STUDY OF ALGORITHMS TO RECTIFY CHALLENGES OF SENTIMENT ANALYSIS

Mily Lal[1], Akanksha Goel[2], Abha Jain[3]
Dr D Y Patil Institute of Engineering Management and Research

Abstract
Sentiment Analysis is an emerging field of research in opinion mining field. It is the computational treatment of opinions, sentiments and subjectivity of text. This paper focuses on a study of various algorithms that can be used for efficient sentiment mining.
Index Terms: Opinion mining, Sentiment analysis, Text Mining.

I. INTRODUCTION

The increase in use of internet and online services led to the importance of analysis of the huge amount of structured and unstructured data. This can be done with the help of various Data Mining, Web Mining and Text Mining techniques. The ever increasing amount of Customer reviews requires an efficient approach to analyze and generate opinion summary [1]. Researchers and academicians are working on various algorithms and technologies to accomplish easy analysis of sentiments. Sentiment analysis can be defined as a computational study of opinions, sentiments, emotions, and attitude expressed in text [2]. It evolves around opinion mining which is the task of detecting, extracting and classifying opinions expressed in textual input [3]. Opinion mining is widely used in applications like market intelligence [4], customer satisfaction [5] etc

II. LITERATURE SURVEY

Pang and Lee [6] throws light on over three hundred papers by covering applications, common challenges for sentiment analysis, and major tasks of opinion mining. Tang et al. [7] discussed about the issues related to opinion mining. NB classifier, Multiple NB classifier, and cut-based classifier can be used for subjectivity classification. O'Leary [8] paper includes techniques for blog mining. Tsytsarau and Palpanas [9] paper focuses on spam detection and contradiction analysis. Comparison of various opinion mining methods was also put forth in their paper. Liu [10] tries to showcase different works using in sentiment analysis and opinion mining. Cambria et al. [11] was successful in laying out the complexities involved in opinion mining and possible future research directions. Most recently, Medhat et al. [12] is survey on feature selection and sentiment classification methods.

This paper proposes a study on various algorithms used in the existing studies on opinion mining.

III. PREPROCESSING

Data corpus collected can be highly related to the type, data format of the source and the type of analysis that is to be performed.

The data obtained needs to be cleaned and preprocessed. Preprocessing steps generally usually done are the following: tokenization - breaks a sentence into words/ phrases, stop word removal – words that do not contribute to analysis phase are to be discarded, stemming- it is the process of bring the word to its root form, parts of speech (POS) tagging [13] – tagging of token on basis of their parts of speech, and feature extraction- selecting the relevant tokens. [14]. Negations-negative words have to be treated separately.

IV. FEATURE EXTRACTION

Opinion words are identified to be adjective, adverb, verb, and noun. These words majorly contribute to the task of sentiment analysis. Different methods have been used in various studies for the extraction of opinion words like machine learning based [15], lexicon based [16] and other hybrid methods[17]. Hu and Liu [10]

had proposed a technique to extract features based on association rule mining. Zhan et al. approach based on frequent word sequences shows better results. Wang and Lee [16] proposed a dictionary based approach like WordNet. Li et al. [6] proposed a Double Propagation-Based Linear Regression with Rules method to rank the product aspects based on the opinions.

Pang Lee *et al.* [6] used supervised learning in sentiment analysis with the aim of determining whether it could be treated as a special case of topic-based categorization with positive and negative topics. Naıve Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) classifiers were tested to achieve this, with all performing well in topic-based categorization. Document words and symbols were used for features as either a unigram or a bigram bag-of features. Unigram features performed better than bigram features. Feature Frequency (FF) and Feature Presence (FP) when tested revealed that by using a SVM with unigram FP better accuracy could be achieved (82.9%) in a 3-fold cross validation.

SentiWordNet (SWN) is a lexical resource of sentiment information for terms in the English language introduced in [15] designed to assist in opinion mining tasks.

PMI (Pointwise Mutual Information) is used in many applications for selecting relevant features. PMI are two different ways of measuring the correlation between terms and categories. Chi-square is better than PMI as it is a normalized value; therefore, these values are more comparable across terms in the same category [18]. LSI is one of the famous feature transformation methods [19]. LSI method transforms the text space to a new axis system which is a linear combination of the original word features. Principal Component Analysis techniques (PCA) are used to achieve this goal [20].

V. SUMMARIZATION

Summarization leads to the final step of sentiment analysis. Opinions from a single customer will not give promising results. It is necessary to analyze opinions from a larger crowd. Summarization of opinions found on aspects and their entities can be of both quantitative and qualitative nature. Aspect based summary is one of the commonly used summarization methods [10]. Opinions are summarized `based on a particular aspect. The user can get the quantitative analysis about sentiments each aspect. It is also possible to drill down to get the actual opinions of the review sentences.

Opinion summary with a timeline helps the user get the exact knowledge of opinions about a target and also helps in future analysis like figuring out what changes people's opinions . Visualization of summaries and ranking of opinionated sentences to show the strongest opinion about an aspect gives a clear knowledge about the opinion

statistics of the target.[10]

VI. CONCLUSION

This paper proposes a study of various algorithms that are used in each stage of Sentiment analysis. This paper will help in selecting and improving the existing Sentiment analysis algorithms. Hence improve opinion mining results in various fields like business intelligence, evaluation of customer satisfaction and many more.

REFERENCES

[1] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis SupportSyst 2011;50:755–68.

[2] C. W. Medhat et al., Sentiment analysis algorithms and applications: a survey, Ain Shams Eng. J. (2014), http://dx.doi.org/10.1016/j.asej.2014.04.011 .

[3] A. Balahur, Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types, PhD Thesis, University of Alicante, Spain,2011, 273p

[4] Y.M. Li, T.-Y. Li, Deriving market intelligence from microblogs, Decis. Support Syst. 55 (2013) 206–217.

[5] D. Kang, Y. Park, Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach, Expert Syst. Appl.

(2013), http://dx.doi.org/10.1016/j.eswa.2013.07.101.

[6] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval 2 (2008) 1–135.

[7] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews, Expert Syst. Appl. 36 (2009) 10760–10773.

[8] D.E. O'Leary, Blog mining-review and extensions: ''From each according to his opinion'', Decis. Support Syst. 51 (2011) 821–830.

[9] E. Cambria, C. Havasi, A. Hussain, SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis, Proc. 25th Int'l Florida Artificial Intelligence Research Society Conf., AAAI, 2012, pp. 202–207.

[10] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool publishers, 2012 (May).

[11] E. Cambria, B. Schuller, Y.-Q. Xia, New avenues in opinion mining and sentiment analysis (extended abstract), in: Proceedings of IJCAI, Buenos Aires 2015.

[12] W. Medhat et al., Sentiment analysis algorithms and applications: a survey, Ain Shams Eng. J. (2014), http://dx.doi.org/10.1016/j.asej.2014.04.011

[13] M. Collins, Discriminative training methods for hidden Markov models:theory and experiments with perceptron algorithms, in: Proc. EMNLP, 2002.

[14] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M.A. Goncalves, W. Meira Jr., Word co-occurrence features for text classification, Inform. Syst. 36 (2011) 843–858.

[15] H. Liu, J. He, T. Wang, W. Song, X. Du, Combining user preferences and user opinions for accurate recommendation, Electron. Commer. Res. Appl. 12(2013) 14–23.

[16] J.H. Wang, C.C. Lee, Unsupervised opinion phrase extraction and rating in Chinese blog posts, in: IEEE Third International Conference on Privacy,

[17] Z. Zhai, H. Xu, B. Kang, P. Jia, Exploiting effective features for Chinese sentiment classification, Expert Syst. Appl. 38 (2011) 9139–9146.

[18] Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data.Springer New York Dordrecht Heidelberg London: _ Springer Science+Business Media, LLC'12; 2012.

[19] Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R. Indexing by latent semantic analysis. JASIS 1990;41:391–407.

[20] Jolliffee IT. Principal component analysis. Springer; 2002.