Opinion Mining and Sentiment Analysis – A Review

Haseena Rahmath P[1] Dr.Tanvir Ahmad[2]
[1,2]Department of Computer Science and Engineering
[1]Al-Falah School of Engineering, Dhauj, Haryana, India. [2]Jamia Millia Islamia, New Delhi, India

Abstract--- The exponential growth of internet usage has created a new platform where people can freely communicate and exchange ideas and opinions. Review web sites, social Medias, blogs, discussion forums, community websites, etc are encouraging users to share their sentiments and opinion towards products and services. Most of these opinions are expressed in unstructured nature. Prevailing these factors such as availability of heterogeneous content and increased expectation of the customers, gives birth to a new area of research called opinion mining. Opinion mining and sentiment analysis is an extension of text mining that extract the sentiments present in the documents and analyze its polarity orientation towards positive, negative and neutral. Both supervised and unsupervised classification approach can be used for sentiment classification. This paper reviews various work carried out in this area.

Keywords: Opinion Mining, Sentiment Analysis, Supervised learning, Unsupervised learning, Machine Learning, POS Tagger

I. INTRODUCTION

In decision making process, people's opinions and experience have very important role as it help to make right decision. Today several internet sites are encouraging users to share their opinion and experience of various online products. This tendency in fact, helps the users to consider other's opinion and experience while purchasing any product. And it also gives a platform for the users to compare the competing brands of the products. The increasing popularity of blogs, social-medias and the internet itself resulted a huge collection such reviews and taking right decision by analyzing those document become a challenging task. These factors necessitate an automated tool that automatically extract and analyze individual opinion from web documents.

The reviews are written in natural language, and most of them have no pre-defined format. The data can be extracted from the unstructured natural language web documents using text mining technique, an extension of data mining technique, but it will not give the sentiments of the people opinions. So a more advanced technique is required to extract the opinions contained in the documents. This opened a new area of research called Opinion Mining and Sentiment Analysis.

In last few years, many researches are focusing on this area. Researches mainly concentrate on fetching data directly from the internet sources such as product reviews, movie reviews, blog entries, forum entries, comments on news, etc. and analyze it automatically using different natural language processing techniques and machine learning methods.

In this context the present paper attempts to cover some popular techniques and approaches that have been used in this field of research. At first, definition of important terms used in this area are defined and further several problems related to opinion mining and sentiment analysis are discussed and finally the important works carried out to solve these problem are presented.

II. OPINION MINING AND SENTIMENT ANALYSIS

Sentiment Analysis is a Natural Language Processing and Information Extraction technique that tries to find out writer's feelings or opinion about some aspect expressed in comments, reviews, questions and requests or the overall contextual polarity of a document, by analyzing a large number of documents. The sentiment may be one's mood, attitude, feeling, judgment or evaluation about some topic. The extraction and analysis of sentiments from web, a huge repository of structured and unstructured data, is a challenging task.

III. SENTIMENT CLASSIFICATIONS

Sentiments can be classified at the word, sentence, or document level. Word-level sentiment classification determines the sentiment polarity of each word towards some aspects and sentence-level sentiment classification determines the sentiment polarity of entire sentence towards some aspects. In document-level sentiment classification, it determines the sentiment polarity of whole document. In other words, it determines whether each document,'d' expresses a positive or negative opinion with respect to some topic.

The prediction of whether the opinion/review is positive or negative is based on the analysis of text in the review document. Most of the customer reviews target a single object such as in the case of product reviews that usually focus on single product, a movie review explains what he/she felt about a particular movie and a restaurant review contains his feeling or experience with the restaurant. Hence the document level sentiment classification takes an assumption that each document contains opinions on a single object.

IV. FEATURE SELECTION METHODS

The key task in sentiment analysis is extracting features from the documents. Widely used feature extraction methods are the following:

(1) Opinion words: most of the opinion words are adjectives and adverb, but sometime nouns and verbs too express opinion. For example, good, fantastic, amazing, bad and boring are all adjective or adverb which express emotions while rubbish as noun, hate and like as verb express emotions in the document.
(2) Terms and their frequency: single words or word n-grams consider for feature. It also considers their frequency of occurrence.
(3) Part of speech (POS) information: Sentiment words can be easily identified by using POS information.
(4) Syntactic dependency: Using a dependency tree word dependency based features are selected.

(5) Negations: Negation words are very important to determine the polarity of the sentence. The Negation altogether change the sentiment of the sentence. For example, the sentence "this phone is not good" has negative orientation.

V. SENTIMENT CLASSIFICATION TECHNIQUES

One of the main tasks in sentiment analysis is text classification. Two types of classification techniques are commonly used in sentiment classification, namely supervised classification and unsupervised classification.

A. Supervised Methods

Supervised Learning approach is machine learning approaches that clarify the sentiments based on training and test sets. In supervised learning, a number of machine learning algorithms can be used to classify text. The major focuses of supervised learning have 3 aspects: constructing appropriate feature spaces, assignment of feature values and choosing appropriate classification algorithms. These factors are very important in classification performance.

Table1 shows some previous studies conducted in supervised document-level sentiment classification and their details are introduced next.

Table 1: Some previous studies conducted in supervised document-level sentiment classification

Paper	Technique	Feature	Dataset
Whitelaw et al. (2009)[1]	SVM	Adjective word, frequency, percentage of appraisal groups	Movie review
Ye et al. (2009)[2]	SVM, Naïve Bayes, character Based N-gram model	Unigram Frequency	Travel destination reviews
Pang et al. (2002)[3]	SVM, Naïve Bayesian,	Unigrams, bigrams, adjective, position of words	Movie review
Prabowo and Thelwall(2009) [4]	SVM, Rule based Classifier	POS tag, Ngrams	movie reviews, product reviews, MySpace comments

the sentiments. To extract features from the document they used various feature extraction framework such as unigrams, bigrams, adjective and position of words. To filter out unimportant features they considered only those features that appear four times in the document in the case of unigrams and in case of bigrams occurrence is at least seven times. They applied SVM, Naïve Bayesian, and Maximum Entropy to the feature spaces they constructed. They found that three machine learning methods outperformed the manual classifications and they got better result when using unigrams with SVM classifier.

Prabowo and Thelwall (2009) [4] considered a hybrid approach for sentiment analysis. They applied a series of classifiers sequentially until an acceptable accuracy is obtained. First they applied one classifier and checked the result if it is not satisfactory then system will passes on to next classifier and so on, until no more classifiers to passes on. In order to achieve this, they integrated rule-based classification with supervised learning and machine learning techniques.

Whitelaw et al. (2009) [1] conducted a study on sentiment analyses using movie review data. It was a document-level supervised learning and they applied SVM classification algorithm for text classification and for feature space construction, bag-of-words classification with shallow parsing and classification of attitude types of words from appraisal theory [11] are used. A manually constructed attitude lexicon also considered for better result. They achieved an accuracy of 90.2%.

Ye et al. (2009) [2] conducted sentiment classification techniques on travel destination review domain. They applied three supervised learning techniques namely support vector machine, NB and the character based N-gram model for review classification. For constructing feature set the information gain (IG) method was used and they used the frequency of words instead of word presence. They compared the result of three supervised machine learning algorithm applied on their study and found SVM outperforms the other two classification algorithms.700 training reviews used as training set and they got an accuracy around 86% .

Pang et al. (2002) [3] used supervised sentiment classification method to solve the sentiment analysis problem. Movie reviews are used as data set and three machine learning algorithm are used to classify

Authors used three different rules from existing research for rule-based classification. Support Vector Machine classifier is used for sentiment classification. They conducted their research on a three domains such as movie reviews, product reviews and MySpace comments. The System achieved an accuracy range of 72.77% to 90% depending up on the testing corpus.

B. Unsupervised Methods

Similar to supervised learning, unsupervised learning is also used often for sentiment analysis. Unsupervised Learning approach is semantic orientation approach to opinion mining as it does not require prior training data sets. Unsupervised learning involves the calculation of the opinion polarities of opinion words, and classifies the documents or sentences by aggregating the orientation of opinion words. Several works have been done in sentiment classification using unsupervised learning methods based on sentiment words and phrases. Table 2 shows some previous studies conducted in unsupervised document-level sentiment classification and their details are introduced next.

Table 2: Selected previous studies conducted in unsupervised document-level sentiment classification

Paper	Technique	Feature	Dataset
Turney(2002) [5]	PMI-IR	adjectives and adverbs	Automobile, bank, movie, travel reviews
Harb et al. (2009)[6]	Association Rule	adjectives and adverbs	Movie review
Taboada et al. (2002)[7]	Dictionary based approach	Unigrams, bigrams, adjective, position of words	Movie Review Camera Review Epinions
Samuel Brody and Noemie Elhadad (2010) [8]	label propagation algorithm	adjectives	Restaurant reviews

Turney(2002)[5] conducted sentiment classification using a simple unsupervised learning algorithm. It classified the entire reviews into two categories: recommended review or not recommended review. He utilized words' point-wise mutual information (PMI) concept to determine whether words has positive orientation or negative orientation. The feature sets are constructed by scanning each review for phrases that match certain part of speech patterns (adjectives and adverbs). Each such phrase's semantic orientation is calculated, and to get semantic orientation of a whole review, take the sum of all the semantic orientation of phrases. The study yielded 74% accuracy in sentiment classification.

Harb et al. [6] performed their sentiment analysis on movie review. They started the classification with the two sets of seed words having positive and negative semantic orientations, and created association rule to find out more seed words using Google's search. The sum of positive adjectives and the sum of negative adjectives are used to classify the documents. They got separate accuracy for identifying negative and positive document. For positive document they

achieved 71% while for negative documents they got only 62%.

Taboada et al. [7] presented a lexicon-based method for sentiment word extraction. They used dictionary based approach for classification. In such approach, dictionaries of positive or negative polarized words are utilized for classification task. Based on these dictionaries and intensifiers and negation of words, a semantic orientation calculator (SO-CAL) was constructed. They conducted their experiment on movie review dataset. They achieved accuracy between 59.6% and 76.4% while conducting experiment on 1900 documents of movie review dataset.

Samuel Brody and Noemie Elhadad [8] developed a fully unsupervised system for sentiment analysis and opinion mining. They extracted all adjectives from the documents for feature space. They used label propagation algorithm. They evaluated their system on restaurant reviews. Their system outperformed the manual sentiment classification.

VI. Conclusion

This review paper discussed some of the supervised and unsupervised classification approaches that are widely used in Sentiment Analysis. The paper explained related work conducted using these two approaches in detail along with the accuracy obtained in each approach. Since the Opinion Mining and Sentiment Analysis is most emerging field in data mining, more researches are necessary to achieve maximum accuracy in sentiment classification. People usually express their emotions in short sentences and in abbreviations, hence the research should be carried out in that direction as well. Anyone can put anything in internet, so the sentiment analysis should be integrated with some spam handling technique too.

References

[1] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," presented at the Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005.

[2] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications, vol. 36, pp. 6527-6535, 2009.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques", presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002.

[4] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach" , Journal of Informetrics, vol. 3, pp.143-157, 2009.

[5] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews" , presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002.

[6] A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, Cergy-Pontoise, France, 2008.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M.Stede, "Lexicon-based methods for sentiment analysis", Comput. Linguist. , vol. 37, pp. 267-307, 2011.

[8] Samuel Brody and Noemie Elhadad(2010)." An Unsupervised Aspect-Sentiment Model for Online Reviews", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 804–812, Los Angeles, California, June 2010.

[9] M. K. Dalal and M. A. Zaveri "Semisupervised Learning based Opinion Summarization and Classification for Online Product Reviews", Applied Computational Intelligence and Soft Computing, Article ID 910706, 2013.

