ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/108465, December 2016* *ISSN (Online) : 0974-5645*

Sentiment Analysis: A Comprehensive Overview and the State of Art Research Challenges

J. Rexiline Ragini[1*] and P. M. Rubesh Anand[2]

[1]Department of Computer Applications, Hindustan University, Chennai – 603103, Tamil Nadu, India; rexilineragini@gmail.com
[2]Department of Electronics and Communication Engineering, Hindustan University, Chennai – 603103, Tamil Nadu, India; rubesh.anand@gmail.com

Abstract

Objectives: To extract the knowledge from social media and other review sites of importance. **Methods/Analysis:** Analyzing such a huge amount of data to summarize the opinion out of that text is a hot research field. In this study, a systematic literature review is done to summarize the various works that are carried out in this field. **Findings:** In this survey, the various methods used for sentiment analysis, its applications and the challenges are summarized in order to give an overall view of sentiment analysis. **Novelty/Improvement:** This valuable information is used in evaluating the opinion that could be used by business organizations and other text mining entities.

Keywords: Big Data, Opinion Polls, Sentiment Analysis, Social Media

1. Introduction

The big data from micro blogging sites attracts many communities to explore the hidden content to get valuable information out of it. Sentiment analysis is one such research area which concentrates on identifying subjective information from a given piece of text. Opinions that are expressed in social media serve as a major input for detecting public outlook across various areas such as buying products, predicting the share market and movie reviews. Such web generated contents play a major role in mining user sentiments for customer relationship management and public opinion tracking. Sentiment analysis is basically a natural language processing technique that uses computational linguistics and text mining to identify the polarity of the text as positive, negative and neutral. Sentiment analysis is defined as the automated knowledge discovery technique which identifies the hidden patterns in reviews, blogs and tweets[1].

In recent days, companies have started to use sentiment analysis as part of their research. Apart from the data received from social networking sites, companies create their own web sites to gather review about their products. Mining these reviews, they are able to build better customer relationship and also create recommendation systems with the help of the positive and negative feedback from customers. Another advantage of sentiment analysis is that the companies are able to develop their marketing strategies by predicting public attitude towards their product. Numerous companies have already developed tools that crawl online information and summarize that information in graphical representation of the recent trends[2].

Sentiment analysis is broadly classified into three categories namely, document level, sentence level and aspect based. The goal of document level sentiment classification is determining the overall sentiment of a given review document[3]. Sentence level[4] analysis focuses on categorizing the text at the level of subjective and objective nature. Aspect based[5] approach is more pinpointed as it splits the entire document into various

** Author for correspondence*

aspects (entities) and sentiment analysis is carried out on each entity to find out the overall polarity.

A detailed study[6] has been carried out on the various techniques for sentiment analysis. The study reviewed the recent work that has been carried out with various techniques. It also discussed about some of the feature selection methods and related fields to sentiment analysis. A detailed survey[7] on the various applications and challenges in sentiment analysis is presented. In [8] has written a survey that covers the latest trends in sentiment analysis. The works discussed above deals with sentiment analysis from the evolution to the multimodal sentiment analysis.

Rest of this paper is divided into various sections. Section 2 briefs some of the common sentiment analysis tasks, section 3 gives an overview of sentiment analysis across various domains, section 4 discusses the various challenges in this field and section 5 concludes the research.

2. Common Sentiment Analysis Tasks

The various tasks involved in sentiment analysis are subjectivity detection, feature selection in sentiment classification and sentiment classification.

2.1 Subjectivity Detection

Subjectivity detection is the process of identifying the subjective sentences. Sentences can be classified as subjective sentence and objective sentence. Subjectivity indicates that the text contains/bears opinion content whereas objectivity indicates that the text is without opinion content. For example, "*This movie is superb*." is a subjective sentence since it has an opinion as it talks about the movie and the writer's feeling about the same. "*Fruits are good for health*" is the sentence that is a fact, general information rather than an opinion or a view of some individual and hence its objective. Sentence level sentiment analysis deals with the process of subjectivity detection. Various approaches like bootstrapping[9,10], Conditional random fields[11], Viterbi algorithm[12] are used for subjectivity detection. SVM (sequential minimal optimization algorithm with poly kernel) is used for classification[13]. Objective words from SentiWordNet are used to improve the sentiment classification.

2.2 Features for Sentiment Classification

Feature engineering is one of the basic and most important steps in sentiment classification. The English sentences should be converted into feature vector in order to perform sentiment classification. The most commonly used features are Term presence and frequency[14], n-gram[15,16], Negation, Adjectives, Adverb-Adjective combination, Gini index[17]. Feature selection methods are divided into two categories they are Lexicon-based and Statistical based. Some of the statistical feature selection methods are Point-wise Mutual Information (PMI)[18,19] Chi-square[20] and Latent Semantic Indexing (LSI).

2.3 Sentiment Classification

The sentiment classification techniques are broadly classified into three categories as Machine learning methods, lexicon based approach and Hybrid approach. Machine learning approach deals with the machine learning algorithms to solve the sentiment analysis problem. Machine learning techniques are broadly classified as Supervised[21,22] and Unsupervised[23,24] algorithms. Machine learning algorithms are widely used for sentiment analysis problems, some of them are Naïve Bayes classifier[25], Support Vector Machine (SVM)[26–28], Neural network[29], Conditional random fields (CRF)[30–32] and Rule based classifier[33,34]. Some of the approaches in Lexicon based approach are Dictionary based[35,36] and Corpus based[37,38]. Hybrid approach is in its early stage and not much work has been done in the topic.

3. Study of Sentiment Analysis Application Across Various Domains

Sentiment analysis is a hot research topic and various works has been done in various domains. Few of them are interpreting public sentiment variation[39], classifying customer reviews as positive and negative, detecting internet hotspots[40], and predicting stock market behavior.

3.1 Movie Reviews

Sentiment analysis has been extensively carried out for Movie review. The analysis has greater impact on the success of the movie as in recent days people watch movies that have got good reviews. The data is taken from benchmark datasets like, IMDB, rottentomatoes.com. Few of the works that are carried out in this domain shows positive results[41–45].

3.2 Product Reviews

Sentiment analysis is mostly used by the marketing companies to increase the sales of their products. Sentiment analysis has been carried out for many products like iPhone, cameras, hardware components, printers and scanners. Apart from just products, many works are carried out for restaurant reviews. Various aspects of the restaurant like food, services have been reviewed. The data for review is mainly taken from social networking sites like twitter, Face book and from other review sites created by the respective companies[46–49].

3.3 Stock Market

Sentiment analysis is more useful in the stock market to predict the performance of shares. The data is collected from Yahoo Finance discussion board and other networking sites. In general, the shares are categorized into five categories and weights are assigned to them accordingly. The five categories are (2) for "Strong Buy", (1) for "Buy", (0) for "Hold", (-1) for "Sell", (-2) for "Strong sell"[50].

3.4 Crime Analysis

A preliminary work[51] has been carried out in predicting crime with sentiment analysis techniques. In the research work, the author has carried out spacio temporal mining to identify the crimes that are happening in various fields. Linguistic analysis and statistical topic modeling is used to automatically identify discussion topics across a major city in the United States, and then incorporated them in the crime prediction model.

3.5 Disaster Recovery

A number of works has been carried out in analyzing the mood of the people during crisis and disasters. Few of the works include, analyzing how social networking sites are used during disasters. Such analyses are helpful in reaching out people in need and help them. Voluntary organizations can read the data and render help to people who are in need. Some of the disasters that are analyzed are earthquakes, typhoons[52].

4. Challenges

Sentiment analysis is a growing field; still there are many research challenges that need to be addressed. Some of the open challenges in text mining are summarized as follows.

- Negation[53] is very important because negation changes the text polarity. Negation terms affect the contextual polarity of words but the presence of a negation word in a sentence does not mean that all of the words conveying sentiments are inverted. Negation is not only conveyed by common negation words (not, never, no) but also by other lexical units.
- Another major hurdle is the handling of anaphora resolution. Anaphora means referring to same meaning but with different phrases. This problem mainly occurs while grouping the entities in aspect based sentiment analysis. For example, "battery life" and "power usage" refer to the same aspect of a phone, sentiments about both of these aspects should be combined in order to produce accurate results.
- Word arrangement in a sentence plays a vital role in identifying the subjective nature of the text. Word order is important in deciding the polarity of a text. In a given piece of text, if the words order changes, the polarity of the text gets affected.
- **Implicit sentiment and Sarcasm:** Without the presence of any sentiment bearing words, sentences may have an implicit sentiment. For example, "How can you do this?" In this sentence, none of the words express negative opinion, but the meaning of the sentence is negative. Thus identifying semantics is very important in semantic analysis.
- **Spam Detection:** Anyone from any location can express their views in social media without disclosing their true identity. By this way, many fake reviews are written in order to promote the sales of the product. Such an activity is called opinion spamming. Apart from individuals, there are also commercial companies that are into this business spreading fake information. It is a challenging task to identify such opinion spams to extract the exact sentiment.
- **Conjunctions:** Presence of conjunctions in a sentence changes the entire meaning of the sentence. For example: "The restaurant was very nice, but the service was poor". This sentence is split into two parts. When we analyze the first part, we get a

positive sentiment. But the presence of the other words reverses the entire meaning of the sentence. So conjunctions should be considered for sentiment analysis.

- Co-reference resolution is one of the biggest research challenges. This has to be done in both aspect level and entity level. This is more applicable in places where comparative texts are used. The reference between the sentences must be effectively resolved in order to produce better analysis. For example, consider the following opinionated text, "Comparing Nikon's Cool pix to its main competitor the Canon, it takes excellent photos and is quite compact". In the above sentence, the pronoun "it" refers to 'Nikon Cool pix'. If this co-reference is not identified correctly, sentiment analysis cannot be carried out effectively[54].
- Domain adaptation is another important aspect of sentiment analysis. Most of the available sentiment lexicons are general-purpose; though these are general-purpose, it is important that to study the ways for adapting to a specific domain. In this regard, there are three main issues. First is the same entity term that has different polarity in different domains. The next issue is assigning a strength marker for each and every sentiment word. Third is the difference in vocabularies across different domains which make sentiment analysis a domain dependent application.

5. Conclusion and Future work

This paper has presented a brief survey on various aspects of sentiment analysis. Naïve Bayes classifier and Support vector machines are the most commonly used methods for sentiment classification. Most of the researches concentrate on English language and they can be extended to other languages to understand the regional trend. In sentiment classification, machine learning and Lexicon based approach are the most widely used approaches but hybrid approach needs to be explored further for better results. Since sentiment analysis is a domain dependent problem, there are only a few domains in which work has been carried out and there are a quiet lot of domains that need to be explored.

6. References

1. Mostafa MM. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications. 2013; 40(10):4241–51.
2. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems. 2013; 28 (2):15–21.
3. Moraes R, Valiati JF, Neto WPG. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications. 2013; 40(2):621–33.
4. Alexandre T, Alias F. Sentence-based sentiment analysis for expressive text-to-speech. IEEE Transactions on Audio, Speech, and Language Processing. 2013; 21(2):223–33.
5. Zheng-Jun Z, Yu J, Tang J, Wang M, Chua TS. Product aspect ranking and its applications. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(5):1211–24.
6. Walaa M, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014; 5(4):1093–113.
7. Bo P, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 2(1–2):1–135.
8. Ronen F. Techniques and applications for sentiment analysis. Communications of the ACM. 2013; 56(4):82–9.
9. António S, Oliveira PHG, Ramos C, Marques NC. A bootstrapping algorithm for learning the polarity of words. Computational Processing of the Portuguese Language, Springer Berlin Heidelberg; 2012. p. 229–34.
10. Svitlana V, Wilson T, Yarowsky D. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Bulgaria; 2013. p. 1–6.
11. Ryan M, Hannan K, Neylon T, Wells M, Reynar J. Structured models for fine-to-coarse sentiment analysis. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic; 2007. p. 432–9.
12. Tetsuji N, Inui K, Kurohashi S. Dependency tree-based sentiment classification using CRFs with hidden variables, Human Language Technologies. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California; 2010. p. 786–94.
13. Chihli H, Lin HK. Using objective words in SentiWordNet to improve word-of-mouth sentiment classification. IEEE Intelligent Systems. 2013; 28(2):47–54.
14. Yelena M, Srinivasan P. Exploring feature definition and selection for sentiment classifiers. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Spain; 2011. p. 546–9.
15. Ahmed A, France S, Zhang Z, Chen H. Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(3):447–62.
16. Cao J, Zeng K, Wang H, Cheng J, Qiao F, Wen D, Gao Y. Web-based traffic sentiment analysis: Methods and appli-

cations. IEEE Transactions on Intelligent Transportation Systems. 2014; 15(2):844–53.

17. Charu CA, Zhai CX. Mining text data. Springer Science and Business Media; 2012. p. 123.
18. Chih YL, Wu JL, Chang PC, Chu HS. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowledge-based Systems. 2013; 41:89–97.
19. Alena N, Aono M. Sentiment word relations with affect, judgment, and appreciation. IEEE Transactions on Affective Computing. 2013; 4(4):425–38.
20. Michael H, Liebmann M, Neumann D. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems. 2013; 55(3):685–97.
21. Tetsuya N, Yi J. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA; 2003. p. 70–7.
22. Bo P, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing- Association for Computational Linguistics. 2002; 10:120–36.
23. Pan XP, Jin HL, Shi HX, Chen W. An unsupervised sentiment information identification approach. Applied Mechanics and Materials. 2013; 263:3330–4.
24. Shamery A, Salih E, Gheni HQ. Plagiarism detection using semantic analysis. Indian Journal of Science and Technology. 2016; 9(1):1–20.
25. Hanhoon K, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2012; 39(5):6000–10.
26. Pérez RV, Mihalcea R, Morency LP. Multimodal sentiment analysis of Spanish online videos. IEEE Intelligent Systems. 2013; 3:38–45.
27. Eiji A, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, UK; 2011. p. 1568–76.
28. Ahmed A, Chen H, Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS). 2008; 26(3):12.
29. Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications. 2013; 40(16):6266–82.
30. Shabnam S, Moghaddam S. Fine-grained opinion mining using conditional random fields. IEEE 11th International Conference on Data Mining Workshops (ICDMW), Vancouver, Canada; 2011. p. 109–14.
31. Yejin C, Cardie C, Riloff E, Patwardhan S. Identifying sources of opinions with conditional random fields and extraction patterns. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, Canada; 2005. p. 355–62.
32. Jian J, Zhou Y. Sentiment polarity analysis based multi-dictionary. Physics Procedia. 2011; 22:590–6.
33. Li S, Wang Z, Lee SYM, Huang CR. Sentiment classification with polarity shifting detection. Proceedings of the International Conference on Asian Language Processing (IALP), Urumqi, China; 2013. p. 129–32.
34. Zhang Y, Jiang Y, Tong Y. Study of sentiment classification for Chinese microblog based on recurrent neural network. Chinese Journal of Electronics. 2016; 25(4):601–7.
35. Charng-Rurng TA, Wu CE, Tsai RTH, Hsu JYJ. Building a concept-level sentiment dictionary based on commonsense knowledge. IEEE Intelligent Systems. 2013; 2:22–30.
36. Alena N, Prendinger H, Ishizuka M. SentiFul: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing. 2011; 2(1):22–36.
37. Yanyan Z, Qin B, Liu T. Creating a fine-grained corpus for Chinese sentiment analysis. IEEE Intelligent Systems. 2015; 30(1):36–43.
38. Albert W, Gindl S, Scharl A. Extracting and grounding context-aware sentiment lexicons. IEEE Intelligent Systems. 2013; 28(2):39–46.
39. Shulong T, Li Y, Sun H, Guan Z, Yan X, Bu J, Chen C, He X. Interpreting the public sentiment variations on twitter. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(5):1158–70.
40. Yanyan Z, Qin B, Liu T, Tang D. Social sentiment sensor: A visualization system for topic detection and topic sentiment analysis on microblog. Multimedia Tools and Applications; 2014. p. 1–18.
41. Cheon NJ, Thet TT, Khoo CSG. Comparing sentiment expression in movie reviews from four online genres. Online Information Review. 2010; 34(2):317–8.
42. Thura TT, Na JC, Khoo SG. Sentiment classification of movie reviews using multiple perspectives. Universal and Ubiquitous Access to Information, Springer Berlin Heidelberg; 2008. p. 184–93.
43. Xiaohui Y, Liu Y, Huang X, An A. Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Transactions on Knowledge and Data Engineering. 2012; 24(4):720–34.
44. Apoorv A, Balasubramanian S, Zheng J, Dash S. Parsing screenplays for extracting social networks from movies. Proceedings of the 3rd Workshop on Computational Linguistics for Literature (EACL), Gothenburg, Sweden; 2014. p. 50–8.
45. Anurag M, Kelkar K. Sentiment analysis on movie reviews based on combined approach. International Journal of Science and Research. 2014; 3(7):1739–42.
46. Garcia-Moya L, Anaya-Sanchez H, Berlanga-Llavori R. Retrieving product features and opinions from customer reviews. IEEE Intelligent Systems. 2013; 3:19–27.
47. Zhu J, Wang H, Zhu M, Tsou BK, Ma M. Aspect-based

opinion polling from customer reviews. IEEE Transactions on Affective Computing. 2011; 2(1):37–49.

48. McGuire M, Kampf C. Using social media sentiment analysis to understand audiences: A new skill for technical communicators? Proceedings of the IEEE International Professional Communication Conference (IPCC), Limerick, Ireland; 2015. p. 1–7.
49. Wollmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, Louis-Philippe M. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems. 2013, 28 (3):46-53.
50. Wu DD, Zheng L, David LO. A decision support approach for online stock forum sentiment analysis. IEEE Transactions on Systems, Man, and Cybernetics. 2014; 44(8):1077–87.
51. Xiaofeng W, Gerber MS, Brown DE. Automatic crime prediction using events extracted from twitter posts. Social Computing, Behavioral-Cultural Modeling and Prediction, LNCS 7227, Springer Berlin Heidelberg; 2012. p. 231–8.
52. Caragea C, Squicciarini A, Stehle S, Kishore N, Tapia A. Mapping moods: Geo-mapped sentiment analysis during hurricane Sandy. Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA; 2014 May. p. 642–51.
53. Wiegand M, Balahur A, Roth B, Klakow D, Montoyo A. A survey on the role of negation in sentiment analysis. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, Uppsala, Sweden; 2010. p. 60–8.
54. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-based Systems. 2015; 89:14–46.