ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 4, April 2018

Opinion Mining On Twitter Data for Sentiment Classification

[1] G. Anuradha, [2] CH. Raga Madhuri, [3] M.Srikanth
[1] Associate Professor, VRSEC, VIJAYAWADA, [2] [3] Assistant professor, VRSEC, VIJAYAWADA

***Abstract*- Opinion mining and Sentiment analysis gives sentiments, opinions and subjectivity of text. Now a days many people express their opinions and ideas through social networking sites like Face book, Google+ and Twitter. These are platforms to allow people to share and express their views about topics, have discussion with different communities. Twitter data is short and continues data suitable for sentiment analysis. This paper focus on sentiment classification (positive, negative and neutral) which is multistep process involves preprocessing phase, parts of tagging (POT), and calculating polarity and apply three classification algorithms that are Decision tree, Naïve bayes and Support vector machine on twitter data in Jupiter. This paper also presents empirical comparison of classification algorithms in which decision tree algorithm is highest accuracy in comparison of all three algorithms considered in this study.**

***Keywords*: Naïve-Bayes, Decision Tree, Support Vector Machine, POT, Polarity.**

I. INTRODUCTION

Micro blogging websites have become a source of various kinds of information. In micro blogs, people post real time messages about their opinions on various topics such as discuss current issues, complain and express positive sentiment for products they use in daily life. Twitter is such a platform where people expressing and sharing their thoughts and opinions on all kinds of information or opinions about products, politicians, events etc. Sentiment detection of tweet is one of the basic research topics over twitter data and it is very useful because it allows feedback to aggregate without manual intervention. From the sentiment analysis many people like consumer, businessman or may be typical person get benefit and many companies study user reactions and reply to users on micro blogs. One challenge is to build technology to detect and summarize an overall sentiment. In this paper, we look at Twitter and build models for classifying "tweets" into positive and negative sentiment. Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm. We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. It includes rich structured information about the individuals involved in the communications. For example it maintains information of who follows whom and re-tweets and tags inside of tweet provide discourse information. Another reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. The response of twitter is more prompt and more general. Sentiment Analysis is process of collecting and analysing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data. Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it analyse whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment. Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or market. The layout of the paper is as follows: Section II discusses about the Design Framework.. Section III presents Experiments and Results. Section IV concludes with the Conclusion.

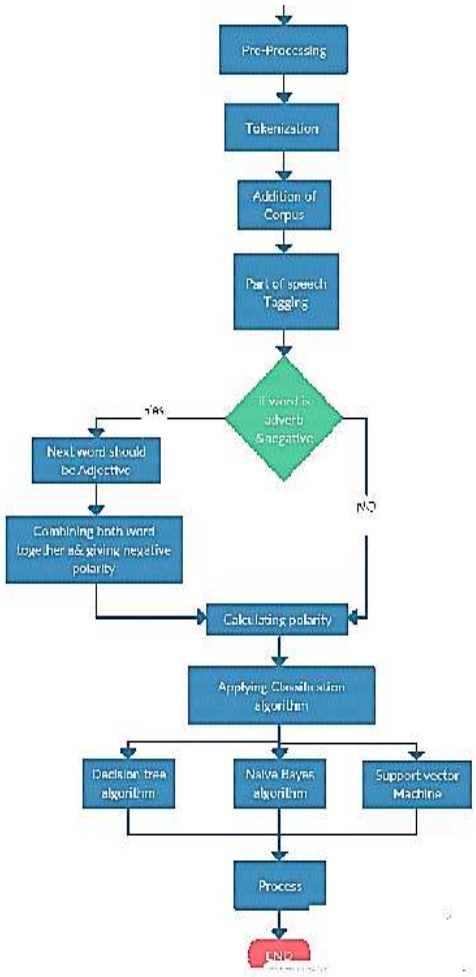
II. DESIGN FRAMEWORK

In this section, we represent an overview of the architecture which resembles process in fig1. The following steps will expand the process of the proposed system,

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 4, April 2018

1 Collect Tweets
2 Preprocessing of Tweets
3 Feature Extraction
4 Data Modeling



The Fig:1 shows the diagram in schematic form the general arrangement of the parts or the components of a complex system or process, such as an industrial apparatus or an electronics circuit. The block diagrams are heavily used in engineering in various hardware designs.

1. Collect Tweets

Twitter data is short and continues data suitable for sentiment analysis. To improve visibility of the content user generally used Emoticons for pictorially facial expressions, "@" symbol used to refer other users and Hashtags. Hashtags are used to mark topics.

The Initial Data or the Raw Data is collected from Twitter for analysis purpose. Overall 1400 tweets are collected.

2. Preprocessing of Tweets

The reviews are pre-processed primarily to make them noise free at the classification part as the review sentences may describe about two or more features which will be difficult to classify if the sentence has positivity and negativity associated[1]. So the features if two or more are present in a sentence are splitted in order to eliminate the above case And we also took into consideration that the reviews scraped from the websites [2] will contain some spelling mistakes which will be serious issue. Example: 'Gud' may mean nothing to the computer which if corrected does make some sense like positive or negative. Thus the pre-processing includes these steps to make review sentences noise-free along with Stemming as it is the standard procedure to make the words cut short to make match with the features.

3. Feature Extraction

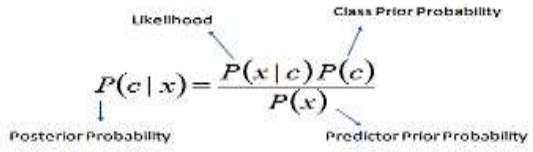
Large number of features without considering un important features [3] are extracted. Thus the sentences are first tokenized to check for the words which are adverbs/adjectives and nouns. Then we apply a POS tagging to select Adverbs/Adjectives and Nouns. Here adverbs/adjectives and nouns are only considered because the word which describes the feature in a review sentence is an adverb or a noun. Example: The body of the mobile is fragile. This sentence when tokenized and POS is applied will look like: [(The,DT),(body,VBP),(of,IN),(the,DT),(mobile, NN), (is,VBZ),(fragile,JJ)] Here we take the sentence to see if a feature is present in it or not to check if the sentence is worth further processing. If it does not have any feature, we continue with other review sentences. Senti-wordnet is a large database consisting of many words associated with its pos and neg score [4]. Example: 'Good' has a positivity score of 0.75 and negativity score of over 0.25

4. Data Modeling

For sentiment classification on twitter data, here we are using decision tree[6], naïve bayes[5] and support vector machine. In the proposed feature selection, a Decision tree induction chooses relevant features using following formula:

$$\text{info}(D) = -\sum_{i=1}^{m} p_i \log_2(p)$$

For naïve bayes following is used:



Above, • P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 4, April 2018

- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor

SVM are supervised machine learning methods used for classification, regression and detection models. SVM are more effective for high dimensional space. SVCs are capable for multi-class classification. SVC and NuSVC are similar whereas, LinearSVC are based on linear kernels.

III. IPLEMENTATION AND TESTING

This is implemented in Jupyter on Windows Download the Jupyter software from the official website. http://jupyter.org/ and click on Download the Jupyter. Fig 2 shows tweets data set. Fig 3shows important features in given data set.

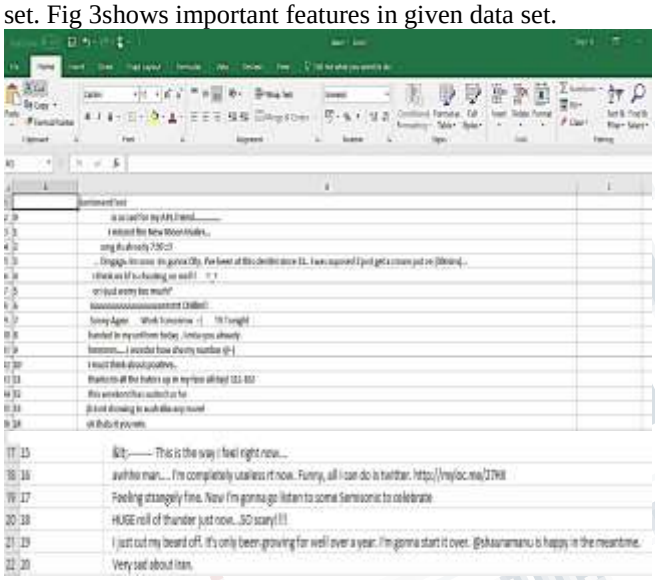


Fig2: Tweets Data set

```
In [15]: words
Out[15]: ['sad',
          'friend',
          'trailer',
          'already',
          'dentist',
          'since',
          'get',
          'crown',
          'put',
          'think',
          'cheating',
          'worry',
          'much',
          'handed',
          'uniform',
          'today',
          'miss',
          'already',
          'wonder',
          'number'
```

Fig3: Extracted features for classification of data set

After extracting important features from tweets, assign polarity to each feature in three levels. These three level sentiment polarities are positive, negative, neutral.
Fig4 shows performance of Naïve bayes, SVM, decision tree algorithms

	Accuracy	Precision	Recall	F-score
Naïve Bayes	0.56128	0.4129	0.25169	0.22907
Support vector machine	0.8006	0.8459	0.62999	0.68586
Decision Tree	0.8961	0.8713	0.8493	0.8596

Fig4: Comparison table

IV CONCLUSION

We presented results for sentiment analysis on Twitter. Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. By using machine learning algorithms i.e naïve bayes, decision tree, SVM, we conclude that it is easier to classify the tweets and more we improve the training data set and more we can get accurate results. We can also implement features like emoticons, neutralization, negation handling and capitalization or internationalization as they have recently become a huge part of the internet. This paper also presents empirical comparison of classification algorithms in which decision tree algorithm is highest accuracy in comparison of all three algorithms considered in this study.

REFERENCES

[1] Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. Elsevier Procedia Computer Science. 2013; 17:26–32.Available from: www.sciencedirect.com

[2] Ceci M, Loglisci C, Macchia L. Ranking sentences for key phrase extraction: A relational data mining approach. Elsevier Procedia Computer Science. 2014; 38:52–9. Available from: www.sciencedirect.com C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[3] Thamizharasi P, Sathiyavathi R. An approach to product rating based on aspect ranking in opinion mining. Indian Journal of Science and Technology. 2016; 9(14):1-6. doi:10.17485/ijst/2016/v9i14/89604

[4] Singh A, Ullah ME. Aspect based sentiment analysis. Available from: http://home.iitk.ac.in/~enayat/files/absa_report.pdf

[5] Kiruthika M, Sanjana Woonna, Priyanka Giri-'Sentiment Analysis of Twitter data'- International Journal of Innovations in Engineering and Technology– Vol.No.6,Issue No.4, April 2016.

[6] Durgesh Patel, Sakshi Saxena, Toran Verma– 'Sentiment Analysis using Maximum Entropy Algorithm in Big Data' – International Journal of Innovative Research in Science, Engineering and Technology –Vol No.5,Issue No.5,May 2016.

[7] Bholane Savita D., Prof.Deipali Gore – 'Sentiment Analysis on Twitter Data Using Support Vector Machine'– International Journal of Computer Science Trends and Technology (IJCST) Volume 4, Issue 3, June 2016.

[8] V.Lakshmi, K.Harika, H.Bavishya, Ch.SriHarsha – 'Sentiment Analysis on Twitter Data'– International Research Journal of Engineering and Technology (IRJET) Volume 4, Issue 2, February 2017.