COMPARISON OF VARIOUS SENTIMENT ANALYSIS TECHNIQUES – A STUDY

[1]Gurshobit Singh Brar (Author) and [2]Ankit Sharma (Mentor/Guide)

[1]Student - Department of Computer Science,

Baba Farid College of Engineering and Technology, Bathinda, India

[2]Assistant Professor - Department of Computer Science,

Baba Farid College of Engineering and Technology, Bathinda, India

ABSTRACT - In Today's World, A huge amount of user data such as reviews, comments, surveys and opinion polls are collected via many resources like social websites, e-commerce websites and blogs. All of the data is used by private industries, government and individuals for analysis of their products, events and advertising campaigns etc. It is very difficult to analyze this huge amount of data that is why we need a solution. We need to develop an artificial intelligent system which can automatically classify data into different categories on the basis of polarity positive, negative or neutral. Sentiment analysis is an automated solution for classifying and mining of user reviews, opinions and emotions gathered from various resources such as text surveys and database containing user opinions using natural language processing (NLP). The objective of this paper is to understand concept of sentiment analysis and compare various type of sentiment analysis techniques.

Index Terms – Opinion Mining, Sentiment Analysis, Natural Language Processing, Sentiment Score, Sentiment Lexicon

1. INTRODUCTION

As Social Media, E-Commerce websites and Blogs are being popular among internet user on daily basis. [11] So, there is a huge amount of data is produced in terms of reviews, comments, opinions and emotions from various resources such as websites, text surveys and events etc. The collected data is very beneficial to both users and content owners. Users can take a decision by looking at other reviews or opinions. Content Owners can come to conclusion whether their product need some improvement or they have to take some other decision on basis of customer opinion for their product. The process of analysis of this huge amount data is very difficult and time consuming too. This process is time consuming and difficult because all of the collected data is unstructured and written in natural language. So, the complexity for extracting relevant information is too high. That's how we end with a solution called Opinion Mining or Sentiment Analysis which are new field of research.

Sentiment analysis is technique of Natural Language Processing for extracting the sentiments, opinions and emotions of a writer from a text and classify into different categories on the basis of text polarity like positive, negative and neutral [2]. There are two major steps involved in opinion mining and sentiment analysis: 1) Information extraction 2) Information classification. For Information extraction and classification various type of machine learning techniques are used.

There are three types of methods used for opinion mining and sentiment analysis: 1) Machine Learning based techniques 2) Lexicon based techniques 3) Rule Based technique. In Machine Learning Approach, both supervised and unsupervised algorithms can be used to classify opinions from a text. In Lexicon Approach, we extract information from text and compare with word-based dictionary and results in polarity like positive, negative or neutral of a sentence. Rule base approach is almost similar to Lexicon technique the only difference is its based-on words only. If word consists in text is of positive polarity then sentiment score will be increased and else it sentiment score will decrease.

Sentiments can be classified at sentence level, document level and feature level. In Sentence level classification, sentiment polarity is considered for a particular topic (Example: A Movie Review). In Document level classification, a whole document can be classified polarized positive or negative. In Feature level classification, extracting features of a product (Example: Smartphones Features, Camera Features or Laptop Features).

This paper is a study to understand concept of sentiment analysis and compare various type of techniques used for sentiment analysis in field of Natural Language Processing (NLP). This paper is organized as follows: various sentiment analysis techniques in Section 2. Analysis and Comparison of techniques in Section 3. In Section 4 concludes the paper.

2. SENTIMENT ANALYSIS TEVHNIQUES

There are three main techniques used for sentiment analysis: 1) Machine Learning based, 2) Lexicon based and 3) Rule Based. In some of research papers they combine these two techniques for better results.

2.1. Machine Learning Based

In this approach, we need two data set one is training dataset and other one is test dataset. Any of machine leaning algorithms supervised (Example: Naïve Bayes) or unsupervised can be used for sentiment analysis. Training dataset is used to train classifier and Test dataset is used to test accuracy of classifier. Machine learning algorithms can be used for classifiers are Support Vector Machine (SVM) and Maximum Entropy (ME).

First step is to select a Classifier. Next step is to train our classifier using our training dataset. Once Classifier is trained, the next step is feature selection. These steps are very important because classifier and feature determine the performance. The most commonly used feature classification techniques.

1. Part of Speech (POS)

POS is used to disambiguate sentence in order to extract features from a sentence [2]. In POS tagging each word is labelled. It is used determine word position in the grammatical context.

2. Opinion words and phrases

Opinion words can be adjectives, adverbs, verbs or nouns. Document polarity can be determines using polarity of opinion words. Word-Net Application Programming Interface (API) is used by [4] for determining polarity of adjectives and adverbs.

3. Taxonomy Presence and their frequency

Taxonomy features are uni-grams or n-grams with occurrence of their presence and their frequency. Pang and Lee [8] used bi-grams on product review dataset and concluded bi-grams perform better than uni-grams. Dave and Lawrence [3] used uni-grams and bi-grams both on movie review dataset and concluded that uni-gram perform better than bi-gram.

4. Negations

Negations also plays important role in sentiment polarity as they totally reserve the meaning of the sentiment [2].

Pang and Lee [8] used three techniques Naïve Bayes, Support Vector Machine (SVM) and Maximum Entropy for text classification. They also compared performance of all the techniques with different feature selection method like Part of Speech (POS), uni-grams, bi-grams and both uni-grams and bi-grams. They concluded that SVM Performed better than other two techniques in case of big dataset. They also showed in their conclusion that Naïve Bayes Performs better than SVM in case of small dataset.

2.2. Lexicon Based

It is an unsupervised learning approach. Lexicon Based Techniques don't require any dataset for training or testing. It requires word dictionary with word polarity as positive, negative and neutral. When features are extracted form a document and they will be compare with dictionary word if they exist than sentiment score increases else sentiment core decreases. Steps followed by Lexicon based approach [5]:

1. Preprocessing: Remove all HTML tags and noisy characters like $, % and # etc.
2. Feature Extraction: Features can be extracted from a document using POS tagging and label them as tokens.
3. Algorithm for Lexicon Based approach:

Initialize sentiment score S = 0

Check if token is present in dictionary and is positive, then S + 1

Check if token is present in dictionary and is negative, then S – 1

Check total sentiment score if sentiment score S is greater than threshold, then classify document as positive

Check total sentiment score if sentiment score S is less than threshold, then classify document as negative.

Most common method used for Lexicon Construction is dictionary-based method. A small dictionary is created using adjectives, adverbs and verbs later on can be expanded using Word-Net by looking for synonyms and antonyms. Drawback of this method is its only used for context which is written only in English because word-net is only available for English language.

S. Moghaddam and Martin Ester [9] used unsupervised technique. Their system was based on Word-Net and called opinion digger. They used aspects or features of a product and rating guidelines (Example: "5" is excellent, "4" is good, "3" is okay, "2" is poor and "1" is very poor). POS is used extract features and result outputs as rating of each aspect. It was based on E-Commerce website dataset. They conclude that accuracy is very low using this technique.

W. Zhang [10] used feature-based technique for extraction of product weakness. So that they can help manufactures to determine their product weakness and improve its quality. These used reviews in Chinese language to collect information. Their system achieved accuracy of 86.6%, general precision of 82.62% and F1-measure of 83.92%.

2.3. Other Techniques

There some other techniques used for sentiment analysis are as follows:

1. Rule-based Technique

Rule-based Approach is similar to lexicon approach and it used by [1] for customer reviews. They Extracted features using POS tagging and it is a domain independent technique. This Techniques works in three steps: first step is feature extraction using POS tagging and store each word. Second step is word extraction which allows determining polarity of sentence based on contextual information and sentence's structure. Their system achieved accuracy of 91% at review level and 86% at sentence level.

2. Hybrid Techniques

Mudinas and Zhang [6] uses two techniques to get better results. Their solution is called pSenti which combines lexicon and learning based techniques. Their system measures and reports the overall sentiment score to be positive, negative, neutral or 1-5 stars classification. This system's accuracy is almost equal to pure leaning based technique and higher than pure Lexicon based technique.

3. ANALYSIS AND COMPARISON

Most of Supervised learning-based technique algorithms performs much better than all other techniques. However, Rule based technique, unsupervised learning techniques and Hybrid Techniques can't be ruled out of options. Because, rule-based technique performed very well not only with movie reviews but also with software reviews and it is a domain independent technique. Unsupervised Learning techniques is also very important because these days most of data obtained in unstructured form. Unsupervised technique can be used with supervised technique to obtain better result like a hybrid technique used by Mudinas and Zhang [6] which resulted in better accuracy.

Support Vector Machine has higher accuracy than other machine learning algorithms like Naïve Bayes (NB) and Maximum Entropy (ME). In Lexicon based approach, performance is totally dependent on size of dictionary. If dictionary size is small than performance will be poor and if dictionary size is large than performance will be good.

Sentiment Analysis is purely based on domain. So, it is a big challenge because it is determined by polarity of words. For Example, "This is a superb Smartphone". This review is positive for smartphone domain and "waste of time" This review is negative for smartphone domain.

Rule based Technique used by Aurangzeb Khan [1] which is SentiWordNet based gives more accuracy than pure lexicon-based technique for sentiment analysis.

Hybrid Techniques used by Mudinas and Zhang [6] gives better performance than lexicon and almost perform like leaning based technique. Hybrid Techniques are stable as lexicon technique and performance as machine learning based techniques. In Tables from 1 to 4 is Summary of sentiment analysis techniques used by researchers in this field along with accuracy as per their evaluation which collected from authors research. Tables also include classification and drawback of techniques.

Table 1

Machine Learning Technique

Technique	Machine Learning	
Text Approach	Document Level	
Classification	Global Rating (Positive, Negative or Neutral)	
Dataset	Movies Review	
Accuracy in Percentage	SVM	82.95 %
	Naïve Bayes	81.5 %
	Maximum Entropy	81 %
Drawback	Based on Large dataset.	
Example	Pang and Lee [8]	

Table 2
Lexicon Technique

Technique	Lexicon
Text Approach	Document Level, Sentence Level
Classification	Global Rating
Dataset	Customer Reviews
Accuracy in Percentage	84%
Drawback	Based on Dictionary. It can be turned into advantage by expanding dictionary size.
Example	Hu and Liu [4]

Table 3
Rule based Technique

Technique	Rule based	
Text Approach	Sentence Level	
Classification	Rating per aspect and Global Rating	
Dataset	Customer Reviews, Software Reviews	
Accuracy in Percentage	Document Level	91 %
	Sentence Level	86 %
Drawback	Based on SentiWordNet	

Table 4
Hybrid Techniques

Technique	Hybrid Techniques (Both Machine Learning and Lexicon)
Text Approach	Document Level
Classification	Rating per aspect and Global Rating
Dataset	Customer Reviews
Accuracy in Percentage	82.3 %
Drawback	Reviews with lot of Noise (irrelevant words for the subject of review) are often assigned neutral score. Because it fails to find any sentiment.
Example	Mudinas and Zhang [6]

4. CONCLUSION

In past few years as per a study, huge amount of unstructured is collected [11] and various types of existing method used to sentiment analysis and opinion mining. But there is still need to create more accurate techniques and algorithms to get better result. Most of businesses are depended on opinion mining because this help business and consumers both for selling and purchasing better products. As mentioned in above table, every technique has a drawback. Researchers have to try for better solutions. But most common method used for sentiment analysis is machine learning approach. If Researchers able to overcomes drawback in techniques than a better sentiment analysis is possible.

5. REFERENCES

[1] 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.

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

[3] 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.

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

[6] 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.

[7] H. Wang, Yue Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 783-792. ACM, 2010.

[8] 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.

[9] S. Moghaddam and Martin Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1825-1828. ACM, 2010.

[10] 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

[11] M. Chen, Huazhong University of Science, China "Big Data: A Survey", Springer, 2014.