A Novel Approach for Text Sentiment Analysis in Social Network

Kunwar Aman Singh and Mr. Mahesh Kumar
[1]*M.Tech Student,* [2]*Assistant Professor*
CSE Department, Ganga Institute of Technology and Management, Kablana, Jhajjar, Haryana, India

Abstract - In daily life, our beliefs and perceptions of reality are largely dependent on how other people analyze the world. People mostly share their opinions about different entities like organization, products, services etc. on social media, forums, and blogs. Microblogging sites have millions of people sharing their thoughts daily because of its characteristic short and simple manner of expression. It's very difficult to analyze this enormous amount of data manually. Thus, sentiment analysis is the process which automatically identifies people's attitudes and emotional states.

Keywords *- Sentiment analysis, Twitter, tweets, Microblogging*

I. INTRODUCTION

Ongoing increase in wide-area network connectivity promise vastly augmented opportunities for collaboration and resource sharing. Now-a-days, various social networking sites like Twitter1 , Facebook2 , MySpace3 , YouTube4 have gained so much popularity and we cannot ignore them. They have become one of the most important applications of Web 2.0 [1]. They allow people to build connection networks with other people in an easy and timely way and allow them to share various kinds of information and to use a set of services like picture sharing, blogs, wikis etc. It is evident that the advent of these real-time information networking sites like Twitter have spawned the creation of an unequaled public collection of opinions about every global entity that is of interest. Although Twitter may provision for an excellent channel for opinion creation and presentation, it poses newer and different challenges and the process is incomplete without adept tools for analyzing those opinions to expedite their consumption. More recently, there have been several research projects that apply sentiment analysis to Twitter corpora in order to extract general public opinion regarding political issues [2]. Due to the increase of hostile and negative communication over social networking sites like Facebook and Twitter, recently the Government of India tried to allay concerns over censorship of these sites where Web users continued to speak out against any proposed restriction on posting of content. As reported in one of the Indian national newspaper [3] "Union Minister for Communications and Information Minister, Kapil Sibal, proposed content screening & censorship of social networks like Twitter and Facebook". Instigated by this the research carried out by us was to use sentiment analysis to gauge the public mood and detect any rising antagonistic or negative feeling on social medias. Although, we firmly believe that censorship is not right path to follow, this recent trend for research for sentiment mining in twitter can be utilized and extended for a gamut of practical applications that range from applications in business (marketing intelligence; product and service bench marking and improvement), applications as subcomponent technology (recommender systems; summarization; question answering) to applications in politics. This motivated us to propose a model which retrieves tweets on a certain topic through the Twitter API and calculates the sentiment orientation/score of each tweet. The area of Sentiment Analysis intends to comprehend these opinions and distribute them into the categories like positive, negative, neutral. Till now most sentiment analysis work has been done on review sites [4]. Review sites provide with the sentiments of products or movies, thus, restricting the domain of application to solely business. Sentiment analysis on Twitter posts is the next step in the field of sentiment analysis, as tweets give us a richer and more varied resource of opinions and sentiments that can be about anything from the latest phone they bought, movie they watched, political issues, religious views or the individuals state of mind. Thus, the foray into Twitter as the corpus allows us to move into different dimensions and diverse applications.

II. SENTIMENT ANALYSIS

Sentiment Analysis (SA) is a process of information gathering task to extract user's feelings conveyed in a positive or a negative way. SA helps to retrieve the opinion of a user about any entity or organization. Due to the exponential increase in the exchange of public opinions at Internet, SA has become a challenging task in today's world. Sentiment analysis can be performed at three levels such as aspect/ feature, sentence and document level. SA mainly finds the polarity and degree of the sentiment in terms of positive, negative or objective (facts) from the text.

a) Application of Sentiment Analysis

SA has many applications in various fields. Most of the ecommerce activities use SA. It helps the companies in development of new products or services by gathering the opinions of customers about their products or services. SA helps in finding functional and non-functional requirements of the products from customer's opinions. It also helps in identifying online recognition of company's products or

services. It can assist government in judging their strengths and weaknesses by analysing sentiments of people.

b) Challenges of Sentiment Analysis

There arise many challenges while performing SA. There may be sentences having no sentiment bearing word but consist of indirect sentiment and it's very difficult to identify it. Identification of semantics and pragmatics is also an important challenge in sentiment analysis. SA also becomes a difficult task as polarity of words changes from domain to domain. Sentences may consist of multiple entries, so it is important to find out the entity about which opinion is being expressed. Subjectivity detection and handling of negation is also an important task of sentiment analysis (Turney, 2002). All these challenges need to be handled to develop an effective sentiment analysis system.

III. IMPLEMENTATION

The flow diagram of our process which we have used is given in the figure 1. In our experiment we had performed four operations which are discussed.

The complete process is carried out in steps given below

1.First we pickup a dataset from the internet.

2.We pre-processed the data to tokenisd the tweet.

3.After pre-processing we pass the pre-processed data to our feature extractor which extract the feature using Bag of Word model.

4.After this we pass the extracted feature to our classifier which is random-forest classifier to classify our model.

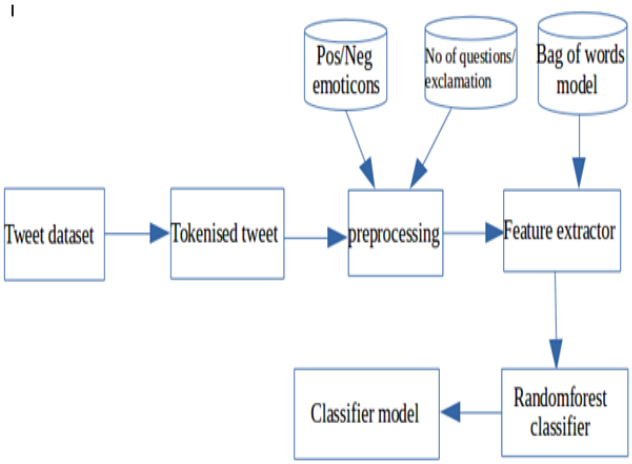

Fig.1 Flow Diagram of the process

IV. RESULTS

We took dataset of 9500 tweets from which we have used 5471 tweets to train our classifier and the rest to test the classifier. Firstly we perform pre-processing of our dataset using natural language processing in which we tokenise the tweet, tag the part of speech, collect all the stopwords etc,After the pre-processing we extract some extra features from the tweets like number of exclamation marks, number of question mark, number of positive emoticons and number of negative emoticons and pass it to the classifier which than classify the data. The result of experiments are discussed below.

Sentiment type distribution in training set

Analysing the above graph we can conclude that out of 5900 tweets in training set:

Number of positive tweets are 2800 approx,

Number of negative tweets are 2200 approx.

Number of neutral tweets are 900 approx.

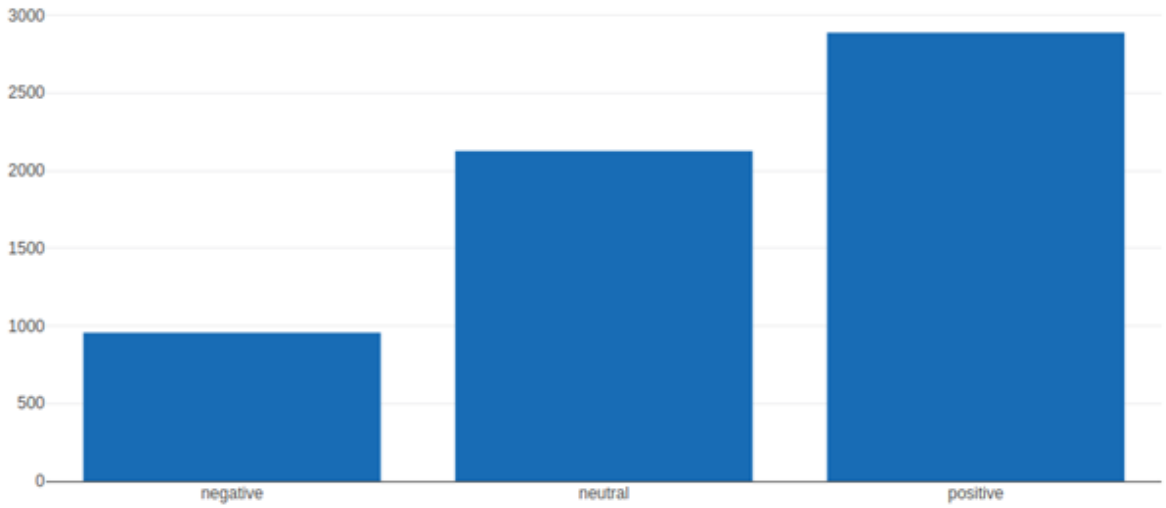

Fig. 2 Graph showing Sentiment distribution

Top words in buid wordlist

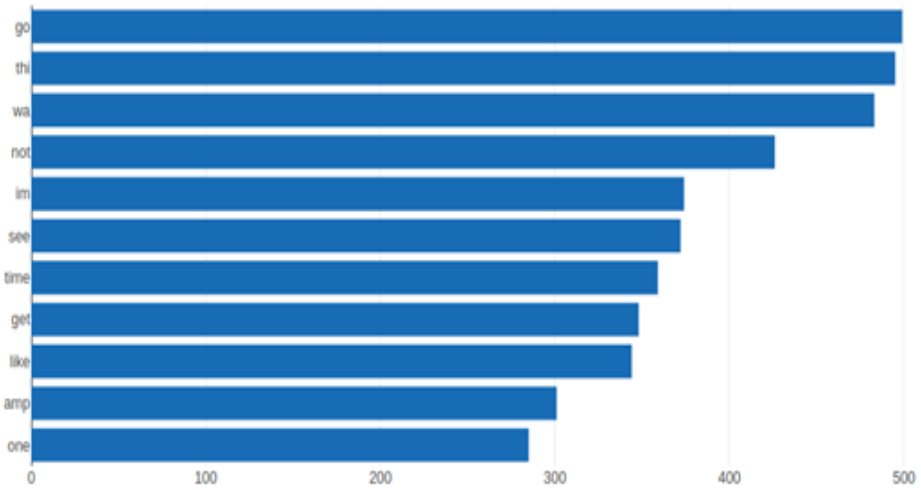

Fig. 3 Top words in build wordlist

Sentiment distribution using extra features:

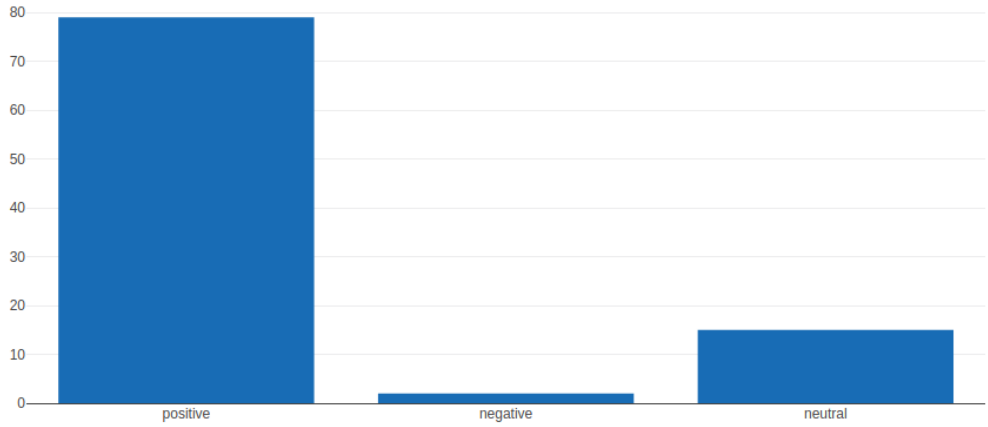

Fig.4 Sentiment distribution using no of positive emoticons

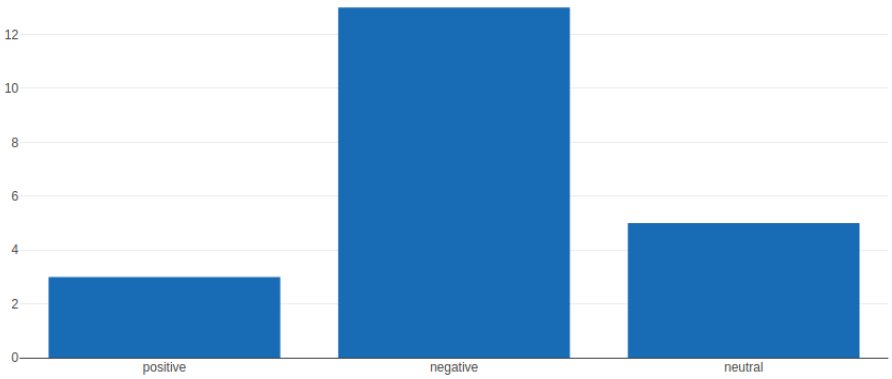

Fig .5 Sentiment distribution using no of negative emoticons

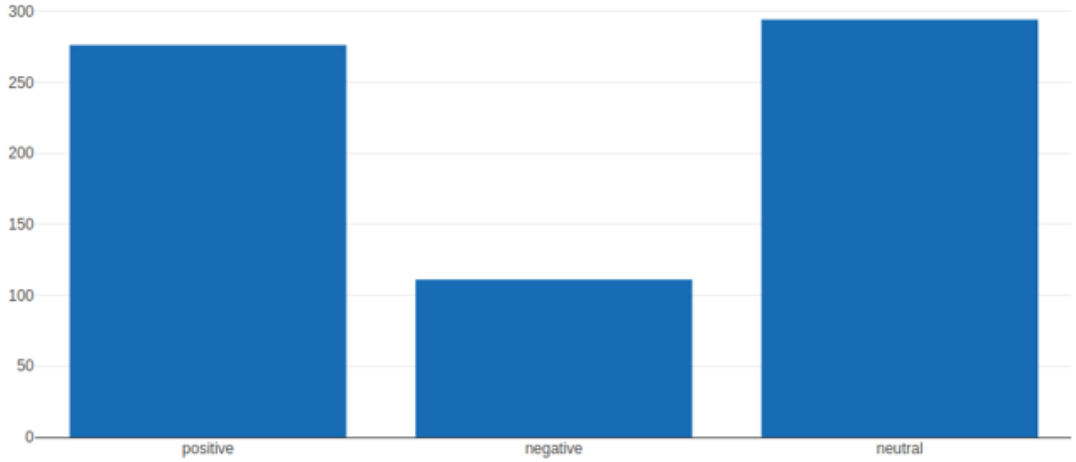

Fig. 6 Sentiment distribution using no of exclamations

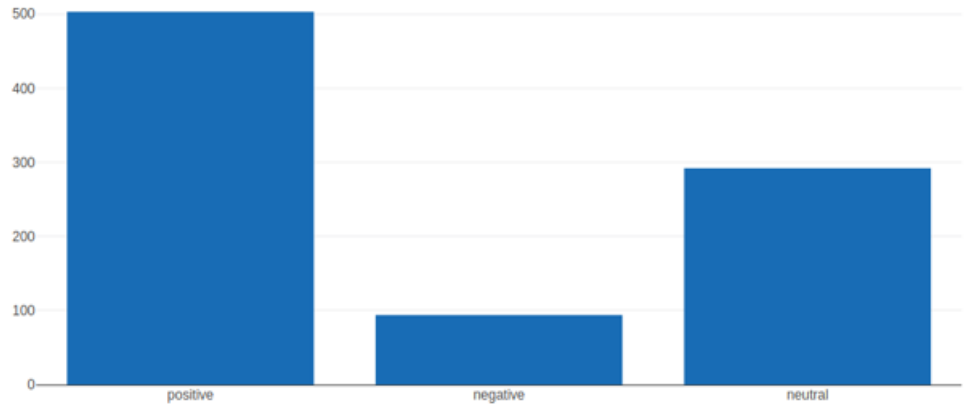

Fig. 7 Sentiment distribution using no of hashtags

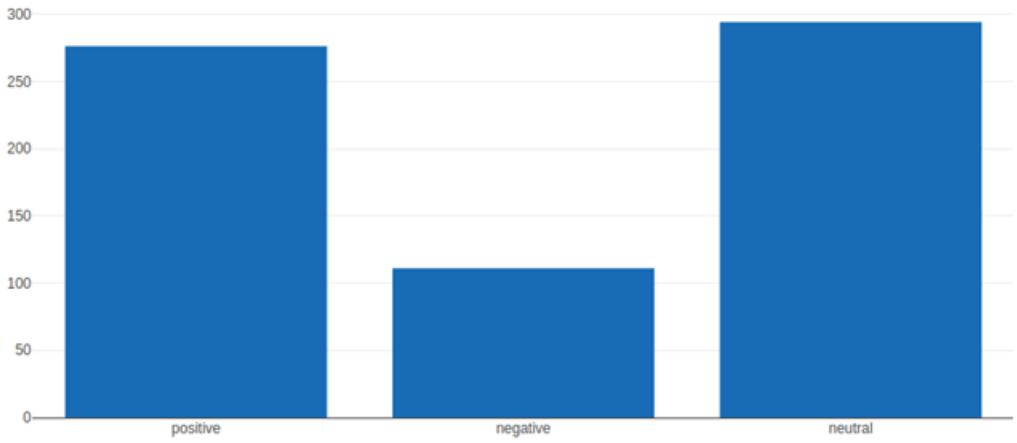

Fig. 8 Sentiment distribution using no of question-marks

```
Predicting time 3.6000521183013916s
=================== Results ===================
            Negative     Neutral     Positive
F1       [0.33595801 0.50823938 0.72674419]
Precision[0.53333333 0.51675485 0.66489362]
Recall   [0.24521073 0.5        0.80128205]
Accuracy 0.6035648432698217
================================================
================================================
```

Fig. 9 Accuracy using random forest classifier

Most important feature:

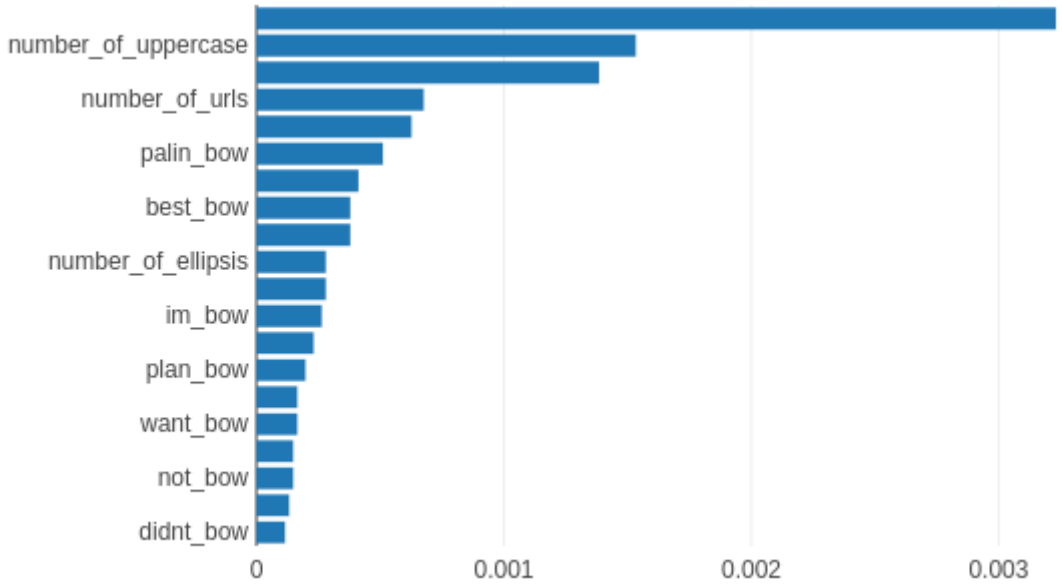

Fig.10 Most important features

V. CONCLUSION

The task of sentiment analysis, especially in the domain of micro-bloging, is still in the developing stage and far from complete. So we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance.

Right now we have worked with only the very simplest models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized ifs effect should be.

VI. REFERENCES

[1]. L. Colazzo, A. Molinari and N. Villa. "Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world", International Conference on Education Technology and Computer, 2009, pp.321-325.
[2]. nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf
[3]. National Daily, Economic Times: articles.economictimes.indiatimes.com › Collections › Facebook
[4]. K. Dave, S. Lawrence, and D.M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews". In Proceedings of the 12th International Conference on World Wide Web (WWW), 2003, pp. 519–528.
[5]. A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320–1326.
[6]. R. Parikh and M. Movassate, "Sentiment Analysis of UserGenerated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

[7]. A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper ,2009
[8]. J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005
[9]. L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp. 36-44.
[10]. S. Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", Stanford University,2010
[11]. A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1–15.
[12]. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 30–38
[13]. A. Kumar. and T. M. Sebastian, "Sentiment Analysis: A Perspective on its Past, Present and Future", International Journal of Intelligent Systems and Applications (IJISA), MECS Publisher, 2012 (Accepted to be published)
[14]. A. Kumar and T. M. Sebastian, "Machine learning assisted Sentiment Analysis". Proceedings of International Conference on Computer Science & Engineering (ICCSE'2012), 2012, pp. 123-130.