Proc. of Int. Conf. on Emerging Trends in Engineering & Technology, IETET

Sentiment Analysis: An Overview

Pallavi Pandey[1], Anil Saroliya[2] and Varun Sharma[3]
[1]Amity University Rajasthan, Jaipur, India
Email: pandeypallavi027@gmail.com
[2-3]Amity University Rajasthan, Jaipur, India
Email: {asaroliya, vsharma}@jpr.amity.edu

***Abstract*—In today's state of affairs, information gathering has become a vital aspect of human behavior. We are always curious to know what others think about a particular thing. With the evolution in web technology, online shopping has become quite popular. It additionally provides a platform to share experiences and feedbacks associated with a selected product. The reviews on these sites reflect whether the product is good or bad and helps the potential customers to take the right decision. Sentiment analysis is a method using which information can be extracted from these reviews, analyzed and classified as positive, negative or neutral. Thus, the opinions gathered will facilitate to predict the online customer's preferences and it can be beneficial for economic and marketing research. This paper presents a survey covering the techniques and methods in sentiment analysis and also the challenges appearing in this field.**

***Index Terms*— marketing strategies, opinion holder, polarity, polarity strength, sentiment analysis**.

I. INTRODUCTION

The advancement of web technology has led to the massive and extraordinary amount of consumer-generated media and has significantly changed the way we control, structure and interact with the information. Due to the massive amount of customer feedbacks, opinions, reviews, comments and suggestions it is important to traverse, scrutinize and order the content for efficient and effective decision-making. In the last decade, sentiment analysis has turned out as one of the most popular techniques for information retrieval and for the analysis of the data available on the web.

Opinion mining can be defined as a subfield of Computational Linguistics that emphasizes on obtaining people's opinion from the web. Blogs, videos, social networking sites, etc platforms provide a huge amount of valuable information that can be analyzed. Given a piece of text, opinion mining systems examine [9]:

- Which part is opinion revealing or exhibiting
- Who is the writer of the opinion
- What is being said about the entity by the writer

Sentiment Analysis, on the other hand, determines the subjectivity, polarity (positive, negative or neutral) and polarity strength of a piece of text that is [5] [9]:

Bing Liu defines an opinion as a quintuple $<o_i, f_{ij}, s_{oijkl}, h_i, t_i>$, where o_i is the target object, f_{ij} is the feature of the target object o_i, h_i is the opinion holder, t_i is the time when the opinion has been expressed and s_{oijkl} is the

Grenze ID: 02.IETET.2016.5.8

sentiment value of the opinion conveyed by the opinion holder h_i about the object o_i at time t_i [5].
Today, many companies are developing their marketing strategies based on the results obtained from the sentiment analysis. They access, scrutinize and make predictions about the public opinion for their brand. Researchers are also working to develop an automated tool for opinion mining.
An attempt has been made to explore different levels of sentiment analysis along with its applications and existing techniques for sentiment analysis have been discussed. Furthermore, a pre-processing model for the dataset has also been proposed.

II. Applications Of Sentiment Analysis

Among the various applications of sentiment analysis some of them have been listed below [8] [9]:

- The applications for sentiment analysis are endless. It is getting more and more used in social media monitoring and to track customer reviews, survey responses, competitors, etc [7].
- Sentiment analysis is in demand because of its efficiency. Thousand of text documents can be processed and analyzed for sentiment in seconds, compared to the hours it would take to manually complete it by a group of people [5].
- It can be used to forecast market movement based on news and by analyzing the blogs and social media to judge the sentiments of the masses.
- Sentiment analysis can be used to identify the clients with negative sentiment in social media or news to increase the margin for transactions with them for default protection.

III. Different Levels Of Sentiment Analysis

Sentiment analysis can be performed on three different levels depending on the granularities which are being considered during the process. The three levels of sentiment analysis are [2] [5]:

A. Document Level Sentiment Analysis

This is the simplest form of classification. The whole document is considered as the basic unit of information to classify the sentiment in that opinionated text. It is assumed that the document is having an opinion about a single entity. Classification of the full document is done as positive or negative. This approach is not suitable if the document contains opinions about different objects. Irrelevant sentences should be eliminated before processing.
Once the data has been processed, each document is encoded and stored as a standard vector of term weights. In the vector space model, there are several variations to attribute the weights to the terms. Two of the most common weights are: TF (Term Frequency) and IDF (Inverse Document Frequency).
Term Frequency: Here, in this approach, each term in a document is assigned a weight for that term, that depends on the number of occurrences of the term in the document. The simplest approach is to assign the weight to be equal to the number of occurrences of the term t in document d. Thus, term frequency can be defined as [3]:

TF (term, document) = the frequency of the term in the document

Inverse Document Frequency: The inverse document frequency of a term can be defined as [3]:

$$IDF(term) = \log \frac{Total\ number\ of\ documents\ in\ a\ collection, N}{Document\ Frequency\ of\ term}$$

Now TF-IDF combines the definition of term frequency and inverse document frequency to produce a composite weight of each term in each document. It can be computed by the following formula:

TF-IDF (term, document) = TF (term, document) * IDF(term)

B. Sentence Level Sentiment Analysis

Sentence level sentiment analysis is a more fine-grained analysis of the document. In this approach, the polarity is calculated for each sentence as each sentence is considered as a separate unit and each sentence can have different opinions. Sentence level sentiment analysis has two tasks [11]:

C. Subjectivity Classification

A sentence can be either a subjective sentence or an objective sentence. Subjective sentences have an opinion about the particular product and the opinions expressed in such sentences can be identified

and classified easily [11]. In the objective sentences, there is either no use or indirect use of sentiment words. The sentiments expressed in objective sentences are hard to identify and thus advanced algorithm is needed for the identification and classification purpose [8].

D. Sentiment Classification

A sentence can be classified as positive, negative or neutral depending on the opinion words present in it. The same document-level classification methods can be applied to the sentence level classification method [11]. The different sentiment classification methods have been discussed in the next section**.**

E. Feature Level Sentiment Classification

In feature level sentiment analysis the piece of text is analyzed as a feature of any product. Both the document level and the sentence level analyzes do not discover what exactly people liked and did not like. Aspect or Feature level performs fine-grained analysis. For example, a hotel can have an exotic location, but the services offered are not up to the mark. This problem involves identification of several sub- problems like identifying relevant entities, extracting their feature/aspects and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral.

IV. EXISTING TECHNIQUES FOR SENTIMENT ANALYSIS

Some of the existing techniques for sentiment analysis which are used frequently is discussed below [2] [4] [5]:

A. Machine Learning Method

Machine learning focuses on the development of computer programs that are capable of training themselves to develop and remodel when exposed to new data. It makes use of the training data to detect patterns in the dataset and adjust program actions accordingly. Machine learning algorithms are often categorized as supervised or unsupervised.

B. Supervised Machine Learning Techniques

In this approach, the training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations. New data is classified based on the training set. This requires the classifier to discern from the training data to veiled situations in a rational and logical way.
Some of the supervised learning methods are Decision Tree, Logistic Regression for classification purpose, Rule Based Classification, Linear Classification like the Support Vector Machine (SVM), and Neural Network, Probabilistic Classifiers like Naïve Bayes, Bayesian Network and Maximum Entropy.

C. Unsupervised Machine Learning Technique

Unsupervised machine learning techniques don't use training dataset for classification. Clustering algorithms like K-means clustering, Hierarchical clustering are used to classify data into categories. Neural Network can be also used for defining threshold values for the words and classify them based on the defined values. Point wise Mutual Information (PMI) is also one of the unsupervised classification methods for sentiment analysis.

D. Lexicon Based Method

The lexicon-based method is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase. Lexicon based method can be further classified as Corpus-based approach and Dictionary based approach [4].

V. PROPOSED METHOD

During the sentiment analysis process, we have to encounter a number of sentences or sentences of documents. All these documents or sentences may convey some opinion or maybe not [2]. In the proposed approach, the following steps have been followed for classification.

A. *First Step*: Access any unstructured data
B. *Second Step*: Apply the pre-processing steps on the given dataset. Some of the pre-processing techniques are [6]:

- *Stemming* is a widely used technique in text analysis. It is the process of removing inflectional affixes of the words, reducing words to their stems.
- *Lowering* converts all terms into lower cases.
- *Tokenization* is the process of breaking a text into tokens. A token is a non-empty sequence of characters, excluding spaces and punctuations.
- *Pruning* discards terms either appearing rarely or too frequently. Terms that rarely appear in document or terms that appear too frequently do not contribute to identify the topic of the document.

C. *Third Step:* After the pre-processing of the data, each document is encoded and stored as a standard vector of term weights [4]. This is done by computing the term frequency and inverse document frequency as discussed in section 3.1.

D. *Fourth Step:* Now apply the various classification methods as discussed in section 4 for classifying the processed dataset as positive, negative or neutral sentiments.

VI. Major Challenges In Sentiment Analysis

There are several open-ended issues and challenges in the field of sentiment analysis that still needs to be addressed. Some of them are [5] [9] [11]:

A Correct meanings of a word based on the context in which it is used needs to be understood as the same word can have different meanings for different domains. For example, *large size* can be positive opinion for hotels but negative for mobiles.

B. To determine the polarity of comparative sentences can be a challenge. For example, Product A is better than Product B. This review has positive word "better" but the author's preferred object is difficult to determine.

C. Handling negations is the most challenging task. If not handled properly, it can give completely wrong results. For example, There is a good chance that this product will not lead to allergic reactions. This review shows positive polarity but the presence of negation changes the effect completely.

D. Sarcasm in the reviews is very challenging to identify and to identify emotions expressed in the text.

The detection of spam and fake reviews

VII. Future Work

Sentiment analysis is very useful to identify current and future trends. In the future, major developments can be done to perform sentiment analysis at aspect/feature level which provides a finer grained analysis [4] [5]. In future, pure semantic, ontology and description logic approach can be applied to classify subjective and objective sentences and then classify them as positive, negative or neutral sentiment.

VIII. Conclusion

Sentiment analysis is an emerging field of data mining. It is an approach which allows using the unstructured data efficiently. This paper discusses about an overview of sentiment analysis, its applications and the levels on which it can be performed in the form of Document, Sentence and Feature level sentiment analysis. Various existing techniques have been discussed and a model has been proposed for sentiment analysis. Various challenges are also discussed which make sentiment analysis a difficult task.

References

[1] Sneha. Y.S, Dr. G.Mahadevan and S.Muthulakshmi "A Critical Review of Recommender Systems in Web Usage Mining Based on User Ratings", Proceedings of International Conference on Artificial Intelligence and Embedded Systems,2012

[2] Jalaj S.Modha Gayatri S. Pandi and Sadip J. Modha "Automatic Sentiment Analysis for Unstructured Data", Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering,2013

[3] http://nlp.stanford.edu/IR-book/html/htmledition/term-frequency-and-weighting-1.html

[4] Xing Fang and Justin Zhan "Sentiment Analysis using Product Review Data" Journal of Big Data a Springer Open Journal

[5] Bing Liu “Sentiment Analysis and Opinion Mining”
[6] Gautami Tripathi and Naganna S. “Feature Selection and Classification Approach for Sentiment Analysis”, Proceedings of Machine Learning and Applications: An International Journal,2015
[7] Pragati Vaidya “Opinion Mining and Sentiment Analysis in Data Mining” Scholars Journal of Engineering and Technology,2015
[8] Bo Pang and Lillian Lee “Opinion Mining and Sentiment Analysis”
[9] David Osimo and Francesco Mureddu “Research Challenges on Opinion Mining and Sentiment Analysis”
[10] Carlos Adriano Goncalves, Celia Talma Goncalves, Rui Camacho, Eugenio Oliveira “The Impact of Pre-Processing on the Classification of MEDLINE Documents”
[11] Seema Kolkur, Gayatri Dantal and Reena Mahe “Study of Different Levels for Sentiment Analysis”, Proceedings of International Journal of Current Engineering and Technology