Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 12, December-2017

A Survey On Techniques and Approaches for Opinion Mining

Allan Dsouza[1], Divay Mohan[2], Neha Bhamare[3], Diksha Rajput[4], Prof.A.M.Jagtap[5]

[1]*Department of Computer Engineering, AISSMS College of Engineering*
[2]*Department of Computer Engineering, AISSMS College of Engineering*
[3]*Department of Computer Engineering, AISSMS College of Engineering*
[4]*Department of Computer Engineering, AISSMS College of Engineering*
[5]*Department of Computer Engineering, AISSMS College of Engineering*

Abstract — ***Opinion mining is a fast-growing field of computer science. In many institutions worldwide and especially democratic systems opinion of people matter a lot. Understanding opinion of people is a very challenging task due to large amount of physical work involved in it. Current manual systems involve large amount of physical work and also causes high cost of operation. With increasing popularity of opinion mining various new techniques are also being developed as well as old techniques are being improved and modified to improve performance and accuracy. In this paper we provide a survey on various techniques available along with comparison between them for accuracy and other parameters.***

Keywords-Sentiment Analysis, Opinion Mining, Natural Language Processing, Computational Linguistics, Text Mining.

I. INTRODUCTION

Opinion Mining or sentiment analysis involves understanding the opinion or sentiment of a person or public as a whole. Understanding the sentiments of a person or public is very important to decide the opinion of people on a particular topic or a decision. Governments can make use of them to understand decision taken by them are being liked or disliked by the public as well as polling organizations can make use of opinion mining in pre-poll election surveys as well as other surveys.

Opinion Mining involves a lot of steps ranging from data set collection to sentiment classification. Each and every step can be carried out using different techniques available. As more and more research is being carried out in this field the techniques available are also increasing. Selection of best techniques depends upon several parameters such as accuracy, available computational capacity, etc.

There are many stages involved in opinion mining. The first step is collection of data. This is a very important step as data collection is the most important step involved. Without proper data collection opinion mining cannot be carried out accurately. The next step involves cleaning and stemming of words present in the data set. Cleaning and stemming are also important to improve the overall accuracy of the opinion mining task as it involves removal of unnecessary words and conversion of words to more relevant words.

The next task is application of the opinion mining algorithm. The main concentration of this paper is on this phase as this is the most time consuming and computationally important phase. Also due to research carried out on extensive scale we have many different techniques available at our disposal. These techniques involve various machine learning techniques like Naïve Bayes classification support vector machine classification and a wide range of corpus-based techniques such as TF (Term Frequency), IDF (Inverse Document Frequency), etc.

The last step involves checking the accuracy of the prescribed algorithm for the given data set. The accuracy may vary across data sets and types of data involved and checking the accuracy is very important to decide the current algorithm is useful for the current data set or not.

II. STUDY OF APPROACHES USED CURRENTLY

Authors of [1] have developed a food dish recommender system based on ingredients. They have developed a application named foodholic and user is asked to provide the name of the ingredient they would like to have in their dish. A crawler is used to crawl food recipes related websites and searches for the relevant dishes on the different websites having the particular ingredient. Now it makes use of reviews of this dish and calculate the score for each dish shortlisted. Based on this score the user is provided with recommendations.

The authors of this paper [2] have made a detailed survey on diff erent techniques used in sentiment analysis and their level of work along with advantages and disadvantages.

Authors of [3] try to analyze twitter posts about electronic products like mobiles, laptops, washing machines, television etc. using machine learning approach. By doing Sentiment analysis in specific domain, it is possible to identify domain information in sentiment classification. They present a new feature vector for classifying the tweets as positive, negative and extract public opinion about the product. Some of the most commonly used machines learning approaches are Naïve Bayes approach and Support Vector Machine(SVM) approach.

Authors of [4] In the paper have proposed a new algorithm. They have named this algorithm as sentiment fuzzy classification algorithm with parts of speech tags also called as POS tags are used to increase accuracy on a benchmarks movie review data set. Parts of speech tags are nothing but tags given to each and every word based on parts of speech it denotes.Altough Parts of speech tagging appears to be very simple, it is rather a complex task, it is because a single word can fall under many parts of speech, for example a single word like silver in the sentence "I won a silver medal" denotes noun, whereas in the sentence "You gave a silver speech" , silver is a adjective, We can understand from this example that parts of speech of a particular word not only depends on the individual word but on the context as well. A software which performs the task of POS tagging is known as POS tagger.

In this paper [5] the authors emphasize on the fact that sentiment analysis or opinion mining plays important role while making a decision towards a particular product or a service. But it is very important to consider certain quality measures like helpfulness, usefulness and utility while analyzing each review. The algorithm used in this paper makes use of the parameters. In this paper there are mainly three sophisticated methods explained which defines the sentiment analysis with respect to different aspects.

Authors in [6] have explained that Sentiment analysis is a very important process as it provides many valuable indictors in different domains such as medical, social and industrial domains. This paper presents a survey about sentiment analysis addressing the different concepts in this area, problems and its solutions, available APIs, tools used and presenting a list of open challenges in this area.

In paper [7] In this system collects data from Twitter social network site and does NLP techniques to extract features out from the tweets. Word Sense Disambiguation and WordNet synsets are also added to the feature vector to increase the accuracy of prediction. Then various Ensemble methods of classification are applied to classify the sentiment of the data as Positive, Negative and Neutral. Twitter is a very useful platform for performing sentiment analysis and opinion mining. Tweets are nothing but views of people in a particular field or domain. These tweets provide a very useful insight about a person's view of a particular topic or domain. These tweets are freely available for study and research work. This has opened the gateways for opinion mining and sentiment analysis and on a large scale. Twitter tweets also come up with challenges. As these tweets are informal, they contain many slang words as well as sarcasm, metaphors etc. Identifying these is a major challenge for data collected from tweets. This data collected should be cleaned properly in order to ensure high accuracy.

The authors have made use of ensemble learning for classification of tweets. The basic idea behind ensemble learning is to combine many hypothesis and make classification based on these hypotheses. The main motive is that a ensemble of many hypothesis is less likely to go wrong compared to a single hypothesis. For example, if we make use of five hypothesis, the likelihood of three out of five hypothesis to go wrong is much less than a single hypothesis to go wrong.

In paper [8] the authors have discussed a simple but efficient method to calculate the sentiments through texts obtained from online reviews of products. This method builds a dictionary of sentiment words with three degrees of comparison. These degrees of comparison are positive, comparative and superlative.

In paper [9] the authors determine whether a review is positive or negative. They make use of movie reviews as data, the authors found that that standard machine learning techniques definitively outperforms human produced

baselines. However, the three machine learning methods they employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization.

In this paper [10] the authors have shown the taxonomy of various sentiment analysis methods. This paper also shows that Support vector machine (SVM) gives high accuracy compared to Naïve Bayes and maximum entropy methods. Support Vector Machine, Naïve Bayes and maximum entropy methods are types of machine learning techniques.

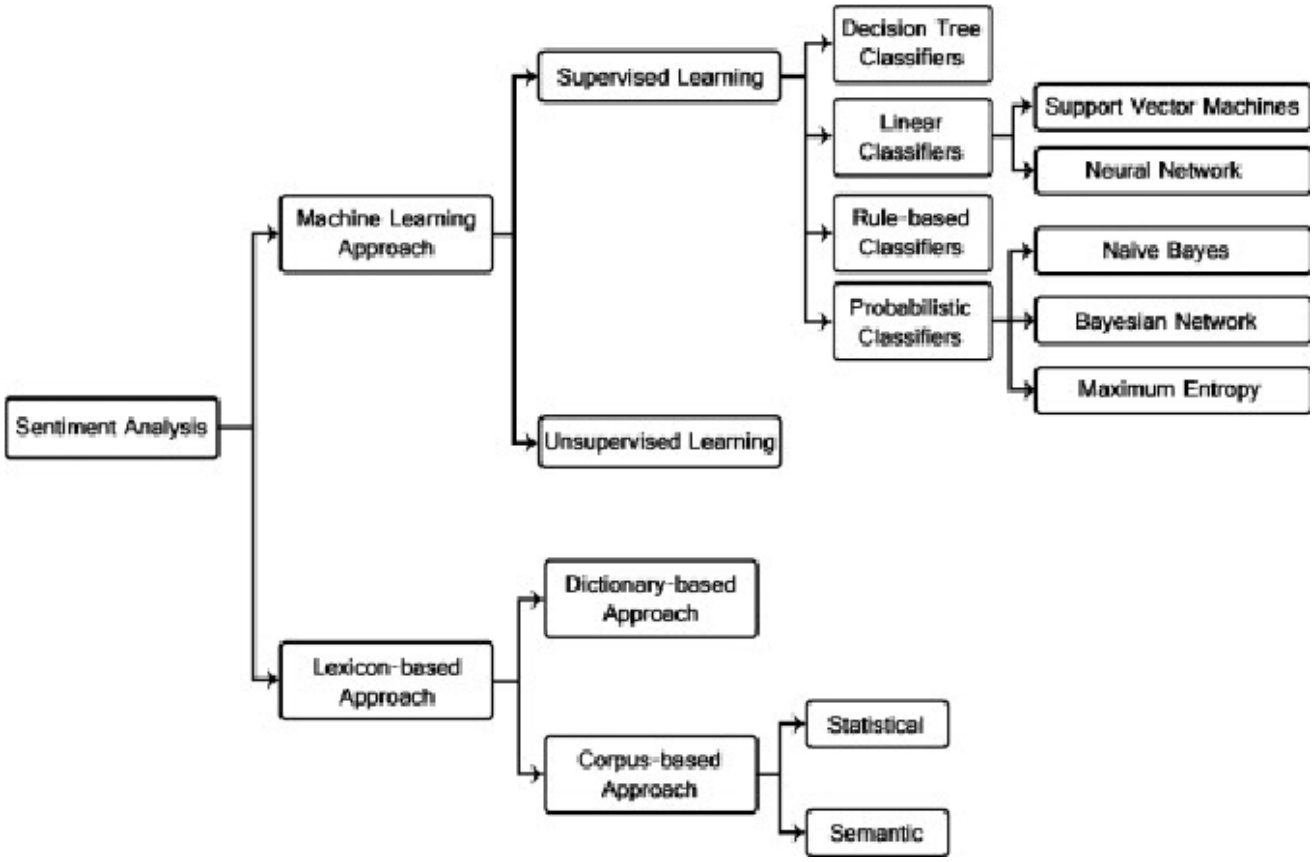

Figure 1. Commonly used techniques for sentiment analysis

III. CHALLENGES

Sentiment Analysis or opinion mining introduces vast range of challenges. These challenges range from data collection to handling sarcasm.

3.1. Sarcasm

Sarcasm poses a major challenge in sentiment analysis or opinion mining. Sarcasm involves usage of words which literally mean different but the context gives the document different meaning to it. Resolving sarcasm is a difficult task as computers cannot directly understand them. Humans make use of their knowledge to understand and resolve sarcasm but computers lack this ability.

There are more and more algorithms and approaches being introduced to identify and resolve sarcasm. Sarcasm is a major concern especially data collected from sources such as twitter, Facebook etc. Sarcasm often causes false positive. Let us understand sarcasm with an example "The food in school canteen was so hygienic that few students got

food poisoning". Here if we go by meaning of words we find positive words such as so hygienic, but by understanding the context and taking into account others words like food poisoning we come to know that the statement is not a positive statement but in fact a negative statement underlining the unhygienic food of the school canteens.

Identifying of such statements are very easy for humans but computers find it very difficult due to lack of knowledge or inability to understand context effectively.

3.2. Incorrect Spellings

The person may not be fluent in the language expressed and may create a lot of spelling mistakes. These words then do not make any sense and due to this the sentences too appear to be meaningless. Tasks such as POS tagging and identification of sentiments becomes very challenging as the misspelled words do not match with any word present in the database.

3.3. Incorrect Grammar

User not well verse with the language may make grammatical mistakes. These grammatical mistakes many times make the sentence meaningless. It makes identifying sentiment behind a particular sentence difficult. The results can be improved if grammatical incorrect words can be mapped to correct words.

3.4. Review Author Segmentation

The people who provide review can be termed as review authors. Depending upon the style of their tweets they can be segmented into different groups. This ensures that credibility evaluation becomes a easy task. In decision making credibility evaluation is very important.

3.5. Refinement or updating Lexicons

The accuracy and performance of sentiment analyzer depends upon the correctness of lexicon. Tuning of lexicons is required to accommodate new words and remove the words which are no longer used to get better results.

3.6. Handling Noise

Data collected from different sources and especially social media sources such as twitter or Facebook contain high amount of noise. This noise is a major problem in sentiment analysis and opinion mining tasks. Identification and removal of noisy data is a challenging task.

IV. CONCLUSION

Opinion mining or sentiment analysis is a widely growing field. Due to its applications in different domains it is truly a multidisciplinary field. Sentiment analysis also poses many challenges which ranges widely. Overcoming these challenges is very important to ensure high accuracy and proper usability of the results obtained through it. Increasing amount of research in this field has introduced new and unique techniques for sentiment analysis, these techniques ensure higher accuracy and simplicity. Availability of these different techniques and their successful usage by different researchers
provides us with wide variety of techniques for our usage and which technique should be used can be decided based on suitability of the techniques to that domain.

REFERENCES

[1] Anshuman, Shivani Rao, Misha Kakkar," Rating Approach Based on Sentiment Analysis",2017, IEEE.

[2] Harshali P Patil, Dr Mohammad Atique," Sentiment Analysis for Social Media: A Survey",2015, IEEE.

[3] Neethu M S, Rajashree R," Sentiment Analysis in Twitter Using Machine Learning Techniques",4 July 2013,4th ICCCNT.

[4] MS K Mouthami, Ms K Nirmala Devi, Dr. V Murli Bhaskaran," Sentiment Analysis and Classification Based on Textual reviews

[5] Zhu Nanli, Zou Ping, LeeWeiguo, Cheng Meng," Sentiment Analysis: A Literature Review",2012, IEEE.

[6] Khaled Ahmed, Neamat El Tazi, Ahmed Hany Hossny," Sentiment Analysis Over Social Network: An Overview,2015, IEEE.

[7] Monisha Kanakaraj, Ram Mohana, Reddy Guddeti," NLP Based Sentimental Analysis on Twitter Data Using Ensemble Classifiers,2015, IEEE.

[8] Santanu Mandal, Sumit Gupta," A Lexicon-Based Text Classification Model To Analyze And Predict Sentiments From Online Reviews".

[9] Bo Pang, Lillian Lee, Shivkumar Vaithyanathan," Thumbs up? Sentiment Classification Using Machine Learning Techniques".

[10] Shivprasad T K, Jyothi Shetty, "Sentiment Analysis of Product Reviews: A Review",2017, IEEE