Sentiment Analysis: An insight into Techniques, Application and Challenges

Deepti G. Aggarwal

CSE Department, JSS Academy of Technical Education, Noida, Uttar Pradesh, India

*Corresponding Author: deeptigaggarwal@jssaten.ac.in

Abstract: Sentiment analysis is a fast-growing field which has gained importance in many sectors. It plays a very major role in the field of Natural Language Processing. With the increase in awareness among people and easy access to internet, massive amount of data is available on internet that can be analysed and hence utilized in a constructive manner. This is a survey paper which gives an overview of sentiment analysis along with an insight into the techniques used for sentiment analysis and various applications and challenges in this area.

Keywords: Sentiment analysis, Machine learning, Lexicon, Classification, Opinion Mining

I. INTRODUCTION

A feedback or a customer review can help any organization/industry in improving its product or services. Giving feedback is also equally important since this is the easiest way for a user to express his views about a particular organization or it's product or service. Feedback taken is of no use until unless it is analysed in a proper manner. Analysing feedback is easier if it is done on objective/factual questions, whose answer is in the form of levels like 0,1,2,3,4 or may be 'strongly disagree' or 'disagree' or 'neutral' or 'agree' or 'disagree'' or 'strongly disagree' as compared to subjective analysis where the user is free to write the feedback in his/her own words. To analyse such a kind of data, sentiment analysis is used to gain an understanding of the attitude, emotions, feelings and opinions of the users about a particular product/service. Lot of research has been done and still going on sentiment analysis in the area of Natural Language Processing.

This paper is divided into 4 parts. First part gives an overview of the classification of sentiment analysis. Second part gives an insight into the techniques used for sentiment analysis. Third section gives an overview of the applications of sentiment analysis. Fourth section tells about the challenges faced in sentiment analysis.

II. CLASSIFICATION

Sentiment analysis is the classification of text based on the orientation of opinions given in the text, hence also called as opinion mining. It is a very important part of Natural Language Processing. Sentiment analysis plays a very important role in industries, social sites, academia and many more. Sentiment analysis can be done at three level [1]s:

- document level
- sentence level
- aspect (feature) level

Document level works best if the analysis is to be done on small data set (like a document written by a single person or the opinions about a single entity).

Sentence level classification is better than document level classification. A sentence can be classified as subjective or objective. Objective sentence basically represents factual information and can be analysed more easily as compared to subjective classification. Subjective sentences contain the emotions, views, or feelings of the opinion holder in his/her own language. Subjective sentences can be classified as having positive, negative or neutral sentiments

Third level of classification is Aspect based classification. Here sentiment analysis is done on the basis of different features/aspects of a particular product. Aspect based classification can be further classified as: aspect extraction and aspect sentiment classification. Aspect extraction includes identifying the various features of an object e.g. the aspect for classification for a mobile phone can be: size, weight, camera, price and RAM. Aspect sentiment classification means identifying the polarity of reviews for

different aspects of a particular object like positive, negative or neutral.

III.　TECHNIQUES

Sentiment analysis is a very lucrative field because of huge amount of data available on the internet. Henceforth to deal with sentiment analysis, multiple techniques have been evolved. Many researches have been done in the field of sentiment analysis and many more are going on. Sentiment analysis can be done basically with three different approaches:

1. Rule Based Approach
2. Lexicon Based Approach
3. Machine Learning Approach
4. Hybrid Approach (Combination of both)

Fig 1 shows the classification of sentiment analysis techniques.

III.I　Rule Based approach

In rule-based approach, certain rules are defined in order to get the opinion. The tokens are generated from the sentence and are checked for the polarity. If the word has positive sentiment, polarity score is incremented and if the word has negative sentiment, the polarity score is decremented. If some words are not in the database, the words are added[2].

III.II　Lexicon Based Approach

Lexicon is an important part of sentiment analysis. In this approach, the text to be analysed is matched with the sentiment dictionary, which contains the opinion words that help to determine polarity of text. [3] This technique is based on unsupervised learning because no prior training is required to analyse the data.

The basic steps pf lexicon-based techniques are summarized below[4]:

(i)　The text to be analysed is pre-processed like removal of HTML tags and noisy characters
(ii)　Text sentiment score(SS) is initialized to 0
　　　SS=0
(iii)　Tokens are generated from the text and checked for its presence in sentiment dictionary. If it is present in the dictionary,
　　a.　If token is positive, then SS is incremented
　　　SS++
　　b.　If token is negative, then SS is decremented
　　　SS--

(iv)　Compare the total SS with the threshold value(T)
　　a.　If SS> T, then the text polarity is positive
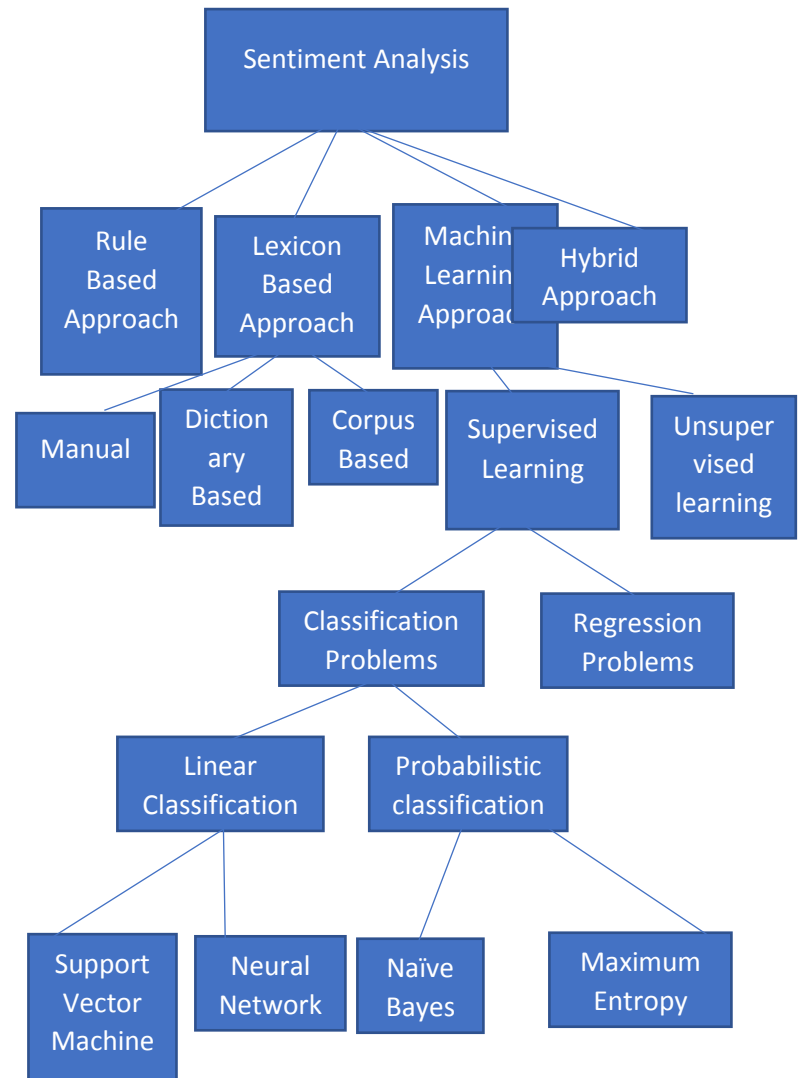　　b.　If SS< T, then the text polarity is negative

Fig 1: Classification of sentiment analysis techniques

To construct a sentiment lexicon, three approaches are used:

(i)　Manual Construction: It is a very tedious and time-consuming task and hence rarely used now a days.
(ii)　Dictionary based method: In this technique, some of the opinion words are collected manually. The orientation of these words is known a-priori.

These words form the seed of a dictionary and this seed list is enhanced by adding the synonyms and antonyms of seed list by looking into the WordNet dictionary [3]. All the synonyms of a seed will have the same orientation and all the antonyms will have the opposite orientation. But, it's limitation is that it doesn't work with domain specific orientation.

(iii) Corpus based method: This technique requires large data set in order to check the orientation of text. It will have access not only to sentiment labels, but also to a context which can be used in a machine learning algorithm. It relies on syntactic pattern in large corpora. This method needs very large training set. It has the advantage over dictionary-based approach that it can find domain specific opinion words and their orientations.

III.III   Machine learning approach

Machine learning technique basically works by training an algorithm on a large data set, also called training set, which has known output. [5] Later the algorithm is tested on a test set to get the results. Machine learning approach can be classified as supervised learning and unsupervised learning.

1.   Supervised learning**:** In supervised learning, all data is labelled and the machine is trained for every input with a particular target. The input variable (X) and the output variable (Y) is known.

$$Y = f(X)$$

The machine has to predict the function f that maps X to Y. If the output variable is a category, like 'costly' or 'cheap', then the problem is called a classification problem. If the output variable is areal value, like 'price', then the problem is a regression problem.

1.1   Classification problem: Classification problem is generally about predicting a label. Classification filters the data in different classes e.g. when given some e-mails, the data can be filtered out as 'spam' or 'not spam'. Classification can be linear as well as probabilistic.

(i)   Linear classification: A linear classifier identifies the class of an object based on the value of linear combination of characteristics. A linear classifier can represent any mapping that is linearly separable. There are many kinds of linear classifiers like Support Vector Machine and Neural Network.

a. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a type of supervised machine learning, which is mainly used to segregate (or classify) multiple type of inputs. Each individual input can be hypothetically represented on a plane as a single coordinate. The SVMs can be used to create a hypothetical plane separating different types of inputs in to different classes.

Consider the below two types of inputs (Fig 2), represented by blue stars and orange squares. The line between them is created by SVM, as identifying a clear plane of separation between different type of inputs.
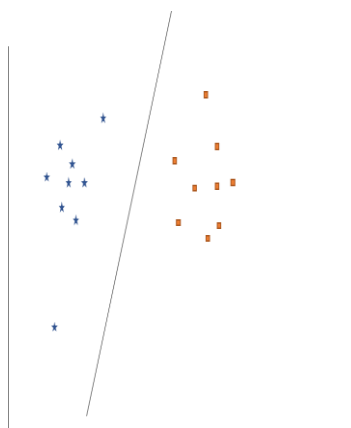

Fig 2: Classification using SVM

There are various possibilities in SVM. It can take of some outlier samples and still result in a better classification. It can be used to classify more than two types of inputs as well.

b. Neural Network(NN)

A neural network is a complex architecture of many layers of interconnected units. Each unit represents a complex function of the input value. It takes input from the previous step and gives output, which is connected to the unit in the next step. Each such unit is called a neuron, and the complete architecture is called a Neural Network (NN). The parameters of the functions in each unit are learned as part of the training step of the neural network.

There is one input layer and one output layer. The design of input and output layers is application specific. For example, in a sentimental analysis, the input can be the sentence or paragraphs to be analysed. The output could be a classifier, which classifies it as neutral, negative or positive. It could be designed for ranking (viz. from 1 to 5) on various aspects like anger or happiness.

The inner layers are called hidden layers. In fig 3, there is one hidden layer. The input and hidden layer each have 4 units. A typical NN can be made of several hidden layers, and several units in each hidden layer. A NN is useful because each unit (neuron) in each layer only, very slightly effects the outcome of the overall output. During the training steps of a NN, it learns very complex features of a given problem. Each unit (neuron) in each layer can affect the overall outcome only extremely slightly. But when that outcome passes through several layers of neurons, the effect is compounded and gives very accurate results. The outcomes of the NN are so accurate that scientists are not sure how NN are able to learn such complex features.
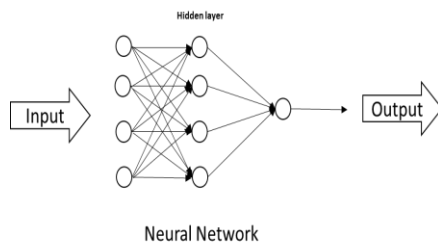

Fig 3: Neural Network

(ii) Probabilistic classification:

Probabilistic classifier is a classifier which works on probability and is able to predict a probabilistic distribution over a set of classes rather than giving a class to which an object belongs to. Probability classification can be done using Naïve Bayes or maximum Entropy Techniques.

a. Naïve Bayes:

This classifier works when the training set is small. It is in the family of sample probabilistic classifier which is based on Bayes theorem. The probability that a sentiment (ST) occurs in a given sentence (SN) is determined by the rule [5]:

$$P(ST/SN) = P(ST)P(SN/ST)/P(SN)$$

b. Maximum entropy classifier:

The maximum entropy classifier is a probabilistic classifier like Naïve Bayes except that it doesn't assume that features act independently. It belongs to the category of exponential models. It is used mainly when we don't know the prior distributions [3]. It is normally used in sentiment analysis for text classification where the words are not independent. Time required for training is more in Maximum entropy as compared to Naïve Bayes.

1.2 Regression Problem

Unlike classification problems, where output is discrete value, regression problems have continuous values. The different regression algorithms are linear regression, regression trees etc. Predicting the temperature in a city is a regression problem because it is a real value whereas predicting the trend in the stock market is not a regression problem because it is a discrete value(rise/fall).

2. Unsupervised Learning: Unsupervised learning is another technique used in the field of machine learning that is used to draw conclusions from the data sets that do not contain labelled responses to input data. Cluster analysis is normally used to identify the patterns in a given data. Several clustering algorithms are used like hierarchical clustering, k-means clustering and so on.

III.IV Hybrid Approach

Now a days, many researchers are using machine learning approach in combination with lexicon based approach, which is called as hybrid approach [4]. This approach combines the advantages of both the approaches – stability and readability of lexicon-based approach and high performance and accuracy from a supervised machine learning approach.

IV. APPLICATIONS

Sentiment analysis is a fast-growing tool used in the world of Natural Language Processing in order to detect and analyse people's emotions, feelings, views and different perspectives about a certain product/service. It is advance stage od data mining which uses various techniques to extract the information from a certain unstructured text. It has gained lot of popularity because of its versatile applications in various fields. Some of which are listed below:

A. Social Media
People use social networking sites now a days to express their views in many contexts. The media may be Facebook, Twitter, or any other. The currents issues are discussed on social networking websites through blogs or different forums, which gives the researchers an opportunity to analyse the text in various forms and come up with even better algorithms.

B. Industries or business organization

Sentiment analysis play a very important role in business world. Industries can take the survey of their products or services from their customers and hence improve the same.

[6] Different customers will have different views about some product/service. And the data set is quite large to manually deal with. Hence a sentiment analysis or opinion mining tool will prove to be very helpful in understanding the average customer reviews.

C. Education

In education field also, sentiment analysis plays a very important role. Since the feedback is taken by the students or parents at the end of session/semester, the analysis of this feedback will give a way out to improve the teaching-learning process. The faculty will come to know the expectations of the students and limitations of self. This analysis may help the faculty to improve on his teaching methods. Not only this, it can be quite helpful for the management also like if they get the summarized feedback of faculty through an automated analysis tool, the appraisal process can be done in an efficient manner.

D. E-Commerce

Buying and selling goods/services online is called e-commerce. Sentiment analysis is very helpful in this context. If a person wants to buy something online, he can take a survey of the reviews given by other persons who bought the same product, before buying one. If all the reviews are analysed and summarized, it will save lot of time and effort for the buyer. Sentiment analysis can help in summarizing the different reviews. Similarly, if a user wants to sell something, he can take a survey of the market before actually quoting the price of his product.

E. Finance Sector

Sentiment analysis plays a very important role in finance sector[9]. A user can look at the market trends before investing in the stock market. Monitoring of financial news on the web can be done with the help of automatic sentiment detection tools.

F. Hospitality industry

People share their experiences about their travel like hotel or travel providers. Many sites which are very popular for this are TripAdvisor, Trivago etc. Before getting the bookings done, people look for the reviews of other users and analyse them manually to reach any conclusion. [6] If some automated analysis tool is available for analysing the reviews, the task becomes much easier. Not only the customers are benefitted from this, but also the hotel management authorities and travel providers can improve their services by looking at the analysis of feedbacks given by customers.

G. Politics

He views of voters about various issues can be analysed by politicians in order to work in a better way. During elections, various surveys take place, which when analysed can help a common man in deciding the right candidate.

H. Entertainment

In the field of entertainment, sentiment analysis plays a vital role. People look at the reviews of a movie or play or any other event before watching them. From the several reviews, if the feature-based extraction is done like story of a movie is good, direction is bad, casting is good, music is very good and so on, then a user can easily decide whether to go for that movie or not based on his preferences about different features of a movie.

I. Medical sector

The views of doctors and patients about a particular treatment or medicine can help the authorities to work in a particular direction e.g. if a particular procedure is reviewed as costly by many patients, then the price factor may be reconsidered for future perspectives.

J. Spam Review Detection

In the world of online reviews, where there are genuine reviews, there exists lot of spam reviews also which can be given by some individuals or may be the organizations themselves for their advertisement. Detection of spam reviews is a big application in sentiment analysis e.g. an e-mail may be considered as 'spam' or 'not spam'. This task has great impact on industrial communities also.

V. CHALLENGES

Even though many researches have taken place in the field of sentiment analysis, lot of scope is still there to improve the accuracy of sentiment analysis. Some of the challenges in this field have been summarized below.

A. Spam review detection

Where on one side, the reviews on a particular product/service are very helpful, on the other side, there are lot of fake reviews also. These fake reviews may be put by a person or some organization in order to upgrade/ degrade the overall rating of a particular product/service. Detecting these fake (spam) reviews is a big challenge in the field of sentiment analysis [6][7]. Even though many researches have taken place in order to overcome this challenge, still there is lot of scope in this area.

B. Sarcasm detection

Sarcasm can be defined as "a sharp, bitter, or cutting expression or remark, or a bitter gibe or taunt." Sentences in a text may be straight forward as well as sarcastic sentences, which means the sentence seems to give one meaning, but actually it has some other meaning e.g. a sentence seems a praise but actually it is a taunt. In other words, the polarity of a sentence can be false positive. Detecting sarcasm and finding the correct sentiments in a sentence is one of the biggest challenge for sentiment analysis researchers[6][8] .

C.   Noise removal

When we analyse some data, it actually contains noise along with the relevant data that needs to be filtered out.

D.   Abbreviations and slangs

To use abbreviations and slangs is a sort of status symbol these days. There are millions of slangs and abbreviations that are used in communication. To analyse the text while considering these slangs and abbreviations is a big challenge that is to be dealt with when working with sentiment analysis.

E.   Grammatical errors

When a big data set is taken, the same view may be written using several sentences of which many may be grammatically wrong also. To find out the polarity of a sentence [6] , the grammatical errors are to be dealt with. Very less work has been done in this direction.

F.   Feature extraction

To identify the features of an object whose sentiment analysis is being done is a challenging task. Until unless the features are correctly identified, the polarity cannot be calculated appropriately.

G.   Real time analysis

Static data can be analysed easily as compared to real time data. With the increase in the use of social networking sites, large data is available for opinion mining which increases the need for an automated system

H.   Sentiment Analysis using audio and video

Many researches are going on for text analysis in the field of opinion mining. But data is available in audio as well as video form also which needs to be analysed. A good scope of research is there in audio and video analytics in the field of sentiment analysis.

I.   Sentiment Analysis of data in mother tongue

The data that is being analysed is normally in English language. The reviews from different users can be in any language. Translators are required to convert the reviews from the people's mother tongue to English language for sentiment analysis.

J.   Ambiguity in word sense

The same word can have different meaning in different contexts. Like 'small size' has positive polarity for mobile phones but negative polarity for a multiplex. [1] To analyse the meaning of a word in proper context is a challenging task.

K.   Negations

Negations in a sentence can reverse the polarity of that sentence. These need to be handled properly otherwise the results can be disastrous.

VI.   CONCLUSION

Sentiment analysis has been discussed in this survey paper along with the classification and various techniques used to implement sentiment analysis in very simple language. A broader view of application of sentiment analysis was also discussed. Many researches have taken place in this field. Still many challenges are there for the researchers, some of which have been discussed in this paper. Analysing the reviews from different blogs help the people in better understanding of the object and good decision making. A big scope is there in the field of sentiment analysis for the researchers to work on.

REFERENCES

[1]   Seema Kolkur†, Gayatri Dantal and Reena Mahe(2015) , "Study of different levels for sentiment analysis", International Journal of Current Engineering and Technology ,Vol.5, Issue 2, pp 768-770, April 2015

[2]   Ankush Sharma1, Aakanksha(2014) "A comparative study of sentiments analysis using rule based and support vector machine" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 3, pp 5898-5900, March 2014

[3]   Walaa Medhat a,, Ahmed Hassan b, Hoda Korashy(2014) "Sentiment analysis algorithms and applications:A survey", Ain Shams Engineering Journal Volume 5, Issue 4, pp 1093-1113, December 2014

[4]   S. M. Vohra, J. B. Teraiya(2013) "A comparative study of sentiment analysis techniques", Journal of information, knowledge and research in Computer engineering, Vol 02, Issue 02, pp 313-317, October 13

[5]   Devika M D, Sunitha C, Amal Ganesha(2016), "Sentiment Analysis-A comparative study on different approaches", Procedia Computer Science 87, pp44 – 49 In Association with Elsevier, Fourth International Conference on Recent Trends in Computer Science & Engineering, Chennai, Tamil Nadu, India, March 2016

[6] Harshali P. Patil1 and Mohammad Atique(2017) "Applications, issues and challenges in sentiment analysis and opinion mining– A User's perspective", International Journal of control theory and Application, Vol 10 ,Issue 19, pp 33-43, 2017

[7] Haseena Rahmath P(2014) "Opinion mining and sentiment analysis -challenges and applications", International Journal of Application or Innovation in Engineering & Management,Volume 3, Issue 5, pp 401-403, May 2014

[8] U. Aggarwal and G. Aggarwal, "Sentiment Analysis : A Survey", International Journal of Computer Sciences and Engineering , Volume-5, Issue-5, pp 222-225, May 2017

[9] Uma Aggarwal, Gaurav Aggarwal , "Sentiment Analysis on Demonetization using SVM", International Journal of Computer Sciences and Engineering, Vol.5, Issue.6, pp.183-187, 2017.

Author's Profile

Deepti G. Aggarwal is M.Tech(CSE), B.Tech(CSE) and pursuing Ph.D. from Jaipur National University, Jaipur. She is working as Assistant Professor at JSS Academy of Technical Education, Noida with a teaching experience of 17 years.