ISSN No: - 2456- 2165

Sentiment Analysis on Twitter in R

[1]S.Sathya, Amith Mathew [2], T. S Karthikraja [3], John Rosy [4]
[1] Assistant Professor, [2] [3] [4] UG Scholars
Department of CSE, RVS College of Engineering and Technology, Coimbatore

***Abstract*—Mining posts in social networks have great potential for new applications. However, the huge amount of data produced every day in these networks makes it impractical for people to manually undertake the task. The Sentiment Analysis returns polarity of tweets written in English about a topic of interest. Two main requirements direct the design and the development of the system: (a) good usability and (b) good precision in determining the sentiments. The topic of interest has a clean interface to enter the keywords that describe the topic of interest and to present the results at several levels of detail. To meet the second requirement, the system uses Naïve Bayes classifier to identify the sentiments of tweets, the literature shows that this algorithm combines good classificatory performance and low response time. The systems also present a good result in measuring its precision in identifying the predominant polarity of Tweets and are related to five different topics of interest.**

Keywords—Sentiment analysis, opinion mining, Twitter, Naive Bayes classifier.

I. INTRODUCTION

Social networks are increasingly present in the daily life of people of all ages and social layers, being used in diverse activities professional, entertaining and socializing. The opinions expressed in these networks are an important source to help understand the collective feeling about various topics. They can provide feedback to companies regarding their brands and products and public figures, such as politicians, artists and athletes, about their reputation. The amount of information circulating through networks is very high. Twitter, for example, in 2014, recorded an average of 500,000 tweets per day. By 2015, the number of its active users rose from 300 million to 400 million.

Twitter users a day who comment, enjoy and share ideas, opinions and criticism. The sheer volume of data produced on social networks makes it impractical for people to analyze it. In this context, the data mining, this aims to extract information to gather people opinion among the various data mining tasks that can perform from text messages in natural language. Analysis of sentiment, called the mining of opinions, aims at identifying polarity (positive/negative) of messages. The terms analysis of feeling and mining of opinions are currently used area in polarity (positive / negative) of messages. The terms and a large range of applications can be developed from the identification of the subjectivity embedded in the messages. The analysis of feeling is a problem of processing Natural language (PNL), in which it only determines the positive or negative feelings related to sentences, entities or topics. Determining the feeling associated with a fragment of essentially consists of classifying it into one of the categories: positive, negative or neutral.

Eventually, the number of categories may be higher when you want to express the intensity of feeling: very positive, positive, not positive, neutral etc. The success of these methods depends heavily on the selection or extraction of an adequate set of characteristics of the text, including terms (n-grams), grammatical categories, and Semantic dependencies.

Mostly focusing on the qualifying performance of the algorithms used for the detection of feelings expressed in texts (blogs, news sites, and social networks) written in English. This analysis processes text messages written and published on the social network Twitter, known as Tweets, in order to identify the feeling. A Naive Bayes (*NB*) classifier for evaluation of the feelings expressed in tweets is used. This algorithm is adopted because it is simple to implement, requires relatively few computational resources and have similar performance to more complex alternatives and computationally expensive. The Naive Bayes classifier is to filter out comments that are interested. Result of the analyses is of feelings is displayed in bars, so that the user can have a clear visualization of the results obtained.

II. THE NAIVE BAYES CLASSIFIER

The Naive Bayes (*NB*) classifier is based on the rule of Bayes for the inversion of conditional probabilities, presented in the context of classification of texts.
(I.e.)

$$p\left(\frac{c}{d}\right) = \left(p\left(\frac{d}{c}\right)p(c)\right) \div (p(d))$$

$P(c \mid d)$ is the posterior probability of Category C given the document (tweet, in this case Context) d, $P(c)$ the prior probability of category c, $P(d)$ the prior probability of the document d And $P(d \mid c)$ a Posterior probability of the document d belonging to the class w. Considered, c is

positive and negative for the analysis of Feelings. The prior probability of a category $P(c)$, is determined from

the number of categories considered. In using the Bayes theorem in the *NB* classifier, the denominator $P(d)$ is seen as a constant, being eliminating, which reduces to the numerator. $P(xi \mid c)$ is the conditional probability of a word (term) *xi* occurs in category *c* and *nd* is the number of terms in document d. The calculation of $P(xi \mid c)$ consists of dividing the number of times that the word *xi* is found in the training list of category *c* by the total of words in this list. The lists associated with categories are built in the training phase of the classifier, from examples of positive tweets and negative results. In this way, the best class to assign to a document by the Naive Bayes classifier *NB*, where *C* is the set of classes considered.

The Naïve Bayes classifier (*NB*) follows that two propositions are assumed:

1) The Positions of the words do not matter.
2) The probabilities of the terms are Independent given a class *c*. *NB* performs well in text classification tasks, comparing to that of more complex algorithms are computationally time-consuming.

III. FUNCTIONALITY AND ARCHITECTURE

A. System Flow Diagram

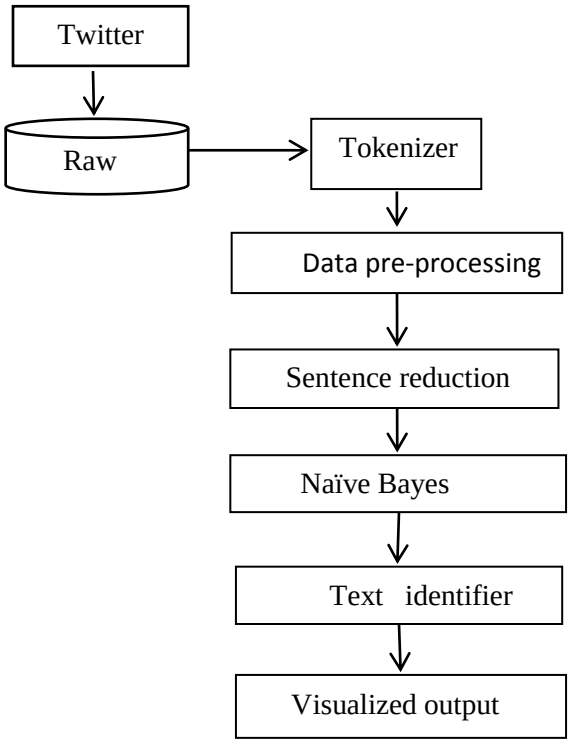

Fig 1. System Architecture

B. Fetching data

In this project, the fetching of raw data from the Twitter and to do its analysis using R language is done.

C. Tokenizing

In a sentimental analysis, tokenization is the process of breaking a stream of text up into words.

D. Data Pre-Processing

Cleaning and removing the unwanted words.

E. Sentence Reduction

After completing the data preprocessing the data are transformed to the understandable format.

F. Visualizing the data

It produces the output of the classifying the data such as positive, negative, neutral and negation are estimated from the twitter.

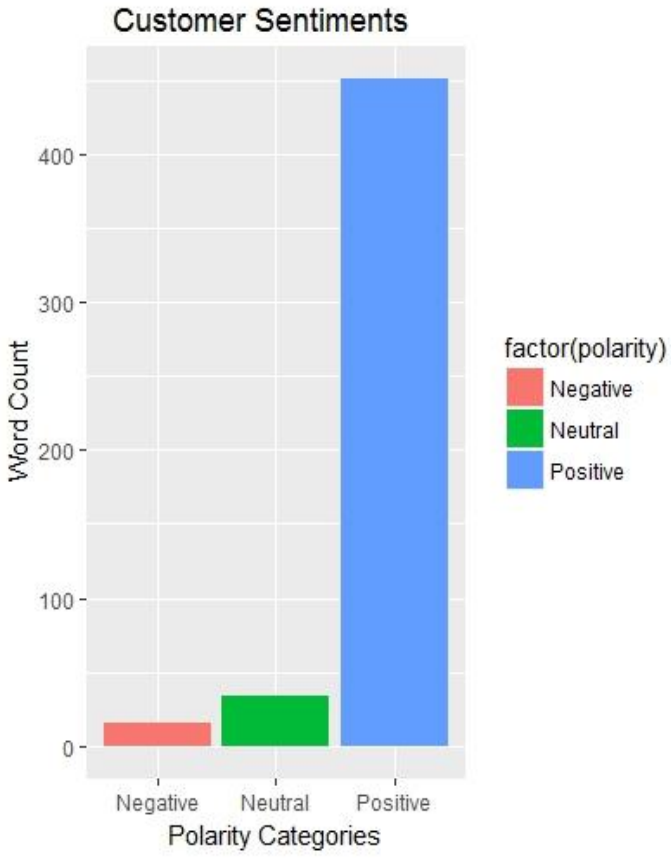

Fig 2. Result for the Sentiment Analysis.

ISSN No: - 2456- 2165

IV. FINAL CONSIDERATIONS

Identify the polarity of comments made about Generic topics on the social network Twitter offers a wide Range of possibilities, but it is a challenging task, even for humans. In addition to the difficulties inherent automatic interpretation of texts in natural language, it is added that tweets are often written using Slang and often do not respect standard language standards. The user who seeks to simplify the process of identifying the topic of interest, collecting tweets, and analyzing polarity, contained in these tweets. Determine the polarity of using a Naive Bayes classifier.

The results obtained in the experimental evaluation described in the article Good performance, particularly considering that the system allows you to analyze tweets on any topic. To train the algorithm with this perspective it is necessary to have a large number of tweets on various subjects, positive and negative, in order to train the Classifier. The strategy adopted was to collect a few thousand of tweets, classifies them into positive or negative according to the polarity associated with the emoticons each contained.

A system focused on a particular vision (a brand, product, person or company) would allow training of the classifier with more specific messages which would tend to increase the accuracy as well as to compare the performance of the classifier when the manual classification of training tweets.

V. CONCLUSION

Nowadays, the use of social networking sites is increasing and the users wish to have a short and quick view of the discussions going on currently. By this, the users can have the analysis quickly, less cost and more accurate. The output is in a graphical form. The application can have a quick view on things like politics during an election, health care, movie review, improved customer service and product review.

REFERENCES

[1] Grandin and Adan “Piegas: A System for Sentiment Analysis of Tweets in Portuguese”, 2016.

[2] Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani” Contextual Semantics for Sentiment Analysis of Twitter”,2015.

[3] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, And V. Stoyanov, " Sentiment analysis in Twitter, "in Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, 2015.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., vol. 17, no. 3, p. 37, 1996

[5] S. Das and M. Chen, “Yahoo! for Amazon: Extracting market sentiment from stock message boards,” in Proceedings of the Asia Pacific finance association annual conference (APFA), 2001, vol. 35, p. 43.

[6] R. Feldman, “Techniques and applications for sentiment analysis,” Commun. ACM, vol. 56, no. 4, pp. 82–89, 2013

[7] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov, “Semeval-2015 task 10: Sentiment analysis in twitter,” in Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, 2015.

[8] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data,” in Proceedings of the Workshop on Languages in Social Media, 2011, pp. 30–38.