007. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, 334–342.
Lloret, E., A. Balahur, M. Palomar, and A. Montoyo. 2009. Towards building a competitive opinion summarization system: challenges and keys. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Companion Volume: S, 72–77.
Louviere, and J.J. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.
Lu, B., and B.K. Tsou. 2010. Cityu-dac: Disambiguating sentiment-ambiguous adjectives within context. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 292–295.
Ma, C., H. Prendinger, and M. Ishizuka. 2005. Emotion estimation and reasoning based on affective textual interaction. In *First International Conference on Affective Computing and Intelligent Interaction (ACII-2005)*, ed. J. Tao, R.W. Picard, Beijing, 622–628.
Màrquez, L., X. Carreras, K.C. Litkowski, and S. Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics* 34(2): 145–159.

Mihalcea, R., C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Mihalcea, R., and H. Liu. 2006. A corpus-based approach to finding happiness. In *AAAI-2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 139–144. AAAI Press.

Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.

Mitchell, J., and M. Loapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8): 1388–1429.

Mohammad, S. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, 587–591.

Mohammad, S., S. Kiritchenko, and X. Zhu. 2013a. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta.

Mohammad, S., S. Kiritchenko, and X. Zhu. 2013b. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'13, Atlanta.

Mohammad, S., and T. Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, Portland, 70–79.

Mohammad, S.M. 2012. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval'12, Stroudsburg, 246–255.

Mohammad, S.M. 2016a. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Mohammad, S.M. 2016b. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*, ed. H. Meiselman. Amsterdam: Elsevier.

Mohammad, S.M., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'16, San Diego.

Mohammad, S.M., P. Sobhani, and S. Kiritchenko. 2016b, In Press. Stance and Sentiment in Tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*.

Mohammad, S.M., and P.D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, California.

Mohammad, S.M., and P.D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3): 436–465.

Mohammad, S.M., X. Zhu, S. Kiritchenko, and J. Martin. 2015a. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management* 51: 480–499.

Mohammad, S.M., X. Zhu, S. Kiritchenko, and J. Martin. 2015b. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management* 51(4): 480–499.

Murakami, Akiko, and Rudy Raymond. 2010. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd international conference on computational linguistics*, Beijing.

Nahar, V., S. Unankard, X. Li, and C. Pang. 2012. Sentiment analysis for effective detection of cyber bullying. In *Web Technologies and Applications*, 767–774. Berlin/Heidelberg: Springer.

Nalisnick, E.T., and H.S. Baird. 2013a. Character-to-character sentiment analysis in Shakespeare's plays. In *Proceedings of the 51st annual meeting of the association for computational linguistics (ACL)*, Short Paper, Sofia, 479–483, Aug 2013.

Nalisnick, E.T., and H.S. Baird. 2013b. Extracting sentiment networks from Shakespeare's plays. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 758–762. IEEE.

Neviarouskaya, A., H. Prendinger, and M. Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-09)*, 278–281, San Jose.

O'Connor, B., R. Balasubramanyan, B.R. Routledge, and N.A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Orme, B. 2009. *Maxdiff analysis: Simple counting, individual-level logit, and HB*. Orem: Sawtooth Software, Inc.

Pak, A., and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, LREC'10, Valletta.

Palmer, M., D. Gildea, and N. Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies* 3(1): 1–103.

Pang, B., and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. 2014a. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'14, Dublin.

Pontiki, M., H. Papageorgiou, D. Galanis, I. Androutsopoulos, J. Pavlopoulos, and S. Manandhar. 2014b. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval'14, Dublin.

Popescu, A.-M., and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT'05, Stroudsburg, 339–346.

Qadir, A. 2009. Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*, eETTs'09, Stroudsburg, 38–43.

Ravaja, N., T. Saari, M. Turpeinen, J. Laarni, M. Salminen, and M. Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments* 15(4): 381–392.

Reyes, A., P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47(1): 239–268.

Rosenthal, S., P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval'15, Denver, 450–462.

Rosenthal, S., P. Nakov, A. Ritter, and V. Stoyanov. 2014a. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, ed. P. Nakov, and T. Zesch, SemEval-2014, Dublin.

Rosenthal, S., A. Ritter, P. Nakov, and V. Stoyanov. 2014b. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, 73–80.

Rudolph, S., and E. Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 907–916.

Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6): 1161.

Salameh, M., S.M. Mohammad, and S. Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the North American Chapter of Association of Computational Linguistics*, Denver.

Schwartz, H., J. Eichstaedt, M. Kern, L. Dziurzynski, R. Lucas, M. Agrawal, G. Park, et al. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Severyn, A., and A. Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 464–469. Denver: Association for Computational Linguistics.

Socher, R., B. Huval, C.D. Manning, and A.Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'12, Jeju.

Socher, R., A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, Seattle.

Somasundaran, Swapna, and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP*, Singapore.

Somprasertsri, G., and P. Lalitrojwong. 2010. Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science* 16(6): 938–955.

Sridhar, Dhanya, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the 52nd annual meeting of the association for computational linguistics*, Baltimore.

Stone, P., D.C. Dunphy, M.S. Smith, D.M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: The MIT Press.

Stoyanov, V., and C. Cardie. 2006. Toward opinion summarization: Linking the sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 9–14.

Strapparava, C., and R. Mihalcea. 2007. Semeval-2007 Task 14: Affective text. In *Proceedings of SemEval-2007*, Prague, 70–74.

Su, Q., K. Xiang, H. Wang, B. Sun, and S. Yu. 2006. Using pointwise mutual information to identify implicit features in customer reviews. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*, ICCPOL'06, 22–30. Berlin/Heidelberg: Springer.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2): 267–307.

Taboada, M., K. Voll, and J. Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Thelwall, M., K. Buckley, and G. Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 62(2): 406–418.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. Sydney: Association for Computational Linguistics.

Tokuhisa, R., K. Inui, and Y. Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1*, COLING'08, 881–888.

Turney, P., and M.L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21(4): 315–346.

Turney, P.D. 2014. Semantic composition and decomposition: From recognition to generation. arXiv preprint arXiv:1405.7908.

Veale, T., and Y. Hao. 2010. Detecting ironic intent in creative comparisons. *ECAI* 215: 765–770.

Velásquez, J.D. 1997. Modeling emotions and other motivations in synthetic agents. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, 10–15. AAAI Press.

Walker, Marilyn A., et al. 2012. A corpus for research on deliberation and debate. In *proceedings of the eighth international conference on language resources and evaluation (LREC)*, Istanbul.
Wan, X. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, 553–561.
Wang, C., and F. Wang. 2012. A bootstrapping method for extracting sentiment words using degree adverb patterns. In *2012 International Conference on Computer Science and Service System (CSSS)*, 2173–2176. IEEE.
Wilson, T., Z. Kozareva, P. Nakov, S. Rosenthal, V. Stoyanov, and A. Ritter. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'13, Atlanta.
Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT'05, Stroudsburg, 347–354.
Xu, G., C.-R. Huang, and H. Wang. 2013. Extracting Chinese product features: Representing a sequence by a set of skip-bigrams. In *Proceedings of the 13th Chinese Conference on Chinese Lexical Semantics*, CLSW'12, 72–83. Berlin/Heidelberg: Springer.
Xu, R., K.-F. Wong, Q. Lu, Y. Xia, and W. Li. 2008. Learning knowledge from relevant webpage for opinion analysis. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08.*, vol. 1, 307–313. IEEE.
Yessenalina, A., and C. Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 172–182.
Zarrella, G., and A. Marsh. 2016. MITRE at SemEval-2016 Task 6: Transfer learning for Stance detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'16, San Diego.
Zhang, C., D. Zeng, Q. Xu, X. Xin, W. Mao, and F.-Y. Wang. 2008. Polarity classification of public health opinions in Chinese. In *Intelligence and Security Informatics*, 449–454. Berlin/Heidelberg: Springer.
Zhang, L., and B. Liu. 2011. Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers – Volume 2*, HLT'11, 575–580.
Zhang, L., B. Liu, S.H. Lim, and O'Brien-E. Strain. 2010. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING'10, Stroudsburg, 1462–1470.
Zhe, X., and A. Boucouvalas. 2002. *Text-to-Emotion Engine for Real Time Internet Communication Text-to-Emotion Engine for Real Time Internet Communication*, 164–168.
Zhu, X., H. Guo, S. Mohammad, and S. Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, 304–313.

Chapter 5
Sentiment Resources: Lexicons and Datasets

Aditya Joshi, Pushpak Bhattacharyya, and Sagar Ahire

Abstract Sentiment lexicons and datasets represent the knowledge base that lies at the foundation of a SA system. In its simplest form, a sentiment lexicon is a repository of words/phrases labelled with sentiment. Similarly, a sentiment-annotated dataset consists of documents (tweets, sentences or longer documents) labelled with one or more sentiment labels. This chapter explores the philosophy, execution and utility of popular sentiment lexicons and datasets. We describe different labelling schemes that may be used. We then provide a detailed description of existing sentiment and emotion lexicons, and the trends underlying research in lexicon generation. This is followed by a survey of sentiment-annotated datasets and the nuances of labelling involved. We then show how lexicons and datasets created for one language can be transferred to a new language. Finally, we place these sentiment resources in the perspective of their classic applications to sentiment analysis.

Keywords Sentiment lexicons • Sentiment datasets • Evaluation • Transfer learning

The previous chapter shows that sentiment analysis (SA) is indeed more challenging than it seems. The next question that arises is, where does the program 'learn' the sentiment from? In other words, where does the knowledge required for a SA system come from? This chapter discusses sentiment resources as means to this requirement of knowledge. We refer to words/phrases and documents as 'textual units'. In sentiment resources, it is these textual units that are annotated with sentiment information.

A. Joshi (✉)
IITB-Monash Research Academy, Mumbai, India
e-mail: adityaj@cse.iitb.ac.in

P. Bhattacharyya • S. Ahire
IIT Bombay, Mumbai, India
e-mail: pb@cse.iitb.ac.in

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_5

5.1 Introduction

Sentiment resources, *i.e.*, lexicons and datasets represent the knowledge base of a SA system. Thus, creation of a sentiment lexicon or a dataset is the fundamental requirement of a SA system. In case of a lexicon, it is in the form of simpler units like words and phrases, whereas in case of datasets, it consists of comparatively longer text. There exists a wide spectrum of such resources that can be used for sentiment/emotion analysis. Before we proceed, we reiterate the definition of sentiment and emotion analysis. We refer to sentiment analysis as a positive/negative/neutral classification task, whereas emotion analysis deals with a wider spectrum of emotions such as angry, excited, etc. A discussion on both sentiment and emotion lexicons is imperative to show how different the philosophy behind construction of the two is.

A sentiment resource is a repository of textual units marked with one or more labels representing a sentiment state. This means that there are two driving components of a sentiment resource: (a) the textual unit, and (b) the labels. We discuss the second component, labels in detail in Sect. 5.2.

In case of a sentiment lexicon, the lexical unit may be a word, a phrase or a concept from a general purpose lexicon like WordNet. What constitutes the labels is also important. The set of labels may be purely functional: task-based. For a simple positive-negative classification, it is often sufficient to have a set of positive and negative words. If the goal is a system that gives 'magnitude' ('The movie was horrible' is more strongly negative than 'The movie was bad'), then the lexicon needs to capture that information in terms of a magnitude in addition to positive and negative words.

An annotated dataset consists of documents labelled with one or more output labels. As in the case of sentiment lexicons, the two driving components of a sentiment-annotated dataset are: (a) the textual unit, and (b) the labels. For example, a dataset may consist of a set of movie reviews (the textual units) annotated by human annotators as positive or negative (the labels). Datasets often contain additional annotation in order to enrich the quality of annotation. For example, a dataset of restaurant reviews annotated with sentiment may contain additional annotation in the form of restaurant location. Such annotation may facilitate insights such as: which restaurant is the most popular, what are the issues with respect to this outlet of a restaurant that people complain the most about, etc.

5.2 Labels

A set of labels is the pre-determined set of attributes that each textual unit in a sentiment resource will be annotated with. The process of assigning a label to a textual unit is called annotation, and in case the label pertains to sentiment, the process is called *sentiment annotation*. The goal of sentiment annotation is to assign labels in one out of three schemes: absolute, overlapping and fuzzy. The first two are shown in Liu (2010).

Absolute labelling is when a textual unit is marked as exactly one out of multiple labels. An example of absolute labelling may be positive versus negative – where each document is annotated as either positive or negative. An additional label 'neutral' may be added. A fallback label such as 'ambiguous'/'unknown'/'unsure' may be introduced. Numeric schemes that allow labels to range between, say, +5 to −5 also fall under this method of labelling.

Labels can be overlapping as well. A typical example of this is emotion labels. Emotions are more complex than sentiment, because there can be more than one emotion at a time. For example, the sentence, "Was happy to bump into my friend at the airport this afternoon." would be labelled as positive as a sentiment-annotated sentence. On the other hand, an emotion annotation would require two labels to be assigned to this text: happiness and surprise. Emotions can, in fact, be thought of arising from a combination of emotions, and their magnitudes. This means that while positive-negative are mutually exclusive, emotions need not be. In such cases, each one of them must be viewed as a Boolean attribute. This means that the word 'amazed' will be marked as 'happy: yes, surprised: yes' for an emotion lexicon, whereas the same 'amazed' will be marked as 'positive' for a sentiment lexicon. By definition, a positive word implies that it is not negative.

Finally, the third scheme of labelling is fuzzy: where a distribution over different labels is assigned to a textual unit. Consider the case where we assign a distribution over 'positive/negative' as a label. Such a distribution implies likelihood of the textual unit to belong to the given label. For example, a word with 'positive:0.8, negative:0.2' means that the word tends to occur more frequently in a positive sense – however, it is not completely positive and it may still be used in the negative sense to an extent.

Several linguistic studies have explored what constitutes basic labels for a sentiment resource. In the next subsections, we look at three strategies.

5.2.1 Stand-Alone Labels

A sentiment resource may use two labels: positive or negative. The granularity can be increased to strongly positive, moderately positive and so on. A positive unit represents a desirable state, whereas a negative unit represents an undesirable state (Liu 2010). Emotion labels are more nuanced. Basic emotions are a list of emotions that are fundamental to human experience. Whether or not there are any basic emotions at all, and whether it is worthwhile to discover these basic emotions has been a matter of disagreement. Ortony and Turner (1990) state that the basic emotion approach (*i.e.,* stating that there are basic emotions and other emotions evolve from them) is flawed, while Ekman (1992) supports the basic emotion theory. Several basic emotions have been suggested. Ekman suggests six basic emotions:

anger, disgust, fear, sadness, happiness and surprise. Plutchik has listed eight basic emotions: six from Ekman's list in addition to anticipation and trust (Plutchik 1980).

5.2.2 Dimensions

Sentiment has been defined by Liu (2010) as a 5-tuple: <sentiment-holder, sentiment-target, sentiment-target-aspect, sentiment, sentiment-time>. This means that sentiment in a textual unit can be captured accurately only if information along the five dimensions is obtained. Similarly, emotions can also be looked at in the form of two dimensions: valence and arousal (Mehrabian and Russell 1974). Valence indicates whether an emotion is pleasant or unpleasant. Arousal indicates the magnitude of an emotion. Happy and excited are two forms of a pleasant emotion, but they differ along the arousal axis. Excitement indicates a state where a person is happy, but aroused to a great degree. On the other hand, calm and content, while still being pleasant emotions, represent a deactivated state. Corresponding emotions in the left quadrant (that indicates unpleasant emotions) are sad, stressed, bored and fatigued. In such a case, overlapping labelling must be used. A resource annotated using dimensional structure will assign a value per dimension for each textual unit.

5.2.3 Structures

Plutchik wheel of emotions (Plutchik 1982) is a popular structure that represents basic emotions, and emotions that arise as a combination of these emotions. It combines the notion of basic emotions, along with arousal as seen in case of emotion dimensions. The basic emotions according to Plutchik's wheel are joy, trust, fear, surprise, anticipation, sadness, disgust, anger and anticipation. The basic emotions are arranged in a circular manner to indicate antonymy. The opposite of 'joy' is placed diametrically opposite to it: 'sadness'. Similarly, 'anticipation' lies diametrically opposite to 'surprise'. Each 'petal' of the wheel indicates the arousal of the emotion. The emotion 'joy' has 'serenity' above it and 'ecstasy' below it. These emotions indicate a deactivated and activated state of arousal respectively. Similarly, an aroused state of 'anger' becomes 'rage'. Thus, the eight emotions in the central circle are the aroused forms of the basic emotions. These are: rage, loathing, grief, amazement, terror, admiration, ecstasy and vigilance. The wheel also allows combination of emotions to create more nuanced emotions. A resource annotated using a structure such as the Plutchik wheel of emotions will place every textual unit in the space represented by the structure.

5.3 Lexicons

We now discuss sentiment lexicons: we describe them individually first, and then show trends in lexicon generation. Words/phrases have two kinds of sentiment, as given in Liu (2010): absolute and relative. Absolute sentiment means that the sentiment remains the same, given the right word/phrase and meaning. For example, the word 'beautiful' is a positive word. Relative sentiment means that the sentiment changes depending on the context. For example, the word 'increased' or 'fuelled' has a positive/negative sentiment based on what the object of the word is. There exists a third category of sentiment: implicit sentiment. Implicit sentiment is different from absolute sentiment. Implicit sentiment is the sentiment that is commonly invoked in the mind of a reader when he/she reads that word/phrase. Consider the example 'amusement parks'. A reader typically experiences positive sentiment on reading this word. Similarly, the phrase 'waking up in the middle of the night' does involve an implicit negative sentiment.

Currently, most sentiment lexicons limit themselves to absolute sentiment words. Extraction of implicit sentiment in phrases forms a different branch of work. However, there exist word association lexicons that capture implied sentiment in words (Mohammad and Turney 2010). We stick to this definition as well, and discuss sentiment and emotion lexicons that capture absolute sentiment.

5.3.1 Sentiment Lexicons

Early development of sentiment lexicons focused on creation of sentiment dictionaries. Stone et al. (1966) present a lexicon called 'General Inquirer' that has been widely used for sentiment analysis. Finn (2011) present a lexicon called AFINN. Like General Inquirer, it is also a manually generated lexicon. To show the general methodology underlying sentiment lexicons, we describe some popular sentiment lexicons in the forthcoming subsections.

5.3.1.1 SentiWordNet

SentiWordNet, described first by Esuli and Sebastiani (2006), is a sentiment lexicon which augments WordNet (Miller 1995) with sentiment information. The labelling is fuzzy, and is done by adding three sentiment scores to each synset in the WordNet as follows. Every synsets has three scores:

1. Pos(s): The positive score of synsets
2. Neg(s): The negative score of synsets
3. Obj(s): The objective score of synsets

Thus, in SentiWordNet, sentiment is associated with the meaning of a word rather than the word itself. This representation allows a word to have multiple sentiments corresponding to each meaning. Because there are three scores, each meaning in itself can be both positive and negative, or neither positive nor negative.

The process of SentiWordNet creation is an expansion of the approach used for the three-class sentiment classification to handle graded sentiment values. The algorithm to create SentiWordNet can be summarized as:

1. *Selection of Seed Set:* A seed set L_p and L_n consisting of 'paradigmatic' positive and negative synsets respectively was created. Each synset was represented using the TDS. This representation converted words in the synset, its WordNet definition and the sample phrases together with explicit labels for negation into vectors.
2. *Creation of Training Set:* This seed set was expanded for k iterations using the following relations of WordNet: Direct antonymy, Similarity, Derived from, Pertains to, Attribute and Also see. These were the relations hypothesized to preserve or invert the associated sentiment. After k iterations of expansion, this gave rise to the sets Tr_p^k and Tr_n^k. The objective set L_o = Tr_o^k was assumed to consist of all the synsets that did not belong to Tr_p^k or Tr_n^k.
3. *Creation of Classifiers:* A classifier can be defined as a combination of a learning algorithm and a training set. In addition to the two choices of learning algorithms (SVM and Rocchio), four different training sets were constructed with the number of iterations of expansion k = 0, 2, 4, 6. The size of the training set increased substantially with an increase in k. As a result, low values of k yielded classifiers with low recall but high precision, while higher k led to high recall but low precision. As a result there were 8 ternary classifiers in total due to all combinations of the 2 learners and 4 training sets. Each ternary classifier was made up of two binary classifiers, positive vs. not positive and negative vs. not negative.
4. *Synset Scoring:* Each synset from the WordNet was vectorized and given to the committee of ternary classifiers as test input. Depending upon the output of the classifiers, each synset was assigned sentiment scores by dividing the count of classifiers that give a label by the total number of classifiers (8).

5.3.1.2 SO-CAL

Sentiment Orientation CALculator (SO-CAL) system (Brooke et al. 2009) is based on a manually constructed low-coverage resource made up of raw words. Unlike SentiWordNet, there is no sense information associated with a word. SO-CAL uses as its basis a lexical sentiment resource consisting of about 5000 words. (In comparison, SentiWordNet has over 38,000 polar words and several other strictly objective words.) Each word in SO-CAL has a sentiment label which is an integer in [−5, +5] apart from 0 as objective words are simply excluded. The strengths of SO-CAL lie in its accuracy, as it is manually annotated, and the use of detailed features that handle sentiment in various cases in ways conforming to linguistic phenomena.

SO-CAL uses several 'features' to model different word categories and the effects they have on sentiment. In addition, a few special features operate outside the scope of the lexicon in order to affect the sentiment on the document level. These are some of the features of SO-CAL:

1. *Adjectives*: A manual dictionary of adjectives was created by manually tagging all adjectives in a 500-document multidomain review corpus, and the terms from the General Inquirer dictionary were annotated added to the list thus obtained.
2. *Nouns, Verbs and Adverbs*: SO-CAL also extended the approach used for adjectives to nouns and verbs. As a result, 1142 nouns and 903 verbs were added to the sentiment lexicon. Adverbs were added by simply adding the -ly suffix to adjectives and then manually altering words whose sentiment was not preserved, such as essentially. In addition multi-word expressions were also added, leading to an addition to 152 multiwords in the lexicon. Thus, while the adjective 'funny' has a sentiment of +2, the multiword 'act funny' has a sentiment of −1.
3. *Intensifiers and Downtoners*: An Intensifier is a word which increases the intensity of the phrase to which it is applied, while a Downtoner is a word which decreases the intensity of the phrase to which it is applied. For instance the word 'extraordinarily' in the phrase 'extraordinarily good' is an intensifier while the word somewhat in the phrase 'somewhat nice' is a downtoner.

5.3.1.3 Sentiment Treebank & Associated Lexicon

This Treebank was introduced in Socher et al. (2013). In order to do create the Treebank, the work also came up with a lexicon called the Sentiment Treebank, which is a lexicon consisting of partial parse trees annotated with sentiment.

The lexicon was created as follows. A movie review corpus was obtained from www.rottentomatoes.com, consisting of 10,662 sentences. Each sentence was parsed using the Stanford Parser. This gave a parse tree for each sentence. The parse trees were split into phrases, i.e., each parse tree was split into its components, each of which was then output as a phrase. This gave rise to 215,154 phrases. Each of these phrases was tagged for sentiment using Amazon's Mechanical Turk's interface. The selection of labels is also described in the original paper. Initially, the granularity of the sentiment values was 25, *i.e.*, 25 possible values could be given for the sentiment, but it was observed from the data from the Mechanical Turks experiment that most responses contained any one of only 5 values. These 5 values were then called 'very positive', 'positive', 'neutral', 'negative' and 'very negative'.

Table 5.1 Summary of sentiment lexicons

	Approach	Lexical unit	Labels	Observation
SO-CAL	Manual	Word	Integer in [−5, +5]	Performance can be improved by incorporating linguistic features even with low coverage
SentiWordNet	Automatic	WordNet Synset	3 fractional values Pos, Neg, Obj in [0, 1]	WordNet captures senses. Different senses may have different sentiment.
Sentiment Treebank	Manual, Crowdsourced	Phrase	5 labels ranging from "very negative" to "very positive"	Crowdsourcing can be beneficial. Tune labels according to the task.
Macquaire semantic orientation lexicon	Semi-supervised	Words	Positive/ negative	Using links in a thesaurus to discover new words.

5.3.1.4 Summary

Table 5.1 summarizes sentiment lexicons described above, and in addition, also mentions some other sentiment lexicons. We compare along four parameters: the approach used for creation, lexical units, labels and some observations. Mohammad et al. (2009) present Macquaire semantic orientation lexicon. This is a sentiment lexicon that contains 76,400 terms, marked as positive or negative. In terms of obtaining manual annotations, Louviere (1991) present an approach called the MaxDiff approach. In this case, instead of obtaining annotations for one word at a time, an annotator is shown multiple words and asked to identify the least positive and most positive word among them.

5.3.2 *Emotion Lexicons*

We now describe emotion lexicons. They have been described in this separate subsection so as to highlight challenges and the approaches specific to emotion lexicon generation.

5.3.2.1 LIWC

Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001) is a popular manually created lexicon. The lexicon consists of 4500 words and word stems

(An example word stem is happ* which covers adjectival and adverbial forms of the word) arranged in four categories. The four categories of words in LIWC are: Linguistic processes (pronouns, prepositions, conjunctions, etc.), Speaking processes (Interjections, Fillers, etc.), personal concerns (words related to work, home, etc.) and psychological processes. The words in the psychological processes category deal with affect and opinion, and are further classified into cognitive and affective processes. Cognitive processes include words indicating certainty ('definitely'), possibility ('likely') and inhibition ('prevention'), etc. Affective processes include words with positive/negative emotion, words expressing anxiety, anger, sadness. LIWC 2001 has 713 cognitive and 915 affective processes words. LIWC was manually created by three linguistic experts in two steps:

(a) Define category scales: The judges determined categories and decided how they can be grouped into a hierarchy
(b) Manual population: The categories were manually populated with words. For each word, three judges manually evaluated whether or not a word should be placed in a category. In addition, they also considered if a word can be moved higher up in the hierarchy.

LIWC now exists in multiple languages, and has been widely used by several applications for analysis of topic as well as sentiment/emotion.

5.3.2.2 ANEW

Affective norms for English words (ANEW) (Bradley and Lang 1999) is a dictionary of around 1000 words where each word is indicated with a three-tuple representation: pleasure, arousal and activation. Pleasure indicates the valence of a word, arousal the intensity while activation indicates whether the emotion expressed in the word is in control or not. Consider the example word 'afraid'. This word is indicated by the tuple (negative, 3, not) indicating that it is a negative emotion, with an arousal of 3, and is a deactivated emotion. ANEW was manually created by 25 annotators separately. Each annotation experiment was conducted in runs of 100–150 words. Annotators are given a sheet called ScanSAM sheet. Each annotator marks values of S, A and M for word. The annotators perform the annotation separately.

5.3.2.3 Emo-Lexicon

Emo-Lexicon (Mohammad and Turney 2013) is a lexicon of 14,000 terms created using crowd-sourcing portals like Amazon Mechanical Turk. Association with positive and negative valence as well as with the eight Plutchik emotions is also available. Although it is manually created, the lexicon is larger than other emotion lexicons – a clear indication that crowdsourcing is indeed a powerful mechanism for large-scale creation of emotion lexicon. However, because the task of lexicon creation has been opened up to the 'crowd', quality control is a key challenge. To mitigate this, the lexicon is created with additional drivers, as follows:

1. A list of words is created from a thesaurus.
2. When an annotator annotates a word with emotion, he/she must first ascertain the sense of the word. The target word is displayed along with four words. The annotator must select one that is closest to the target word.
3. Only if the annotator was able to correctly determine the sense of the word is his/her annotation for emotion label obtained.

5.3.2.4 WordNet-Affect

WordNet-Affect (Strapparava and Valitutti 2004) like SentiWordNet, is a resource that annotates senses in WordNet with emotions. WordNet Affect was created using a semi-supervised method. It consists of 2874 synsets annotated with affective labels (called a-labels). WordNet-Affect was created as follows:

1. A set of core synsets is created. These are synsets whose emotion has been manually labelled in the form of a-labels.
2. These labels are projected to other synsets using WordNet relations.
3. The a-labels are then manually evaluated and corrected, wherever necessary.

5.3.2.5 Chinese Emotion Lexicon

A Chinese emotion lexicon (Xu et al. 2010) was created using a semi-supervised approach, in absence of a graphical structure such as WordNet. There are two steps of creation:

1. Select a core set of labelled words.
2. Expand these words using a similarity matrix. Iterate until convergence.

   The similarity matrix takes three kinds of similarity into account:

1. *Syntagmatic similarity*: This includes co-occurrence of two words in a large text corpus.
2. *Paradigmatic similarity*: This includes relations between two words in a semantic dictionary.
3. *Linguistic peculiarity*: This involves syllable overlap, possibly to cover different forms of the same word.

5.3.2.6 SenticNet

SenticNet (The most recent version, being SenticNet 4) by Cambria et al. (2016) is a rich graphical repository of concepts. The resource aims to capture semantic,

Table 5.2 Summary of emotion lexicons

	Approach	Labels	Observation
LIWC	Manual	Hierarchy of categories	Decide hierarchy of categories; have judges interacting with each other
ANEW & ANEW for Spanish	Manual	Valence, arousal, dominance	ScanSAM lists; have a set of annotators annotating in parallel
Emo-Lex	Manual	Eight emotions, two valence categories	Use crowd-sourcing. Attention to quality control.
WordNet affect	Semi-supervised	Affective labels	Annotate a seed set. Expand using WordNet relations.
Chinese emotion lexicon	Semi-supervised	Five emotions	Annotate a seed set. Expand using similarity matrices
NRC Hashtag emotion lexicon	Automatic	Eight emoticons	Use hashtag based supervision of tweets
SenticNet 4	Semi-supervised	A larger structure	Semi-supervised graphical structure, created using techniques such as agglomerative clustering

and sentic properties of words and phrases. The sentic properties are related to connotations of words. A detailed discussion of SenticNet forms a forthcoming chapter of this book.

5.3.2.7 Summary

Table 5.2 shows a summary of emotion lexicons discussed in this section. We observe that manual approaches dominate emotion lexicon creation. Key issues in manual emotion annotation are: ascertaining the quality of the labels, deciding hierarchies if any. Additional useful lexicons are available at: http://www.saifmohammad.com/WebPages/lexicons.html. On the other hand, automatic emotion annotation is mostly semi-supervised. To expand a seed set, structures like WordNet may be used, or similarity matrices constructed from large corpora can be employed. Mohammad (2012) present a hashtag emotion lexicon that consists of 16,000+ unigrams annotated with eight emotions. The lexicon is created using emotion-denoting hashtags present in tweets. Mohammad and Turney (2010) is also an emotion lexicon created using a crowdsourcing platform.

5.4 Sentiment-Annotated Datasets

This section describes sentiment-annotated datasets, and is organized as follows. We first describe sources of data, mechanisms of annotation, and then provide a list of some sentiment-annotated datasets.

5.4.1 Sources of Data

The first step is to obtain raw data. The following are candidate sources of raw data:

1. **Social networking websites** like twitter are a rich source of data for sentiment analysis applications. For example, Twitter API (Makice 2009) is a publicly available API that allows you to download tweets based on a lot of interesting search criteria such as keyword-based-search, download-timelines, download-tweet-threads, etc.
2. **Competitions such as SemEval** have been regularly conducting Sentiment analysis related tasks. These competitions release a training dataset followed by a test dataset. These datasets can be used as benchmark datasets.
3. **Discussion forums** are portals where users discuss topics, often in the context of a central theme or an initial question. These discussion forums often arrange posts in a thread-like manner. This allows discourse nature to sentiment. However, this also introduces an additional challenge. A reply to a post could mean one out of three possibilities: (a) The reply is an opinion with respect to the post, offering an agreement or disagreement (example: Well-written post), (b) The reply is an opinion towards the author of the post (example: Why do you always post hateful things?), or (c) The reply is an opinion towards the topics being discussed in the post. (Example: You said that the situation is bad. But do you think that....). Reddit threads have been used as opinion datasets in several past works.
4. **Review websites**: Amazon and other review websites have reviews on different domains. Each kind of reviews has unique challenges of its own. In case of movie reviews, the review often has a portion describing 'what' the movie is about. It is possible to create subjective extracts before using them as done by Mukherjee and Bhattacharyya (2012). In case of product reviews, the review often contains sentiment towards different 'aspects'. ('Aspects' of a cell phone are battery, weight, OS, etc.).
5. **Blogs** are often long text describing an opinion with respect to a topic. They can also be crawled and annotated to create a sentiment dataset. Blogs tend to be structured narratives analyzing the topic. They may not always contain the same sentiment throughout but can be useful sources of data that looks at different aspects of the given topic.

5.4.2 Obtaining Labels

Once raw data has been obtained, the second step is to label this data. There are different approaches that can be used for obtaining labels for a dataset:

1. **Manual labelling**: Several datasets have been created by human annotators. The labelling can be done through crowd-sourcing applications like Amazon

Mechanical Turk. They allow obtaining large volumes of annotations by employing the ‘power of the crowds’ (Paolacci et al. 2010). To control the quality of annotation, one way is to use a seed set of gold labels. Human annotators within the controlled setup of the experiment create a set of gold labels. If a crowd-sourced annotator (known as ‘worker’ in the crowd-sourcing parlance) gets a sufficient number of gold labels right, only then is he/she permitted to perform the task of annotation.

2. **Distant supervision**: Distant supervision refers to the situation where the label or the supervision is obtained without an annotator – hence the word ‘distant’. One way to do so is to use annotation provided by the writer themselves. However, the question of reliability arises because not every data unit has been manually verified by a human annotator. This has to be validated using the approach used to obtain distant supervision. Consider the example of Amazon reviews. Each review is often accompanied by star ratings. These star ratings can be used as labels provided by the writer. Since these ratings are out of 5, a review with 1 star is likely to be strongly negative, whereas a review with 5 stars is likely to be strongly positive. To improve the quality of the dataset obtained, Pang and Lee (2005) consider reviews that are definitely positive and definitely negative – i.e. reviews with 5 and 1 stars respectively.

   Another technique to obtain distant supervision is the use of hashtags. Twitter provides a reverse index mechanism in the form of hashtags. An example tweet is ‘Just finished writing a 20 page long assignment. #Engineering #Boring’. ‘#Engineering’ and ‘#Boring’ as hashtags – since they are phrases preceded by a hashtag symbol. Note that a hashtag is created by the author of the tweet and hence, can be anything – topical (i.e. identifying what the tweet is about. Engineering, in this case) or emotion-related (i.e. expressing an opinion through a hashtag. In this case, the author of the tweet is bored). Purver and Battersby (2012) emotion-related hashtags to obtain a set of tweets containing emotion-related hashtags. Thus, hashtags such as ‘#happy’, ‘#sad’, etc. are used to download tweets using the Twitter API. The tweets are then labelled as ‘#happy’, ‘#sad’, etc. Since hashtags are user-created, they can be more nuanced than this. For example, consider the hypothetical tweet: ‘Meeting my ex-girlfriend after about three years. #happy #not’. The last hashtag ‘#not’ inverts sentiment expressed by the preceding hashtag ‘#happy’. This unique construct ‘#not’ or ‘#notserious’ or ‘#justkidding’/‘#jk’ is popular in tweets and must be handled properly when hashtag-based supervision is used to create a dataset.

5.4.3 Popular Sentiment-Annotated Datasets

We now discuss some popular sentiment-annotated datasets. We divide them into two categories: sentence-level annotation, discourse-level annotation. The latter points to text longer than a sentence. While tweets may contain more than a sentence, we group them under sentence-level annotation because of limited length of tweets.

Sentence-Level Annotated Datasets

Niek Sanders released a dataset at http://www.sananalytics.com/lab/twitter-sentiment/. It consists of 5513 manually labelled tweets, classified as per four topics.

SemEval is a competition that is run for specific tasks. Sentiment analysis and related tasks have featured since 2013 (Nakov et al. 2013; Rosenthal et al. 2014, 2015). The datasets for these tasks are released online, and can be useful for sentiment applications. SemEval 2013 dataset is at: http://www.cs.york.ac.uk/semeval-2013/semeval2013.tgz SemEval 2014 dataset is at: http://alt.qcri.org/semeval2014/task9/ SemEval 2015 dataset is at: http://alt.qcri.org/semeval2015/task10/index.php?id=subtaske-readme

Darmstadt corpus consists of consumer reviews annotated at sentence and expression level. The dataset is available at: https://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus/ Sentence annotated polarity dataset from Pang et al. (2002) is also available at: https://www.cs.cornell.edu/people/pabo/movie-review-data/ Sentiment140 (Go et al. 2009) is a corpus made available by Stanford at http://help.sentiment140.com/for-students. The dataset is of tweets and contains additional information such as timestamp, author, tweet id, etc.

Deng et al. (2013) released *a goodFor/badFor corpus* that is available at: http://mpqa.cs.pitt.edu/corpora/gfbf/. goodFor/badFor indicates positive/negative sentiment respectively. This corpus uses a five-tuple representation for opinion annotation. Consider this example sentence from their user manual: 'The smell stifled his hunger.' This sentence is marked as: 'span: stifled, polarity: badFor, agent: the smell, object: his hunger'.

Discourse-Level Annotated Datasets

Many *movie review* datasets and lexicons are available at: https://www.cs.cornell.edu/people/pabo/movie-review-data/. These datasets include: sentiment annotated datasets, subjectivity annotated datasets, and sentiment scale datasets. These have been released in Pang and Lee (2004, 2005), and widely used.

A Congressional speech dataset (Thomas et al. 2006) annotated with opinion is available at: http://www.cs.cornell.edu/home/llee/data/convote.html The labels indicate whether the speaker supported or opposed a legislation that he/she was talking about.

A corpus consisting of *Amazon reviews* from different domains such as electronics, movies, etc. is available at: https://snap.stanford.edu/data/web-Amazon.html (McAuley and Leskovec 2013). This dataset spans a period of 18 years, and contains information such as: product title, author name, star rating, helpful votes, etc.

The Political Debate Corpus by Somasundaran and Wiebe (2009) is a dataset of political debates that is arranged based on different topics. It is available here: http://mpqa.cs.pitt.edu/corpora/product_debates/.

MPQA Opinion Corpus (Wiebe et al. 2005) is a popular dataset that consists of news articles from different sources. Version 2.0 of the corpus is nearly 15,000 sentences. The sentences are annotated with topics and labels. The topics are from different countries around the world. This corpus is available at http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/.

5.5 Bridging the Language Gap

Creation of a sentiment lexicon or a labelled dataset is a time/effort-intensive task. Since English is the dominant language in which SA research has been carried out, it is only natural that many other languages have tried to leverage on resources developed for English by adapting and/or reusing them. Cross-lingual SA refers to use of systems and resources developed for one language to perform SA of another. The first language (where the resources/lexicons/systems have been developed) is called the source language, while the second language (where a new system/resource/lexicon needs to be deployed) is called the target language. The basis of cross-lingual SA is availability of a lexicon or an annotated dataset in the source language. It must be noted that several nuanced methodologies to perform cross-lingual SA exist, but have been left out due to the scope of this chapter. We focus on cross-lingual sentiment resources.

The fundamental requirement is a mapping between the two languages. Let us consider what happens in case we wish to map a lexicon in language X to language Y. For a lexicon, this mapping can be in the form of a parallel dictionary where words of one language are mapped to another. ANEW For Spanish (Redondo et al. 2007) describes the generation of a lexicon called ANEW. Originally created for English words, its parallel Spanish version is created by translating words from English to Spanish, and then manually validating them. It can also be in the form of linked WordNets, in case the lexicons involve concepts like synsets. Hindi SentiWordNet (Joshi et al. 2010) map synsets in English to Hindi using a WordNet linking, and generate a Hindi SentiWordNet from its English variant. Mahyoub et al. (2014) describe a technique to create a sentiment lexicon for Arabic. Based on a seed set of positive and negative words, and Arabic WordNet, they present an expansion algorithm to create a lexicon. The algorithm uses WordNet relations in order to propagate sentiment labels to new words/synsets. The WordNet relations they use are divided into two categories: the ones that preserve the sentiment orientation, and the ones that invert the sentiment orientation.

How is this process of mapping words in one language to another any different for datasets? In case a machine translation (MT) system is available, this task is simple. A dataset in the source language can be translated to the target language. This is a common strategy that has been employed (Mihalcea et al. 2007; Duh et al. 2011). It follows that translation may introduce additional errors into the system, thus causing a degradation in the quality of the dataset. This is particularly applicable to translation of sentiment-bearing idioms. Salameh et al. (2015) perform their experiments for Arabic where a MT system is used to translate documents, following which sentiment analysis is performed. An interesting observation that the authors make is that although MT may result in a poor translation making it difficult for humans to identify sentiment, a classifier performs reasonably well. However, MT systems may not exist for all language pairs. Balamurali et al. (2012) suggest a naive replacement for a MT system. To translate a corpus from Hindi to Marathi (and vice versa), they obtain sense annotations for words in the dataset. Then, they use a WordNet linking to transfer the word from the source language to the target language.

An immediate question that arises is the hypothesis at the bottom of all cross-lingual approaches: sentiment is retained across languages. This means that if a word has a sentiment s in the source language, the translated word in target language (with appropriate sense recorded) also has sentiment s. How fair is the hypothesis that words in different languages bear the same emotion? This can be seen from linear correlations between ratings for the three affective dimensions, as was done for ANEW for Spanish. ANEW for Spanish (Redondo et al. 2007), as described above, was a lexicon created using ANEW in English. The correlation values for valence, arousal and dominance are 0.916, 0.746 and 0.720 respectively. This means that a positive English word is very likely to be a positive Spanish word. The arousal and dominance values remain the same to a lower extent.

Thus, we have two options now. The first option is cross-lingual SA: use resources generated for the source language and map it to the target language. The second option is in-language SA: create resources for the target language on its own. Balamurali et al. (2013) weighs in-language SA against cross-lingual SA based on Machine Translation. The authors show for English, German, French and Russian that in-language SA does consistently better than cross-lingual SA relying on translation alone.

Cross-lingual SA also benefits from additional corpora in target language:

1. *Unlabeled corpus in target language*: This type of corpus is used in different approaches, the most noteworthy being the co-training-based approach. Wan (2009). The authors assume that a labelled corpus in the source language, unlabeled corpus in target language and a MT system to translate back and forth between the two languages are available.
2. *Labelled corpus in target language:* The size of this dataset is assumed to be much smaller than the training set.
3. *Pseudo-parallel data:* Lu et al. (2011) describe use of pseudo-parallel data for their experiments. Pseudo-parallel data is the set of sentences in the source language that are translated to the target language and used as an additional polarity-labelled data set. This allows the classifier to be trained on a larger number of samples.

5.6 Applications of Sentiment Resources

In the preceding sections, we described sentiment resources in terms of labels, annotation techniques and approaches to creation. We will now see how a sentiment resource (either a lexicon or a dataset) can be used.

A lexicon is useful as a knowledge base for a rule-based SA system. A rule-based SA system takes a textual unit as input, applies a set of pre-determined rules, and produces a prediction. Joshi et al. (2011) present C-Feel-It, a rule-based SA system for tweets. The workflow is as follows:

1. A user types a keyword. Tweets containing the keyword are downloaded using the Twitter API
2. The tweets are pre-processed to correct extended words (e.g. 'happpyyyyy' is replaced with two occurrences of happy. Two, because the extended form of the word 'happy' has a magnified sentiment)
3. The words in a tweet are looked up individually in four *lexical resources*. The sentiment label of a tweet is calculated as a sum of positive and negative words – with rules applied for conjunctions and negation. In case of negation, the sentiment of words within a window is inverted. In case of conjunctions such as 'but', the latter part of a tweet is considered.
4. The resultant prediction of a tweet is a weighted sum of prediction made by the four lexical resources. The weights are determined experimentally by considering how well the resources perform on an already labelled dataset of tweets.

The above approach is a common framework for rule-based SA systems. Levallois (2013) also use lexicons and a set of rules to perform sentiment analysis of tweets. The goal, as stated by the authors, is to design it as 'fast and scalable'. LIWC provides a tool which also uses the lexicon, applies a set of rules to generate a prediction. Typically, systems that use SA as a sub-module of a larger application can benefit greatly from a lexicon and simple hand-crafted rules.

Lexicons have also been used in topic models (Lin and He 2009) to set priors on the word-topic distributions. A topic model takes as input a dataset (labelled or unlabeled) and generates clusters of words called topics, such that a word may belong to more than one topic. A topic model based on LDA (Blei et al. 2003) samples a latent variable called topic, for every word occurrence in a document. This results in two types of distributions over an unlabeled dataset: topic-document distributions (the probability of seeing this topic in this document, given the words and the topic-word assignments), and word-topic distributions (the probability of seeing this word belonging to the topic in the entire corpus, given the words and the topic-word assignments). The word-topic distribution is a multinomial with a Dirichlet prior. Sentiment lexicons have been commonly used as Dirichlet Hyperparameters for the word-topic distribution. Consider the following example. In a typical scenario, all words have symmetric priors over the topics. This means that all words are equally likely to belong to a certain topic. However, if we wish to have ‘sentiment coherence’ in topics, then, setting Dirichlet Hyperparameters appropriately can adjust priors on topic. Let us assume that we wish to have the first half of topics to represent ‘positive’ topics, and second half of topics to represent ‘negative’ topics. A ‘positive’ topic here means a topic with positive words corresponding to a concept. More complex topic models which model additional latent variables (such as sentiment or switch variables) also use lexicons to set priors (Mukherjee and Bhattacharyya 2012). Lexicons have also been used to train deep learning-based neural networks (Socher et al. 2013). A combination of datasets and lexicons has also been used. Tao et al. (2009) propose a three-pronged factorization method for sentiment classification. They factor in information from sentiment lexicons (in the form of word level polarities), unlabeled datasets (in the form of word co-occurrence) and labelled datasets (to set up the correspondences). Lexicons can also be used to determine values of frequency-based features in a statistical classification system. Kiritchenko et al. (2014) use features derived from a lexicon such as: number of tokens with non-zero sentiment, total and maximal score of sentiment, etc. This work also presents a set of ablation tests to identify value of individual sets of features. When the lexicon-based features are removed from the complete set, the maximum degradation is observed. Such lexicon-based features have been used for related tasks such as sentiment annotation complexity prediction (Joshi et al. 2014), thwarting detection (Ramteke et al. 2013) and sarcasm detection (Joshi et al. 2015).

Let us now look at how sentiment-labelled datasets can be used, especially in machine learning (ML)-based classification systems. ML-based systems model sentiment analysis as a classification problem. A classification model predicts the label of a document as one among different labels. This model is learnt using a labelled dataset as follows. A document is converted to a feature vector. The most common form of a feature vector of a document is the unigram representation with the length equal to the vocabulary size. The vocabulary is the set of unique words in the labelled dataset. A Boolean or numeric feature vector of length equal to the vocabulary size is constructed for each document where the value is set for the words present in the document. The goal of the model is to minimize error on training documents, with appropriate regularization for variance in unseen documents. The labelled documents serve as a building block for a ML-based system. While the unigram representation is common, several features such as word sense based features (Balamurali et al. 2011), qualitative features such as POS sequences (Pang et al. 2002), have been used as features for ML-based systems. The annotated datasets form the basis for creation of feature vectors with the documents acting as observed instances. Melville et al. (2009) combine knowledge from lexicons and labelled datasets in a unique manner. Sentiment lexicon forms the background knowledge about words while labelled datasets provide a domain-specific view of the task, in a typical text classification scenario.

5.7 Conclusion

This chapter described sentiment resources: specifically, sentiment lexicons and sentiment-annotated datasets. Our focus was on the philosophy and trends in the generation and use of sentiment lexicons and datasets. We described creation of several popular sentiment and emotion lexicons. We then discussed different strategies to create annotated datasets, and also presented a list of available datasets. Finally, we add two critical points in the context of sentiment resources: how a resource in one language can be mapped to another, and how these resources are actually deployed in a SA system. The diversity in goals, approaches and uses of sentiment resources highlights the value of good quality sentiment resources to sentiment analysis.

References

Balamurali, A.R., Aditya Joshi, and Pushpak Bhattacharyya. 2011. Harnessing wordnet senses for supervised sentiment classification. In *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics.

———. 2012. Cross-lingual sentiment analysis for Indian languages using linked WordNets. In: *COLING*.

Balamurali, A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. 2013. Lost in translation: Viability of machine translation for cross language sentiment analysis. In *Computational linguistics and intelligent text processing*. Berlin/Heidelberg: Springer.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Bradley, M.M., and P.J. Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1. The Center for Research in Psychophysiology, University of Florida.

Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. In: *RANLP*.

Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *The 26th International conference on computational linguistics (COLING)*, Osaka, 2666–2677.

Deng, Lingjia, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/Malefactive event and writer attitude annotation. *ACL* (2).

Duh, Kevin, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification?. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers-volume 2*. Association for Computational Linguistics.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Emotion* 6 (3–4): 169–200.

Esuli, Andrea, and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC,* vol. 6.

Finn, Arup. 2011. *AFINN*. Informatics and Mathematical Modelling, Technical University of Denmark.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford 1 (2009): 12.

Joshi, Aditya, A.R. Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: A case study. In *Proceedings of the 8th ICON*.

Joshi, Aditya, A.R. Balamurali, and Pushpak Bhattacharyya, and Rajat Mohanty. 2011. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of the 49th annual meeting of the association for computational linguistics*.

Joshi, Aditya, Abhijt Mishra, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *Conference for association of computational linguistics*.

Joshi, Aditya, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for Sarcasm detection. In *Conference for association of computational linguistics*.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50: 723–762.

Levallois, Clement. 2013. Umigon: Sentiment analysis for tweets based on lexicons and heuristics. In *Proceedings of the international workshop on semantic evaluation.* SemEval, vol. 13.

Lin, Chenghua, and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM.

Liu, Bing. 2010. Sentiment analysis and subjectivity. In *Handbook of natural language processing*, vol. 2, 627–666.

Louviere, Jordan J. 1991.*Best-worst scaling: A model for the largest difference judgments*. Technical report, University of Alberta.

Lu, Bin, et al. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics.

Mahyoub, Fawaz H.H., Muazzam A. Siddiqui, and Mohamed Y. Dahab. 2014. Building an Arabic sentiment Lexicon using semi-supervised learning. *Journal of King Saud University-Computer and Information Sciences* 26 (4): 417–424.

Makice, Kevin. 2009. *Twitter API: Up and running: Learn how to build applications with the Twitter API*. Beijing: O'Reilly Media, Inc.

McAuley, Julian, and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on recommender systems*. ACM.

Mehrabian, Albert, and James A. Russell. 1974. *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Mihalcea, Rada, Carmen Banea, and Janyce M. Wiebe. 2007. *Learning multilingual subjective language via cross-lingual projections*.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38 (11): 39–41.

Mohammad, Saif. 2012. #Emotional tweets. In *Proceedings of the first joint conference on lexical and computational semantics (*Sem)*, June 2012.

Mohammad, Saif M., and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics.

———. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29 (3): 436–465.

Mohammad, S., C. Dunne, and B. Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *EMNLP*, 599–608.

Mukherjee, S., and P. Bhattacharyya. 2012. Wikisent: Weakly supervised sentiment analysis through extractive summarization with wikipedia. In *Machine learning and knowledge discovery in databases*, 774–793. Berlin/Heidelberg: Springer.

Nakov, Preslav, et al. 2013. Semeval-2013 task 2: Sentiment analysis in twitter.

Ortony, Andrew, and Terence J. Turner. 1990. What's basic about basic emotions? *Psychological Review* 97 (3): 315–331.

Pang, Bo, and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on association for computational linguistics*. Association for Computational Linguistics.

———. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing-volume 10*. Association for Computational Linguistics.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical turk. *Judgment and Decision Making* 5 (5): 411–419.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*, vol 71. Mahwah: Lawrence Erlbaum Associates.

Plutchik, Robert. 1980. *Emotion: A psychoevolutionary synthesis*. New York: Harpercollins College Division.

———. 1982. A psychoevolutionary theory of emotions. *Social Science Information/sur les sciences sociales* 21: 529–553.

Purver, Matthew, and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics*. Association for Computational Linguistics.

Ramteke, Ankit, Pushpak Bhattacharyya, and J. Saketha Nath. 2013. Detecting Turnarounds in sentiment analysis: Thwarting. In *Conference for association of computational linguistics*.

Redondo, Jaime, et al. 2007. The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods* 39 (3): 600–605.

Rosenthal, Sara, et al. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of SemEval,* 73–80.

Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation*, SemEval.

Salameh, Mohammad, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*.

Socher, Richard, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP),* vol. 1631.

Somasundaran, Swapna, and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 1-volume 1*. Association for Computational Linguistics.

Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. 1966. *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Strapparava, Carlo, and Alessandro Valitutti. 2004. WordNet affect: An affective extension of WordNet. LREC, vol. 4.

Tao, L., Y. Zhang, and V. Sindhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP Association for Computational Linguistics.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics.

Wan, Xiaojun. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 1-volume 1*. Association for Computational Linguistics.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39 (2–3): 165–210.

Xu, Ge, Xinfan Meng, and Houfeng Wang. 2010. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics.

Chapter 6
Generative Models for Sentiment Analysis and Opinion Mining

Hongning Wang and ChengXiang Zhai

Abstract This chapter provides a survey of recent work on using generative models for sentiment analysis and opinion mining. Generative models attempt to model the joint distribution of all the relevant data with parameters that can be interpreted as reflecting latent structures or properties in the data. As a result of fitting such a model to the observed data, we can obtain an estimate of these parameters, thus "revealing" the latent structures or properties of the data to be analyzed. Such models have already been widely used for analyzing latent topics in text data. Some of the models have been extended to model both topics and sentiment of a topic, thus enabling sentiment analysis at the topic level. Moreover, new generative models have also been developed to model both opinionated text data and their companion numerical sentiment ratings, enabling deeper analysis of sentiment and opinions to not only obtain subtopic-level sentiment but also latent relative weights on different subtopics. These generative models are general and robust and require no or little human effort in model estimation. Thus they can be applied broadly to perform sentiment analysis and opinion mining on any text data in any natural language.

Keywords Generative model • Probabilistic topic model • Topic-sentiment mixture • Latent aspect rating analysis • Latent variable analysis

There are many approaches to performing sentiment analysis and opinion mining. At a high level, we can distinguish two main families of approaches. The first is rule-based approaches where human expertise is leveraged to create rules (e.g., sentiment lexicon) for determining sentiment of a text object (Ding and Liu 2007; Ding et al. 2008; Esuli and Sebastiani 2006; Taboada et al. 2011; Cambria et al. 2016). The second is statistical model based approaches, where statistical models

H. Wang (✉)
Department of Computer Science, University of Virginia, 22903, Charlottesville, VA, USA
e-mail: hw5x@virginia.edu

C.X. Zhai
Department of Computer Science, University of Illinois at Urbana-Champaign, 61801, Urbana, IL, USA
e-mail: czhai@illinois.edu

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*, Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_6

are estimated on labeled data or domain-specific priors generated by humans to essentially learn “soft” rules for sentiment prediction (Dave et al. 2003; Kim and Hovy 2004; Leskovec et al. 2010; Pang et al. 2002; Poria et al. 2015), a.k.a, learning based methods. Learning based approaches usually require labeled data for parameter estimation, while rule based approaches have less dependence on manual annotation but they also suffer from limited generalization capability. The rules can also be treated as high-level features to be used in a statistical model so as to combine the two families of approaches (Hu et al. 2013; Lu et al. 2011; Rao and Ravichandran 2009; Melville et al. 2009).

Among the statistical approaches, we may further distinguish generative models from discriminative models (Bishop 2006). Generative models focus on modeling the joint probability between class labels (e.g., sentiment labels) and data instances (e.g., text documents). Latent variables can be introduced in generative models to capture the unobservable or missing structures, e.g., latent topics (Blei et al. 2003; Blei 2012; Hofmann 1999). As a result, a generative model is a full probabilistic model of both observed and unobserved variables. In general, generative models attempt to model the joint distribution of all the relevant data with parameters that can be interpreted as reflecting latent structures or properties in the data. As a result of fitting such a model to the observed data, we can obtain an estimate of these parameters, thus “revealing” the latent structures or properties of the data to be analyzed.

In contrast, discriminative models, such as support vector machines (Hearst et al. 1998; Joachims 1998), directly model the decision boundaries, e.g., the conditional probability of class labels given data instances. Thus, a discriminative model provides a model only for the target variables conditioned on the observed variables. Flexible feature representations can be exploited in discriminative models, and empirically they often result in better classification performance than generative models (Jordan 2002). This category of statistical solutions for sentiment analysis have been well discussed in Liu’s and Pang’s survey book (Liu 2012, 2015; Pang and Lee 2008), and therefore we will not cover it in our book.

In addition to supporting sentiment classification, one major advantage of generative models over discriminative models is the ability of expressing complex relationships between the observed and target variables, even when such relationships are not directly observable. This property is of particular importance in sentiment analysis and opinion mining, when formalizing the subtle dependency between sentiment and text document content for more accurate modeling of opinions.

Promising progress in exploring generative models for sentiment analysis and opinion mining has been achieved in recent studies (Lin and He 2009; Mei et al. 2007; Titov and McDonald 2008a; Jo and Oh 2011; Wang et al. 2010, 2011; McAuley and Leskovec 2013; Moghaddam and Ester 2011). Previously, generative models have already been widely used for analyzing latent topics in text documents, e.g., topic models (Blei et al. 2003; Blei 2012; Hofmann 1999). Some of the models have been extended to model the sentiment of a topic, thus enabling sentiment analysis at the topic level (Lin and He 2009; Mei et al. 2007; Titov and McDonald 2008a; Jo and Oh 2011). Moreover, new generative models have also been developed to model both opinionated text data and their companion numerical sentiment ratings, enabling deeper analysis of sentiment and opinions to not only obtain subtopic-level sentiment but also latent relative weights on different subtopics (Wang et al. 2010, 2011; McAuley and Leskovec 2013; Moghaddam and Ester 2011). This chapter provides a survey of these recent works on using generative models for sentiment analysis and opinion mining, and discusses various applications of such models.

The rest of this chapter is organized as follows. In Sect. 6.1, we provide essential background about language models and topic models, which is the basis of the generative models that we will review in this chapter. We then present a detailed review of the major generative models for sentiment analysis in Sect. 6.2. We will discuss their applications in Sect. 6.3. To facilitate application development using such models, in Sect. 6.4, we also provide a brief review of the relevant resources on the Web.

6.1 Background: Language Models and Probabilistic Topic Models

As a background, we first introduce generative models for modeling text data, starting from the N-gram language models, proceeding to introducing the probabilistic topic models. We will introduce two most typical topic models, i.e., probabilistic latent semantic indexing model (Hofmann 1999) and latent Dirichlet allocation model (Blei et al. 2003). We will also briefly discuss the model estimation procedure for these generative models.

6.1.1 Language Models for Text

The simplest generative model for modeling text data is the N-gram language models, which were first introduced in speech recognition for distinguishing between words and phrases that sound similar (Katz 1987; Rabiner and Juang 1993) and later introduced to information retrieval for matching keyword queries with text documents (Ponte and Croft 1998; Hiemstra and Kraaij 1998; Zhai and Lafferty 2001a).

A statistical language model specifies a probability distribution over sequences of words. For example, with a language model estimated on a collection of computer science research papers, one can make statistical assertions about which text sequence is more likely to be generated by a computer scientist, e.g., P("generative models for sentiment analysis") > P("the flight to Chicago is cancelled"). Formally, a language model $P(w_1, w_2, \ldots, w_n)$ specifies the joint probability of observing the word sequence $w_1, w_2, \ldots, w_n$. Using the chain rule of probability, it can be written as,

$$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\ldots P(w_n|w_1w_2\ldots w_{n-1})$$
$$= \prod_{k=1}^{n} P(w_k|w_1, \ldots, w_{k-1}) \quad (6.1)$$

where $P(w_k|w_1, \ldots, w_{k-1})$ is a multinomial distribution over words in the vocabulary given the word sequence of $w_1, \ldots, w_{k-1}$.

The chain rule shows the link between computing the joint probability of a sequence of words and computing the conditional probability of a word given all preceding words. Intuitively, Eq. (6.1) defines the generation process of a word sequence: repeatedly select the next word with regard to all the words in front of it until meeting the predefined sequence length. For this reason, such a model is often called a generative model.

Although Eq. (6.1) suggests that one can compute the joint probability of an entire sequence of words by multiplying together a number of conditional probabilities, it does not reduce the computational complexity. The bottleneck is that we do not have any efficient way to compute the exact probability of a word given a long sequence of preceding words. For example, with a vocabulary size of V, to compute $P(w_k|w_1, \ldots, w_{k-1})$ one needs in total $(V-1)V^{k-1}$ elements in the probability table (minus one because the probabilities sum up to one). And this complexity is in the same order as that to directly compute $P(w_1, \ldots, w_k)$, which is $V^k - 1$. Since in general these probabilities must be estimated based on empirically observed data, and in practice, we almost never have so much data to observe all these different sequences, we must make simplification assumptions about the model to make it tractable and actually useful in an application.

N-gram language models provide a practical solution to this computation complexity challenge: instead of computing the probability of a word given the entire preceding sequence, we can *approximate* the preceding sequence by just a finite number of previous words, i.e., $P(w_k|w_1, \ldots, w_{k-1}) \doteq P(w_k|w_{k-N+1}, \ldots, w_{k-1})$. The assumption that the conditional probability of a word depends only on the previous N-1 words is called a Markov assumption. Unigram model is the simplest N-gram language model, in which one assumes the current word is totally independent of any other words in the sequence, i.e., $P(w_k|w_1, \ldots, w_{k-1}) = P(w_k)$; as a result,

$$P(w_1, w_2, \ldots, w_n) = \prod_{k=1}^{n} P(w_k) \quad (6.2)$$

In literature, unigram language model is also referred as bag-of-words model (Harris 1954), since the order between words is totally ignored. To capture the local dependency between words, bigram and trigram models are usually exploited.

One fundamental problem in applying the N-gram language models is to estimate the N-gram probabilities of $P(w_k|w_{k-N+1}, \ldots, w_{k-1})$. The simplest and most intuitive way for estimating such probabilities is the maximum likelihood estimation (Bishop 2006), in which one looks for the configuration of those unknown probabilities to maximize the likelihood function over a given set of training data. For the general case of maximum likelihood estimation for N-gram language models, one estimates the conditional probability as follows,

$$P(w_k|w_{k-N+1}, \ldots, w_{k-1}) = \frac{C(w_{k-N+1}, \ldots, w_{k-1}w_k)}{C(w_{k-N+1}, \ldots, w_{k-1})} \tag{6.3}$$

where $C(w_{k-N+1}, \ldots, w_{k-1})$ is the frequency of word sequence $w_{k-N+1}, \ldots, w_{k-1}$ in the training corpus.

One important concept in maximum likelihood estimation for N-gram language models is called "smoothing." Due to the sparse observations in the training data, zero probability is assigned to some word sequences, which makes any sequence containing such sequences has a zero probability in the estimated model. Various types of techniques have been developed to smooth a language model, e.g., Laplace Smoothing, Good-Turing discounting and linear interpolation. Since this topic is beyond the scope of this book, we refer the audiences to the following literature for more details (Jurafsky and Martin 2009; Chen and Goodman 1996; Zhai and Lafferty 2001b).

6.1.2 Probabilistic Topic Models

Topic models are a class of generative models for uncovering the underlying semantic structure of a document collection. The very original idea of topic modeling roots in Deerwester et al.s' seminal work in latent semantic indexing (LSI) (Deerwester et al. 1990), in which singular value decomposition is performed to discover inter- and intra-document statistical structures in a lower dimensional space. However, this approach is not a generative model, making it unclear how to interpret the latent topics discovered. A significant step forward in this direction was made by Hofmann (1999), who solved the problem of latent semantic indexing in a probabilistic fashion (the pLSI model). In pLSI, words and documents are modeled in a generative perspective: a document is modeled as a mixture of latent topics and each topic is modeled as a multinomial distribution over words. However, pLSI model is not a complete generative model, which does not specify the generation process at the document level. To address this problem, a full Bayesian probabilistic model, latent Dirichlet allocation (LDA) model (Blei et al. 2003), was introduced, in which the topic proportion in each document is assumed to be drawn from a shared Dirichlet distribution in the same corpus. LDA is an important milestone which opened up many possibilities for further development of various generative models for modeling topics. It has served as a springboard for many other topic models in analyzing different types of text data, including scientific literature (Steyvers et al. 2004; Blei and Lafferty 2007; Wang and Blei 2011), social media (Zhao et al. 2011; Hong and Davison 2010) and opinionated text reviews (Titov and McDonald 2008a; Lin and He 2009; Mei et al. 2007; Jo and Oh 2011; Wang et al. 2011).

In this section, we will briefly introduce these two basic probabilistic topic models for text modeling, i.e., pLSI and LDA. We will focus on the basic notations, generative assumptions, graphical model representation, and model estimation procedure for each model.

6.1.2.1 pLSI

Probabilistic latent semantic indexing (pLSI), also known as probabilistic latent semantic analysis (pLSA) is a generative model for document modeling. It models a text document as a mixture over a set of latent topics, and each topic is modeled as a probabilistic distribution over a fixed vocabulary. To formally describe the pLSI model, and later other more advanced topic models, we will first introduce some notations and terminologies.

Formally, a word w is the basic unit defined in a fixed size vocabulary, indexed from 1 to V. A document is a length-N sequence of words, denoted as $d = (w_1 w_2 \dots w_N)$. A corpus is a collection of M documents, denoted as $D = \{d_1, d_2, \dots, d_M\}$. In pLSI, a corpus is assumed to contain a set of k latent topics, each of which is modeled as a multinomial distribution over the vocabulary, i.e., $p(w|\beta_i)$, where β_i is the distribution parameter for topic i. Thus a document is modeled as a composition of those k topics: each word in a document is generated from a single topic indexed by z, and different words in a document may be generated from different topics.

An important assumption made in the pLSI model is that given the topic assignments $\mathbf{z} = (z_1 z_2 \dots z_N)$ for the words in a document d, the words are independent of the document index. As a result, the joint probability of document d and its words $w_1 w_2 \dots w_N$ can be computed as,

$$P(d, w_1 w_2 \dots w_N) = P(d) \prod_{i=1}^{N} \sum_{z_i} P(w_i|z_i) P(z_i|d) \tag{6.4}$$

The decomposition of joint probability of a document and its words in pLSI can be described by the following generative process:

1. For each $d \in D$, sample d by $d \sim p(d)$;
2. To generate each word $w_i \in d$,
   a. Sample topic assignment z_i by $z_i \sim p(z|d)$;
   b. Sample word w_i by $w_i \sim p(w|\beta, z_i)$;

Using the graphical model presentation, the above generation process of a text document defined by pLSI model can be illustrated in Fig. 6.1.

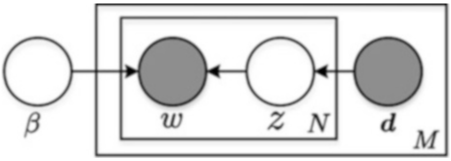


Fig. 6.1 Graphical model representation of probabilistic latent semantic indexing (pLSI) model. The plates represent replicates, where the index on the *bottom right corner* indicates the number of repetitions. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. The *circles* represent random variables, where *shaded circle* indicates observable variables and *light circle* indicates latent variables

Consider the unigram language model described in Eq. (6.2), which assumes the whole corpus only contains one topic and every word in documents is sampled from that topic. pLSI relaxes this assumption by introducing k latent topics in a given collection, and allows each document to be a mixture over those k topics. Hence, in pLSI each document is represented as a list of mixing proportions for these mixture components (i.e., $p(z|d)$) and thereby reduced to a probability distribution on a fixed set of topics. Those mixing proportions can be considered as a lower dimensional representation of a document, which can also be regarded as useful knowledge about coverage of topics in each document.

pLSI model has served as building blocks in many other generative model for text documents. Brants et al. used pLSI model to perform topic-based document segmentation (Brants et al. 2002), Mei et al. utilized it to model the facets and opinions in weblogs (Mei et al. 2007) and discover evolutionary theme patterns from text (Mei and Zhai 2005), Zhai et al. used it for cross-collection comparative text mining (Zhai et al. 2004), and Lu et al. exploited it for rated aspect summarization of short comments (Lu et al. 2009).

pLSI model has two parameters to be estimated, i.e., the word distribution under a given topic i, $p(w|\beta_i)$, and the topic proportions in a given document d, $p(z|d)$. Due to the existence of latent variables in pLSI (i.e., the topic assignments of words), maximum likelihood estimation is no longer applicable. Expectation maximization (EM) algorithm (Dempster et al. 1977) is popularly used to estimate those two parameters. Briefly, the EM algorithm approximates the lower bound of data likelihood function (i.e., $p(d, \boldsymbol{w}) = \sum_z p(d, \boldsymbol{w}, z)$) by computing the expectation of complete data likelihood over the latent variables (i.e., $E_z[p(d, \boldsymbol{w}, z)]$). Two steps are alternatively executed in EM algorithm: in E-step, the expectation of complete data likelihood over the latent variables is computed; in M-step, the optimal model parameters are found to maximize this expectation. Since a principled derivation of EM algorithm and the proof of its convergence are beyond the scope of this book, interested readers can refer to Dempster et al. (1977), McLachlan and Krishnan (2007), and Wu (1983) for more details.

The EM iterations are guaranteed to stop at a local maximum. However, there is no guarantee for an EM algorithm to find the global optimal. As a result, pLSI is prone to overfitting the data and good initialization in pLSI becomes very important.

Another source of overfitting in the pLSI model is its incomplete generative process: the document variable d is simply modeled as an index in the corpus, and there is no generative assumption about it. As a result, the number of parameters in the model grows linearly with the size of the corpus (each document has its own k-dimensional topic proportion vector), and it is not clear how to assign probability to a document outside of the training set.

To address these limitations, latent Dirichlet allocation model was introduced later to impose a full generative assumption about the document generation process. We will introduce the LDA model in the next section.

6.1.2.2 LDA and Advanced Topic Models

Latent Dirichlet allocation model (LDA), proposed by Blei et al. in (Blei et al. 2003), introduces a shared Dirichlet distribution over the topic proportions in each document to control the number of parameters in a topic model. As shown in Fig. 6.2, the topic proposition $p(z|\theta, d)$ in document d is modeled as a multinomial distribution parameterized by a k-dimensional vector θ, which is assumed to be drawn from a Dirichlet distribution with α as the concentration parameter,

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{6.5}$$

where $\Gamma(\cdot)$ is the Gamma function.

According to Fig. 6.2, the generative process of documents specified by a LDA model can be described as follows,

1. For each $d \in D$, sample θ by $\theta \sim Dir(\alpha)$;
2. For each $w_i \in d$,
   a. Sample topic assignment z_i by $z_i \sim p(z|\theta, d)$;
   b. Sample word w_i by $w_i \sim p(w|\beta, z_i)$;

The corresponding joint probability of words $\boldsymbol{w}$, latent topic assignments $\boldsymbol{z}$, and latent topic proportion θ in document d specified by a LDA model can be computed as,

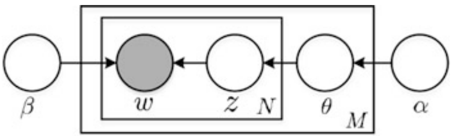


Fig. 6.2 Graphical model representation of latent Dirichlet allocation (LDA) model. α and β are corpus-level parameters for the distribution of topic proportion in documents and word distribution under topics (Blei et al. 2003)

$$p(\boldsymbol{w}, \boldsymbol{z}, \theta|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(w_n|\beta, z_n)p(z_n|\theta) \tag{6.6}$$

LDA model postulates a two-layer hierarchal Bayesian assumption in the document generation process: the topic proposition θ is drawn from a Dirichlet distribution, and the specific topic assignment of each word is drawn from a multinomial distribution specified by θ. The conjugacy between Dirichlet distribution and multinomial distribution provides additional computational advantage, which facilitates posterior inference. Compared to the pLSI model, the topic proposition θ is now modeled as a latent variable, rather than a model parameter. It thus makes the number of parameters in LDA model independent from the training corpus, and provides a principled way to estimate the topic proposition in unseen test documents, i.e., via statistical posterior inference.

Many extensions of LDA have been made. Blei and Lafferty replaced the Dirichlet prior for the topic proportion in documents with a log-normal distribution to model the covariance of topics in a corpus (Blei and Lafferty 2007). Temporal dynamics of word distribution under topics in a given corpus are modeled in Blei and Lafferty (2006). Both continuous supervision (Mcauliffe and Blei 2008), e.g., opinion ratings, and discrete supervision (Zhu et al. 2009; Ramage et al. 2009), e.g., sentiment class, are introduced into LDA. Teh et al. introduced another layer of Bayesian hierarchy over the generation of Dirichlet parameter α (Teh et al. 2006), such that the clustering property of documents can be captured.

Because of the coupling between the continuous variable θ and discrete variable $\boldsymbol{z}$ in a document, the posterior inference in LDA model becomes more challenging than that in pLSI model. Two most popularly used inference methods are Gibbs sampling (Griffiths and Steyvers 2004) and variational inference (Blei et al. 2003). Both inference methods take advantage of the conjugacy between Dirichlet distribution and multinomial distribution to facilitate the computation, e.g., θ can be integrated out in Gibbs sampling and a closed form solution for θ exists in variational inference. Further details about those two inference procedures can be found in Andrieu et al. (2003) and Wainwright and Jordan (2008). Parallel implementation of LDA model for large-scale document collection can be found in Smola and Narayanamurthy (2010), Andrieu et al. (2003), Zhai et al. (2012), and Wang et al. (2009). And the parameter estimation in a LDA model can also be achieved via EM algorithms (Blei et al. 2003).

6.2 Generative Models for Sentiment Analysis

With the basic concepts about generative modeling of text documents introduced in the previous section, we are now ready to discuss how to utilize the generative models for sentiment analysis. Before diving into the details of specific models, we will first define some categorizations of generative models for sentiment analysis

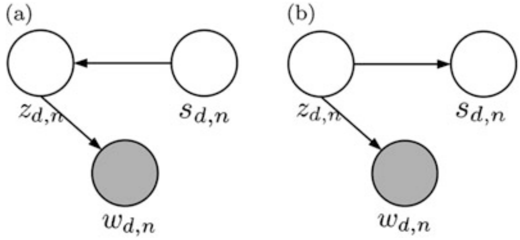


Fig. 6.3 Basic categorization of generative models for sentiment analysis (Mimno and McCallum 2008). (**a**) Upstream model. (**b**) Downstream model

to facilitate our later discussions. According to the notion proposed in Mimno and McCallum's work (Mimno and McCallum 2008), we can categorize most of existing generative models for sentiment analysis as upstream models and downstream models, according to their particular dependency assumption among the sentiment label s, topic assignment z and observed word w in a given document. Using the language of graphical models, we can illustrate these two classes of generative models for sentiment analysis in Fig. 6.3.

Upstream models assume that in order to generate a word $w_{d,n}$ in a given document d, one needs to first decide the sentiment polarity $s_{d,n}$ of this word, and $s_{d,n}$ then determines the topic assignment $z_{d,n}$ for this word. Upstream models usually model sentiment as discrete labels and assume there are different topic proportions under different sentiment labels. In contrast, downstream models assume the sentiment label $s_{d,n}$ is determined by the topic assignment $z_{d,n}$, in parallel to the word $w_{d,n}$. Therefore, downstream models are more flexible in modeling the sentiment, e.g., continuous ratings can also be modeled (Mcauliffe and Blei 2008; Wang et al. 2011). The key difference between the two kinds of models lies in the way we specify the dependency.

Intuitively, in the upstream models, topics and words are potentially dependent on the sentiment variable, thus it can be regarded as in the "up stream" with its influence on other variables directly captured in the model. In the downstream model, the sentiment variable is assumed to depend on topics, thus the sentiment variable can be regarded as in the "down stream", and the model attempts to capture how other variables (mostly topics) influence the sentiment variable. Since we treat sentiment as a response variable of topic variable, it opens up many different ways to model sentiment, and can easily model numerical ratings, which would be hard to model with an upstream model.

One thing we need to emphasize about the graphical representation illustrated in Fig. 6.3 is that we do not explicitly distinguish the scope of sentiment label s, e.g., a document-level label v.s., a word-level variable. In some existing models, s is considered as a document-level variable, such that all $s_{d,n}$ is forced to share the same value (Mcauliffe and Blei 2008; Wang et al. 2011); while some models treat s as a word-level or sentence-level variable, so that different words or sentences in the same document might be associated with different sentiment (Jo and Oh 2011; Lin and He 2009; Mei et al. 2007). Another factor not specified in Fig. 6.3 is whether $s_{d,n}$ is observable or latent. In most of downstream models, $s_{d,n}$ is considered as an observable random variable, e.g., sentiment class label for the documents (Mcauliffe and Blei 2008). Some upstream models treat $s_{d,n}$ as latent variables and sentiment prior is introduced to guide the corresponding model learning process, e.g., in Mei et al. (2007), Lin and He (2009), and Jo and Oh (2011); while some consider it as document-level observable variables (Ramage et al. 2009, 2011).

Following this categorization, we will introduce the basic modeling assumptions, model specifications and interesting findings and results from upstream and downstream models for sentiment analysis in the following sections.

6.2.1 Upstream Models for Sentiment Analysis

Upstream models assume that to generate a word in a text document, one needs to first sample a latent sentiment label, then sample a topic label with respect to this sentiment category, and finally sample the word from this chosen topic. One typical upstream generative model for sentiment analysis is the Topic-Sentiment Mixture model (TSM) proposed in Mei et al. (2007). TSM is constructed based on the pLSI model: in addition to assuming a corpus consists of k topics with neutral sentiment, TSM introduces two additional sentiment models, one for positive and one for negative opinions. In TSM, the sentiment models are assumed to be orthogonal to topic models in the sense that they would assign high probabilities to general words that are frequently used to express sentiment polarities whereas topical models would assign high probabilities to words representing topical contents with neutral opinions. For example, for a collection of MP3 player reviews, the words "nano," "price" and "mini" are supposed to be observed more often in the neutral topic models, "awesome," "love" are more likely to be found in positive sentiment models, and "hate," "bad" are more likely to be found in negative sentiment models. A new concept called "theme" is then introduced in TSM and it is modeled as a compound of these three components: neutral words, positive words and negative words, in each document. The combination of topic models and sentiment models creates a theme about a particular aspect with certain sentiment polarity in a given document. And such combination varies across different documents to reflect users' distinct sentiment polarities toward the same aspect. Once the themes are determined, a document is modeled as a mixture over the themes, and the rest generation process follows what in the pLSI model.

We followed the representation used in Mei et al. (2007) to depict the TSM model in Fig. 6.4. We should note this representation does not follow the conventional graphical model representation of probabilistic models. According to the figure, the generation of words from the document-specific themes follows the same assumption as that in a pLSI model. The themes in a particular document are modeled as another mixture over the corpus-level neutral, positive and negative topics. As a result, a TSM model can be considered as a three-layer Bayesian model of documents.

Since TSM model is based on the pLSI model, EM algorithm with a closed form posterior inference is possible. TSM is unsupervised and it does not directly

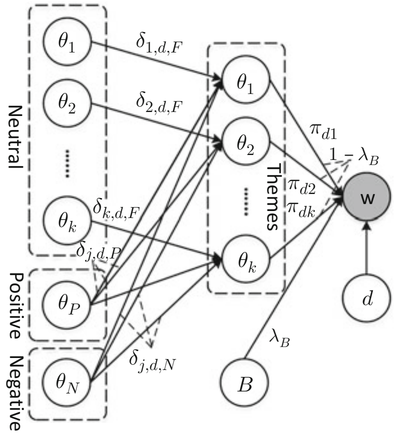


Fig. 6.4 Illustration of Topic-Sentiment Mixture model. $\{\theta_1, \theta_2, \ldots, \theta_k\}$, θ_P and θ_N labeled with "Neutral," "Positive" and "Negative" in the *dash round box* denote the neutral, positive and negative topics in the corpus accordingly. $\{\theta_1, \theta_2, \ldots, \theta_k\}$ located in the *dash round box* labeled with "theme" denote the themes of a particular document. A theme is modeled as a mixture over the latent neutral, positive and negative topics; and the mixing weights are denoted as $\{\delta_{i,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$ for each specific theme i. B represents the background topic model, and words in a given document are sampled from a mixture of the themes and background topic (Mei et al. 2007)

model sentiment labels. In TSM, sentiment prior extracted from external corpus was introduced to the EM algorithm to guide the parameter estimation of sentiment models. Thus a collection of text data with sentiment labels is needed to induce priors for effective separation of positive and negative topics, but the sample text data does not have to be related to the opinionated text data to be analyzed. With the learned topic models and sentiment models in TSM, topic life cycles and sentiment dynamics can be extracted from text documents. These mining results provide unique insights about the latent sentiment conveyed in unstructured text data.

Because TSM model is based on the pLSI model, it also suffers from its limitations, e.g., overfitting and can hardly generalize to unseen documents. Several follow-up work tries to address the limitations with LDA's modeling assumptions.

In (Lin and He 2009), Lin and He proposed a joint sentiment and topic (JST) model for sentiment analysis. In JST model, a corpus is assumed to contain $S \times k$ topics, where S is the number of sentiment categories, e.g., positive, negative and neutral. As a result, in JST the combination of topics and sentiments is modeled as a Cartesian product between topic models and sentiment models, similarly to the linear interpolation combination assumed in the TSM model.

As an upstream model, JST model first samples a sentiment label and then samples topic assignment and the word from corresponding distributions. To generate a document with the JST model, one needs to first sample a sentiment mixture for that document from a shared Dirichlet distribution; and under each sentiment category, sample a topic mixing proportion from another corpus-level Dirichlet distribution.

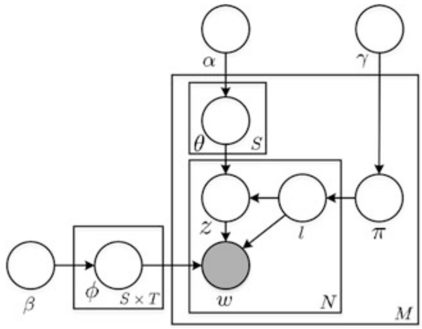


Fig. 6.5 Graphical model representation of Joint Sentiment and Topic (JST) model. ϕ is a S-by-T matrix controlling the word distribution under each sentiment-topic combination. π is the sentiment mixture proportion in a given document, and it is assumed to be drawn from a Dirichlet distribution with parameter γ. l is a specific sentiment assignment for word w, and it also controls the topic assignment z of this word. θ is S k-dimensional vectors, which denote the topic proportion under each sentiment class in this document (Lin and He 2009)

Specifically, the topic proportion in each document is modeled as S k-dimensional vectors, which allow different topic mixtures under different sentiment categories. Gibbs sampling is used to perform the posterior inference of latent variables in JST, e.g., latent topic assignments, sentiment and topic mixture. The graphical model representation of JST model is illustrated in Fig. 6.5.

Given JST model is also an unsupervised model, sentiment prior is vital for it. Sentiment seed words are injected as the prior for the word distribution under different topics in JST. The authors reported that without sentiment prior, JST's performance in sentiment categorization is close to random (Lin and He 2009).

Jo and Oh's Aspect and Sentiment Unification Model (ASUM) employs the same generative assumption as that in JST model. But to enforce the topic and sentiment coherence inside a document, they further assumed all the words in one sentence share the same topic and sentiment assignment. The same posterior inference procedure as that in JST model is applied in ASUM, which takes sentence as the basic unit for inference. Because ASUM is based on the same generation assumption as that JST, it also heavily depends on sentiment seed words to differentiate different types of sentiments.

A different variant of upstream generative model for sentiment analysis is proposed in Zhao et al.'s work in Zhao et al. (2010). In particular, a Maximum Entropy (ME) model is introduced into LDA model to control the selection of words from background topic, aspect-specific topics and opinion-specific topics. In the proposed ME-LDA model, a given word can be generated from five different types of topics: background topic, general aspect topic, aspect-specific topics, general opinion topics and aspect-specific topics. And a particular word's assignment to those five topics is controlled by a Maximum Entropy model based on discriminative features extracted from previous, current and next words' POS tags, and word content. The authors used a set of training sentences with labeled background, aspect and opinion words to estimate the ME model beforehand. With this pretrained ME model on a separately labeled corpus, ME-LDA should really be regarded as a hybrid of generative and discriminative model.

The generative topic models have been used as building blocks in many other sentiment analysis tasks. Lu et al., used pLSI model to integrate opinions expressed in a well-written expert review with lots of opinions scattering in various sources such as blogspaces and forums (Lu and Zhai 2008). Sentiment prior is given to the pLSI model to identify sentiment-oriented aspects from expert reviews. Such sentiment-oriented aspects are then used to retrieve the most relevant sentences from various sources of opinionated text data. Later on, they used topics learned from pLSI models as lower dimensional representation of documents for clustering (Lu et al. 2009). In each aspect-specific document clusters, the overall sentiment rating is aggregated to predict aspect-level opinions.

From the discussion above, we can observe that most of the typical upstream generative models for sentiment analysis treat sentiment label as latent variable over each word, and sentiment prior is used to inject sentiment polarity into the models. Although such a modeling approach provides flexibility of identifying distinct opinions on individual words, strong knowledge about sentiment is required to ensure satisfactory analysis results. As an alternative solution, Ramage et al.'s Labeled-LDA model provides a different perspective of modeling sentiment with topics in an upstream model (Ramage et al. 2009). Specifically, in Labeled-LDA model, sentiment can be modeled as document-level variables, which is directly observable. And the choice of document sentiment labels affects the topic mixing proportion in this document. Later on, partially Labeled-LDA model was developed to handle the situation, in which some of labels are not directly observable in a document (Ramage et al. 2011).

6.2.2 Downstream Models for Sentiment Analysis

Downstream models reverse the generation assumption between the sentiment labels and latent topic assignments: to generate a text document, one needs to first select the topic assignments in this document, and sample the words and sentiment labels with respect to those topics. One typical downstream generative model for sentiment analysis is Blei and McAuliffe's supervised LDA (sLDA) model (Mcauliffe and Blei 2008). The graphical model representation of sLDA model is illustrated in Fig. 6.6.

The assumed generation process of text content in sLDA model is identical to that assumed in LDA model. In addition to document generation, sLDA assumes the document-level response variable y is drawn from a Gaussian distribution with mean $\eta^{\mathsf{T}}\bar{z}$ and standard deviation σ, in which $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$, i.e., the mean vector of topic assignments in document d. With this continuous assumption about the response variable y, sLDA can be used as a regression model to model the opinion ratings in text documents. The generation of y can be further modeled with a generalized

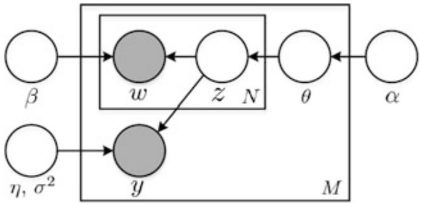


Fig. 6.6 Graphical model representation of supervised Latent Dirichlet Allocation (sLDA) model. *y* is the response variable observed in document *d* (Mcauliffe and Blei 2008)

linear model, e.g., a logistic model, to model discrete sentiment classes. Variational inference similar to that used in LDA model can be applied in sLDA model for posterior inference. Later on, Zhu et al. introduced the idea of maximum margin training in sLDA model for better predictive performance (Zhu et al. 2009). Blei and Wang extended sLDA to a collaborative setting (Wang and Blei 2011), where collaborative filtering based on users' opinion ratings can be achieved in the latent topic space.

Boyd-Graber and Resnik further generalized sLDA model to perform holistic sentiment analysis across languages (Boyd-Graber and Resnik 2010). In their proposed MLSLDA model, topics organized according to some shared semantic structure that can be represented as a tree, and the sentiment label in a given document is modeled as a regression response variable with respect to the topic assignments. As a result, MLSLDA simultaneously identifies how multilingual concepts are clustered into thematically coherent topics and how topics associated with text connect to the sentiment ratings.

In (Lin et al. 2012), Lin and He performed an interesting reparameterization of JST to turn their original upstream JST model into a new downstream joint sentiment-topic model, named Reverse-JST. In Reverse-JST, it is assumed that to generate the word sequence in a given document, one needs to first sample topic assignment, then sample sentiment category with respect to the selected topic, and select a word under this topic sentiment combination. Without the sentiment seed words being specified, the JST model and Reverse-JST model are essentially the same, since both of them model the combination of topics and sentiments with Cartesian product. The authors' empirical evaluation indicates JST performs consistently better than Reverse-JST when sentiment seed words are available.

One important line of research in downstream generative models for sentiment analysis focuses on aspect-level understanding of opinions. Those aspect ratings can be understood as users' sentiment polarities over the latent topics in a given document. This line of research exploits and analyzes user-generated opinionated text content at the detailed topical aspect level and enables a deeper and more detailed understanding of user opinions.

Titov and McDonald developed a LDA-based generative model called Multi-Aspect Sentiment (MAS) model for joint modeling of text content and aspect ratings for sentiment summarization (Titov and McDonald 2008b). In their solution, two types of topics, i.e., global and local topics, are explicitly modeled; and each fraction inside a document (modeled as a moving window of sequential words in the document) is assumed to be a mixture over those global and local topics. Based on the latent topic assignments, aspect ratings are assumed to be determined by a logistic regression model, which takes the topic assignments and the word sequence in that window as input. Comparing to sLDA model, which only captures the document-level sentiment, MAS enables the understanding of sentiment at finer granularity, in which the detailed prediction of aspect-level opinions is possible.

However, in MAS the aspect-level sentiment labels are assumed to be known to the model during the training phase. This limits the application of this type aspect-level sentiment analysis, when such detailed annotations are not available. Wang et al.'s work in latent aspect rating analysis (LARA) (Wang et al. 2010, 2011) alleviates the dependency on the fully annotated data and enables in-depth understanding of users' opinions at the aspect-level. In the LARA model, the overall rating is assumed to be observable in a given document and it provides guidance for estimation of corresponding latent aspect ratings. Moreover, in addition to analyzing opinions expressed in text document at the level of topical aspects to discover each individual user's latent opinion on each aspect, the LARA model also identifies the relative preference users have placed onto those different aspects when forming the overall judgment.

A two-stage approach based on bootstrapping aspect segmentation and latent rating regression model was first proposed to solve the problem of LARA in Wang et al. (2010). This solutions assumes that a set of predefined keywords specifying the latent topical aspects are available. The overall sentiment rating in a document is assumed to be drawn from a mixture of the latent aspect ratings. Via posterior inference, the overall rating can be decomposed into aspect ratings, the inferred mixing weights reflect users' preference over those latent aspects.

However, this two-step solution is not a fully generative model, because it does not specify the generation of text content in a document. Later on, a unified solution based on LDA model is introduced to jointly identify the latent topical aspects, and infer the latent aspect weights/ratings from each user's opinionated review article (Wang et al. 2011). As shown in Fig. 6.7, in the unified LARA model, each latent aspect rating in a given document is assumed to be drawn from a Gaussian

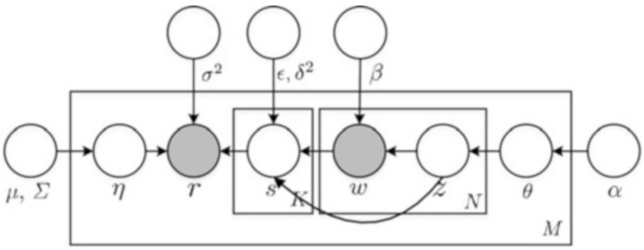


Fig. 6.7 Latent aspect rating analysis (LARA) model. s is a K-dimensional vector indicating the aspect-level latent opinion ratings. r denotes observable document-level opinion rating. Specifically, the LARA model assumes the overall rating r is determined by the weighted average of aspect ratings, i.e., $r \sim N(\eta^\mathsf{T} s, \sigma^2)$ (Wang et al. 2011)

distribution with mean determined by the linear combination of words assigned to that aspect, e.g., $s_i \sim N(\sum_{n=1}^{N} w_n \epsilon_{ij} \Delta[w_n = v_j, z_n = i], \delta^2)$. Intuitively, the latent topic assignments $\boldsymbol{z}$ segment the text content into different aspects, and the observed words in each aspect segment contribute to the sentiment polarity of corresponding aspect rating. Then the observable overall rating is assumed to be drawn from another linear combination of these latent aspect ratings, i.e., $r \sim N(\eta^{\mathsf{T}} s, \sigma^2)$. Variational inference is used to infer the latent topic assignments, aspect ratings and weights in a given document simultaneously.

Clearly distinct from all previous work in opinion analysis that mostly focuses on integrated entity-level opinions, LARA reveals individual users' latent sentiment preferences at the level of topical aspects in an unsupervised manner. Discovering such detailed user preferences (which are often hard to obtain by a human from simply reading many reviews) enables many important applications. First, such analysis facilitates in-depth understanding of user intents. For example, by mining the product reviews, LARA recognizes which aspect influences a particular user's purchase decision the most. Second, by identifying each user's latent aspect preference in a particular domain (e.g., hotel booking), personalized result ranking and recommendation can be achieved. Third, discovering the general population's sentiment preferences over different aspects of a particular product or service provides a more effective way for businesses to manage their customer relationship and conduct market research.

Follow up work extended LARA model in different directions. Diao et al. introduced collaborative filtering into LARA modeling to uniformly model different users' rating preferences in a generative manner (Diao et al. 2014). Wu and Ester also combined the LARA model with collaborative filtering method to predict the latent aspect ratings even when the users have not generated the review content (Wu and Ester 2015). Both of these two models enable aspect-based recommendation.

6.3 Applications of Generative Models for Sentiment Analysis

In the above discussions, we have summarized the most representative works in modeling opinionated text documents with generative models. In this section, we review the landscape of application opportunities of such models.

6.3.1 Sentiment Lexicon Construction

A sentiment lexicon can be directly used for sentiment tagging or suggesting useful features for supervised learning approaches to sentiment analysis. One major challenge in constructing a sentiment lexicon is that the polarity of a word such as "long" highly depends on the context; for example, "long battery life" is positive, while "long rebooting time" is negative in the same review of a laptop. Thus a lexicon must incorporate context when specifying the polarity of a word.

A generative model can capture context by using appropriate latent variables, and thus be useful for constructing a topic-specific sentiment lexicon. The sentiment polarity of a word can be modeled in two different ways in a generative model. In the first, we may explicitly have a positive or negative topic represented as a word distribution. In such a case, the probability of a word can be regarded as an indicator of polarity, thus a word with very high probability according to a positive model would be tagged as a positive word and the probability can be used as a measure of confidence which may be useful to include in the lexicon. In the second, the sentiment of a term is modeled with a real number, which can be positive or negative, depending on the sentiment of the word. In such a case, a high positive weight would indicate a very positive word (for the corresponding topic).

One example of work in the first category is the topic-sentiment mixture model (Mei et al. 2007). In this work, the authors demonstrated a list of positive and negative words specific to the topics of "movies" and "cities": "beautiful," "love" and "awesome" are automatically identified as positive for "cities" while "hate," "traffic" and "stink" are identified as negative for this topic. The authors in Lin and He (2009) also reported a similar list of learned sentiment lexicon from JST model on a movie review data set. However, as we discussed before, upper stream models depend on sentiment priors to determine the sentiment polarity of learned topics. The bias in those sentiment seed words determine the qualify of learned sentiment lexicon.

Another example of the first category is the downstream model sLDA (Mcauliffe and Blei 2008). In general, the downstream models can resolve the dependency on sentiment prior by directly learning from the given sentiment labels. In (Mcauliffe and Blei 2008), the authors applied sLDA on a set of labeled movie reviews, where the learned topics are directly aligned with numerical sentiment polarities, e.g., a topic represented by the words of "least," "problem" and "unfortunately" is strongly correlated with negative opinion while the topic represented by the words of "motion," "simple" and "perfect" is strongly correlated with negative opinion.

An example of the second category is the LARA model (Wang et al. 2010), which is also a downstream model, but in contrast with sLDA, LARA uses numerical weights to model the sentiment of a word, and thus can learn a topic-specific lexicon in the form of positive and negative weights for words. Table 6.1 illustrates an sample output from the LARA model (Wang et al. 2010), where the aspect specific word sentiment polarity was learned from a collection of hotel reviews.

As shown in the table, words "linen", "walk" and "beach" do not have opinion annotations in general sentiment lexicons, e.g., SentiWordNet (Esuli and Sebastiani 2006), since they are nouns, while the LARA model automatically assigns them positive sentiment likely because "linen" may suggest the "cleanliness" condition is good and "walk" and "beach" might imply the location of a hotel is convenient.

In general, one can potentially design a generative model to embed a particular perspective of topical context as needed for an application to automatically construct

Table 6.1 Estimated word sentiment polarities under different aspects. The numbers to the right of listed words indicate their learned sentiment weight from a LARA model (Wang et al. 2010)

Value	Rooms	Location	Cleanliness
Resort 22.80	View 28.05	Restaurant 24.47	Clean 55.35
Value 19.64	Comfortable 23.15	Walk 18.89	Smell 14.38
Excellent 19.54	Modern 15.82	Bus 14.32	Linen 14.25
Worth 19.20	Quiet 15.37	Beach 14.11	Maintain 13.51
Quality 18.60	Spacious 14.25	Perfect 13.63	Spotlessly 8.95
Bad 24.09	Carpet 9.88	Wall 11.70	Smelly 0.53
Money 11.02	Smell 8.83	Bad 5.40	Urine 0.43
Terrible 10.01	Dirty 7.85	MRT 4.83	Filthy 0.42
Overprice 9.06	Stain 5.85	Road 2.90	Dingy 0.38
Cheap 7.31	Ok 5.46	Website 1.67	Damp 0.30

a topic-specific lexicon that would capture the desired dependency of sentiment on context. Such a lexicon may itself be used directly as knowledge about people's opinions about a topic, thus facilitating comparative analysis of opinions across opinion holders or other interesting context variables.

6.3.2 Sentiment Annotation and Pattern Discovery

Another direct application of the generative models for sentiment analysis is sentiment annotation and pattern discovery. Sentiment annotation is to tag a text object with sentiment labels which can be categorical (e.g., positive vs. negative vs. neutral) or numerical (i.e., ratings). Once tagging is done, we can easily examine patterns of opinions by associating sentiment labels with context variables such as time, location, and sources of opinions to reveal patterns of opinions such as spatiotemporal trends of opinions.

In (Lin and He 2009), the JST model is reported to achieve comparable performance as supervised statistical algorithms in binary sentiment classification. And sLDA is reported to have better predictive power than the supervised lasso least-square regression model trained on LDA model's topic output (Mcauliffe and Blei 2008). With maximum margin estimation method, further improved classification performance is achieved in MedLDA model (Zhu et al. 2009). The aspect-level sentiment model, e.g., MAS (Titov and McDonald 2008b) and LARA (Wang et al. 2010, 2011), can also predict aspect-level sentiment ratings, which might be unobservable during the training process, thus enabling discovery of latent patterns of opinions at the level of subtopics.

Based on the identified sentiment polarity from text content, temporal dynamics of opinions in user-generated content is studied in TSM model (Mei et al. 2007). A hidden Markov model is built based on the TSM model's identified neutral, positive and negative opinions over time to capture the topic life cycles and sentiment dynamics. Similar idea has been explored in Si et al. (2013) to leverage topic based sentiments from Twitter to help predict the stock market. A continuous Dirichlet Process Mixture model is developed to estimate the daily topic set, which is mapped to a sentiment time series according to predefined sentiment lexicon. A regression model is build to predict the stock index with respect to this Twitter sentiment time series.

6.3.3 Topic-Specific Sentiment Summarization

Yet another interesting application of the generative sentiment analysis models is to generate topic-specific sentiment summaries. Summarization of opinions facilitates digestion of opinions by users and also provides entry points for a user to navigate into detailed information about a specific aspect of opinion. In (Jo and Oh 2011), review text content can be summarized according to its topic and sentiment. Table 6.2 illustrated the aspect-specific sentiment summarization reported in Wang et al. (2010). Such detailed aspect-level sentiment analysis and summarization provide flexibility for ordinal users to navigate through the opinionated text corpus.

6.3.4 Deep Analysis of Latent Preferences of Opinion Holders

An important application enabled by generative models is deep analysis of *latent* preferences of opinion holders. While the applications discussed above can all

Table 6.2 Aspect-based comparative summarization (Hotel Max in Seattle as an example) (Wang et al. 2010)

Aspect	Summary	Rating
Value	Truly unique character and a great location at a reasonable price Hotel Max was an excellent choice for our recent three night stay in Seattle	3.1
	Overall not a negative experience, however considering that the hotel industry is very much in the impressing business there was a lot of room for improvement	1.7
Room	We chose this hotel because there was a Travelzoo deal where the Queen of Art room was $139.00/night	3.7
	Heating system is a window AC unit that has to be shut off at night or guests will roast	1.2
Location	The location ‘a short walk to downtown and Pike Place market’ made the hotel a good choice	3.5
	When you visit a big metropolitan city, be prepared to hear a little traffic outside!	2.1

Table 6.3 User rating behavior analysis (Wang et al. 2010)

	Expensive hotel		Cheap hotel	
Aspect	5 Stars	3 Stars	5 Stars	1 Star
Value	0.134	0.148	**0.171**	0.093
Room	0.098	0.162	0.126	0.121
Location	0.171	0.074	0.161	0.082
Cleanliness	0.081	**0.163**	0.116	**0.294**
Service	**0.251**	0.101	0.101	0.049

be potentially supported by other approaches to sentiment analysis, the deep analysis of latent preferences of opinion holders cannot be easily supported by other approaches, and thus represents a unique advantage of generative models for sentiment analysis. This unique benefit comes from the explicit use of meaningful latent variables in a generative model to model and capture the latent information about an opinion holder.

For example, the aspect-level sentiment analysis enabled by LARA model enables the in-depth understanding of users' sentiment preference in their decision making process. In (Wang et al. 2010), the authors demonstrated the learned aspect weights in a hotel data set (see in Table 6.3), and such latent weights unveil reviewers' detailed sentiments preference over those aspects.

It is interesting to note that according to the learned aspect preference weights in Table 6.3, reviewers give the "expensive hotels" high ratings mainly due to their nice services and locations, while they give low ratings to such hotels because of undesirable room condition and overprice. In contrast, reviewers give the "cheap" hotels high ratings mostly because of the good price/value and good location, while giving low ratings for its poor cleanliness condition. Such analysis can be performed for different groups of hotels, or different groups of consumers, or different time periods, etc, thus enabling potentially many interesting applications. Note that such a deep understanding of reviewers cannot be easily achieved by other approaches to sentiment analysis; indeed, it cannot even be easily achieved by humans even if they read all the reviews, thus representing an important benefit of using generative models for sentiment analysis.

Such a deep understanding of latent preferences would further enable many applications, particularly those requiring better understanding people's behavior and preferences and finding groups of people with shared preferences. Examples include market research where we want to understand consumer's preferences, business intelligence where we want to understand the relative strength and weakness of a product with respective to another product for a particular group of consumers, and targeted advertising where the goal is to discover groups of consumers that may potentially find a product appealing.

6.3.5 Entity Ranking and Recommendation

Generative models enable detailed understanding of opinions about entities such as products as well as detailed understanding of preferences of people such as reviewers. Thus they can be used to generate more informative representations for both entities and users, which further helps improving the ranking and recommendation of entities for users.

For example, based on the identified aspect preferences, collaborative filtering can be performed. In (Wang and Blei 2011), scientific article recommendation is performed based on the learned latent topics in each individual user from their rating history. Comparing to the tradition collaborative filtering solutions, which can only provide item-level recommendations, the collaborative topic model enables topic-specific recommendations. Diao et al.'s JMARS model identifies users' aspect-level sentiment preference and the content distribution in their generated review content (Diao et al. 2014). Improved recommendation performance is reported comparing to traditional collaborative filtering solutions.

In LARA (Wang et al. 2010), the inferred reviewer preferences can be leveraged to support personalized entity recommendation. Specifically, a user can specify his or her preferences (e.g., price is much more important than service or location), and the system can selectively use only those reviewers that are written by reviewers with similar preferences to recommend hotels, instead of using the generic set of all reviewers, making the recommendation more accurately reflect the specific preferences of this particular group of users. Such a personalized recommendation is only possible because of the inferred latent preference information, which enabled us to know which reviewers have put more weight on price than on location and service.

6.3.6 Social Network and Social Media Analysis

The generative model based solutions for sentiment analysis have also been explored in the context of social networks. Liu et al. explore topic modeling technique to study topic-level influence in heterogeneous networks (Liu et al. 2010). Rao et al. developed a supervised topic model to analyze emotion based on social media content (Rao et al. 2014). Xu et al. developed a pLSI-based generative model to analyze users' posting behaviors on Twitter: via generative modeling, the motivation of a user's posting behavior is decomposed into the factors of breaking news, posts from social friends and user's intrinsic interest.

6.4 Resources on the Web

Most of aforementioned generative models for sentiment analysis have open implementations online and there are also publicly available sentiment data sets on the Web. In this section, we will briefly summarize some resources for this line of research.

David M. Blei maintains a page for topic modeling, where implementations of many LDA-based generative models (e.g., the LDA (Blei et al. 2003) and sLDA (Mcauliffe and Blei 2008) models) are provided: http://www.cs.princeton.edu/~blei/topicmodeling.html. The Stanford Natural Language Processing group provides a Topic Modeling Toolbox, which can easily import and manipulate text from cells in Excel and other spreadsheets. This toolbox focuses on helping social scientists and others who wish to perform analysis on datasets that have a substantial textual component. Implementations of LDA and Labeled-LDA (Ramage et al. 2009) models are provided in this toolbox. Andrew McCallum and David Mimno developed a Java-based package for statistical text document modeling named MALLET (McCallum 2002), which provides implementations of several aforementioned topic models, e.g., LDA model. Besides those generic implementation of standard topic models, there are also implementations of those specific generative models for sentiment analysis introduced above. The authors of JST model (Lin and He 2009) provide their implementation on GitHub at: https://github.com/linron84/JST. And the authors of LARA model (Wang et al. 2010) provide their implementation of two-step solution at: http://www.cs.virginia.edu/~hw5x/Codes/LARA.zip.

Besides those open implementation of generative models, there are also public sentiment data sets available on the Web. The Stanford Network Analysis Project provides a large collection of Amazon reviews, spanning a period of 18 years, including around 35 million reviews up to March 2013. The data can be found at http://snap.stanford.edu/data/web-Amazon.html. The authors of book "Sentiment Analysis and Opinion Mining" (Liu 2012) also provide a large collection of amazon reviews at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html, where additional sentence-level positive and negative annotations are possible in a subset of reviews. Yelp.com hosts an annual "Yelp Dataset Challenge," which provides more than 1.6 million Yelp reviews from more than 366k users. Besides the text content and opinion ratings, this Yelp data set also includes the social connections among those reviewers. In addition to those user review data sets, twitter data sets with sentiment annotations are also available. Go et al. manually created a collection of 40,216 tweets with polarity sentiment labels (Go et al. 2009). This data set can be found at http://help.sentiment140.com/for-students. Shamma et al. used Amazon Mechanical Turk to annotate sentiment polarities in 3,269 tweets posted during the presidential debate on September 26, 2008 between Barack Obama and John McCain (Shamma et al. 2009). The data set can be found at https://bitbucket.org/speriosu/updown/src/5de483437466/data/. Saif et al. provided a survey of datasets for twitter sentiment analysis (Saif et al. 2013).

6.5 Summary

In this chapter, we provide an introduction and systematic review of generative models for sentiment analysis, which represent an important family of (mostly unsupervised) approaches to sentiment analysis that can be potentially applied to any opinionated text data due to their generality and robustness. They are especially powerful in inferring latent variables about opinion holders or detailed opinions about specific subtopics and can very effectively perform joint analysis of both opinionated text data and the companion numerical ratings. Besides supporting common applications of sentiment analysis such as sentiment classification, sentiment lexicon construction, and sentiment summarization, they also enable many other interesting new applications such as topic-specific lexicon construction, detailed opinion pattern discovery in association with context variables such as time, location, and sources, personalized entity ranking and recommendation, and deep analysis of latent preferences of opinion holders. When using appropriate latent variables, such generative models can discover latent opinion patterns from large amounts of data that are hard to discovery by humans even if they have time to read all the opinionated text data, thus are essential tools for building intelligent systems for opinion understanding and its related applications, as well as for research in computational social science.

References

Andrieu, C., N. De Freitas, A. Doucet, and M.I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine Learning* 50(1–2): 5–43.

Bishop, C.M. 2006. *Pattern recognition and machine learning*. New York: Springer.

Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

Blei, D.M., and J.D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. ACM.

Blei, D.M., and J.D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1(1): 17–35.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Boyd-Graber, J., and P. Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, 45–55, Stroudsburg. Association for Computational Linguistics.

Brants, T., F. Chen, and I. Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 211–218. ACM.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: *COLING*, 2666–2677.

Chen, S.F., and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, 310–318. Association for Computational Linguistics.

Dave, K., S. Lawrence, and D.M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, 519–528. New York: ACM.

Deerwester, S.C., S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs* 41(6): 391–407.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 39: 1–38.

Diao, Q., M. Qiu, C.-Y. Wu, A.J. Smola, J. Jiang, and C. Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 193–202. ACM.

Ding, X., and B. Liu. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 811–812. ACM.

Ding, X., B. Liu, and P.S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*, 231–240. New York: ACM.

Esuli, A., and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, vol. 6, 417–422. Citeseer.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford, 1: 12.

Griffiths, T.L., and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5228–5235.

Harris, Z.S. 1954. Distributional structure. *Word*.

Hearst, M.A., S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf. 1998. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4): 18–28.

Hiemstra, D., and W. Kraaij. 1998. Twenty-one at TREC7: ad-hoc and cross-language track. In *Proceedings of The Seventh Text REtrieval Conference (TREC 1998)*, Gaithersburg, 174–185, 9–11 Nov 1998.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. ACM.

Hong, L., and B.D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, 80–88. ACM.

Hu, X., J. Tang, H. Gao, and H. Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd International Conference on World Wide Web*, 607–618. International World Wide Web Conferences Steering Committee.

Jo, Y., and A.H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 815–824. ACM.

Joachims, T. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Berlin/New York: Springer.

Jordan, A. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems* 14: 841.

Jurafsky, D., and J.H. Martin. 2009. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Katz, S.M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35(3): 400–401.

Kim, S.-M., and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367. Association for Computational Linguistics.

Leskovec, J., D. Huttenlocher, and J. Kleinberg. 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, 641–650. ACM.

Lin, C., and Y. He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 375–384. ACM.

Lin, C., Y. He, R. Everson, and S. Rüger. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering* 24(6): 1134–1145.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1): 1–167.

Liu, B. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press.

Liu, L., J. Tang, J. Han, M. Jiang, and S. Yang. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, 199–208. New York: ACM.

Lu, Y., M. Castellanos, U. Dayal, and C. Zhai. 2011. Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, 347–356. ACM.

Lu, Y., and C. Zhai. 2008. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th International Conference on World Wide Web*, 121–130. ACM.

Lu, Y., C. Zhai, and N. Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, 131–140. ACM.

McAuley, J., and J. Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, 165–172. ACM.

Mcauliffe, J.D., and D.M. Blei. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems*, 121–128.

McCallum, A.K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McLachlan, G., and T. Krishnan. 2007. *The EM algorithm and extensions*, vol. 382. John Wiley & Sons.

Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, 171–180. ACM.

Mei, Q., and C. Zhai. 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 198–207. ACM.

Melville, P., W. Gryc, and R.D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, 1275–1284. New York: ACM.

Mimno, D., and A. McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *The 24th Conference on Uncertainty in Artificial Intelligence*, 411–418.

Moghaddam, S., and M. Ester. 2011. ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 665–674. ACM.

Pang, B., and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, vol. 10, 79–86. Association for Computational Linguistics.

Ponte, J.M., and W.B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, 24–28 Aug 1998, Melbourne, 275–281.

Poria, S., E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4): 26–36.

Rabiner, Lawrence R., and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Upper Saddle River: Prentice-Hall.

Ramage, D., D. Hall, R. Nallapati, and C.D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 1, 248–256. Association for Computational Linguistics.

Ramage, D., C.D. Manning, and S. Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 457–465. ACM.

Rao, D., and D. Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 675–682. Association for Computational Linguistics.

Rao, Y., Q. Li, X. Mao, and L. Wenyin. 2014. Sentiment topic models for social emotion mining. *Information Sciences* 266: 90–100.

Saif, H., M. Fernandez, Y. He, and H. Alani, 2013. Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. In *Proceedings, 1st workshop on emotion and sentiment in social and expressive media (ESSEM) in conjunction with AIIA conference*, Turin.

Shamma, D.A., L. Kennedy, and E.F. Churchill. 2009. Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the First SIGMM Workshop on Social Media*, 3–10. ACM.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, 24–29.

Smola, A., and S. Narayanamurthy. 2010. An architecture for parallel topic models. *Proceedings of the VLDB Endowment* 3(1–2): 703–710.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 306–315. ACM.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307.

Teh, Y.W., M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476).

Titov, I., and R. McDonald 2008a. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120. ACM.

Titov, I., and R.T. McDonald 2008b. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, vol. 8, 308–316. Citeseer.

Wainwright, M.J., and M.I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1–2): 1–305.

Wang, C., and D.M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 448–456. ACM.

Wang, H., Y. Lu, and C. Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 783–792. ACM.

Wang, H., Y. Lu, and C. Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 618–626. ACM.

Wang, Y., H. Bai, M. Stanton, W.-Y. Chen, and E.Y. Chang. 2009. PLDA: Parallel latent Dirichlet allocation for large-scale applications. In *Algorithmic Aspects in Information and Management*, 301–314. Springer.

Wu, C.J. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics* 11: 95–103.

Wu, Y., and M. Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 199–208. ACM.

Zhai, C., and J. Lafferty 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 403–410. ACM.

Zhai, C., and J. Lafferty 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 334–342. ACM.

Zhai, C., A. Velivelli, and B. Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 743–748. ACM.

Zhai, K., J. Boyd-Graber, N. Asadi, and M.L. Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st International Conference on World Wide Web*, 879–888. ACM.

Zhao, W.X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, 338–349. Springer.

Zhao, W.X., J. Jiang, H. Yan, and X. Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, Stroudsburg, 56–65. Association for Computational Linguistics.

Zhu, J., A. Ahmed, and E.P. Xing. 2009. Medlda: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1257–1264. ACM.

Chapter 7
Social Media Summarization

Vasudeva Varma, Litton J. Kurisinkel, and Priya Radhakrishnan

Abstract Social media is an important venue for information sharing, discussions or conversations on a variety of topics and events generated or happening across the globe. Application of automated text summarization techniques on the large volume of information piled up in social media can produce textual summaries in a variety of flavors depending on the difficulty of the use case. This chapter talks about the available set of techniques to generate summaries from different genres of social media text with an extensive introduction to extractive summarization techniques.

Keywords Social media summarization • Extractive summarization • Conversational summarization • Event summarization • Sentiment analysis • Attribute extraction semantic similarity • Topic modeling

7.1 Introduction

Text Summarization is one of the prominent areas in the domain of Computational Text Processing. The relevance of the field is of particular interest in the prevailing era of social media than ever before, given the enormous amount of data available in diverse styles and formats, from tweets, blogs to articles and news reports. Some of these data such as tweets and posts of social media stand apart from the conventional formal-styled texts, due to their highly informal, often non-grammatical usage. Nevertheless, their prominence in terms of content are no less than any formal document because of social media data are instantaneous, temporally and topically relevant and sensitive to affairs of the world. This precisely makes the idea of social media summarization interesting, despite the challenges posed by the data. In this chapter we talk about the psychological perspectives about social media usage, then discuss at length a wide range of issues pertinent to the field, present a coherent description of various methodologies in prevalence and list out the variability in the choice of summarization technique with the variability in data.

V. Varma (✉) • L.J. Kurisinkel • P. Radhakrishnan
International Institute of Information Technology-Hyderabad, Hyderabad, India
e-mail: vv@iiit.ac.in; litton.jKurisinkel@research.iiit.ac.in; priya.r@research.iiit.ac.in

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_7

Section 7.2 presents an overview of general approaches to automated text summarization with more emphasis on extractive summarization techniques. We go on to describe the recent works on extractive summarization in Sect. 7.2.1 and subsequently the nature of scoring function for candidate summary is discussed. Section 7.3 is the final part which outlines the challenges involved in social media summarization, General Approaches to Social Media Summarization, event summarization, sentiment analysis and summarization, conversational summarization and emerging trends in social media summarization under Sects. 7.3.1, 7.3.2, 7.3.3, 7.3.4, 7.3.5 and 7.3.6.

7.1.1 Expressiveness of Social Media

According to Erikson's psycho-social theory, the phases which characterize the process of adolescent and adult development include the formation of identity and the development of intimate relationships. Social networking sites allow people to engage in activities that reflect their identity. Friendships, romantic relationships, and ideology remain as key aspects of adolescent development. These identity challenges of adulthood is addressed through self-disclosure, particularly with peers.

Since online interactions offer a level of anonymity and privacy, which are quite uncommon in actual interactions, people tend to express themselves more openly in the relatively safe environment. (Kang 2000) has noted that, 'Cyberspace makes talking with strangers easier'. People with stigmatized social identities (homosexuality or fringe political beliefs) may be inspired to join and participate in online groups devoted to that particular identity, because of the relative anonymity, safety in internet and the shortage of such groups in offline world (Bargh and McKenna 2004).

Polarization of political opinions, social support groups for various causes, intimate relationships are expressed by people more openly online than in the offline world. This is due to relative insulation from identity disclosure, implicit trust in the privacy of communication and disruption the reflexive operation of racial stereotypes etc. (Kang 2000). Usage of online networks requires deep faith within. It depicts about our trust that the information which we share will not be used in unlawful or deceitful ways. We write open and confidential messages to our friends and colleagues and believe that it will remain confidential. Due to these reasons, data obtained from social media are more expressive of the people's actual opininons than in most offline interactions.

7.1.2 Need for Text Summarization on Social Media Data

Social Media interactions are instrumental in massive production and sharing of data in the form video, images and text. This enormous amount of data can be utilized to identify implicit patterns in social behavior which can be utilized for social surveys, business decisions or framing governmental policies.

The majority of data shared and produced by social media applications is in the form of text prevalent widely as posts, comments or messages.[The data produced by social media as a consequence of a particular pattern of social behaviour, can be huge in size and noisy]. This data needs to be summarized and converted into intrepretable forms so that the information contained can be utilized for practical purposes.

The information can be reported in graphical forms like Histograms or Pie charts which analyze the data on various parameters and present them statistically. But laymen who are searching for the opinion of masses about a movie, an incident or a retail product may be ignorant or impatient to interpret these representations. In such a context, a noise-free textual summary, generated out this huge volume of data makes it possible to leverage the information for the benefit of a layman end user who can afford only a 'skimming' to grasp the information conveyed.

In other words, while statistical representations can effectively capture the information pertaining to various specific parameters from a large social media data, a text summary aims to capture the information pertaining to contents of various topics and present a coherent overview of those topics. For example a statistical representation may rate the cinematography of a movie as good with 4 on a scale of 5. But a textual summary may actually give an overview of what is good in the cinematography: say, 'the veteran cinematographer Rajiv Menon has displayed sheer brilliance in the climax which received critical acclamation'.

7.2 An Overview of Autmated Text Summarization

> A summary is a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. (Radev et al. 2002)

In an era of information explosion where a large number of sources of information co-exist and produce a significantly huge content overlap, there is an immense necessity for auomatic means for summarizing this information so that a noise-free essence of the entire information available can be brought out. Automated Text summarization techniques provide means for summarizing textual content and are broadly classified into Extractive and Abstractive methods. Abstractive summarization Techniques convert the source text into an internal semantic representation which in turn is utilized by Natural Language Generation techniques to generate a summary which is equivalent to a human created summary. Due to the complexity constraints of abstractive techniques, research community has been overwhelmingly inclined towards extractive techniques. We will focus on extractive summarization techniques in the remaining part of this section.

Extractive summarization approaches try to identify from the original corpus of textual data, a proper subset of linguistic units, which can be the best representative of the original corpus within the constraints of a stipulated summary size. The linguistic units can be sentences, phrases or a short textual entity like a tweet. The research community of the field has approached the problem of auto- mated summarization in a variety of ways, but most of them can be generalized to follow three steps given below.

1. Creating an intermediate representation for the target text such that the key textual features are captured. Possible approaches are Topic Signatures, Word-frequency count, Latent Space Approaches using Matrix Factorisations or Baysian approaches.
2. Using the intermediate representation to assign scores for individual linguistic units within the text.
3. Selecting a set of linguistic units which maximises the total score as the summary for target text.

Candidate summaries are those subsets of linguistic units in the original corpus whose total size falls within the stipulated targeted summary size. The quality of a candidate summary is estimated with a scoring function and the maximum scoring candidate summary is chosen as the summary of the corpus. The scoring function for candidate summaries for a generic summarization purpose is of the form:

$$F(S) = \lambda \ * \ Coverage(S) - (1 - \lambda) \ * \ Redundancy(S) \tag{7.1}$$

Or

$$F(S) = \lambda * Coverage(S) + (1 - \lambda) * Diversity(S) \tag{7.2}$$

where λ is a constant, S is a candidate summary. Coverage function positively rewards the summary which covers maximum information from the original text, Redundancy function penalizes a candidate summary for carrying a redundant information and Diversity function encourages candidate summaries with diverse information with higher values.

7.2.1 Recent Developments in Extractive Summarization

Extensive work has been done on extractive summarization which tries to achieve a proper content coverage by scoring and selection of sentences. Typically these methods extract candidate sentences to be included in the summary and then reorder them separately. Most of the extractive summarization researches aim to increase the total salience of the sentences while reducing redundancy. Approaches include the use of Maximum Marginal Relevance (Carbonell and Goldstein 1998), Centroid-based Summarization (Radev et al. 2002), Summarization through Keyphrase Extraction (Qazvinian et al. 2010) and Formulation as Minimum Dominating Set problem (Shen and Li 2010). Graph centrality has also been used to estimate the salience of a sentence (Erkan and Radev 2004). Approaches to content analysis include generative topic models (Haghighi and Vanderwende 2009; Celikyilmaz and Hakkani-Tur 2010; Li et al. 2011a) and Discriminative models (Aker et al. 2010).

ILP2 (Galanis et al. 2012) is a system that uses Integer Linear Programming (ILP) to jointly optimize the importance of the summary's sentences and their diversity (non-redundancy), while also respecting the maximum allowed summary length. They use a Support Vector Regression model to generate a scoring function for the sentences. Woodsend and Lapata (2012) arrived at a scoring function which holds linear components to quantify the salience of bi-grams, salience of parse tree nodes and a component based on a language model which penalises the unlikely sentences. An approach based on the distribution of some important concepts in the summary was done by (Berg-Kirkpatrick et al. 2011). The concepts are bi-grams in the corpus to be summarised. They formulated an ILP objective function in the space of candidate summaries that maximizes the total concept weight score of the summary to be chosen.

Takamura and Okumura (2009) have treated multidocument summarization as a maximum concept coverage problem with knapsack constraint (MCKP). They have also exploited the possibility of decoding algorithms in solving MCKP in the summarization task. Lin and Bilmes (2011) formulated summarization as a submodular function maximization problem in the possible set of candidate summaries with due respect to the space constraint. The primary goal of all thes above methods is to achieve maximum content coverage.

As far as sentence ordering is concerned, Li et al. (2011b) used context inference to achieve better sentence ordering while (McKeown et al. 2001) used majority ordering algorithm to sort sentences. (Lapata 2013) provided an unsupervised probabilistic model for sentence ordering while (Ji and Yu 2013) used a cluster adjacency based approach. One disadvantage in these approaches is that though the sentence ordering approaches can achieve a topical order of sentences, the local structural relations of the sentences are never captured.

The work which pioneered a holistic approach towards multi-document summarization by bringing sentence selection and coherence under a single umbrella is G-Flow by (Janara et al. 2013). They built a graph which stored dis- course relations with proper edge weights to quantify coherence. This value was linearly combined along with salience and redundancy in the scoring function of sentences to formulate multi-document summarization as a constraint optimization problem. The system has taken into consideration the readability of the extracted sentences in output summary by quantifying its coherence by means of discourse graph. This has ensured the optimal content coverage with readability and coherence of the sentences taken care of in the resultant summary.

Varma et al. (2011) and Jagadeesh et al. (2007b) utilized Hyperspace Analog to language model to create a semantic space of words from word co-occurance based statistics and effectively leverage this information for summarization. Chandan et al. (2008) created a scheme for genearting personalised summaries on web documents by utilizing user specific information according to the user's subjective information need. Chandan et al. (2009) formulated summarization as a decision making problem where a risk associated with the selection of sentence in terms of information loss is estimated and the set sentences inducing minimum total risk of selection generate the summary. Rahul et al. (2009) approached summarization sentence position policy with an assumption that key sentences are present at specific locations of the text.

7.2.2 Expected Nature of Scoring Function for Candidate Summary

The scoring function of candidate summaries designed for an extractive summarization can be formalized as follows.

For a given corpus containing set of sentences V $= \{v_1, v_2, \ldots, v_n\}$,

$F : 2^V \rightarrow R$ is a function that returns a real value for any subset $S \subseteq V$. And the summarization function traces out a subset of bounded size which maximises F. i.e.

$$S_{sum} = \arg\max_{S \subseteq V} F(S) \tag{7.3}$$

where $|Ssum| \leq k$ and $k \rightarrow$ Targeted summary size.

And this optimization is obviously NP-complete. An automated multi-document summarization approach is expected to be scalable on large document set to produce a reliable summary. Lin and Bilmes (2011) observed the importance of monotone, submodular functions for extractive summarization process. It has been shown by (Nemhauser et al. 1978) that if F is monotone, non-decreasing, submodular function there exists a greedy approach which approaximates sum- maryS_{sum} such that

$$F\left(S_sum\right) >= (e - 1/e) * F\left(S_{opt}\right) \tag{7.4}$$

where

$$S_{opt} = \arg\max_{S \subseteq V} F(S).$$

(Minoux 1978) has come up with a version of this algorithm which scales to very large dataset. Submodular functions possess an interesting property of 'diminishing returns' which can be formalised as follows.

For any $A \subseteq B \subseteq V,\ and\ (v \in V, v \notin A\ and\ v \notin B)$, if F is submodular,

$$F(A + v) - F(A) \geq F(B + v) - F(B) \tag{7.5}$$

i.e. the value addition induced by v decreases as A grows to B. And F is non-decreasing if,

$$\forall A \subseteq B, F(A) \leq F(B) \tag{7.6}$$

A monotone, non-decreasing submodular functions (MND) has an additional property that a resultant function formulated as a weighted sum of several MND submodular functions, will in turn, be a monotone submodular function if weights used are positive real numbers. i.e. $F = \sum_i (\alpha_i * F_i)$ is submodular if each of the F_i is a monotone, non-decreasing function. For all i, $\alpha_i > 0$. This is of significant importance to summarization, as in most of the cases, the scoring function of sentences utilised for extractive summarization is a weighted sum of a function which estimates Topical Coverage and another function which maximises topical diversity.

For a generic summarization purpose, Lin and Bilmes (2011) used the following function

$$F(S) = L_1(S) + R_1(S) \tag{7.7}$$

Here $L_1(S)$ and $R_1(S)$ are given by

$$L_1(S) = \sum_{i \in V} \min \left\{ \sum_{j \in S} w_{i,j} \alpha \sum_{k \in V} w_{i,k} \right\} \tag{7.8}$$

where

$w_{i,j} \rightarrow$ TF-IDF cosine similarity between sentences i and j
V $\rightarrow$ set of all sentences in the corpus
S $\rightarrow$ candidate summary
$\alpha \rightarrow$ A learned parameter

$$R_1(S) = \sum_{k=1}^{K} \sqrt{\sum_{j \in S \cap P_k} \frac{1}{N} \sum_{i \in V} w_{i,j}} \tag{7.9}$$

where

$P_1, P_2, \ldots . P_k$ are sentence clusters formed out of applying k-means clustering on the set of sentences in the corpus with TF-IDF cosine similarity as the similarity metric. Description of all other variables are the same as mentioned above for Eq. 7.8.

7.3 Social Media Summarization

Social media platforms have a large number of users and the interactions among these users produce an enormous amount of data every day. Summarizing such large amount of user-generated content produced by the social media platforms without disturbing its essence can provide useful insights for various purposes. In this section on social media summarization we discuss about the type of social media data, the nature of interactions contributing to the data are introduced and then we talk about the challenges involved in the information extraction in social media summarization.

7.3.1 Challenges of Information Extraction in Social Media

The text in user-generated content occurring in social interactions is usually not well-formed in terms of Natural Language grammar, structure and formality. It also disagrees with other conventions of language in notably different ways like: usage of inconsistent cases of alphabets while dealing with named entities, missing punctuations, repetitions and reduplications, lack of good sentence structure, false starts, non-standard words, filler words like "uuumm", "uhh" and other texting disfluencies. It is indeed so because the social media texts are more likely to express emotional and context-specific content.

Unstructured noisy text data is found in informal environmental settings such as online chat, text messages, e-mails, message boards, user reviews, blogs, wikis and social networking posts. Hence while carrying out tasks specific to Information Extraction, the contemporary research is facing a lot of problems with unstructured text, as the standard natural language processing tools such as Parts of speech tagging (POS), Parsing and Named Entity Recognition (NER) exhibit poor performance on such unstructured data.

Most of the textual content in social media are in the form of tweets, comments and foot notes for images/videos and are relatively short in size. Traditional text similarity metrices expect a reasonable amount of contextual information for an accurate similarity estimation. The short textual content in social media data exhbits sparsity of contextual information and this can result in improper text classification, clustering or infomation extraction when traditional text-similarity metrices are used.

Another distinguishing characteristic of social media data is its sensitivity to chronological recency. Social media is always volatile on its views towards persons or commercial products as it relies only on the most recent outcomes related to the same. For instance, all media build-ups and public image about a politician can collapse once a video tape on his private affair goes viral in social media. When a summarization technique needs to be applied on a particular scenario, the system needs to ensure that it makes use of the most chronolgically significant set of social media interactions.

7.3.2 General Approaches to Social Media Summarization

Microblogging sites such as twitter generate and share data with users at an unprecedented rate. Such raw data from these social media are informative but overwhelming, given the sheer volume of data along with all the noise and redundancies contained within. Unlike conventional summarization systems which focus on short static data, social media summarization involves dynamic, quick to change, large-scale streams of information. Some of the approaches attempted are as follows. Online tweet clustering algorithm to cluster tweets and distilled statistics called Tweet Cluster vectors (Shou et al. 2013). They have implemented a prototype called Sumblr. An exploratory search application for twitter called as Tweetmotif was implemented by (Shou et al. 2010). TweetMotif groups messages by frequent significant terms which facilitate navigation and drilldown through a faceted search interface. The idea of User influence models, which project user interaction information onto a Twitter context tree, to help in twitter context summarization within a supervised learning framework (Chang et al. 2013). Using a scheme called as Location Centric Word Cooccurrence that uses the content of the tweets and the network information of the twitters to identify tweets that are location specific (Rakesh et al. 2013). Word graph along with optimization techniques such as decaying windows and pruning is attempted by (Olariu 2014).

There are different genres of social media interactions which can be categorised into friendship-driven, interest-driven, expert discussions etc. The approaches to social media summarization vary with respect to the purpose of summarization and the genre of data. Below explained are some special cases of social media summarization.

7.3.3 Event Summarization

Most of the events that pop up in any part of the world, whether natural, ad-hoc or planned is subject to global attention and people from near and far from the scene of event report, analyse and share their opinions through social media. Social media also gives an easy accesible platform for similar minds across the world to group together and organize events. Generally event can be viewed as a sequence of incidents/sub-events evolving at different points in event timeline or as a set of contributions of different entities that realise a single event.

Deepayan and Kunal (2011) came up with an approach to achieve event summarization in tweets which segments the event time-line into different segments where each segment corresponds to a sub-event which is a semantically distinct portion of the full-event and pick up required number of tweets from each. A specifically designed HMM is employed, which will take care of burstiness of the tweet stream and the word distribution used in the tweets, to segement the tweet stream. The approach accommodates representative tweets from both low activity periods and bursty periods along the event time-line. This ensures that certain inherently bursty sub-events which produce more tweets in comparison with other sub-events, do not undeservedly occupy more summary space. For example, a terrorist attack during an athletic meet grabs the attention of large section of audience, even of those who are not interested in athletics. A one shot summarization of tweet data without due consideration to its burstiness can result in a summary on only ‘terrorist attack’ for throughout the week. The use of automatically learned language models can ensure the seperation of sub-events which are not temporally far apart.

F. Chuan and S. Asur (2013a) proposed a Search and Summarize framework which executes event summarization in twitter in a bootstrapping manner. A normal key-word based event search can fetch a large and highly heterogenous set of tweets as the result, which makes the task of subevent detection a herculean deal. The twitter stream D_e for an event exihibits a high-level temporal topical relation i.e. If tweets $d_1, d_2, d_3 \in D_e$ are written respectively at times t_1, t_2, t_{13}, and if $t_1 \ll t_2 \ll t_3$, then the topical similarity between d_1 and d_2 will be higher than that between d_1 and d_3. They have also formulated a Decay Topic Model which takes the temporal significance of a latent topic in the tweet stream by quantifying it with an exponential decay function along with conventional word co-occurance based estimations. Initially, the system starts querying for tweets with a set of key words related to the event and apply the topic model on the resultant set of tweets. Each latent topic identified corresponds to one of the sub-events and the top ranked words from each topic is used to query for tweets again. The new result set is merged with the older set of tweets and the topic model is updated. The final set of topics are utilised to summarize the sequence of tweets by selecting the tweets for to each latent topic that give lowest perplexity.

Chao Shen and Tao Li (Chua and Asur 2013b) had come up with a participant based approach for event summarization in twitter where participants are entities that play a key role in shaping up the event. They trigger the process by tagging proper nouns using *CMU TweetNLP tool* (Gimpel et al. 2011), followed by a hierarchical clustering where resultant clusters contain the different mentions of same entity. The similarity metric used in the clustering process is represented as follows.

$$sim\left(c_i, c_j\right) = lexSim\left(c_i, c_j\right) * contSim\left(c_i, c_j\right) \tag{7.10}$$

where lexSim evaluates the lexical similarity between two mentions on the basis of *Edit distance* between the two, while *contSim* quantifies contextual similarity between two entity mentions. Context is defined by a temporal segment in the tweet sequence surrounding an entity mention and the calculation of contextual similarity between contexts relies on the term distribution in respective contexts. The global tweet set corresponding to the event is divided into different partcipant streams where each of the individial stream contains tweets holding atleast one mention of the participant entity. The major sub-events corresponding to each of the partcipant stream are identified through mixture model approach incorpoarting both time and content aspects. Such sub-events that are identified for each of the participant are merged to create a global list of sub-events and the summary is finally generated by extracting a representative tweet for each of the listed sub-event.

7.3.4 Sentiment Analysis and Summarization

Internet has brought about revolutionary change in the way people across the globe can communicate with each other, melting away the geographical rifts between them to a considerable extent. This has tremendously increased the visibility of incidents from across the world which would otherwise have been treated as only locally signifant in earlier days. More powerful modes of communication have also paved the way for a cultural penetration between communities which are far apart geographically and they have started experiencing something new which was alien to their previous generations. This enabled larger chunks of humanity to get sensitized on same issues like [Israel enchroachments in Palestine, US Presidential Election, a newly released music album, or students' agitations in Gulf countries for democracy] etc. This phenomenon of shared sensitization among larger groups of people enables people to exhibiting their emotions of support, empathy, hatred and aggressions through social media. The huge textual data piling up in social media due to debates and discussions among people belonging to different cultural, regious, economic and racial backgrounds can be utilised for comprehensive opinion surveys motivated for a large variety of purposes. Mass opinion about an entity can be broken down into different subsets where each subset brings out and highlights some aspect of the entity. A well-formed opinion summary should provide a fine-grained view of popular opinions on different aspects of the entity. Keeping this fact under consideration, (Hyun Duk Kim et al. 2010) in their detailed survey on opinion summarization, abstracted the opinion- summarization techniques, which generate a textual summary that holds opinion distribution of each aspect, into 3 major steps as follows.

1. Identify the various aspects (features/subtopic)
2. Sentiment prediction for each occurrence of an aspect
3. Extract sentences that represent the popular sentiment on each aspect

NLP techniques devised for feature identification include a combination of POS tagging and syntactic tree parsing, as most of the features are noun phrases (Lu et al. 2009; Popescu and Etzioni 2005; Hu and Liu 2004a, b). Hu and Liu (2004a, b) devised association rule mining for feature extraction to learn rules of the form $A_1, A_2, \ldots\ldots, A_n \rightarrow F_s$ where F_s stands for the feature and the approach based on other words and their POS taggs in a sentence.

The problem of product attribute and opinion extraction has been handled as a sequence labeling task in the papers (Jin and Ho 2009; Qi and Chen 2010; Zhang et al. 2010. Somprasertsri and Lalitrojwong use maximum entropy models to address the issue (Somprasertsri and Lalitrojwong 2008). By making use of a lexicalized HMM-based method Jin and Ho (2009) have proposed to perform opinion mining at the level of attributes. In (Miao et al. 2010), Miao et al. introduced a novel method to do opinion mining, with very fine-grained granularity, by utilizing Conditional Random Field models (CRFs) (Lafferty et al. 2001) and domain knowledge. To extract information about both the products and their opinions at the same time, Qi and Chen have made use of a linear-chain CRFs in (Qi and Chen 2010).

Given a source text and a context, sentiment prediction deals with ideintifying the sentiment orientation or inclination of sentiment towards some aspect of the text in the given context. For e.g. ‘Story line of the movie is bad’ holds a negative polarity towards story line of some movie, while ‘cinematography is excellant’ holds a highly biased positive polarity. There were some methods for sentiment prediction based on numerical information associated with an opinion text such as product rating associated with a product review comment. But it cannot be generalized for all the textual opinions appearing in web as many of these lack the privilege of having a user-given numerical information. Lexicon based methods exhibit a more flexible and generalisable approach for assessing sentiment polarity. SentiWordNet (Baccianella et al. 2010) is a such a lexical resource which is devised to support sentiment classificatiion and is evolved out of an automatic annotation of WordNet synsets[1] with their degrees of positivity, negativity and neutrality. Sentiment polarity of a word appearing within the context of occurrence of a specific aspect provides a reliable clue about the sentiment orientation towards that aspect in the particular context. Sufficient works have been done which trace out the sentiment polarity distribution on each aspect and generate a statistical summary which when transformed gives rise to an easily intrepretable graphical representation.Having said that, there are indeed certain contexts where a textual summary carrying more specific information, including reasons of polarity, is inevitable.

Opinion Summaries can be generated to convey different levels of granularity of opinions. Popular opinionated terminologies (e.g. excellent, boringe etc), relevant to various aspects of a particular topic, are used to retrieve the word-level opinions, as shown by Popescu and Etzioni (2005). A summary which is based on word-level popularity gives a coarser level of information about the opinion, say ‘Direction: Good’. More granularity and deeper level of understanding can be achieved by sentence level summary. e.g. ‘Technically brilliant attempt from director Vasantabalan’.

Along with popular sentiments, a sentence in an opinion summary should also convey the reason for the sentiment so that the user will get a reliable insight. A summary generated by picking up sentences carrying popular sentiment polarity may not hold the reason for the sentiment.

Glaser and Schutze (2012) have come up with an approach to identify ‘supporting sentence’ that represents the overall sentiment of a product and carries a convincing reason for the sentiment. As an initial step, they apply a sentiment classification on the entire set of sentences and classify the sentences into positive and negative sentences and pick up ‘n’ sentences which exhibit the highest probability of conforming to the overall sentiment of the document. In the succeeding step, they filter out a sentence which contains enough supporting reason for the orientaion of polarity. The quantity of suppoting information contained in a sentence is quantified

[1] https://wordnet.princeton.edu/

by weighting function based on the frequency of domain specific noun-phrases, the intution being that a supporting information cannot be conveyed without noun-phrases.

Hyun Duk Kim et al. (2013) rank the explanativeness of a sentence based on the following heuristics namely.

1. Sentence Length → a lengthier sentence can be more explanatory
2. Popularity and representativeness → a sentence is more explanatory if it contains more terms that are frequent in source text
3. Discriminativeness relative to background → A sentence is expected to be explanatory if it can discriminate source text O which is to be summarised from the background set B which is a superset of O. The set O consists of sentences satisfying the constraints that they cover aspect A of topic T with sentiment polarity P. The background set B can be constructed by relaxing any of these constrains adopted to create O.

They have come with two schemes for measuring explanativeness. First one is a modified version of BM25 (Jones et al. 2000) ranking function for information retrieval. It treats sentence as a query and ranks the explanativeness of a sentence based on the frequency of the words of the sentence in O and B.

$$BM25_E\,(S,O,B) = \sum_{\mathrm{w} \in S} IDF\,(w,B) \quad \frac{c\,(w,O)\ \ (k_1+1)}{c\,(w,O) + k_1\,(1-b) + b\frac{|O|}{avgdl}\Big)}$$

$$IDF\,(\mathrm{w},B) = log\frac{|B| - c\,(w,B) + 0.5}{c\,(w,B) + 0.5} \tag{7.11}$$

where

c(w, O)→ count of w in data set O,
$|O|$→ total number of term occurrences in data set O
$|B|$→ total number of term occurrences in data set B
avgdl → average no. of total term occurrences of sub-clusters in T from which O is extracted.
k_1 and b are parameters which can be set empirically

The second scheme measures the explanativeness of a sentence as the sum of explanativeness of each word in it. The explanativeness of each word is modelled probabilistically as follows:

$$ES(S) = \sum_{w \in S} \frac{p\,(w|E=1)}{p\,(w|E=0)} \tag{7.12}$$

Here $E = 1$ implies that the word w is observed from an explanatory sentence and $E = 0$ implies that the word w is observed from a non-explanatory sentence.

7.3.5 Conversational Summarization

Besides catering to people's need of expressing themselves and providing them with a platform to address their instantaneous emotional reflexes, social media also provides a lot of sophisticated venues for expert discussions and provisions for seeking expert advices for almost all domains including healthcare, IT and finance. Such discussion forums produce a considerably vast description of expert opinions about latest updates in different fields, ranging across different perspectives and effects a concurrent knowledge creation in the form of user generated content. An expert dialog summary created out of such data can satisfy many academic queries. Usually such discussions will be surrounding a primary topic of interest, but liable to frequent topic shifts due to a relatively large number of participants, consequently leading to data sparseness. This data sparseness can be countered by incorporating web documents of related content so that unsupervised topic modelling techniques can be employed. One among the latent topics can be traced out as the primary topic of dicussion.

Arpit et al. (2013) define the primary topic as the most prevalent topic in the longest sentence in each conversation element. A search engine is queried for each word in such a 'topic sentence' and the first web document obtained as result is fetched out. The document obtained is considered to be a description of the particular word. Such documents obtained for words belonging to the same sentence are concatenated to form a single document and the set of such documents obtained for all topic sentences constitute the input corpus for topic modelling schemes like Latent Dirichletete Allocation (LDA). The latent topic whose topic terms have a popular presence in longest sentence of each of the conversation element is treated as the primary topic of the on-going conversation and a sentence's presence in the summary is decided by quantifying its relation with the primary topic. Statistical measures based on word co-occurrence can reliably quantify the relation of word with the primary topic word. For this purpose we use HAL model which constructs the dependencies of a word w on other words based on their occurrence in the context of w within a sufficiently large corpus. HAL model creates a *term*term* matrix where each element represents a co-occurrence score between two words within a predefined window of length K.

$$HAL\left(w'|w\right) = \sum_{k=0}^{K} W(k) * matrix\left(w, k, w'\right) \tag{7.13}$$

where

matrix(w',k,w) $\rightarrow$ number of times word w' occurs k distance away from w,
K $\rightarrow$ Window length,
W(k) $\rightarrow$ K $-$ k $+$ 1 denotes the strength of coccurrence between two words

pHAL is given by

$$pHAL\left(w'|w\right) = \frac{HAL\left(w'|w\right)}{n(w) * K} \tag{7.14}$$

Here pHAL is the probability of associating a word $w^{'}$ with another word w in a window of size K. n(w) is word frequency of w.

Given the topic terms t_1 , t_2 ... t_k of the primary topic, the salience of a sentence to be present in summary is given by,

$$\text{Score}\,(S) = \prod_{\text{w} \in S} \left(P\,(w_i)_* \prod_{t_k} \text{pHAL}\,(t_k/w_i) \right) \tag{7.15}$$

A scenario where a user raises a question for an expert or peer opinion and receives more than one answers, deserves a seperate treatment compared to the one that is discussed above where a many-to-many interaction happens. Wang et al. (2014) try to attend this problem by incorporating a Ranking function which is used to quantify the revelance of a sentence to the posted query, along with other linear components for Topical Coverage and Diversity in Scoring function. They also encourage the contributions from more number of authors with an author coverage component. The sentence scoring function adopted, includes a linear component to assess the relevence of a candidate summary S which can be illustrated as follows

$$r(S) = \sum_{i}^{|S|} \sqrt{rank_i^{-1}} \tag{7.16}$$

$rank_i$ is the rank of sentence i in V, the set of all sentences in the source corpus to be summarized. $rank_i$ is calculated using ListNet (Cao et al. 2007) ranker.

Along with other coverage functions, they have introduced an author coverage function which will encourage the particpation of all the authors in the summary. Authorship coverage involves clustering the sentences based on authorship. It is given by authorship score a(S).

$$a(S) = \sum_{A \in \boldsymbol{\Lambda}} \sqrt{\left(|S \cap \text{A}| \right)} \tag{7.17}$$

Λ is the clustering induced by the sentence to author relation.

7.3.6 *Future Trends*

A summary is expected to be a representative of original corpus and is intended to convey the information contained in the original corpus without any incorrect reading. More than extracting the content, a lot of work needs to be done in re-organizing the extracted information to a presentable output which creates the right inference. The quality of the summaries can be better advanced by applying abstractive summarization techniques on user-generated content which treat user-generated content just as a source of information and generates summary in an intrepretable good language. In future, it is quite possible to extract more specific information about users and their interests to generate insightful summaries highlighting aspects relevant to the user's interests. User activity network and summarization based on those activities can provide such meaningful insights.

The virtual world of social media provides a lot of opportunities for a user to find new friends and expand his circle of closeness. The strength of each friendship can be evaluated based on different parameters such as frequency of wall posts shared, number of messges sent or the number of comments made on each other's posts (Viswanath et al. 2009). The macroscopic view of all active, user-to-user links can bring into focus the existence of larger user-activity networks within social media. Such user-activity networks share a lot attributes like geographical location, age, batchmates in college, or people having similar tastes and interests. A textual summary of interactions happening in a user-activity network can offer granular data, based on specific fine-tuned attributes of the network. For e.g., 'Interactions of photography enthusiasts in the district'. Such a precise summary is more insightful than the conventional generic summary on all the social media interactions on a particular topic. But it should be remembered that the data being dealt with, belongs to the domain of inter-personal social interactions and the individuals are naturally endowed with their privacy settings on what should be shared or not shared. Due to such intrinsic limitations on the data disclosure, such fine-tuned granular summaries are practically constrained, if not infeasible.

Acknowledgement We extend our sincere thanks to people of SIEL lab, IIIT Hyderabad for giving us the suggestion in organizing the chapter and to Vigneshwaran M, LTRC, IIIT for helping us in editing the content. We also thank Sangeetha Thomas, MA Psychology, University of Hyderabad for her insightful inputs on psychological aspects of social media usage. We received grants from DIETY, NOKIA (Microsoft Mobile) and acknowledge their contribution towards the research activities at SIEL lab, IIIT Hyderabad.

References

Aker, Ahmet, Trevor Cohn, and Robert Gaizauskas. 2010. Multi-document summarization using A * search and discriminative training. In *EMNLP*.

Arpit, Sood, Thanvir P. Mohamed, and Vasudeva Varma. 2013. Topic-focused summarization of chat conversations. In *ECIR*.

Baccianella, S, A. Esuli, and F. Sebastiani. 2010. SENTIWORDNET 3.0, An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on international language resources and evaluation. (LREC'10).*

Bargh, J.A., and K.Y. McKenna. 2004. The Internet and social life. *Annuual Review of Psychology* 55: 573–590.

Berg-Kirkpatrick, Taylor, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th annual meeting of the association for computational linguistics*, vol. 1, 481–490.

Cao, Zhe, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th international conference on machine learning, ICML 07.*

Carbonell, Jaime, and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*.

Celikyilmaz, Asli, and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Annual meeting- association for computational linguistics 2, no. conf,* 48, 815–824.

Chandan, Kumar, Prasad Pingali, and Vasudeva Varma. 2008. Generating personalized summaries using publicly available Web documents. In *Ternational conference on Web intelligence and intelligent agent technology*.

———. 2009. Estimating risk of picking a sentence for document summarization. *CICLing*.

Chang, Yi, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. 2013. Towards Twitter context summarization with user influence models. In *WSDM'13.*

Chua, Freddy Chong Tat, and Sitaram Asur. 2013a. Automatic summarization of events in social media-Freddy Chong. In *ICWSM.*

———. 2013b. A participant-based approach for event summarization using Twitter streams. In *NAACL.*

Deepayan, Chakrabarti, and Punera Kunal. 2011. Event summarization using Tweets. In *Association for the advancement of artificial intelligence.*

Erkan, Gunes, and Dragomir R. Radev. 2004. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (1): 457–479.

Galanis, Dimitrios, Gerasimos Lampouras, and Ion Androutsopoulo. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In Proceedings of the ... International conference on computational linguistics, vol. 1, 911–926.

Gimpel, Kevin, Nathan Schneide, Brendan O. Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL.*

Glaser, Andrea, and Hinrich Schutze. 2012. Automatic generation of short informative sentiment summaries. In *ACL 2012.*

Haghighi, Aria, and Lucy Vanderwende. 2009. Exploring content models for multi- document summarization. In *NAACL.*

Hu, M., and B. Liu. 2004a. Mining and summarizing customer reviews. In *KDD 04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.*

———. 2004b. Mining opinion features in customer reviews. In *AAAI04: Proceedings of the 19th national conference on Artifical intelligence.*

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, Chenxiang Zhai. 2010. Comprehensive review of opinion summarization. In *Survey paper, 2010.*

Hyun Duk Kim, Malu G. Castellanos, Meichun Hsu, ChenXiang Zhai, Umeshwar Dayal, and Riddhiman Ghosh. 2013. Ranking explanatory sentences for opinion summarization. In *SIGIR.*

Jagadeesh, J., Prasad Pingali and Vasudeva Varma. 2007a. Capturing sentence prior for query-based multi-document summarization. In *Conference RIAO2007.*

Jagadeesh, Jagarlamudi, Prasad Pingali, and Vasudeva Varma. 2007b. Capturing sentence prior for query-based multi-document summarization. In *Document understanding conferences.*

Janara, Christensen, Stephen Soderland Mausam, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of NAACL-HLT 2013,* 1163–1173.

Ji, Donghong, and Nie Yu. 2013. Sentence ordering based on cluster adjacency in multi- document summarization. In *ACL 2013.*

Jin, W., and H.A. Ho. 2009. Novel lexicalized HMM based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning.*

Jones, K.S., S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. In *Information Processing and Management.*

Kang, J. 2000. Cyber-race. *Harvard Law Review* 113 (5): 1130–1208.

Karamanis, Nikiforos, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2009. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the 42nd annual meeting of the association for computational linguistics*, 391–398, Barcelona, Spain.

Lafferty, J., A. McCallum, F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning.*

Lapata, Mirella. 2013. Probabilistic text structuring: Experiments with sentence ordering. In *ACL 2003*, 545–552.

Li, Peng, Yinglin Wang, Wei Gao, and Jing Jiang. 2011a. Generating aspect-oriented multi-document summarization with event-aspect model. In *EMNLP*.

Li, Peifeng, Guangxi Deng, and Qiaoming Zhu. 2011b Multi-document Summarization. In *Proceedings of the 5th international joint conference on natural language processing,* 1055–1061.

Lin, Hui, and Jeff Bilmes. 2011. A class of submodular functions for document summa- rization. In *Proceedings of the association for computational linguistics*, vol. 1, 510–520.

Lu, Y., C. Zhai, and N. Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web.*

Mckeown, Kathleen R., Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, and Simone Teufel. 2001. Multi-document summarization: Approach and evaluation. In *DUC01.*

McKoon, Gail, and Roger Ratcliff. 1992. Inference during reading. *Psychological Review* 99 (3): 440–446.

Miao, Q., Q. Li, Z. Daniel. 2010. Mining fine grained opinions by using probabilistic models and domain knowledge. In *Proceedings of the IEEE/WIC/ACM international conference on Web intelligence.*

Minoux, M. 1978. Accelerated greedy algorithms for maximizing submodular set functions.In *Optimization techniques,* 234–243.

Nemhauser, G.L., L.A. Wolsey, and M.L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming* 14 (1): 265–294.

Olariu, Andrei. 2014. Efficient online summarization of microblogging streams. In *ACL, 2014*.

Popescu, A.M., and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on human language technology and empirical methods in natural language processing., HLT'05.*

Qazvinian, Vahed, Dragomir R. Radev, and Arzucan Ozgur. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the ... International conference on computational linguistics*, vol 2, 895–903.

Qi, L., and L. Chen. 2010. A linear-chain CRF-based learning approach for web opinion mining. In *Proceedings of the 11th international conference on Web information systems engineering.*

Radev, D., A. Winkeil, and M. Topper. 2002. Multi-document centroid based text summarization. In *Meeting of the association for computational linguistics,* 112–113.

Rahul, Katragadda, Prasad Pingali, and Vasudeva Varma. 2009. Sentence position revisited: A robust light-weight update summarization baseline Algorithm *CLIAWS3.* In *Third international cross lingual information access workshop.*

Rakesh, Vineeth, Chandan K. Reddy, Dilpreet Singh, and M.S. Ramachandran. 2013. Location-specific tweet detection and topic summarization in Twitter. In *IEEE.*

Shen, Chao, and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of COLING*, 984–992.

Shou, Lidan, Zhenhua Wang, Ke Chen, and Gang Chen. 2010. TweetMotif: Exploratory search and topic summarization for Twitter. In *AAAI.*

Shou, Lidan, Zhenhua Wang, Ke Chen, amd Gang Chen. 2013. Sumblr: Continuous summarization of evolving tweet streams. In *SIGIR'13.*

Sipos, R., A. Swaminathan, P. Shivaswamy, and T. Joachims. 2012. Temporal corpus summarization using submodular word coverage. In *Proceedings of CIKM.*

Somprasertsri, G., and P. Lalitrojwong. 2008. Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In *Proceedings of the 2008 IEEE international conference on information reuse and integration.*

Takamura, Hiroya, and Manabu Okumura 2009. Text summarization model based on maximum coverage problem and its variants. In *Associations for Computational Linguistics,* Stroudsburg.

Varma, Vasudeva, Sudheer Kovelamudi, Jayant Gupta, Nikhil Priyatam, arpit.soodug08@students.iiit.ac.in, Harshit Jain, Aditya Mogadala Mogadala, and Srikanth Reddy Vaddepally. 2011. IIIT Hyderabad in summarization and knowledge base population. In *TAC 2011.*

Viswanath, Bimal, Alan Mislove, Meeyoung Cha, and P. Krishna. 2009. Gummadi on the evolution of user interaction in Facebook. In *WOSN09.*

Wang, Lu, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-focused opinion summarization for user-generated content. In *COLING.*

Woodsend, Kristian, and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the … Joint conference on EMNLP and computational natural language learning*, 233–243.

Zhang, S., W. Jia, Y. Xia, Y. Meng, and H. Yu. 2010. Product features extraction and categorization in Chinese reviews. In *Proceedings of the 6th international multi-conference on computing in the global information technology.*

Chapter 8
Deception Detection and Opinion Spam

Paolo Rosso and Leticia C. Cagnina

Abstract In this chapter we first introduce the reader to the problem of deception detection in general, describing how lies may be detected automatically using different methods. Later we address the specific problem of deception detection in predatory communication. We make emphasis especially on those approaches using affective resources as categorical and psychometric information provided by natural language processing tools. Finally, we focus on the problem of opinion spam whose detection is very important for reliable opinion mining. In fact, nowadays a large number of opinion reviews are posted on the Web. Such reviews are a very important source of information for customers and companies. Unfortunately, due to the business behind it, there is an increasing number of deceptive opinions on the Web. Those opinions are fictitious and have been deliberately written to sound authentic in order to deceive the consumers promoting a low quality product (positive deceptive opinions) or criticizing a potentially good quality one (negative deceptive opinions). Then, we summary some interesting approaches to detect spam opinion on the Web.

Keywords Deception detection • Opinion spam • Lie detection • Online sexual predators detection

8.1 Lie Detection

It has been demonstrated that deception is frequently present in computer-mediated communication (CMC)[1] in everyday human communication (Hancock et al. 2004). Verbal deception, defined as (Buller and Burgoon 1996): "a message knowingly

[1]The term CMC was proposed in Wolz et al. (1997).

P. Rosso (✉)
PRHLT Research Center, Universitat Politècnica de València, Valencia, España
e-mail: prosso@dsic.upv.es

L.C. Cagnina
CONICET – LIDIC, Universidad Nacional de San Luis, San Luis, Argentina
e-mail: lcagnina@unsl.edu.ar

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_8

transmitted by a sender to foster a false belief or conclusion by the receiver" is a concept that can be perfectly applied to CMC. Deception detection is a well-known challenging problem in any research area, basically because the human ability to detect deception is poor. Maybe for that reason, there is not a reliable and robust approach that is able to automatically perform that kind of detection.

Particular studies on social psychology and communications show that the accuracy rates of people abilities for detecting deception are in the range of 55–58% (Frank et al. 2004), that is, slightly better than chance. Many works point out how to find patterns to help to solve this task (Fitzpatrick et al. 2015; Poria et al. 2015). In Newman et al. (2003) the authors present the results of an experiment with participants who either lied or told the truth. They observed that liars use less frequently first person singular pronouns (*I, me, my*) maybe due to the lack of personal experience. In Burgoon et al. (2003) the authors suggest that liars use more emotional terms and concretely, negative and positive emotions (*hate, happy, sad*) than true tellers.

According to Zuckerman et al. (1981) the non-verbal behavior of liars include: emotional reactions (guilt, fear and delight are usually associated to deception (Ekman 2001)), cognitive effort in formulating their stories for avoiding contradictions (Vrij et al. 2008), and behavioral control (verbal and not) to result convincing. Such is the case of strong emotions that can activate facial muscles in almost the same moment in which the deception occurs. The work presented in Ekman (2001) shows that observing those facial micro-expressions, the deception could be detected. In fact, the author was able to classify correctly deception with an accuracy of 80% using micro-expressions observations. Then, that value was further improved in a 7% only incorporating the tone of voice (Frank and Ekman 1997) of the person who is lying.

Some computational approaches include computer vision methods used to distinguish expressions of genuine and posed pain (Littlewort et al. 2007), and facial expressions (Zhang et al. 2007; Valstar et al. 2006, 2007; Cohn and Schmidt 2004). Particularly in Littlewort et al. (2007), the authors trained 20 Support Vector Machine (SVM) classifiers with 5500 images of posed and spontaneous facial expressions (fake versus real pain). The proposed method considered 20 facial action units and obtained 72% of accuracy in differentiating fake from real pain expressions. A different strategy is the one presented in Zhang et al. (2007) to detect real facial expressions which arise from internal emotions versus those deceptive which are simulated. The system uses facial action units related to emotions (anger, enjoyment, fear and sadness), considering distance and texture based features. The results obtained with this strategy were good enough, that is, accuracy values in the range of 73–90%. In a first attempt to detect posed from spontaneous facial behavior, in Valstar et al. (2006) the authors proposed a semi automated system to discriminate brow actions. They used the speed, intensity, duration and occurrence order of each brow movement. The system obtained 90.7% of classification rate on 189 samples of spontaneous and volitional facial data. Later, in Valstar et al. (2007) the authors proposed a method for automatic multi-cue discrimination between posed and spontaneous smiles in videos. For that, they considered head, face and shoulder movements. The classification was performed with kernel methods combined with ensemble learning techniques. The obtained results reached a rate of 94% of correctly classified videos. In Cohn and Schmidt (2004) the authors used a linear classifier with timing and amplitude measures of smiles for discriminate spontaneous from deliberate similes. They obtained results of classification with 93% of accuracy using 81 young-adults videos. A different approach is the one proposed in Mihalcea et al. (2013) in which the verbal component of videos was used to detect deception. The authors used a collection of 140 fake and truthful recordings represented with the unigrams of words model. SVM and Naïve Bayes obtained accuracies in the range of 52–73% with and without considering stop words.

More sophisticated features were used in Newman et al. (2003), in which 568 texts with true and false statements were analyzed considering 29 variables of the Linguistic Inquiry and Word Count (LIWC) tool[2] like word count, amount of pronouns, positive and negative emotions, motion verbs, etc. The statements were obtained recording on videos opinions about abortion and then, these were transcribed. Besides, a group of persons were asked to write about their feeling on abortion and some others were asked about friends and a fictitious crime. With the texts, the authors trained a logistic regression method and obtained 67% of correct classification between truth tellers and liars. The authors also concluded that the five most significant variables over the 29 were: self-reference terms, negative emotions, motion verbs, references to others and exclusive words. A similar study was carried out in Mihalcea and Strapparava (2009), in which the authors constructed three corpora of 100 texts with truths and 100 with lies, considering topics as abortion, death penalty and best friends. In order to answer if the separation between both classes of texts is possible, Naïve Bayes and SVM classifiers were used, considering only the bag of word (BOW) representation with tokenization and stemming preprocessing (no feature selection was performed). The results obtained with each corpus show accuracy values around 70% in identifying true from deceptive texts. The authors also were interested in knowing which features are the distinctive of deceptive statements. For this, they calculated a dominance score of a given word class considering the collection of deceptive texts. This score is a measure of saliency for the word classes used, in this case, the 70 categories defined in LIWC. According to this score, the conclusion indicates that the word usage in deceptive texts includes detachment from the self (*you, other*) and words related to certainty (*always, very*) while belief-oriented words (*feel, believe, think*) are present in truthful statements.

It is important to detect linguistic patterns in verbal deception in order to differentiate deceptive from truthful CMC (email, instant messaging, chat, etc.). In Hancock et al. (2008) the analysis of 242 CMCs showed that liars use more words in general, more sense-based words (*touch, listen*), and few self-oriented pronouns (*I, me*) particularly when they attempt to increase the distance between themselves and the

[2]LIWC (Tausczik and Pennebaker 2010) is a tool able to analyze the positive and negative emotions (among other characteristics) contained in the text. http://www.liwc.net

deception. The authors also revealed that liars ask more questions, use more negative emotions and avoid causation words (*because, hence*) in deceptive conversations for reducing the possibility of contradictions. In Warkentin et al. (2010) the authors examined the effect of three particular warrants (pieces of information): name, photo and acquaintance on 562 CMCs including emails, instant messagings, forums, chat rooms and social networking sites. They aimed to know if the use of warrants could reduce the frequency of deception and constraint its seriousness. This issue can be used to determine if the online identity of a person matches with the real world one. The authors claimed that warrants affect the perception of the information about others, then consequently, this could propitiate deceptive practices in CMC. For that proposal, they analyzed the data collected through a survey using a mixed model approach. The conclusions stated that people lie frequently in chats but least in emails and social networking sites. Also, the authors found that there exists a negative and linear relationship between warrants and deception, with exception of real world acquaintances which constrain deception in emails and social networking sites. With similar characteristics, in Smith et al. (2014) the authors studied the effect of lies in text messaging, a particular form of CMC. A total of 164 participants filled in a short questionnaire with information about demographics and text messaging behavior. Then, the participants completed a Web survey with information related to the last 15 messages sent to two selected persons. After analyzing the obtained results, the authors concluded that deception in text messaging is not very common although there are some prolific liars (people who lie in a day more than the average). Deception in this kind of CMC seems to be less frequent among close people and has to do with concerns about coordinating activities and plans.

As we have described previously, automatic deception detection has been studied considering psychological (Zuckerman et al. 1981; Zhu et al. 2007; Vrij et al. 2008; Tsiamyrtzis et al. 2006) and psycholinguistic (DePaulo et al. 2003; Newman et al. 2003; Burgoon et al. 2003; Hancock et al. 2008) traits. In Hauch et al. (2014) some general linguistic cues related to deception were analyzed as well as those that can be detected using automatic tools. The authors considered 79 linguistic cues extracted from published articles on deception. They determined that around 60% of the studies used the LIWC tool, while less than 25% of the total of works used other general tools. Only 18.6% used specific tools developed for deception detection. The results reported in the work concluded that people who lied could experience greater cognitive load regarding the true-tellers, they demonstrated more negative emotions, and used frequently negations, first person pronoun and present tense verbs. Additional conclusions claimed that liars express fewer sensory and contextual detail words and refer less often to cognitive process in comparison with true-tellers. Besides the results obtained regarding the relationship between the language used by liars and the act of deception, the authors concluded that these linguistic cues can be applied in computational methods to detect deception.

Some proposals considered other languages besides English, as Italian and Spanish. In Fornaciari and Poesio (2012) a corpus named DECOUR of 3015 utterances transcribed from hearings held in Italian Courts, was used. The authors represented an Italian utterance by a feature vector considering: the length of the utterance (with and without the punctuation), the number of words with length longer than 6 letters, 80 linguistic variables obtained from LIWC for Italian language and, frequencies of lemmas and n-grams (with $n = 1 \ldots 5$). SVM was used for the experimental study. Three different experiments were performed: one considering the whole corpus divided into train and test sets, while the remaining two experiments used smaller subsets of utterances for training the models. These subsets were obtained performing two kinds of clusterings, each one considering: (1) distances between hearings for detecting outliers, and (2) the gender of the speakers. According to the Monte Carlo simulation applied to the test sets, values as 59.60%, 61.26% and 63.19% of correct predictions were obtained for each experiment respectively. This suggests that the models are effective in detecting deceptive texts. Later, in Fornaciari et al. (2013) a combination of personality traits was used as set of features to guide the classification of deceptive communication. The authors used 5 different classifiers to perform the experiments on a subset of DECOUR corpus. The best results were obtained with a decision tree technique, and the features used were emotional stability/neuroticism and openness to experience. The F-measure obtained outperformed the baseline with a value of 0.55. In Almela et al. (2013) the authors studied deception in Spanish written communication. They collected 100 true and 100 false statements from Spanish-speakers considering three different topics. The speakers were asked for opinions related to homosexual adoption and bullfighting, and the feeling for the best friend. A linear SVM classifier was trained with the LIWC categories in Spanish of each collected corpus. Combinations of the standard LIWC dimensions were used: linguistic dimension, psychological processes, relativity and personal concerns. The results obtained with each combination showed that F-measure scores are in the range [0.50, 0.72] for the homosexual adoption opinion corpus. In the case of the bullfighting corpus, the scores of F-measure are in the range [0.52, 0.68], and [0.63, 0.84] for the best friend feeling corpus. The conclusions showed that the fourth dimension is the least discriminant and, the first and second one are the most relevant.

8.2 Lies in Predatory Communication: Online Sexual Predators Detection

Pedophilia is a problem that has gained relevance in the past decade due the massive use of social media like *facebook*, *myspace*, *Hi5* and micro-blogs like *Twitter*, *Plurk*, *Tumblr*, etc. New ways of meeting people are offered through the use of chat rooms like `chatroulette.com` and `omegle.com` where often the identification of the user is not needed. The anonymity, the lack of information and the poor parental control promote the pedophilia as a great social problem. Pedophilia is a clinical diagnosis defined as World Health Organization (2012):

"A sexual preference for children, boys or girls or both, usually of prepubertal or early pubertal age". It is a particular case of a disorder of sexual preference of an adult individual, commonly named pedophile. From a computational point of view, a pedophile using social media for gaining access to young victims could be named "online sexual predator" or "cyberpedophile". In Guo (2008) an online sexual predator is defined like "someone who uses internet to sexually exploit vulnerable individuals, typically underaged youths" and it is characterized as a person who talks about sex as soon as he can, usually each three or four message exchanged with the young. The personality of online sexual predators is friendly because they try to detect vulnerabilities in the victims (which use to "understand" them and become thus "a friend"). On the other hand, cyberpedophiles have feelings of inferiority, isolation, loneliness, low self esteem, emotional immaturity that prevents to have adequate interpersonal interaction with people, and experiment high levels of passive aggressiveness (Hall and Hall 2007) that express through the text in the chat conversations. The offenders deceive the young making promises of love and romance but their intentions are primarily sexuals. Cyberpedophiles often create false profiles, pretend to be younger or of the opposite sex and try to copy child's behaviour.

The phenomenon of pedophilia has been studied from different research perspectives. From the law enforcement view, through the reforms in the criminal codes in order to create a new offence for persons who use Internet to procure an underage to commit a sexual act or expose him to pornography. This, in conjunction with programs and non-profit organizations to investigate, control, detect and prevent sexual exploitations of underage in social media, pretend to address the problem. Psychology and forensic psychiatry study the interactions between the offender and the victims in order to establish a behavioral pattern. From the natural language processing (NLP) perspective, the research points out to provide reliable tools to automatically detect pedophilia in online social media.

One of the problems involved with the computational approaches for the detection of sexual predators is the manual monitoring of chat conversations. Usually these are impossible to analyze due to the massive amount of data to processing and some privacy issues. Besides, the characteristics of this kind of text prevent to use general tools to evaluate their content. The texts of the chats generally are informal, quite different from the regular written texts (news, abstracts, monographs, etc.) even blogs. In chat conversations it is possible to find large amount of mistakes and misspellings caused by the fast typing, the use of emoticons, abbreviations, character flooding and specific slangs. For general NLP processing tools, the latter characteristics can be considered as very noise data, but for specific approaches for detecting possible pedophiles, can be valuable information to process.

The detection of certain emotions in the text could help to detect possible pedophiles. An initial work used categorical and psychometric information provided by LIWC as features, besides the traditional term-based features for the representation of the chat conversations (Rahman Miah et al. 2011). Typical words as "friend", "family", "sex", "anger", "happy", "sad" and "anxiety" are indicative of emotional and cognitive components. Then, using standard text categorization techniques as Naïve Bayes, J48 decision tree and classification via regression, the authors classified the chat logs in three categories: (a) underage exploiting: an adult offender chats with a minor, (b) sex fantasy: chats between two adults with sexual content and, (c) general chats: without sexual subject matter. The study showed that the representation enriched with the category of the words and the psychometric information, improved the performance of some classifiers used to predict the class of underage exploiting chats. Other works have used LIWC to extract useful information for the pedophile detection in social media. Such is the case of Gupta et al. (2012) in which LIWC was used to create psycholinguistic profiles for finding patterns related to the six online grooming stages (O'Connell 2003): friendship forming, relational forming, risk assessment, exclusivity, sex and conclusion. These profiles can be used in automatic classifiers to detect possible stages of grooming in chat conversations. From the study, the authors concluded that relationship forming (personal information exchange about family, friends, school, etc.) is the most characteristic stage. However, the pedophile generally does not wait for the ending of grooming to produce the meeting with the underage. Then, the conclusion stage should be identified early in order to detect a possible attack. In Parapar et al. (2012) the authors studied three different strategies to extract a feature-based representation for chat conversations: the standard term-based tf-idf, eleven chat-based features with information about the activity of the person in chatrooms (number of lines in a chat, number of users participating in a conversation, time between consecutive line messages, etc.), and LWIC features for analyzing aspect as deception versus honesty through the category of words (psychological constructs: affect, cognition; personal concern: home, leisure, etc.). The three sets of features were used independently for performing the classification with SVM. The results obtained were not good enough for identifying sexual predators. Then, combinations of these sets were used obtaining the best results with the tf-idf and chat-based features. In a later work (Parapar et al. 2014), the authors proposed additional LIWC features (80 in total) based on psycho-linguistic evidence. They argued that those features are markers of emotional states and provide valuable clues about deception and honesty. A deeper analysis of the best performing classifiers and the most discriminative features concluded that the set of features utilized and the relative weighting of the misclassification costs in the SVM algorithms, are important factors that affect the performance of the system. They identified that the word categories more implicated in deception are: use of pronouns, emotion words, markers of cognitive complexity, and motion verbs. To similar conclusions arrived the authors of Cano et al. (2014) in which the chat conversations were classified considering features of sentiment polarity, content and, psycholinguistic and discourse patterns. The interesting proposal focuses on the behavior of predators in each underage grooming stage (as was proposed in Gupta et al. 2012) classifying the lines into such stages. Basically, the authors in Cano et al. (2014) make a profile of the predator considering six different types of features: BOW (1, 2 and 3 grams), syntactical (POS tags), sentiment polarity (extracted from a sentence with

Sentistrength[3]), content (complexity, readability, length), psycholinguistic (62 in total obtained with LIWC) and discourse patterns (semantic frame in which a word sense is used). Then, a supervised approach for automatic classification of online grooming stages was proposed. The results obtained showed that the discourse 'label' feature outperformed the baseline in terms of precision for the three stages. When combined features were used, the results improved in terms of precision and F-measure for grooming and approach stages. Regarding the analysis of features, the authors found that sentiment polarity characteristics used in the study, were not discriminatory. On the contrary, discourse frames as emotional_state, desiring and stimulus_focus (fine-grained emotions) were useful in the classification task. For evaluating the differences and similarities of the grooming stage in both online and face-to-face environments, the authors in Black et al. (2015) used the transcripts of 44 convicted online offenders. They also used LIWC and content analysis of strategies in order to study the texts. The considered strategies involved situations as friendship forming, risk assessment, exclusivity and sexual stages. The results indicated that many strategies as talk about plans, use of flattery, the assessing of parents activities and the mention of past relationships, are common practices for both environments. Besides, the timing and the order of the considered strategies seem to be different in online communication; for example, the deceiver in CMC uses the strategies faster and, the assessing risk particularly, is more frequent than in face-to-face communication.

The change of mood could be indicative of the level of emotional instability of pedophiles. There is an interesting publicly available resource named SenticNet (Cambria et al. 2016) which associates semantics and sentics to many common and common-sense concepts for the analysis of concept-level sentiments. Another useful resource to obtain information about the emotion contained in the words is WordNet Affect (Strapparava and Valitutti 2004), an additional hierarchy of "affective domain labels" as part of WordNet Domains. WordNet Affect was used for the identification of emotions in Bogdanova et al. (2014) such as positive and negative words related to basic emotions such as joy, sadness, anger, disgust, surprise and fear. Content and stylistic based features as: approach words (meet, car), family words (mum, dad), relationship nouns (boyfriend, date), personal pronouns (I, you) and obligation verbs (must, have to), were also considered in the same work. The results obtained with a SVM classifier concluded that the use of high-level features achieves the 97% of accuracy discriminating cyberpedophiles from cybersex chats in comparison with the use of low-level features (50–64%).

Other works have used only the content of the chat conversations directly. In Kucukyilmaz et al. (2008), Egan et al. (2011), and Barber and Bettez (2014) the authors investigated the feasibility of predicting the author of a chat by the extraction of the information contained in the text. In Kucukyilmaz et al. (2008) the authors stated that chat messaging has evolved in order to transfer emotions. A

[3]http://sentistrength.wlv.ac.uk/

clear example of it is the use of emoticons[4] for representing feelings typing only a sequence of punctuation marks. Also, the repetition of some characters or the use of uppercase letters in a word, are used to transfer emotions. In Egan et al. (2011) the authors analyzed the written language in chats in order to identify recurrent topics that cyberpedophiles usually use. In Barber and Bettez (2014) the authors identified online sexual predators patterns of behavior for a potential use in pattern recognition. Their study concluded that characteristics as fantasizing (cyber sexual elements in text), sexuality assessment (to obtain information about the sexual skill of the youth), domination (over the acts of the victim), enticement, and the intention to have a face-to-face meeting, could be used to improve automated detection software and educational tools.

Due the important challenge involved with the detection of predatory communication, a shared task on sexual predator identification was organized at PAN-2012 (Inches and Crestani 2012). The objective of the task was twofold: identifying the predators among all the users in the different chat conversations and identifying the most distinctive lines of the predator bad behaviour. The 16 teams participating in the contest made possible the recognition of common pattern for the predators identification. The winner method used a two step approach (Villatoro-Tello et al. 2012) for distinguish predators conversations between normal chats. The authors performed a preprocessing step for removing conversations with just one participant, less than 6 interventions per user and containing text with 3 long sequence of unrecognized characters. The best result, with F-measure of 0.87, was obtained with a neural network classifier using BOW with boolean scheme. Only two proposals considered characteristics that go beyond shallow lexical features. In Vartapetiance and Gillam (2012) the authors discovered that sexual intentions can be detected from the text, although not explicitly, considering activities that the pedophile tries to share with the victim as watching TV, listening to music, meeting and having fun. Also some spelling combination of words as "go down on you" and "make you come" are usually used to express the wished of sexual intentions. The latter was used as feature in Vartapetiance and Gillam (2012) along with the identification of words related to age ("you are young", "wish you were"), parents ("your mom", "Ur dads car") and address ("ur address", URLs). Using these four features the authors obtained a F-measure score of 0.47. In Morris and Hirst (2012) the authors proposed behavioral features besides lexical, in order to model the actions of a possible predator. Features as tendency to initiate a conversation, number of times asking the same question, attempts to keep a conversation going, response time, repeated messages and dominance of the conversation, contributed to obtain a F-measure score of 0.72.

[4] *emot(ion) + icon.* "A sideways facial glyph used in e-mail to indicate an emotion or attitude, as to indicate intended humor" (Pickett 2000).

8.3 Lies in Opinions: Deceptive Opinions Detection

With the increasing availability of review sites and blogs, consumers rely more than ever on online reviews to make their purchase decisions. A recent survey[5] found that 68% of them have reinforced the decision to purchase a product or service by positive online reviews and 92% of consumers read online reviews to judge a local business or a product. Therefore, detecting lies in opinions is a very important problem as well as challenging since opinions expressed on the Web are typically short texts, written by unknown people using different styles and for different purposes.

The detection of opinion spam, i.e., the identification of fake reviews that try to deliberately misleading human readers, is just another face of the problem of the detection of lies on the Web (Lau et al. 2012). Nevertheless, the construction of automatic detection methods for this task is complex since manually gathering labeled reviews, and particularly truthful opinions, is difficult (Mukherjee et al. 2011). Due to the lack of reliable labeled data, most initial works on the detection of opinion spam considered unsupervised approaches which relied on meta-information from reviews and reviewers. For example, in Jindal and Liu (2008) the authors proposed detecting opinion spam by identifying duplicate content. This method showed good precision using a logistic regression classifier with a reviews dataset from Amazon but it failed detecting original fake reviews. In a subsequent paper (Jindal et al. 2010), the authors proposed to detect spammers by searching for unusual review patterns. They classified a reviewer as spam suspect if the person wrote negative reviews about all the products of a brand but wrote positive reviews about a competing brand. The duplication of content was also considered in Lin et al. (2014), in which several features based on similarities were presented. The authors measured the similarity of a review regarding other reviews of the same author and other reviews about the same product, reviews frequency of the product, and comments frequency. Then, those features were used to determine if a review is spam or not considering a threshold. Considering also a similarity score, a probabilistic language model detects similar content between two reviews (Lai et al. 2010). The authors tested the model with a SVM classifier and obtained a precision of 81% in detecting spam reviews. A lower precision value of 43.6% was obtained with an analogous approach but considering the conventional cosine function to measure conceptual features (Algur et al. 2010). Similarly, in Wu et al. (2010) the authors presented a method to detect hotels which are more likely to be involved in spamming. They proposed a number of criteria that might be indicative of suspicious reviews and then, they evaluated alternative methods for integrating these criteria to produce a suspiciousness ranking. Their criteria mainly derive from characteristics of the network of reviewers and also from the impact and ratings of reviews. It is worth mentioning that they did not take advantage of reviews' content for their analysis. In the same category of unsupervised approaches, in Mukherjee et al.

[5]Local Consumer Review Survey 2015 (visited: January 3, 2016): https://www.brightlocal.com/learn/local-consumer-review-survey/

(2011) the authors proposed a method for detecting groups of opinion spammers based on criteria such as the number of products for which the group work together and a high content similarity of their reviews. Finally, in (Xie et al. 2012), it has been demonstrated that a high correlation between the increase in the volume of singleton reviews and a sharp increase or decrease in the ratings is a clear signal that the rating is manipulated by possible spam reviews. Supported by this observation the authors proposed an opinion spam detection method based on temporal pattern discovery.

It was only after the release of the gold-standard datasets (Ott et al. 2011, 2013), which contain examples of positive and negative deceptive opinion spam, that it was possible to conduct supervised learning and a reliable evaluation of the task.[6] In Ott et al. (2011) the authors employed a SVM classifier to distinguish between positive deceptive and truthful reviews using different stylistic, syntactic and lexical features. Then, in Ott et al. (2013) they applied the same approach to classify negative opinions. The main conclusion from these works is that standard text categorization techniques using unigrams and bigrams word features are effective at detecting deception in text, and that their results significantly outperform those from human judges. Following this research direction, in Feng et al. (2012a,b) the authors extended Ott et al.'s n-gram feature set by incorporating deep syntax features, i.e., syntactic production rules derived from probabilistic context free grammar parse trees. Their experimental results consistently find statistical evidence that deep syntactic patterns are helpful in discriminating deceptive writing. Similarly, in Feng and Hirst (2013) the authors extended previous Ott et al. and Feng et al.'s works by incorporating features that characterize the degree of compatibility between the personal experience described in a test review and a product profile derived from a collection of reference reviews about the same product. This idea was supported by the hypothesis that since the writer of a deceptive review usually does not have any actual experience with that product, the resulting review might contain some contradictions with facts about the product. This approach significantly improved the performance of identifying deceptive reviews.

Although supervised text classification techniques have demonstrated to be very robust if they are trained using large sets of labeled instances from both deceptive and truthful opinions – some works have reported F_1 measures around 0.90 (Ott et al. 2011, 2013; Feng and Hirst 2013) – in real application scenarios it is very difficult to compile such large training sets and maybe, it is almost impossible to determine the authenticity of the opinions, i.e., to assemble a set of verified truthful reviews (Mukherjee et al. 2011). To overtake this restriction, in PU-learning (Liu et al. 2002) has been applied to detect deceptive opinion spam learning only from a few examples of deceptive opinions and a set of unlabeled data, under the consideration that deceptive opinion spam can be accurately generated via crowdsourcing as suggested in Ott et al. (2011).

[6] http://myleott.com/op_spam

In Li et al. (2014) the authors present a study on Chinese fake review detection. First they considered two classes of reviews: fake and unknown. However, since the unknown data set may contain many fake reviews, it was treated as an unlabeled set. Therefore, the PU-learning model was employed in order to learn from positive and unlabeled examples. Experimental results showed that PU learning not only outperforms supervised learning significantly, but also detects a large number of potentially fake reviews hidden among the unlabeled examples.

In Hernández Fusilier et al. (2015) the authors proposed a PU-learning variant for detecting opinion spam. The evaluation of the proposed method was carried out using the set of hotel reviews gathered in Ott et al. (2013) containing positive and negative deceptive opinion spam. The results are encouraging: on the one hand, they indicate that using only a hundred of examples of deceptive opinions for training it is possible to reach F_1 measures of 0.8 and 0.7 for positive and negative opinions, respectively. On the other hand, they demonstrate the appropriateness of the proposed PU-learning variant for detecting opinion spam, since its results significantly outperformed those from the original PU-learning approach in both kinds of opinion spam. Moreover, the authors analysed the role of opinions' polarity in the detection of deception. Their results confirm that negative deceptive opinions are more difficult to detect than positive ones, but they also show that having one single classifier for analysing both kinds of opinions is better than using two separate classifiers, suggesting that there are common characteristics in the way people write positive and negative opinion spam. In Ren et al. (2014) the same authors proposed a semi-supervised model. Firstly, some reliable negative examples were identified from the unlabeled dataset. Secondly, some representative positive examples and negative examples were generated with Latent Dirichlet Allocation. Thirdly, a SVM classifier was feeded with the remaining unlabeled examples and their similarity weights. Experiments on gold-standard Ott's dataset showed very interesting results obtaining accuracy values above 80%. Better results on detecting whether a review is spam or not were obtained with the framework presented in Sharma and Lin (2013). Criteria as rating consistency, questions, capital letters, comparative sentences and links were used to calculate a rating of a review. Considering that rating the framework could determine whether a review is spam or not, with a high accuracy value.

In Hernández Fusilier et al. (2015) the detection of opinion spam was considered as a stylistic classification task. That is, although deceptive and truthful opinions given a particular domain are similar in content, they differ in the way opinions are written. The authors proposed to use character n-grams as features since they have shown to capture lexical content as well as stylistic information. They evaluated their approach on the standard-de-facto Ott's corpus composed of 1600 hotel reviews, considering positive and negative reviews. They compared the results obtained with character n-grams against the ones with word n-grams. The results obtained show that character n-grams are good features for the detection of opinion spam; they seem to be able to capture better than word n-grams the content of deceptive opinions and the writing style of the deceiver. In particular, the results show an improvement of 2.3% and 2.1% over the word-based representations in the detection of positive and negative deceptive opinions respectively. Furthermore, character n-grams allow to obtain a good performance also with a very small training corpus. Using only 25% of the training set, a Naïve Bayes classifier showed F_1 values up to 0.80 for both opinion polarities. A similar study was presented in Cagnina and Rosso (2015) in which the authors studied the performance of Naïve Bayes and SVM classifiers using character n-grams in tokens, the sentiment score and LIWC linguist features such as pronouns, articles and verbs (present, past and future tenses). The Ott's corpus cited previously was used to test the proposed features and the results were compared with those obtained with state-of-the-art methods. From the experimental study the authors concluded that character n-grams in tokens capture correctly content and the writing style of the reviews, the sentiment-based feature does not provide useful information for detecting deception in this kind of text, and LIWC variables as pronouns, articles and verbs are meaningful. In fact, character 4-grams in tokens combined with LIWC variables performed the best with a SVM classifier reaching a F-measure of 0.87. Regarding the comparison with the results of Hernández Fusilier et al. (2015), the statistical significance test showed that both approaches performed similarly although the proposal in Cagnina and Rosso (2015) used a lower dimensionality representation (95% reduction of features) compared with the one presented in Hernández Fusilier et al. (2015).

8.4 Conclusions

From the point of view of psychological, linguistic and computational processes, the deception detection presents constant challenges to be addressed. In this work different approaches to automatically detect deception have been described, although we have focused mainly on those that considered emotional and cognitive aspects of the problem. Verbal deception detection has been also addressed in online sexual predators communications. Special attention has been given to the problem of the detection of deceptive opinions.

Acknowledgements This work is the result of the collaboration in the framework of the WIQ-EI IRSES research project (grant no. 269180) within the EC FP7 Marie Curie. The work of the first author was also in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO research project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. The first author would like also to thank the co-authors of the papers on deceptive opinions detection. The research work of the second author has been partially funded by CONICET (Argentina).

References

Algur, S.P., A.P. Patil, P.S. Hiremath, and S. Shivashankar. 2010. Conceptual level similarity measure based review spam detection. In *2010 International Conference on Signal and Image Processing (ICSIP)*, 416–423.

Almela, A., R. Valencia-García, and P. Cantos. 2013. Seeing through deception: A computational approach to deceit detection in Spanish written communication. *Linguistic Evidence in Security, Law and Intelligence* 1(1): 3–12.

Barber, C.S., and S.C. Bettez. 2014. Deconstructing the online grooming of youth: Toward improved information systems for detection of online sexual predators. In *Proceedings of the International Conference on Information Systems (ICIS 2014)*.

Black, P.J., M. Wollis, M. Woodworth, and J.T. Hancock. 2015. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect* 44: 140–149.

Bogdanova, D., P. Rosso, and T. Solorio. 2014. Exploring high-level features for detecting cyberpedophilia. *Computer Speech and Language* 28(1): 108–120.

Buller, D.B., and J.K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory* 6(3): 203–242.

Burgoon, J.K., J.P. Blair, T. Qin, and J.F. Nunamaker Jr. 2003. Detecting deception through linguistic analysis. In *Intelligence and security informatics*, ed. H. Chen, R. Miranda, D.D. Zeng, C. Demchak, J. Schroeder, and T. Madhusudan, Lecture notes in computer science, vol. 2665, 91–101. Berlin/Heidelberg: Springer.

Cagnina, L., and P. Rosso. Classification of deceptive opinions using a low dimensionality representation. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Lisboa, 58–66, Sep 2015. Association for Computational Linguistics.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2666–2677, Osaka, Japan.

Cano, A.E., M. Fernandez, and H. Alani. 2014. Detecting child grooming behaviour patterns on social media. In *Social informatics*, ed. L. Aiello and D. McFarland, Lecture notes in computer science, vol. 8851, 412–427. Springer.

Cohn, J.F., and K.L. Schmidt. 2004. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing* 2(2): 121–132.

DePaulo, B.M., J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. 2003. Cues to deception. *Psychological Bulletin* 129(1): 74–118.

Egan, V., J. Hoskinson, and D. Shewan. 2011. Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex. In *Antisocial behavior: Causes, correlations and treatments*, ed. R.M. Clarke, 119–134. New York: Nova Science Publishers.

Ekman, P. 2001. *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. New York: W. W. Norton & Company.

Feng, S., R. Banerjee, and Y. Choi. 2012a. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, 171–175. The Association for Computer Linguistics.

Feng, S., L. Xing, A. Gogar, and Y. Choi. 2012b. Distributional footprints of deceptive product reviews. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ed. J.G. Breslin, N.B. Ellison, J.G. Shanahan, and Z. Tufekci, 98–105. The AAAI Press.

Feng, V.W., and G. Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 338–346. Asian Federation of Natural Language Processing.

Fitzpatrick, E., J. Bachenko, and T. Fornaciari. 2015. *Automatic detection of verbal deception*, Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Fornaciari, T., F. Celli, and M. Poesio. 2013. The effect of personality type on deceptive communication style. In *European Intelligence and Security Informatics Conference*, 1–6. IEEE.

Fornaciari, T., and M. Poesio. 2012. On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection (EACL 2012)*, 39–47. Association for Computational Linguistics.

Forner, P., J. Karlgren, and C. Womser-Hacker, ed. 2012. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, Rome.

Frank, M.G., and P. Ekman. 1997. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology* 72(6): 1429–1439.

Frank, M.G., N. Paolantonio, T.H. Feeley, and T.J. Servoss. 2004. Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation* 13(1): 45–59.

Guo, R.M. 2008. Stranger danger and the online social network. *Berkeley Technology Law Journal* 23(1): 617–644.

Gupta, A., P. Kumaraguru, and A. Sureka. 2012. Characterizing pedophile conversations on the internet using online grooming. arXiv:1208.4324v1.

Hall, R.C.W., and R.C.W. Hall. 2007. A profile of pedophilia: Definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. In *Mayo Clinic Proceedings*, 457–471. Elsevier.

Hancock, J.T., L.E. Curry, S. Goorha, and M. Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45(1): 1–23.

Hancock, J.T., J. Thom-Santelli, and T. Ritchie. 2004. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 129–134. ACM.

Hauch, V., I. Blandon-Gitlin, J. Masip, and S.L. Sporer. 2014. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review* 41: 1–36.

Hernández Fusilier, D., M. Montes y Gómez, P. Rosso, and R. Guzmán Cabrera. 2015. Detecting positive and negative deceptive opinions using pu-learning. *Information Processing and Management* 51(4): 433–443.

Hernández Fusilier, D., M. Montes y Gómez, P. Rosso, and R. Guzmán Cabrera. 2015. Detection of opinion spam with character n-grams. In *16th International Conference on Intelligent Text Processing and Computational Linguistics*, Lecture Notes in Computer Science, vol. 9042, 285–294. Springer.

Inches, G., and F. Crestani. 2012. Overview of the international sexual predator identification competition at pan-2012. In Forner et al. (2012).

Jindal, N., and B. Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*, 219–230. ACM.

Jindal, N., B. Liu, and E. Lim. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ed. J. Huang, N. Koudas, G.J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, 1549–1552. ACM.

Kucukyilmaz, T., B.B. Cambazoglu, C. Aykanat, and F. Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management* 44(4): 1448–1466.

Lai, C.L., K.Q. Xu, R.Y.K. Lau, Y. Li, and L. Jing. 2010. Toward a language modeling approach for consumer review spam detection. In *2010 IEEE 7th International Conference on e-Business Engineering (ICEBE)*, 1–8.

Lau, R.Y.K., S.Y. Liao, R.C. Kwok, K. Xu, Y. Xia, and Y. Li. 2012. Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems* 2(4): 1–25.

Li, H., B. Liu, A. Mukherjee, and J. Shao. 2014. Spotting fake reviews using positive-unlabeled learning. *Computación y Sistemas* 18(3): 467–475.

Lin, Y., T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou. 2014. Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 261–264.

Littlewort, G.C., M.S. Bartlett, and K. Lee. 2007. Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI '07)*, 15–21. ACM.

Liu, B., W.S. Lee, P.S. Yu, and X. Li. 2002. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML '02)*, 387–394. Morgan Kaufmann Publishers Inc.

Mihalcea, R., V. Pérez-Rosas, and M. Burzo. 2013. Automatic detection of deceit in verbal communication. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*, 131–134. ACM.

Mihalcea, R., and C. Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics (ACL-IJCNLP 2009) Conference Short Papers (ACLShort '09)*, 309–312. Association for Computational Linguistics.

Morris, C., and G. Hirst. 2012. Identifying sexual predators by SVM classification with lexical and behavioral features. In Forner et al. (2012).

Mukherjee, A., B. Liu, J. Wang, N. Glance, and N. Jindal. 2011. Detecting group review spam. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*, 93–94. ACM.

Newman, M.L., J.W. Pennebaker, D.S. Berry, and J.M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29(5): 665–675.

O'Connell, R. 2003. A typology of child cybersexploitation and online grooming practices, http://netsafe.org.nz/Doc_Library/racheloconnell1.pdf.

Ott, M., C. Cardie, and J.T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, 497–501. The Association for Computational Linguistics.

Ott, M., Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 309–319.

Parapar, J., D. Losada, and A. Barreiro. 2012. A learning-based approach for the identification of sexual predators in chat logs. In Forner et al. (2012).

Parapar, J., D.E. Losada, and A. Barreiro. 2014. Combining psycho-linguistic, content-based and chat-based features to detect predation in chatrooms. *Journal of Universal Computer Science* 20(2): 213–239.

Pickett, J.P. 2000. *The American Heritage Dictionary of the English Language*. Number 2 in The American Heritage Dictionary of the English Language. Houghton Mifflin.

Poria, S., E. Cambria, A.F. Gelbukh, F. Bisio, and A. Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4): 26–36.

Rahman Miah, M.W., J. Yearwood, and S. Kulkarni. 2011. Detection of child exploiting chats from a mixed chat dataset as text classification task. In *Proceedings of the Australian Language Technology Association Workshop*, 157–165. ALTA.

Ren, Y., D. Ji, and H. Zhang. 2014. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 488–498. Association for Computational Linguistics.

Sharma, K., and K-I. Lin. 2013. Review spam detector with rating consistency check. In *Proceedings of the 51st ACM Southeast Conference (ACMSE '13)*, 34:1–34:6. ACM.

Smith, M.E., J.T. Hancock, L. Reynolds, and J.P. Birnholtz. 2014. Everyday deception or a few prolific liars? The prevalence of lies in text messaging. *Computers in Human Behavior* 41: 220–227.

Strapparava, C., and A. Valitutti. 2004. WordNet-affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1083–1086. ELRA.

Tausczik, Y.R., and J.W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.

Tsiamyrtzis, P., J. Dowdall, D. Shastri, I.T. Pavlidis, M.G. Frank, and P. Ekman. 2006. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision* 71(2): 197–214.

Valstar, M.F., H. Gunes, and M. Pantic. 2007. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI '07)*, 38–45. ACM.

Valstar, M.F., M. Pantic, Z. Ambadar, and J.F. Cohn. 2006. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06)*, 162–170. ACM.

Vartapetiance, A., and L. Gillam. 2012. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. In Forner et al. (2012).

Villatoro-Tello, E., A. Juárez-González, H.J. Escalante, M. Montes y Gómez, and L. Villaseñor Pineda. 2012. A two-step approach for effective detection of misbehaving users in chats. In Forner et al. (2012).

Vrij, A., ed. 2008. *Detecting lies and deceit: Pitfalls and opportunities*, 2nd ed. Chichester: Wiley.

Warkentin, D., M. Woodworth, J.T. Hancock, and N. Cormier. 2010. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*, ed. K.I. Quinn, C. Gutwin, and J.C. Tang, 9–12. ACM.

Wolz, U., J. Palme, P. Anderson, Z. Chen, J. Dunne, G. Karlsson, A. Laribi, S. Männikkö, R. Spielvogel, and H. Walker. 1997. Computer-mediated communication in collaborative educational settings. In *The Supplemental Proceedings of the Conference on Integrating Technology into Computer Science Education: Working Group Reports and Supplemental Proceedings (ITiCSE-WGR '97)*, 51–69. ACM.

World Health Organization. 2012. *International Statistical Classification of Diseases and Related Health Problems (ICD-10)*. Canadian Institute for Health Information.

Wu, G., D. Greene, and P. Cunningham. 2010. Merging multiple criteria to identify suspicious reviews. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*, 241–244. ACM.

Xie, S., G. Wang, S. Lin, and P.S. Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, 823–831. ACM.

Zhang, Z., V. Singh, T.E. Slowe, S. Tulyakov, and V. Govindaraju. 2007. Real-time automatic deceit detection from involuntary facial expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007*, 1–6.

Zhu, Z., P. Tsiamyrtzis, and I. Pavlidis. 2007. Forehead thermal signature extraction in lie detection. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007*, 243–246.

Zuckerman, M., B.M. DePaulo, and R. Rosenthal. 1981. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology* 14: 1–59.

Chapter 9
Concept-Level Sentiment Analysis with SenticNet

Federica Bisio, Claudia Meda, Paolo Gastaldo, Rodolfo Zunino, and Erik Cambria

Abstract SenticNet is a publicly available resource for opinion mining that exploits AI, linguistics, and psychology to infer the polarity associated with commonsense concepts and encode this in a semantic-aware representation. In particular, SenticNet uses dimensionality reduction to calculate the affective valence of multi-word expressions and, hence, represent it in a machine-accessible and machine-processable format. This chapter presents an overview of the most recent sentic computing tools and techniques, with particular focus on applications in the context of big social data analysis.

Keywords SenticNet • Sentic computing • Concept-level sentiment analysis • Big social data analysis

9.1 Introduction

Sentic computing (Cambria and Hussain 2015) is a multi-disciplinary approach to sentiment analysis that exploits both computer and social sciences to better recognize, interpret, and process opinions and sentiments over the Web. The approach specifically brings together lessons from both affective computing and commonsense computing because, in the field of opinion mining, not only commonsense knowledge, but also emotional knowledge is important to grasp both the cognitive and affective information (termed semantics and sentics) associated with natural language opinions and sentiments.

F. Bisio (✉) • C. Meda • P. Gastaldo • R. Zunino
DITEN, University of Genoa, Via Opera Pia 11A, Genova, I-16145, Italy
e-mail: federica.bisio@edu.unige.it; claudia.meda@edu.unige.it; paolo.gastaldo@unige.it; rodolfo.zunino@unige.it

E. Cambria
School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore, Singapore
e-mail: cambria@ntu.edu.sg

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8_9

During most of the last century, research on emotions was conducted by philosophers and psychologists, whose work was based on a small set of emotion theories that continue to underpin research in this area. The first researchers to try linking text to emotions were actually social psychologists and anthropologists who tried to find similarities on how people from different cultures communicate (Osgood et al. 1975). This research was also triggered by a dissatisfaction with the dominant cognitive view centered around humans as 'information processors' Lutz and White (1986).

Later on, in the 1980s, researchers such as Turkle (1984) began to speculate about how computers might be used to study emotions. Systematic research programs along this front began to emerge in the early 1990s. For example, Scherer (1993) implemented a computational model of emotion as an expert system. A few years later, Picard's landmark book affective computing (Vesterinen 2001) prompted a wave of interest among computer scientists and engineers looking for ways to improve human-computer interfaces by coordinating emotions and cognition with task constraints and demands. Picard described three types of affective computing applications:

- Systems that detect the emotions of the user;
- Systems that express what a human would perceive as an emotion;
- Systems that actually 'feel' an emotion.

Although touching upon HCI and affective modeling, sentic computing primarily focuses on affect detection from text. Affect detection is critical because an affect sensitive interface can never respond to users' affective states if it cannot sense their affective states. Affect detection need not be perfect, but must be approximately on target. Affect detection is, however, a very challenging problem because emotions are constructs (i.e., conceptual quantities that cannot be directly measured) with fuzzy boundaries and with substantial individual difference variations in expression and experience. To overcome such a hurdle, sentic computing builds upon a biologically inspired and psychologically-motivated affective categorization model (Cambria et al. 2012) that can potentially describe the full range of emotional experiences in terms of four independent but concomitant dimensions, whose different levels of activation make up the total emotional state of the mind.

In sentic computing, whose term derives from the Latin "sentire" (root of words such as sentiment and sentience) and sensus (intended both as capability of feeling and as commonsense), the analysis of natural language is based on affective ontologies and commonsense reasoning tools, which enable the analysis of text not only at document, page, or paragraph level, but also at sentence and clause level. In particular, sentic computing involves the use of AI and SemanticWeb techniques, for knowledge representation and inference; mathematics, for carrying out tasks such as graph mining and multi-dimensionality reduction; linguistics, for discourse analysis and pragmatics; psychology, for cognitive and affective modeling; sociology, for understanding social network dynamics and social influence; finally ethics, for understanding related issues about the nature of mind and the creation of emotional machines.

Sentic computing tackles the crucial issues of analysis of sentiments and feelings by exploiting affective commonsense reasoning, i.e., the intrinsically human capacity to interpret the cognitive and affective information associated with natural language. In particular, sentic computing leverages on a commonsense knowledge base built through crowdsourcing (Cambria et al. 2012). Commonsense is useful in many different computer-science applications including data visualization (Cambria et al. 2010), text recognition (Wang et al. 2013), and human-computer interaction (Poria et al. 2016). In this context, commonsense is used to bridge the semantic gap between word-level natural language data and the concept-level opinions conveyed by these (Cambria et al. 2015).

To perform affective commonsense reasoning (Bisio et al. 2015), a knowledge database is required for storing and extracting the semantic and affective information associated with word and multi-word expressions. By applying semantic multidimensional scaling (Cambria et al. 2015) on the matrix representation of this knowledge base, we obtain SenticNet (Cambria et al. 2016), a RDF/XML repository of natural language concepts specifically designed for sentiment analysis.

This chapter presents an overview of the most recent and advanced technologies of sentic computing, with particular focus on the applications related to the SenticNet framework. The main result consists in a review of the most interesting methods employed to compare, classify and visualize affective information. The chapter is organized as follows: Sect. 9.2 provides a description of SenticNet and sentic computing techniques; Sect. 9.3 describes several applications which employ sentic computing and the SenticNet framework; finally, Sect. 9.4 sets up conclusions and final remarks.

9.2 SenticNet

SenticNet is a publicly available resource for sentiment analysis that provides the semantics and sentics associated with 30,000 natural language concepts by leveraging on an ensemble of graph mining and multi-dimensional scaling techniques (Fig. 9.1).

The last release, SenticNet 4 (Cambria et al. 2016), exploits 'energy flows' to connect different parts of both common and commonsense knowledge representations to one another, unlike standard graph-mining and dimensionality-reduction techniques. SenticNet 4, therefore, models semantics and sentics, that is, the conceptual and affective information associated with multi-word natural language expressions. To this aim, SenticNet 4 employs an energy-based knowledge representation to provide the semantics and sentics associated with 30,000 concepts, thus enabling a fine-grained analysis of natural language opinions. SenticNet 4 contains both unambiguous adjectives as standalone entries (like 'good' and 'awful') and non-trivial multi-word expressions such as 'small room' and 'cold bed'. This is due to the fact that while unambiguous adjectives convey positive or negative polarities

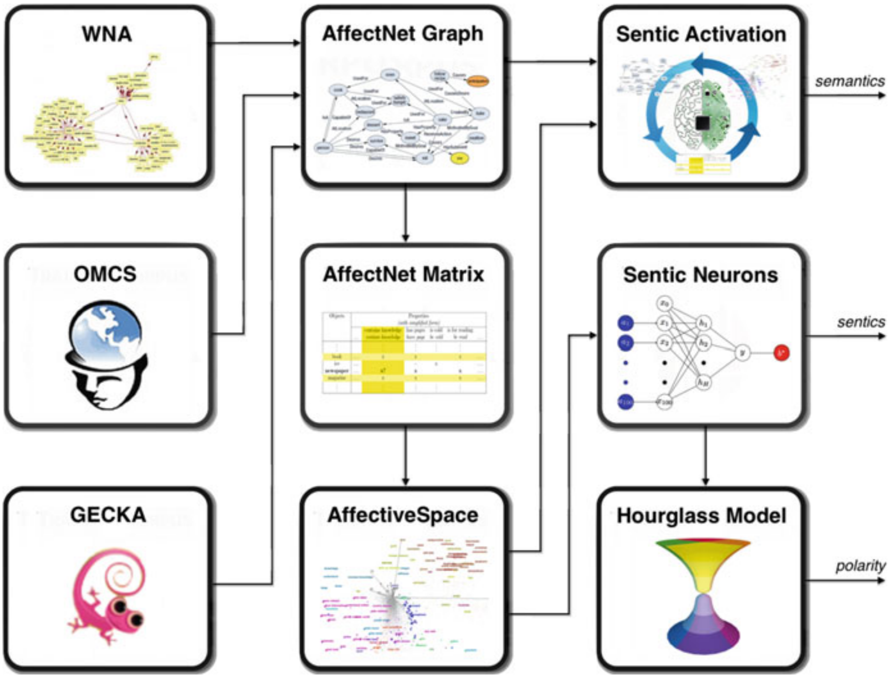


Fig. 9.1 SenticNet construction framework

(whatever noun they are associated with), other adjectives are able to carry a specific polarity only when coupled with certain nouns.

SenticNet 4 focuses on the use of 'energy' or information flows to connect various parts of common and commonsense knowledge representations to one another. Each quantum of energy possesses a scalar magnitude, a valence (binary positive/negative), and an edge history, defined as a list of the edge labels that a particular quantum of energy has traversed in the past. Essentially, common and commonsense knowledge is broken down into 'atoms', thus allowing the fusing of data from different knowledge bases without requiring any ontology alignment.

9.2.1 Knowledge Sources

SenticNet mainly leverages on the general commonsense knowledge extracted from Open Mind Common Sense (OMCS), the affective knowledge coming from WordNet-Affect (WNA) and the practical commonsense knowledge crowdsourced from a game engine for commonsense knowledge acquisition (GECKA).

OMCS (Singh 2002) is a second-generation commonsense database. It differs from previous attempts to build a commonsense database for the innovative way to collect knowledge and represent it. Knowledge, in fact, is represented in natural language, rather than using a formal logical structure, and information is not hand-crafted by expert engineers but spontaneously inserted by online volunteers. The reason why Lenat decided to develop an ad hoc language for Cyc (Lenat and Guha 1989) is that vagueness and ambiguity pervade English and computer reasoning systems generally require knowledge to be expressed accurately and precisely. However, as expressed in the Society of Mind (Minsky 1986), ambiguity is unavoidable when trying to represent the commonsense world.

WNA (Strapparava and Valitutti 2004) is an extension of WordNet Domains, including a subset of synsets suitable to represent affective concepts correlated with affective words. Similarly to the method used for domain labels, a number of WordNet synsets is assigned to one or more affective labels (a-labels). In particular, the affective concepts representing emotional state are individuated by synsets marked with the a-label emotion. There are also other a-labels for those concepts representing moods, situations eliciting emotions, or emotional responses. The resource was extended with a set of additional a-labels (termed emotional categories), hierarchically organized, in order to specialize synsets with a-label emotion. The hierarchical structure of new a-labels was modeled on the WordNet hyperonym relation.

GECKA (Cambria et al. 2015) implements a new game with a purpose (GWAP) concept that aims to overcome the main drawbacks of traditional data-collecting games by empowering users to create their own GWAPs and by mining knowledge that is highly reusable and multi-purpose. In particular, GECKA allows users to design compelling serious games for their peers to play and, while doing so, gather commonsense knowledge useful for intelligent applications in any field requiring in-depth knowledge of the real world, including reasoning, perception and social systems simulation. Besides allowing for the acquisition of knowledge from game designers, GECKA enables players of the finished games to be educated in useful ways, all while being entertained. The knowledge gained from GECKA is later encoded in SenticNet in the form <concept-relationship-concept>. The use of this natural language based (rather than logic-based) framework allows GECKA players to conceptualize the world in their own terms, at an ideal level of semantic abstraction. Players can work with knowledge exactly as they envision it, and researchers can access data on the same level as players' thoughts, greatly enhancing the usefulness of the captured data.

9.2.2 SenticNet Structure

The aggregation of common and commonsense knowledge bases is designed as a 2-stage process in which different pieces of knowledge are first translated into RDF triples and then inserted into a graph. Considering as an example 'Pablo Picasso is an artist', we obtain the RDF triple <Pablo Picasso-isA-artist> and, hence, the entry [artist - SUBSUME → Pablo Picasso].

```xml
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  ▼<rdf:Description rdf:about="http://sentic.net/api/en/concept/celebrate_special_occasion">
     <rdf:type rdf:resource="http://sentic.net/api/concept"/>
     <text xmlns="http://sentic.net/api">celebrate special occasion</text>
     <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_holiday"/>
     <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_occasion"/>
     <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_birthday"/>
     <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_wedding">
     <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/express_appreciation"/>
     <pleasantness xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.93</pleasantness>
     <attention xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.724</attention>
     <sensitivity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</sensitivity>
     <aptitude xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</aptitude>
     <polarity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.551</polarity>
   </rdf:Description>
 </rdf:RDF>
```


Fig. 9.2 A sample concept in SenticNet

In this way, we obtain a shared representation for common and commonsense knowledge, thus performing a conceptual decomposition of relation types, i.e., the unfolding of relation types that are usually opaque in natural-language-based resources.

After low confidence score trimming and duplicates removal, the resulting semantic network (built out of about 25 million RDF statements) contains 2,693,200 nodes. Of these, 30,000 affect-driven concepts (that is, those concepts that are most highly linked to emotion nodes) have been selected for the construction of SenticNet 4 (Fig. 9.2).

SenticNet 4 conceptualizes the information as 'energy' and sets up pathways upon which this energy may flow between different semantic fragments. In this way, complex concepts can be built upon simpler pieces by connecting them together via energy flows. Once an element is reached by a certain quantum of energy flow, it is included in a wider concept representation, thus enabling simple elements to deeply affect larger conceptual connections. Such a representation is optimal for modeling domains characterized by nuanced, interconnected semantics and sentics (including most socially-oriented AI modeling domains).

Each quantum of energy possesses a scalar magnitude, a valence (binary positive/negative), and an edge history, defined as a list of the edge labels that a particular quantum of energy has traversed in the past. These three elements describe the semantics and sentics of the quantum of energy and they are extracted for each concept of the semantic network.

In particular, the extraction of semantics and sentics is achieved through multiple steps of spreading activation with respect to the nodes representing the activation levels of the Hourglass of Emotions (Cambria et al. 2012), a brain-inspired model for the representation and the analysis of human emotions.

9.2.3 *The Hourglass of Emotions*

The Hourglass of Emotions is an affective categorization model developed starting from Plutchik's studies on human emotions (Plutchik 2001). The main advantage over other emotion categorization models is that it allows emotions to be decon-

Table 9.1 The sentic levels of the Hourglass model

Interval	Pleasantness	Attention	Sensitivity	Aptitude
[G(1), G(2/3))	Ecstasy	Vigilance	Rage	Admiration
[G(2/3), G(1/3))	Joy	Anticipation	Anger	Trust
[G(1/3), G(0))	Serenity	Interest	Annoyance	Acceptance
(G(0), G(-1/3)]	Pensiveness	Distraction	Apprehension	Boredom
(G(-1/3), G(-2/3)]	Sadness	Surprise	Fear	Disgust
(G(-2/3), G(-1)]	Grief	Amazement	Terror	Loathing

structed into independent but concomitant affective dimensions, whose different levels of activation make up the total emotional state of the mind. Such a modular approach to emotion categorization allows different factors (or energy flows) to be concomitantly taken into account for the generation of an affective state.

The model can potentially synthesize the full range of emotional experiences in terms of four affective dimensions, Pleasantness, Attention, Sensitivity, and Aptitude, which determine the intensity of the expressed/perceived emotion as a *float* $\in [-1, +1]$. Each affective dimension is characterized by six levels of activation, termed 'sentic levels', which are also labeled as a set of 24 basic emotions (six for each affective dimension) (Table 9.1). Previous works (Cambria et al. 2015) already proved that a categorization model based on these four affective dimensions is effective in the design of an emotion categorization architecture.

The transition between different emotional states is modeled, within the same affective dimension, using the function $G(x) = -\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$, for its symmetric inverted bell curve shape that quickly rises up towards the unit value. In particular, the function models how valence or intensity of an affective dimension varies according to different values of arousal or activation, spanning from null value (emotional void) to the unit value (heightened emotionality). Mapping this space of possible emotions leads to a hourglass shape (Fig. 9.3).

9.2.4 Sentic Patterns

Sentic patterns (Poria et al. 2015) are a novel paradigm for concept-level sentiment analysis that blends computational intelligence, linguistics, and commonsense computing in order to improve the accuracy of computationally expensive tasks such as polarity detection from big social data. The algorithm assigns contextual polarity to concepts in text and flows this polarity through the dependency arcs in order to assign a final polarity label to each sentence. Analyzing how sentiment flows from concept to concept through dependency relations allows for a better understanding of the contextual role of each concept in a text.

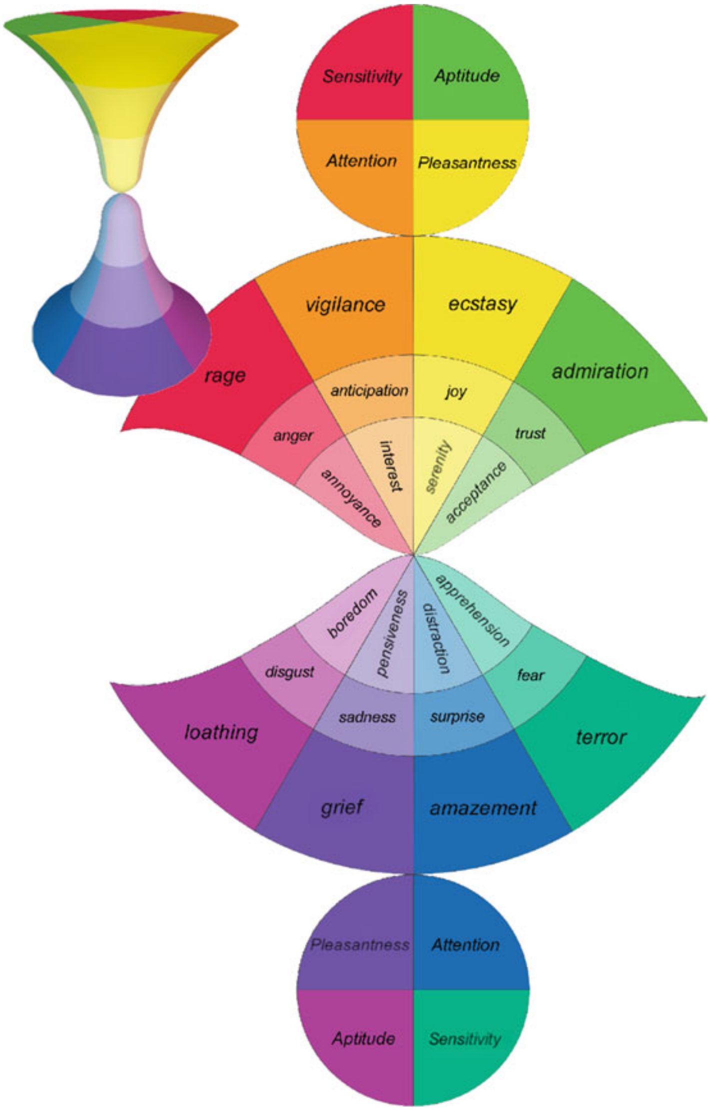


Fig. 9.3 The 3D model of the Hourglass of Emotions

The polarity detection algorithm employs SenticNet to retrieve the polarity scores of concepts. The procedure can be considered as a tree painting algorithm operating on the nodes and arcs of the syntactic dependency tree. For those words or relations (concepts, or multiword expressions) for which the polarity can be determined directly from the existing lexical resources, the algorithm assigns it directly. Then, it gradually extends the labels to other arcs and nodes, with the

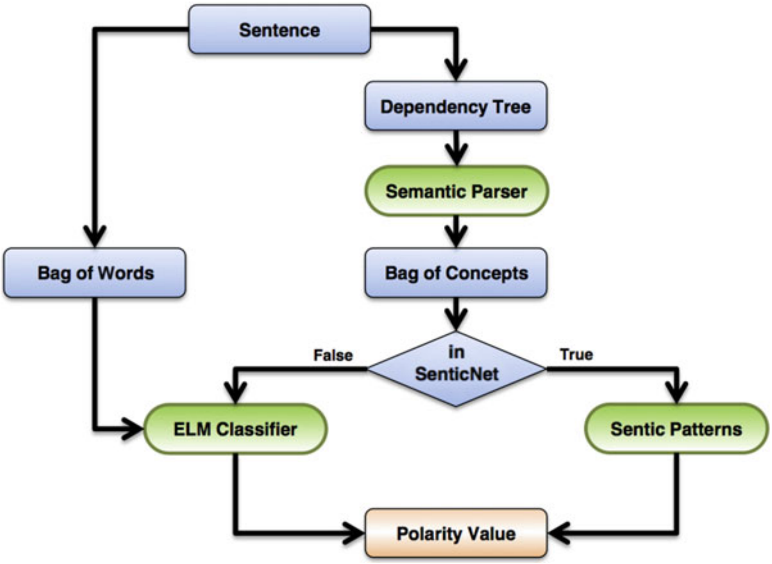


Fig. 9.4 Sentic patterns

necessary transformations determined by sentic pattern rules (Poria et al. 2014), until it obtains the final label for the root element, which is the desired output. The extending of the polarity labels is termed the flow of the sentiment.

The success of this rule-based algorithm crucially relies on the completeness of the knowledge base used, in this case, SenticNet. Namely, for the concepts that are absent in SenticNet, an ELM classifier (Cambria et al. 2013; Huang et al. 2006) is employed (Fig. 9.4).

9.3 Applications of the SenticNet Framework

SenticNet is freely available both as an API[1] and as a RDF/XML standalone resource.[2] The SenticNet framework can be tried at SenticNet demo page.[3] More advanced functionalities are available at SenticNet Ltd. website.[4] Besides many companies using SenticNet services for tasks such as brand positioning, customer relationship management, and social media marketing, there is a good number of research works exploiting it for different sentiment analysis tasks. Xia et al. (2016),

[1]http://sentic.net/api
[2]http://sentic.net/downloads
[3]http://sentic.net/demo
[4]http://business.sentic.net

for example, used SenticNet for contextual concept polarity disambiguation. In their approach, SenticNet was used as a baseline and contextual polarity was detected by a Bayesian method.

Other works Poria et al. (2012, 2014) focused on extending or enhancing SenticNet. Poria et al. (2012), for example, developed a fuzzy based SVM semi-supervised classifier to assign emotion labels to the SenticNet concepts. Several lexical and syntactic features as well as SenticNet based features were used to train the semi-supervised model. Qazi et al. (2014) used SenticNet for improving business intelligence from suggestive reviews. They built a supervised system where sentiment specific features were grasped from SenticNet.

SenticNet can also be used for extracting concepts and discover domains from sentences. This use of SenticNet was studied by Dragoni et al. (2014), who proposed a fuzzy based framework which merges WordNet, ConceptNet and SenticNet to extract key concepts from a sentence. iFeel (Araújo et al. 2014) is a system which allows its users to create their own sentiment analysis framework by combing SenticNet, SentiWordNet and other sentiment analysis methods.

SenticNet was adopted in the context of e-health to mine the opinions of patients about their experience with healthcare providers and to compare these with official ratings (Cambria et al. 2011). Some approaches (Wu et al. 2011) focused on developing the multilingual concept level sentiment lexicon using the way SenticNet was built.

SenticNet was also used to develop several supervised baseline methods (Xia et al. 2016; Duthil et al. 2012; Gezici et al. 2013). Among other supervised approaches using SenticNet, the work by Chenlo and Losada (2014) is notable. They used SenticNet to extract bag of concepts and polarity features for subjectivity and sentiment analysis tasks. Chung et al. (2014) used SenticNet concepts as seeds and proposed a method of random walk in the ConceptNet to retrieve more concepts along with polarity scores. Their method indeed aimed to expand SenticNet containing 265,353 concepts. After expanding SenticNet they formed Bag-of-Sentimental-Concepts features which is similar to Bag of Concepts features. Each dimension in the feature vector represents a concept and each concept was assigned a value by multiplying tf-idf and polarity value of the concept. SenticNet has also been adopted for enhancing Twitter sentiment classification accuracy. The approach by Bravo-Marquez et al. (2014) used both SenticNet and SentiWordNet to improve the baseline Twitter classification system. SenticNet was also used for informal short text message (SMS) classification (Gezici et al. 2013) and within a domain independent unsupervised sentiment analysis system termed Sentilo (Recupero et al. 2014).

The SenticNet framework is optimized for binary polarity classification on sentences in formal English. However, the system can be applied also to document-level sentiment classification and micro-text analysis (as shown in the next two sections, respectively).

9.3.1 Document-Level Sentiment Analysis

An example of how the SenticNet framework can be adapted to document-level classification is provided by Bisio et al. (2016), a work that aims to study and identify the best similarity metric able to describe the sentiment distribution of several types of books, establishing a different point of view on the interpretation of feeling extraction: the classification of documents based on an emotional distance.

In particular, Bisio et al. (2016) employed a text miner application (Meda et al. 2015), in which the word *'document'* is used to denote any source of data able to carry information, e.g., text written in natural language, web pages, images (Bisio et al. 2013). The tool normalizes input documents into an internal representation and applies several metrics to compute distances between pair of documents; the document-distance used takes into account a conventional content-based similarity metric, a stylistic similarity criterion and a semantic representation of the documents, in order to apply machine learning algorithms (Oneto et al. 2016) for both clustering and classification purposes.

After a pre-processing phase, in which language identification, stemming and stopword removal steps are carried out, a text document becomes a *'docum object'*, deprived of useless information (e.g., articles, prepositions, punctuation, special characters). At this level the 'SenticNet semantic descriptor' is applied.

The SenticNet framework allows one to retrieve four different sentiment experiences associated with a specific word; then, the aim is the development of a sentiment semantic descriptor made up of a vector of four affective dimensions (Pleasantness, Attention, Sensitivity and Aptitude). Thus, the 'SenticNet Semantic Descriptor' extracts the list of words that compose the document and submits each single word to SenticNet. After the semantic descriptor step, the distance between two document can be calculated.

In order to test the approach, Bisio et al. (2016) selected books between five distinct literary genres and applied three different distance metrics (Manhattan, Euclidean and Maximum norm). The experiments underline the fact that it is possible to notice a similarity between different literary genres, because, from an affective point of view, even though different novels may be set in different environments, mentality and social constraints, they can still convey similar types of feelings.

9.3.2 Micro-text Sentiment Analysis

Supervised learning classifiers often misclassify tweets containing conjunctions like 'but' and conditionals like 'if', due to their special linguistic characteristics. Moreover, tweets often contain misspelled words, slangs, URLs, elongations, repeated punctuations, emoticons, abbreviations and hashtags. To tackle such

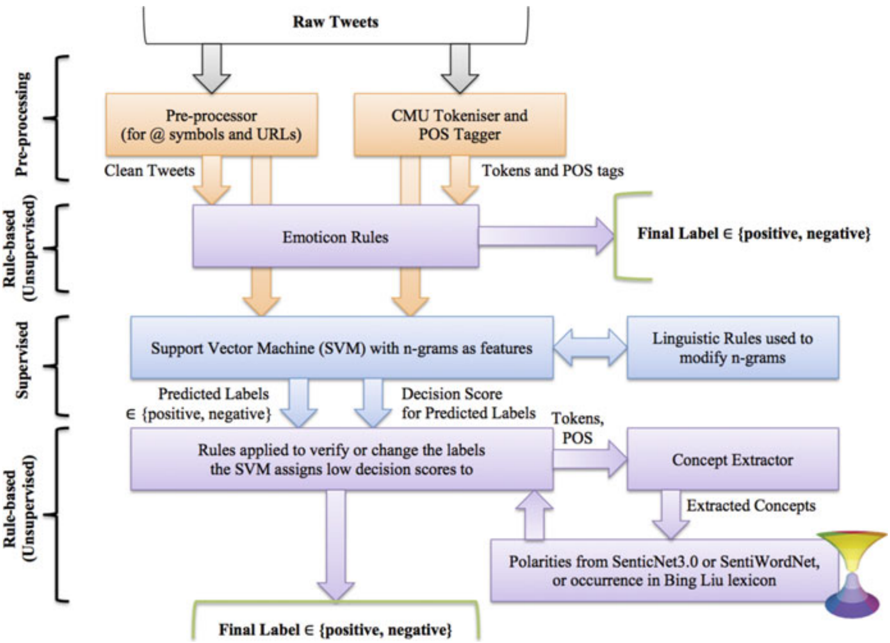


Fig. 9.5 Flowchart of the Twitter sentiment analysis system

challenges, the SenticNet framework can be adopted to enhance supervised learning for polarity classification (Chikersal et al. 2015). The general scheme of the system is presented in Fig. 9.5. This system first considers the number of positive and negative emoticons of the tweet and the following rules are applied:

- If a tweet contains one or more positive emoticons and no negative emoticons, it is labeled as *positive*.
- If a tweet contains one or more negative emoticons and no positive emoticons, it is labeled as *negative*.
- If neither one of the two rules above can be applied, the tweet is labeled as *unknown*.

If these emoticon-based rules label a tweet as *positive* or *negative*, this is considered the final label outputted by the system. Otherwise, all tweets labeled as *unknown* are passed into a supervised learning classifier.

To this end, each tweet is represented as a feature vector of case-sensitive n-grams (unigrams, bigrams, and trigrams). These n-grams are frequencies of sequences of 1, 2 or 3 contiguous tokens in a tweet. After handling negation, all tweets containing the conjunction ‘but’ and the conditionals ‘if’, ‘unless’, ‘until’, and ‘in case’ are considered, and specific linguistic rules are formulated in order to enable removal of irrelevant or oppositely oriented n-grams from the tweet’s feature vector.

Finally, a SVM classifier is trained in order to obtain the tweet's label. For tweets with an absolute decision score or confidence below 0.5, the class labels assigned by SVM are discarded and an unsupervised classifier is employed. The rules used by this classifier are based on a linguistic analysis of tweets, and leverage on sentiment analysis resources that contain polarity values of words and phrases; the primarily resource used for this purpose is SenticNet.

This unsupervised classification process works as follows:

1. Single-word and multi-word concepts are extracted from the tweets in order to fetch their polarities from SenticNet.
2. If a single-word concept is not found in SenticNet, it is queried in SentiWordNet (Esuli and Sebastiani 2006), and if it is not found in SentiWordNet, it is searched in the list of positive and negative words from the Bing Liu lexicon (Liu et al. 2005).
3. Based on the number of positive and negative concepts, and the most polar value occurring in the tweet, the following rules are applied:

   - If the number of positive concepts is greater than the number of negative concepts and the most polar value occurring in the tweet is greater than or equal to 0.6, the tweet is labeled as positive.
   - If the number of negative concepts is greater than the number of positive concepts and the most polar value occurring in the tweet is less than or equal to -0.6, the tweet is labeled as negative.
   - If neither one of the two rules stated above can be applied, the tweet is labeled as unknown by the rule-based classifier, and the SVM's low confidence prediction is taken as the final output of the system.

9.4 Conclusion

With the advent of the Social Web, the way people express their views and opinions has dramatically changed. Reviews, forums and blogs now represent huge sources of information with many practical applications. However, finding opinion sources and monitoring them can be a formidable task because there are a large number of diverse sources and each source may also have a huge volume of opinionated text. Thus, automated opinion discovery and summarization systems are needed.

Due to its tremendous value for practical applications, there has been an explosive growth of sentiment analysis techniques in both research in academia and applications in the industry. However, most of the existing approaches still rely on syntactical structure of text, which is far from the way human mind processes natural language.

This chapter showed how sentic computing techniques can be employed for the development of several sentiment analysis tasks. In order to assess the capability of sentic computing to tackle real-world NLP tasks, we considered several applications in different domains and different text formats.

All such applications demonstrate how SenticNet represents a useful resource for the analysis of social data, as it goes beyond the use of domain-dependent keywords by using an ensemble of commonsense computing tools and linguistics.

References

Araújo, M., P. Gonçalves, M. Cha, and F. Benevenuto. 2014. iFeel: A system that compares and combines sentiment analysis methods. In *WWW*, 75–78.

Bisio, F., P. Gastaldo, C. Peretti, R. Zunino, and E. Cambria. 2013. Data intensive review mining for sentiment classification across heterogeneous domains. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 1061–1067. IEEE.

Bisio, F., P. Gastaldo, R. Zunino, and E. Cambria. 2015. A learning scheme based on similarity functions for affective common-sense reasoning. In *IJCNN*, 2476–2481.

Bisio, F., C. Meda, P. Gastaldo, R. Zunino, and E. Cambria. 2016. Sentiment-oriented information retrieval: Affective analysis of documents based on the senticnet framework. In *Sentiment analysis and ontology engineering*, Studies in Computational Intelligence, eds. W. Pedrycz and S.-M. Chen, vol. 639, 175–195.

Bravo-Marquez, F., M. Mendoza, and B. Poblete. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems* 69: 86–99.

Cambria, E., J. Fu, F. Bisio, and S. Poria. 2015. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In *AAAI*, Austin, 508–514.

Cambria, E., P. Gastaldo, F. Bisio, and R. Zunino. 2015. An ELM-based model for affective analogical reasoning. *Neurocomputing* 149: 443–455.

Cambria, E., G.B. Huang, et al. 2013. Extreme learning machines. *IEEE Intelligent Systems* 28(6): 30–59.

Cambria, E., and A. Hussain. 2015. *Sentic computing: A common-sense-based framework for concept-level sentiment analysis*. Cham: Springer.

Cambria, E., A. Hussain, and C. Eckl. 2011. Bridging the gap between structured and unstructured health-care data through semantics and sentics. In *WebSci*, Koblenz.

Cambria, E., A. Hussain, C. Havasi, and C. Eckl. 2010. SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space. In *Knowledge-based and intelligent information and engineering systems*, ed. R. Setchi, I. Jordanov, R. Howlett, L. Jain, Lecture Notes in Artificial Intelligence, vol. 6279, 385–393. Berlin: Springer.

Cambria, E., A. Livingstone, and A. Hussain. 2012. The hourglass of emotions. In *Cognitive behavioral systems*, ed. A. Esposito, A. Vinciarelli, R. Hoffmann, V. Muller, Lecture Notes in Computer Science, vol. 7403, 144–157. Berlin/Heidelberg: Springer.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: *COLING*, 2666–2677.

Cambria, E., S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi. 2015. The CLSA model: A novel framework for concept-level sentiment analysis. In *LNCS*, vol. 9042, 3–22. Springer.

Cambria, E., D. Rajagopal, K. Kwok, and J. Sepulveda. 2015. GECKA: Game engine for commonsense knowledge acquisition. In *FLAIRS*, 282–287.

Cambria, E., Y. Xia, and A. Hussain. 2012. Affective common sense knowledge acquisition for sentiment analysis. In *LREC*, Istanbul, 3580–3585.

Chenlo, J.M., and D.E. Losada. 2014. An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences* 280: 275–288.

Chikersal, P., S. Poria, E. Cambria, A. Gelbukh, and C.-E. Siong. 2015. Modelling public sentiment in Twitter: Using linguistic patterns to enhance supervised learning. In *CICLing*, 49–65.

Chung, J.K.C., C.E. Wu, and R.T.H. Tsai. 2014. Improve polarity detection of online reviews with bag-of-sentimental-concepts. In *Proceedings of the 11th ESWC*, Semantic web evaluation challenge. Crete: Springer.

Dragoni, M., A.G. Tettamanzi, and C. da Costa Pereira. 2014. A fuzzy system for concept-level sentiment analysis. In *Semantic web evaluation challenge*, 21–27. Springer.

Duthil, B., F. Trousset, G. Dray, J. Montmain, and P. Poncelet. 2012. Opinion extraction applied to criteria. In *Database and expert systems applications*, 489–496. Springer.

Esuli, A., and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *LREC*

Gezici, G., R. Dehkharghani, B. Yanikoglu, D. Tapucu, and Y. Saygin. 2013. Su-sentilab: A classification system for sentiment analysis in twitter. In *International Workshop on Semantic Evaluation*, 471–477.

Huang, G.B., Q.Y. Zhu, and C.K. Siew. 2006. Extreme learning machine: Theory and applications. *Neurocomputing* 70(1): 489–501.

Huang, G.B., E. Cambria, K.A. Toh, B. Widrow, and Z. Xu. 2015. New trends of learning in computational intelligence. *IEEE Computational Intelligence Magazine* 10(2):16–17.

Lenat, D., and R. Guha. 1989. *Building large knowledge-based systems: Representation and inference in the Cyc project*. Boston: Addison-Wesley.

Liu, B., M. Hu, and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *WWW*, Chiba.

Lutz, C., and G. White. 1986. The anthropology of emotions. *Annual Review of Anthropology* 15: 405–436.

Meda, C., F. Bisio, P. Gastaldo, R. Zunino, R. Surlinelli, E. Scillia, and A.V. Ottaviano. 2015. Content-adaptive analysis and filtering of microblogs traffic for event-monitoring applications. In *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, vol. 1, 155–170. Springer.

Minsky, M. 1986. *The society of mind*. New York: Simon and Schuster.

Oneto, L., S. Ridella, and D. Anguita. 2016. Tikhonov Ivanov and Morozov regularization for support vector machine learning. *Machine Learning* 103: 103–136

Osgood, C., W. May, and M. Miron. 1975. *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

Plutchik, R. 2001. The nature of emotions. *American Scientist* 89(4): 344–350.

Poria, S., E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4): 26–36.

Poria, S., I. Chaturvedi, E. Cambria, and A. Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *ICDM*, 439–448.

Poria, S., E. Cambria, G. Winterstein, and G.B. Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69: 45–63.

Poria, S., A. Gelbukh, E. Cambria, D. Das, and S. Bandyopadhyay. 2012. Enriching SenticNet polarity scores through semi-supervised fuzzy clustering. In *IEEE ICDM*, Brussels, 709–716.

Poria, S., A. Gelbukh, E. Cambria, A. Hussain, and G.B. Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69: 108–123.

Qazi, A., R.G. Raj, M. Tahir, E. Cambria, and K.B.S. Syed. 2014. Enhancing business intelligence by means of suggestive reviews. *The Scientific World Journal* 2014: 1–11.

Recupero, D.R., V. Presutti, S. Consoli, A. Gangemi, and A. Nuzzolese. 2014. Sentilo: Frame-based sentiment analysis. *Cognitive Computation* 7(2): 211–225.

Scherer, K. 1993. Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion* 7: 325–355.

Singh, P. 2002. The open mind common sense project. KurzweilAI.net.

Strapparava, C., and A. Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *LREC*, Lisbon, 1083–1086.

Turkle, S. 1984. *The second self: Computers and the human spirit*. New York: Simon & Schuster.

Vesterinen, E. 2001. Affective computing. In *Digital media research seminar*. Helsinki.

Wang, Q., E. Cambria, C. Liu, and A. Hussain. 2013. Common sense knowledge for handwritten Chinese recognition. *Cognitive Computation* 5(2): 234–242.

Wu, H.H., A.C.R. Tsai, R.T.H. Tsai, and J.Y.J. Hsu. 2011. Sentiment value propagation for an integral sentiment dictionary based on commonsense knowledge. In *2011 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, 75–81. IEEE.

Xia, R., F. Xu, J. Yu, Y. Qi, and E. Cambria. 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing and Management* 52: 36–45.

Index

E. Cambria et al. (eds.), *A Practical Guide to Sentiment Analysis*,
Socio-Affective Computing 5, DOI 10.1007/978-3-319-55394-8