International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 3, May-June-2018, ISSN (Online): 2395-566X

A Survey on Various Techniques and Features of Sentiment Analysis

M. Tech. Scholar Arunes Pratap Singh
Department of Computer Science and Engineering,
Bansal Institute of Research and Technology
Bhopal, M.P, India

Asst. Prof. Manisha Patel
Department of Computer Science and Engineering,
Bansal Institute of Research and Technology
Bhopal, M.P, India

***Abstract* – As the digital data increases on servers different researcher have focused on this field. The field of sentiment analysis, in which sentiment is gathered, analyzed, and aggregated from text, has seen a lot of attention in the last few years. The corresponding growth of the field has resulted in the emergence of various subareas, each addressing a different level of analysis or research question. In this paper text sentiment analysis was detailed by various techniques of researchers. Here diverse attributes for the text sentiment analysis is explained in detailed with their necessities as attribute vary as per text study.**

***Keywords*: Supervised Classification, Text Mining, , Text Feature, Text Ontology, Un-supervised Classification**.

I. INTRODUCTION

Sentiment or opinion mining refers to the type of natural language processing used to understand the moods, opinions and sentiments of the public regarding a particular product or a movie or an event. The availability of large amounts of data and the human tendency to always manipulate what other people think has been influential in a decision making process. This unique feature plays a vital role in deciding on matters that have financial, medical, social or other implications. Seeking second or third or many more opinions have fuelled the interest of researchers in the field of sentiment mining. With multiple reviews available for a single product and the enormous growth in the number of internet users it has become indispensable to develop a system that collects, builds, analyzes, and classifies the comments or a review posted online. Usually these kinds of reviews are written by customers who have used the particular product or service. An individual□s interests, opinions and perceptions greatly influence the nature of the review. There are instances where people are biased in their opinions and automatically that has an impact on the content they contribute to the forum as review or blog posts or tweets. As the number of such people contributing content surges it has become a huge challenge to classify and organize the real problems and prospects of the product which makes the user to doubt the reliability of the content. Big companies rely on personal review of customers to improve the scope of their product and deem it to be of great importance in placing content based ads on sites that easily aid a prospective buyer. The same applies to movie enthusiasts and voters as more and more people are using the social networking sites, online shopping and trend analysts who after reading the reviews available decide on various issues. For example placing the ad of a Kitchen Aid Mixer on a food blog not only influences purchase decisions but also goes a long way in modifying the marketing strategy.The marketing division of a company enthusiastically promotes reviewers by sending samples of product to be reviewed or sponsoring giveaways in blogs or in social networking sites like Facebook and Twitter. This has lead to the increase in the volume of data available and the need to classify the available information efficiently as these have a larger impact. The subjective nature of opinion makes a single opinion insufficient in decision making [6]. Also, the writing skills and choice of words by contributors largely depend on the language proficiency and the temperament of the writer. Online reviews that are usually the voice of the customer are written from their angle of interests and preferences can be a combination of a positive and negative opinion which may not help in deciding whether it is a positive or a negative review. For example, consider the sentence "This restaurant□s Chinese dishes are not as good as their Thai dishes". These kind of comparative opinions are different in natural language processing. When a positive word „good□ is negated like „not as good as□ a reader will also find it difficult to comprehend on how good the Thai dishes were as this decides the taste of the Chinese dishes too. When treating negation, one must be able to correctly determine what part of the meaning expressed is modified by the presence of the negation [8,2]. There are different types of opinions like regular, implicit, direct, indirect and comparative. The freedom of expression and anonymity also comes with a price. People with hidden agendas or malicious intentions to easily game the system to give people the impression that they are independent members of the public and post fake opinions to promote or to discredit target products, services, organizations, or individuals without disclosing their true intentions, or the person or organization that they are secretly working for. Such individuals are called opinion spammers and their activities are called opinion spamming [1].

II. FEATURES of TEXT MINING

1 Title feature: The word in sentence that also occurs in title gives high score. This is determined by counting the

number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2. Sentence Length: This features is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belongs to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3. Term Weight: The frequency of the term occurrence with a documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

4. Sentence Position: Whether it is the first 5 sentence in the paragraph, sentence position in text gives the importance of the sentences. This features can involve several items such as the position of the sentence in the documents, section and the paragraph, etc, proposed the first sentence of highest ranking. The score for this features in [6] consider the first 5 sentence in the paragraph.

5. Sentence to sentence similarity: This feature is a similarity between sentences for each sentence S , the similarity between S and each other sentence is computed by the cosine similarity measure with a resulting value between 0 and 1 [6]. The term Weight wi and wj of term t to n term in sentences Si and Sj are represented as the vector. The similarity of each sentence pair is calculated based on similarity.

III. CLASSIFICATION TECHNIQUES

The main task of document-level sentiment classification is to identify the polarities of UGC. Two type of classification techniques have been used in document-level sentiment classification, supervised method and unsupervised method.

1. Supervised Methods: Sentiment classification can be formulated as a supervised [4] learning problem with four classes, positive, negative, neutral and constructive. User generated contents mostly are used as training and testing data. Any existing supervised learning techniques can be used to sentiment classification, such as naïve Bayes and support vector machines (SVM).

2. Sentence level Sentiment Classification The task of classifying a sentence as subjective or objective is often called subjectivity classification. The resulting subjective sentences are also classified as expressing positive or negative opinions, which is called sentence-level sentiment classification. In the sentence level sentiment analysis, the polarity of each sentence is calculated. This is similar to a document level sentiment analysis but done at a sentence level [3]. It assumes each sentence contains an opinion for one entity and aspect, and some of the sentences may not be opinionated (objective). The subjective sentences contain opinion words which help in determining the sentiment about the entity. A two stage inference is done for each sentence: first, each sentence is classified as subjective or objective and then the polarity of each of the subjective sentences are inferred. There may be complex sentences also in the opinionated text. In such cases, sentence level sentiment classification is not useful.

3. Aspect level Sentiment Classification In a typical opinionated document, the author writes both positive and negative aspects of the entity, although the general sentiment on the entity may be positive or negative. Document and sentence sentiment classification does not provide such information. To obtain these details, we need to go to the aspect level. It assumes that a document contains opinion on several entities and their aspects. Aspect level classification requires discovery of these entities, aspects, and sentiments for each of them.

IV. RELATED WORK

In [1] The proposed system develops an innovated micro blog specific sentiment lexicon which is based on data driven approach. Sentiment lexicon is considered to be one of the most important components of sentiment analysis. Existing sentiment lexicons are not performing well for micro blogs because all the reviews in the blogs contains a user specific words such as “Thnx”,”gud”. These types of words can’t be correctly recognized by the existing framework. The proposed framework for handling micro blog based sentiment lexicon is constructed by integrating 3 types of sentiment knowledge such as word opinion knowledge for sentiment score, opinion similarity knowledge for expressing sentiment similarity and primary knowledge which is extracted from the traditional lexicons. The proposed framework also develops a new word detection method by using a proposed new word detection algorithm and that new word will be added to the dictionary. The proposed framework was validated using a Chinese micro blog of 17.2 million messages and the results were compared with the existing sentiment lexicons in terms of subjectivity identification and opinion polarity classification in both sentence and document level opinion mining.

In [2] The Author proposes a dictionary based technique for domain specific sentiment analysis on the movie review dataset. The author make use of lexicon known as

SentiWordNet (SWN-publically available dictionary) including adjectives, adverbs, and verbs. Document level analysis involves by using linguistic features ranging from adverb+adjective to adverb+adjective+verb combination.

In [4] paper proposes a advanced framework for opinion mining that correlates all the merits of semantic web guided solutions to tremendously improve the overall results of traditional NLP (Natural Language Processing). The proposed framework makes use of domain ontology at feature extraction stage. This enhancement makes huge changes in the feature based sentiment classification. Existing machine learning techniques classify the words into limited category such as positive/negative. Existing system also performs sentiment classification at the document level (i.e.) if the document includes huge no of positive than negative terms, then it will be considered to be a positive document otherwise negative document. Dataset of Movie Reviews is used to check the performance of proposed model.

In [5] paper the proposed framework provide a clear understanding about the polarity shift problem. Sentiment Analysis is affected by many factors. Among that polarity shift problem is considered to be very dangerous factor that destroys the complete classification performance of traditional machine learning based sentiment classification. Usually the review data is represented in the form of Bag of Words (BOW) that entirely terminates the semantic correlation between the texts. The existing system makes use of term counting method addressing the polarity shift problem. The proposed polarity shift Detection, Elimination and Ensemble (PSDEE) performs detection of hybrid polarity shifts. To perform hybrid polarity shift detection it makes use of 3 levels of cascading model. Polarity shift problem arises if there is a polarity shifters or valance shifters such as negation, contrast, sentiment inconsistency in the text review. Proposed methodology make use of Rule based Method is used for detecting negations and contrast polarity shift and statistical methods are used for detecting implicit inconsistency. The proposed PSDEE was examined in four domains which are extracted from the Amazon website.

In [6] paper proposes a framework for aspect/feature based sentiment analysis along with the sentence compression technique. Aspect based sentiment analysis is performed based on syntactic features which poses a chance for over natural problem. This type of issue makes the sentiment analysis too difficult to handle the syntactic parsers used in the opinion mining technique. The proposed framework develops an innovated sentence compression technique before the sentiment analysis. For compressing a text for sentiment analysis 2 schemes are used. That is syntactic compression and extractive compression technique. Compared to extractive compression technique syntactic is considered to be more efficient because it compress the text by removing the unimportant words. The proposed technique makes use of Aspect-Polarity (A-P) collection based sentiment analysis. Most of the aspect based sentiment analysis focus on the relationship between the aspects and the polarity words which extremely affects the efficiency. To solve this problem the proposed framework makes use of syntactic patterns.

In [12] propose an innovative method to do the sentiment computing for news events. More specially, based on the social media data (i.e., words and emoticons) of a news event, a word emotion association network (WEAN) is built to jointly express its semantic and emotion, which lays the foundation for the news event sentiment computation. Based on WEAN, a word emotion computation algorithm is proposed to obtain the initial words emotion, which are further refined through the standard emotion thesaurus. With the words emotion in hand, we can compute every sentence's sentiment.

V. TEXT PREPROCESSING

As document is collection of paragraphs. Paragraphs are collection of sentences. While sentences are collection of words. So whole preprocessing focus on word in the document without any punctuations. So in pre-processing of document there are two common steps first is stop word removal, and second is stem word removal. [8]

Stop List Removals: As sentence is frame with number of words but some of those words are just use to contruct a proper sentence although it does not make any information in the sentence. So identification of those words then removing is term as Stop word removal. So a list of words is store by the researcher which help in identifying of stop words. This removal of stop words help in reduce the execution time of the algorithm, at the same time noisy words which not give any fruitful information is also removed. Stop words are like {a, the, for, an, of, and, etc.}. So text document is transform into collection of words which is then compare with these words and then each match word is removed from the document.
Inorder to understand this assume an sentence {India is a great country in the world} then after pre-processing it become {India, great, country, world} while stop words {is, a, in, the} in the sentence are removed.

Stem Word Removal: In this words which are almost similar in prefix are replace by one word. This can be said collection of words share same word is term as stem. So there occurrence in the document make same effect but while processing in text mining algorithm it make different so update each word from the collection into single word is done in this stem word removal pre-processing step. Let us assume an collection of words for better understanding of this work. Collection of word is {play, plays, playing} then replace each with word {play}.

VI. CONCLUSION

As the writing work of different blogs, articles from organization, press media, institutes are increasing day by day. Then publishing their work is also increase which is done by most of the news paper, organizations. Here paper has cover an important issue of text sentiment retrieval. Various techniques with there required features are discussed in detailed. Here paper related work of researchers done in this field. So it can be concluded that one strong algorithm is required that can effectively classify and retrieve document on the basis of author sentiment.

REFERENCES

[1] Fangzhao Wu, Yongfeng Huang, Yangqiu Song, Shixia Liu," Towards building a high quality micro blog-specific Chinese sentiment lexicon", Decision Support Systems-2016.

[2] V.K. Singh, R. Piryani, A. Uddin, P. Waila," Sentiment Analysis of Movie Reviews", conference on IEEE-2013.

[3] Farman Alia, Kyung-Sup Kwaa,Yong-Gi Kimb," Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification", Applied Soft Computing-2016.

[4] Isidro Peñalver-Martinez, Francisco GarciaSanchez, Rafael Valencia-Garcia," Featurebased opinion mining through ontologies", Expert Systems with Applications-2014.

[5] RuiXia,FengXu,JianfeiYu," Polarity shift detection, elimination and ensemble: A three stage model for document-level sentiment analysis" Information Processing and Management 52 (2016) 36– 45.

[6] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu," Sentence Compression for spect-Based Sentiment Analysis" IEEE/ACM transactions on audio, speech, and language processing, vol. 23, no. 12, December 2015

[7] Afef Walha, Faiza Ghozzi and Faïez Gargouri "A Lexicon Approach to Multidimensional Analysis of Tweets Opinion" (2016)

[8] Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1/13/$31.00 ©2013 Ieee.

[9] Shrilakshmi Prasad, B. S. Mamatha." Retrieving documents from encrypted cloud data in a secured way using cosine similarity search with multiple keyword search support. " International Journal of Advance Research in Computer Science and Management Studies. Volume 4, Issue 5, May 2016

[10] H. Yu and V. Hatzivassiloglou. "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03, pages 129-136.

[11] . 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.

[12] Dandan Jiang1, Xiangfeng Luo1, Junyu Xuan, And Zheng Xu ."Sentiment Computing for the News Event Based. on the Social Media Big Data". Digital Object Identifier 10.1109/ACCESS.2016.2607218 IEEE Acess 2017.