A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES

[1] MR. S. M. VOHRA, [2] PROF. J. B. TERAIYA

[1, 2] Department Of Computer Engineering And IT, Marwadi Education Foundation Group Of Institutes, Rajkot, Gujarat, India

saifeevohra@gmail.com, jay.teraiya@gmail.com

ABSTRACT: *The growth of social web contributes vast amount of user generated content such as customer reviews, comments and opinions. This user generated content can be about products, people, events, etc. This information is very useful for businesses, governments and individuals. While this content meant to be helpful analyzing this bulk of user generated content is difficult and time consuming. So there is a need to develop an intelligent system which automatically mine such huge content and classify them into positive, negative and neutral category. Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP). The objective of this paper is to discover the concept of Sentiment Analysis in the field of Natural Language Processing, and presents a comparative study of its techniques in this field.*

Keywords— Natural Language Processing, Sentiment Analysis, Sentiment Lexicon, Sentiment Score.

1. INTRODUCTION

Human decision making is always influence by others thinking, ideas and opinions. The growth of social web contributes large amount of user generated content such as comments, reviews and opinions about products, services and events. This content is useful for consumers as well as manufacturer. While making any purchase online consumer usually check opinions of others about the product. Manufacturer can gain insight into its products strength and weaknesses based on the sentiment of the customers. Though these opinions are helpful for both business organizations and individuals, the huge amount of such opinionated text data becomes overwhelming to users. How to analyze and summarize the opinions expressed in these huge opinionated text data is a very interesting domain for researchers. This new research domain is usually called Sentiment Analysis or Opinion Mining.

Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's often referred to as subjectivity analysis, opinion mining, and appraisal extraction [11]. Customers want to see the opinion of other about a product before buying it. Business organizations want to know what customers are saying about their product or service that an organizations is providing, to make future decisions. It would provide powerful functionality for voice of customer and brand reputation management.

Main fields of research in Sentiment analysis are Subjectivity Detection, Sentiment Prediction, Aspect Based Sentiment Summarization, Text summarization for Opinions, Contrastive Viewpoint Summarization, Product Feature Extraction, detecting opinion spam. Subjectivity Detection is a task of determining whether text is opinionated or not. Sentiment Prediction is about predicting the polarity of text whether it is positive or negative. Aspect Based Sentiment Summarization provides sentiment summary in the form of star ratings or scores of features of the product. Text Summarization generates a few sentences that summarize the reviews of a product. Contrastive Viewpoint Summarization puts an emphasis on contradicting opinions. Product Feature Extraction is a task that extracts the product features from its review. Detecting opinion spam is concern with identifying fake or bogus opinion from reviews.

Sentiment classification can be done at Document level, Sentence level and Aspect or Feature level. In Document level the whole document is classify either into positive or negative class. Sentence level sentiment classification classifies sentence into positive, negative or neutral class. Aspect or Feature level sentiment classification concerns with identifying and extracting product features from the source data.

There are two main approaches for sentiment analysis: machine learning based and lexicon based. Machine learning based approach uses classification technique to classify text. Lexicon based method uses sentiment dictionary with opinion words and match

them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are.

The objective of this paper is to discover the concept of Sentiment Analysis in the field of Natural Language Processing and presents a comparative analysis of its techniques in this field. The paper is organized as follows: Section 2 provides the overview of the most commonly used techniques in Sentiment analysis. Section 3 discusses the analysis and comparison of sentiment analysis techniques. Section 4 concludes the manuscript.

2. SENTIMENT ANALYSIS TECHNIQUES

There are two main techniques for sentiment analysis: machine learning based and lexicon based. Few research studies have also combined this two methods and gain relatively better performance.

1) Machine learning based techniques

The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification. In a machine learning based techniques, two sets of documents are needed: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to check how well the classifier performs. 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) have achieved great success in sentiment analysis. Machine learning starts with collecting training dataset. The next step is to train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is feature selection. They can tell us how documents are represented. The most commonly used features in sentiment classification are introduced below.

- Term presence and their frequency:

These features include uni-grams or n-grams and their frequency or presence. These features have been widely and successfully used in sentiment classification. Pang et al. [1] claim that uni-grams gives better results than bi-grams in movie review sentiment analysis, but Dave et al. [6] report that bi-grams and tri-grams give better product-review polarity classification.

- Part of speech information:

POS is used to disambiguate sense which in turn is used to guide feature selection [11]. In POS tagging each term in sentences will be assigned a label, which represents its position/role in the grammatical context. For example, with POS tags, we can identify adjectives and adverbs which are usually used as sentiment indicators [2].

- Negations:

Negation is also an important feature to take into account since it has the potential of reversing a sentiment [11].

- Opinion words and phrases:

Opinion words and phrases are words and phrases that express positive or negative sentiments. The main approaches to identify the semantic orientation of an opinion word are statistical-based or lexicon-based. Hu and Liu et al. [4] use WordNet to determine whether the extracted adjective has a positive or negative polarity.

Pang et al. [1] compared the performance of three classifiers Naïve Bayes, Maximum Entropy and Support Vector Machines in Sentiment classification at document level on different features like considering only unigrams, bigrams, combination of both, combining unigrams and parts of speech, taking only adjectives and combining unigrams and position information. The result has shown that feature presence is more important than feature frequency and when the feature set is small, Naïve Bayes performs better than SVM. But SVM's perform better when feature space is increased. When feature space is increased, Maximum Entropy may perform better than Naïve Bayes but it may also suffer from over fitting. Abbasi et al. [12] proposed sentiment analysis techniques for classification of hate/extremist web forum postings in multiple languages (English and Arabic) by utilizing of stylistic and syntactic features. They introduced new algorithm entropy weighted genetic algorithm (EWGA) which is hybrid genetic algorithm that uses the information gain heuristic to improve feature selection. They use 14 categories of English and Arabic Feature Sets as initial set. All features with an information gain greater than 0.0025 are selected. They used Support Vector Machine (SVM) with 10-fold cross-validation and bootstrapping to classify sentiments in all experiments. When using both syntactic and stylistic features they achieved 95.55% accuracy in 10 crosses validation.

2) Lexicon based techniques

In unsupervised technique, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative.The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data.

The basic steps of the lexicon based techniques are outlined below [9]:

1. Preprocess each text (i.e. remove HTML tags, noisy characters).

2. Initialize the total text sentiment score: s ← 0.
3. Tokenize text. For each token, check if it is present in a sentiment dictionary.
   (a) If token is present in dictionary,
      i. If token is positive, then s ← s + w.
      ii. If token is negative, then s ← s − w.
4. Look at total text sentiment score s,
   (a) If s > threshold, then classify the text as positive.
   (b) If s < threshold, then classify the text as negative.

There are three methods to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. The manual construction of sentiment lexicon is a difficult and time-consuming task.

In dictionary based techniques the idea is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet dictionary for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration starts. The iterative process stops when no more new words are found [8]. Opinion words share the same orientation as their synonyms and opposite orientations as their antonyms. Hu and Liu [4] use this technique to find semantic orientation for adjectives. They used 30 adjectives as seed list. The dictionary based approach have a limitation is that it can't find opinion words with domain specific orientations [8].

Corpus based techniques rely on syntactic patterns in large corpora. Corpus-based methods can produce opinion words with relatively high accuracy. Most of these corpus based methods need very large labeled training data. This approach has a major advantage that the dictionary-based approach does not have. It can help find domain specific opinion words and their orientations.

The most prominent work done using unsupervised methods for opinion mining and sentiment detection is by Turney [2]. He uses "poor" and "excellent" seed words as they are appear more in web for calculating the semantic orientation of phrases, where orientation is measured by pointwise mutual information.

SO(phrase) = PMI (phrase, "excellent") - PMI (phrase, "poor")

The sentiment of a document is calculated as the average semantic orientation of all such phrases. He was able to achieve 66% accuracy for the movie review domain. Ting-Chun Peng and Chia-Chun Shih [13] uses part-of-speech (POS) patterns for extracting the sentiment phrases of each review, they used unknown sentiment phrase as a query term and get top-N relevant phrases from a search engine. Next, sentiments of unknown sentiment phrases are computed based on the sentiments of nearby known relevant phrase using lexicons. Gang Li & Fei Liu [10] developed an approach for clustering documents into positive group and negative group based on the k-means clustering algorithm. After applying TF-IDF (term frequency – inverse document frequency) technique on the raw data, a voting mechanism is used to extract a more stable clustering result. This result is obtained by multiple implementations of the clustering process then the term score is used to further improve the clustering result. A. Khan et al. [5] proposed rule based domain independent method of sentiment classification at the sentence level. They first classify sentences into objective and subjective and check their semantic scores using the SentiWordNet. The final weight of each individual sentence is calculated after considering the whole sentence structure, contextual information and word sense disambiguation. Their method achieves an accuracy of 86.6% at the sentence level. Zhang et al. [7] proposed weakness finder system which can help manufacturers' find their product weakness from Chinese reviews by using aspects based sentiment analysis. In their method they first identify the implicit and explicit features for each aspect, and then they determine the sentiments about the aspects. They found the product weaknesses by comparing the result of each aspect of a specific product and the aspects of different products. They found aspects of each review by using the explicit and implicit features grouping method, and the sentiments of them can be found via sentence based sentiment analysis. They use PMI method to find implicit feature words, for explicit features sharing morpheme method and the similarity measure of Hownet are used. They achieved general precision of 82.62%, and the recall is 85.26%, F1-measure is about 83.92%.

3) **Hybrid Techniques**

Few research techniques have indicated that the combination of both the machine learning and the lexicon based approaches improve sentiment classification performance. Mudinas et al. [15] presents concept-level sentiment analysis system, pSenti, which is developed by combining lexicon-based and learning-based approaches. The main advantage of their hybrid approach using a lexicon/learning symbiosis is to attain the best of both worlds-stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm. Their system uses a sentiment lexicon constructed using public resources for initial sentiment detection. Currently the sentiment lexicon consists of 7048 sentiment words including words with wildcards and sentiment values are marked in the range from − 3 to +3. They used sentiment words as features in machine learning method. The weight of such a feature is the sum of the sentiment value in the given review. For those adjectives which are not in sentiment lexicon, their occurring frequencies are used as their initial values. Their hybrid approach pSenti achieved 82.30% accuracy,

Fang et al. [16] incorporate not only a general purpose sentiment lexicon but also Domain Specific Sentiment Lexicons into SVM learning, and use this method for identifying both product aspects and their associated polarities. Experiment results show that while a general purpose sentiment lexicon provides only minor accuracy improvement, incorporating domain specific dictionaries leads to more significant improvement. Their system performed a two step classification. In step 1, a classifier is trained to predict the camera aspect being discussed. In step 2, a classifier is trained to predict the sentiment associated with that camera aspect. Finally, the two step prediction results are aggregated together to produce the final prediction. In both steps, the lexicon knowledge is incorporated into conventional SVM learning. They achieved 66.8% polarity accuracy. Zhang et al. [14] employ an augmented lexicon-based method for entity level sentiment analysis. First extract some additional opinionated indicators (e.g. words and tokens) through the Chi-square test on the results of the lexicon-based method. With the help of the new opinionated indicators, additional opinionated tweets can be identified. Afterwards, a sentiment classifier is trained to assign sentiment polarities for entities in the newly identified tweets. The training data for the classifier is the result of the lexicon-based method. Thus, the whole process has no manual labeling. They achieved accuracy of 85.4%.

3. ANALYSIS AND COMPARISON

Supervised machine learning techniques have shown relatively better performance than the unsupervised lexicon based methods. However, the unsupervised methods is important too because supervised methods demand large amounts of labeled training data that are very expensive whereas acquisition of unlabelled data is easy. Most domains except movie reviews lack labeled training data in this case unsupervised methods are very useful for developing applications.

Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms. The main limitation of supervised learning is that it generally requires large expert-annotated training corpora to be created from scratch, specifically for the application at hand, and may fail when training data are insufficient.

The opinion words that are included in the dictionary are very important for the lexicon based approach. If the dictionary contains less words or thorough, one risks the chance of over or under analyzing the results, leading to a decrease in performance. Another significant challenge to this approach is that the polarity of many words is domain and context dependent. For example, 'funny movie' is positive in movie domain and 'funny taste' is negative in food domain. Such words are associated with sentiment in a particular domain. Current sentiment lexicons do not capture such domain and context sensitivities of sentiment expressions. Without a comprehensive lexicon, the sentiment analysis results will suffer. The lexicon-based approach can result in low recall for sentiment analysis.

The main advantage of hybrid approach using a lexicon/learning combination is to attain the best of both worlds, high accuracy from a powerful supervised learning algorithm and stability from lexicon based approach. Table 1 presents summary of Precision of sentiment analysis using different techniques according to the data reported by authors.

Paper	Dataset	Technique (precision, %)
Pang et al. [1]	IMDB	NB (81.5), ME (81.0), SVM (82.9)
Turney [2]	Epinions	PMI(66)
Dave et al.[6]	Amazon, CNET	SVM (85.8-87.2), NB (81.9-87.0)
Hu and Liu [4]	Amazon, CNET	Lexicon (84.0)
Abbasi et al. [12]	U.S. & Middle Eastern web forum postings	SVM(95.55)
A. Khan et al. [5]	IMDB,Skytrax, Tripadvisor	Lexicon(86.6)
Zhang et al. [7]	Luce, Yoka	Lexicon(82.62)
Fang et al. [16]	Multi-Domain Sentiment Dataset	ML + Lexicon (66.8)
Zhang et al. [14]	Twitter	ML + Lexicon (85.4)
Mudinas et al. [15]	CNET, IMDB	ML + Lexicon (82.30)

Table 1: Precision of sentiment analysis using different techniques according to the data reported by authors.

4. CONCLUSION

Applying Sentiment analysis to mine the huge amount of unstructured data has become an important research problem. Now business organizations and academics are putting forward their efforts to find the best system for sentiment analysis. Although, some of the algorithms have been used in sentiment analysis gives good results, but still no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has limitations. More future work is needed on further improving the performance of the sentiment classification. There is a huge need in the industry for such applications because every company wants to

know how consumers feel about their products and services and those of their competitors. Different types of techniques should be combined in order to overcome their individual drawbacks and benefit from each other's merits, and enhance the sentiment classification performance.

5. REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.

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

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

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177.

[5] A. Khan, B. Baharudin, K. Khan; "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp. 317-331, 2011.

[6] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proceedings of WWW, 2003, pp. 519–528.

[7] W. Zhang, H. Xu, W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," Expert Systems with Applications, Elsevier, vol. 39, 2012, pp. 10283-10291.

[8] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.

[9] M. Annett, G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie Blogs", In Canadian Conference on AI, pp. 25–35, 2008.

[10] G. Li, F. Liu, "A Clustering-Based Approach on Sentiment Analysis," IEEE International Conference on Intelligent System and Knowledge Engineering, Hangzhou, China, vol. 2010 / 1, pp. 331-337.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

[12] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," In ACM Transactions on Information Systems, vol. 26 Issue 3, pp. 1-34, 2008.

[13] T. Peng, C. Shih, "An Unsupervised Snippet-Based Sentiment Classification Method for Chinese Unknown Phrases without Using Reference Word Pairs." Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp.243-248.

[14] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011.

[15] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.

[16] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classication", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.