ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A Study on Various Techniques and Challenges in Sentiment Analysis

[1.] P. Bavithra Matharasi
Lecturer, Dept. of MCA,
Mount Carmel College,
Bangalore.
bavithramatharasi@gmail.com

[2.]N. Vijayalaksmi
Lecturer, Dept. of MCA,
Mount Carmel College,
Bangalore.
vijinitya@yahoo.com

[3.] Dr. A.Senthilrajan
Director
Computer Centre
Karaikudi
agni_senthil@yahoo.com

Abstract: ***Social media dataset are unstructured in nature and provide various challenges and opportunities for researchers across different domains. The rich source of information available in social media has inclined businesses. Various analysis are carried out to derive valuable insight from social media dataset. Whether it is researchers finding new algorithm or industries increasing their profit, Sentiment analysis is the fastest growing field. There are numerous applications in performing sentiment analysis. This paper provides an overview of different techniques used to perform sentiment analysis and how it will be more effective when the analysis is semantically interpreted. The limitations and the future challenges to the researchers is also discussed.***

Keywords: Social media; sentiment analysis; machine learning;

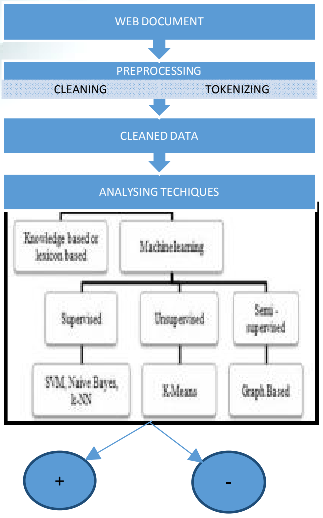
I.INTRODUCTION

Text mining is a specialized domain that applies data mining techniques over text.Some of the early works on text mining were extraction, classification, cleaning and analyzing [1].Sentiment Analysis (SA) aims to identify and extract opinions, moods and attitudes of individuals and communities. Authors in the paper [2] provided a technical survey and early work on sentiment analysis. Sentiment analysis or opinion mining is the field of research that started much early in 1990. It was first introduced in the year 2003 [3][4]. Social media has become a platform for users to share their views, ideas, emotions, comments. They offer their feedback and critiques by posting their comments online. They can also post reviews and opinions about a product, services, policies, celebrities or current news through personal blogs and social networks. There are numerous social media platforms such as Twitter, Facebook and Google+ available in the market today [5] [6] [7]. Analyzing Twitter data is considered much more complex in comparison to the general sentiment analysis task as it offers many challenges like size limit of 140 characters on tweets, usage of informal language, slangs, and emoticons to express opinions. Studies on Twitter data have demonstrated that aggregating millions of messages can provide valuable insights into a population [8].

II.APPLICATIONS OF SENTIMENT ANALYSIS

Since there are variety of dataset available in social media, the applications or the insight acquired is also vast. Applications of sentiment analysis include crime surveillance [9].Sentiment analysis of social media data has also been applied for tracking disease outbreaks; the authors of [10] described a method for extracting tweets for early warning and outbreak detection during a Swine Flu pandemic demonstrating a strong contribution for alerting relevant stake holders for prompt action. The authors of [11] applied text mining techniques to investigate consumer attitudes towards global brands, they also reported that Twitter can be used as a reliable method in analyzing attitudes towards global brands.

III.ARCHITECTURE OF SA



Sentiment analysis is not a single step method,there are various steps in the process of analysing. The web document initially may contain raw and noisy data. Further this has to be tokenized into basic words. This phase is called pre-processing. Various tools are used for this purpose. Lexicon-based, Machine learning based algorithms are used upon these data to finally categorize the opinions.Lexicon based techniques are also called as Knowledge based technique. Machine learning algorithms may be supervised or semi-supervised.This is shown in figure 1.

IV LEVELS OF SENTIMENT ANALYSIS

Though there are different datasets available in social media, we consider only text based data for our survey. Basically there are three different levels of sentiment analysis is carried out for textual data.

A. Document Level

The document level sentiment analysis aims at classifying the entire document as positive or negative [2][12]. In this type of analysis, For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This level of analysis assumes that each document expresses opinions on a single entity. Thus, it is not applicable to documents which evaluate or compare multiple entities.

B. Sentence level

The task at this level goes to the sentences and determine whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is called *subjectivity classification*[13][14], in this level of analysis it distinguishes the sentences (called *objective sentences*) that express factual information from the sentences (called *subjective sentences*) that express subjective views and opinions. This is clearly given in the paper [15].

C. Entity and aspect level

In this level, the lexicon or words are considered as entity. The entity level granular analysis takes into consideration of each opinion expressed in the content. The opinions on the reports, entities and/or their various aspects, blog, and the trial are considered for positive and negative sentiment [16].

IV.TECHNIQUES IN SENTIMENT ANALYSIS

Text based sentiment analysis can be classified from different point of views. The technique used for analysing, the view of text, the required level of detail of text analysis and the rating level. But from a technical point of view, we can classify as below

A. Knowledge Based

This is also known as lexicon-based approach. In this approach, the overall sentiment of a document is calculated as the average aggregate of the semantic orientations (polarity) of its words and phrases[12]. This approach requires construction of dictionary of words i.e. lexicon annotated with the words polarity that tells whether a word belongs to negative or positive class. The semantic orientation of the words in the lexicon is calculated prior to the sentiment analysis work[17]. The sentiment lexicon can either be created manually or expanding automatically from seed of words[18][19]. There are also open lexicons available to support sentiment analysis task, for example WordNet, SentiWordNet[20][21]. Lexicon-based methods are commonly used techniques, but the performance of such systems are limited by semantic ambiguity[22]. For instance, Rao et al. developed an algorithm with three pruning strategies to automatically build a word-level emotional dictionary for social emotion detection[23][24].

B. Machine Learning

This techniques treats the problem of sentiment analysis as text classification task[2], and different machine learning methods are used to build classifiers and train them on the available datasets using various extracted features like term presence, term frequency, unigrams, bigrams, n-grams and Part Of Speech-POS tags, micro blogging features etc.[25]. A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) are few to mention. Although Machine-learning algorithms can outperform simple lexicon-based methods[26][27], they require large training databases to be effective[26][28]. For most real world social media contexts which involve huge datasets, it is difficult to obtain the effective size of a sufficient training dataset because the diversity of the social discussion is often not known *a priori*. Rui and Whinston trained a support vector machine (SVM) with a training dataset and the precision for detecting positive, negative, and neutral tweets are 75%, 65% and 75%, respectively [29]. Feldman noted that many of the commercial sentiment analysis systems continue to use simplistic techniques and their performance leaves a lot to be desired[30]. Although there are reports of hybrid methods, which combined rule based classification, supervised learning and machine learning[31], they suffer from the same limitations as machine learning methods, including the issue on insufficient training data to be effective. Three machine learning based classifiers, Naive-Bayes, Maximum Entropy and Support Vector Machines and a hybrid technique, label propagation, were investigated in [32] and a number of issues that need to be addressed before using thesetechniques for sentiment classification were outlined.

V. SEMANTICS AND SENTIMENT

Sentiment analysis rely on scoring sentiment words based on their *polarity i.e.* their position on a scale between positive and negative. But accurate sentiment analysis should not be limited at only keywords or individual words, whereas the semantics or context of analysis should also be considered. This is called as semantic sentiment analysis. Semantics is the study of relation of words, phrases, signs and their denotation.Semantic sentiment analysis is the level that allows for semantic interpretation.Users can search and navigate by semantic roles, exploring sentences and documents by the functional role played by each concept[33].

VI. CHALLENGES IN SENTIMENT ANALYSIS

1. The majority of existing methods are tested using limited annotated corpus. Such datasets can be a cleaned sample that consists of either positive or negative text, or manually filtered and annotated. This is different from real-world sentiment classification task (e.g., on tweets), which may also contain spam, advertisement, and bipolar texts.
2. Existing methods have to deal with the requirements of topic domain-specific adaptation and training database. It is important to develop high accuracy domain-specific adaptation methods that can function without the need of customized training datasets.
3. There is a lack of sentiment analytics for non-English languages. This research gap is evident and hence multilingual sentiment analysis is to be considered very importantly.
4. Detection of spam and fake reviews: The web contains both authentic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer[39].
5. Sarcastic sentences,text may have sarcastic and ironic sentences. For example, “What a great car, it stopped working in the second day.” In such case, positive words can have negative sense of meaning. Sarcastic or ironic sentences can be hard to identify which can lead to erroneous sentiment analysis.
6. Requirement of Universal Knowledge about text, facts, events, people are often required to correctly classify the text.Consider the following example, “Casablanca and a lunch comprising of rice and fish: a good Sunday” The system without world knowledge classifies above sentence as positive due to the word “good”, but it is an objective sentence because Casablanca is the name of the famous movie.
7. Often people use abbreviations, lack of capitals, poor spelling, poor punctuation, poor grammar and icons. Example Time – tym, Come – cme, ☺ - happy

VII. CONCLUSION

Social media has become the platform for users to share their knowledge, view, opinion, comment or criticism. They post their comments and opinion about anything, whether it is product or a celebrity or a crisis. They even share their emotions, which reflects their attitude and behaviour. This rich set of data paved the way for industries and researchers to bring the insight on sentiment analysis. In this paper, we have presented the different techniques of sentiment analysis and their limitations.In future, we plan to refine the sentiment analysisthrough semantic way of interpretation.

REFERENCE

[1] K. Dégardina, Y. Roggoaand and P. Margot, "Understanding and fighting the medicine counterfeit market," Journal of Pharmaceutial and Biomedical Analysis, vol. 87, pp. 167-175, January 2014.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, January 2008.

[3] Nasukawa, Tetsuya and Jeonghee Yi, 2003, Sentiment analysis: capturing favourability using natural language processing, Proceedings of the K-CAP03, 2nd International Conference on knowledge capture.

[4] Dave et al, 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 2003, 519-528.

[5] M. Naaman, “Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications,” Multimedia Tools and Applications, vol. 56, no. 1, pp. 9-34, May 2010.

[6] W.-ying S. Chou, Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse, “Social media use in the United States: implications for health communication.,” Journal of medical Internet research, vol. 11, no. 4, p. e48, Jan. 2009.

[7] M. Salathé and S. Khandelwal, “Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control,” PLoS computational biology, vol. 7, no. 10, p. e1002199, Oct. 2011

[8] M.J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for Public Health," in 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, 2011.

[9] M.D. Sykora, T.W. Jackson, A. O'Brien, and S. Elayan, "National Security and Social Media Monitoring: A Presentation of the EMOTIVE and Related Systems," in 2013 European Intelligence and Security Informatics Conference (EISIC), Uppsala, 2013, pp. 172-175.

[10] E. de Quincey and P. Kostkova, "Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter," in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Istanbul, Turkey: Springer Berlin Heidelberg, 2010, ch. 3, pp. 21-24.

[11] M. Mostafa Mohamed, "More than words: Social networks’ text mining for consumer brand sentiments," Expert Systems with Applications, vol. 40, no. 10, pp. 4241-4251, August 2013.

[12] P.D. Turney, 2002, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), 417–424

[13] Riloff, E &Wiebe, J., 2003, Learning extraction patterns for subjective expressions, EMNLP‟03.

[14] Loren Terveen et al, 1997, PHOAKS: A system for sharing recommendations, Communications of the Association for Computing Machinery (CACM), 40(3):59–62.

[15] Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? Finding strong and weak opinion clauses, in Proceedings of National Conference on Artificial Intelligence (AAAI-2004). 2004.

[16] Mejova, Y. “Sentiment analysis: An overview,” Comprehensive exam paper, available on http: //www. cs. uiowa. edu/˜ymejova/publications/CompsYelenaMejova. pdf, 2009.

[17] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. “Lexicon-based methods for sentiment analysis.” Computational linguistics,volume 37, number2, 267–307, MIT Press, 2011

[18] Kanayama, H., and Nasukawa, T. “Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis.” Proceedings of the 2006

[19] Saif, M., and Turney, P. “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, pages 26–34, Los Angeles, CA, 2010.

[20] Strapparava, C., and Valitutti, A. “WordNet-Affect: an affective extension of WordNet.” Proceedings of LREC 1083–1086, 2004.

[21] Bruno, O. “Opinion Minning with the SentWordNet Lexical Resource.” Dublin Institute of Technology. 2009.

[22] A. Balahur, R. Mihalcea, and A. Montoyo, “Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications,” Computer Speech & Language, vol. 28, no. 1, pp. 1-6, Jan. 2014.

[23] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, “Building emotional dictionary for sentiment analysis of online news,” World Wide Web, vol. 17, pp. 723-742, Jun. 2014.

[24] V. Loia and S. Senatore, “A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content,” Knowledge-Based Systems, vol. 58, pp. 75-85, Mar. 2014.

[25] Forman, G. “An extensive empirical study of feature selection metrics for text classification”. The Journal of Machine Learning Research, 3, pp. 1289-1305, 2003.

[26] Z. Wang, V. J. C. Tong, and H. C. Chin, “Enhancing machine learning methods for

sentiment classification of Web data," in 10th Asia Information Retrieval Society Conference, 2014.

[27] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," Association for Computational linguistics, vol. 35, no. 3, 2009.

[28] S. Kim, N. W. Cho, B. Kang, and S.-H. Kang, "Fast outlier detection for very large log data," Expert Systems with Applications, vol. 38, no. 8, pp. 9587-9596, Aug. 2011

[29] H. Rui and A. Whinston, "Designing a social-broadcasting-based business intelligence system," ACM Transactions on Management information Systems, vol. 2, no. 4, p. 22:1, 2011.

[30] R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, vol. 56, no. 4, p. 82, Apr. 2013

[31] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," Journal of Informetrics, vol. 3, no. 2, pp. 143-157, Apr. 2009.

[32] S. Bhuta, A. Doshi, U. Doshi, and M. Narvekar, "A review of techniques for sentiment analysis Of Twitter data," in International Conference on Issues and Challenges in Intelligent Computing Techniques, Ghaziabad, 2014, pp. 583-591.

[33] Neri, F., Geraci, P.: "Mining Textual Data to boost Information Access in OSINT", Open Source Intelligence and Web Mining Symposium, 13th International Conference on Information Visualization, IV09, Proceedings, ISBN: 978-0-7695-3733-7, pp. 427-432, IEEE Computer Society, Barcelona (ES), 16-17/07/2009.