Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 4, April 2020, pg.40 – 43

Political Tendency Identification in Twitter Using Naive Bayes Classification

Sushma R V[1]; Nishkala L K[1]; Rakshitha H P[1]; Rakshitha K S[1]; Mrs. Shruthi T R[2]
[1]Final year B.E. pursuing in Computer Science & Engineering, Malnad College of Engineering, Hassan, India
[2]B.E, MTech, Asst Professor, Malnad College of Engineering, Hassan, India
[1] sushmavbharadwaj99@gmail.com, [1] pinbk99@gmail.com, [1] raksharakshi0868@gmail.com, [1] ksvanishiva8@gmail.com, [2] shruthi7129@gmail.com

Abstract- The generation of social media in the recent past has provided end users a powerful dias to voice their opinions. Businesses need to analyze the polarity of these opinions in order to understand user orientation and thereby make smarter judgement. One such application is in the field of politics, where political parties need to understand public opinion and thus determine their campaigning strategy. Sentiment analysis on social media data has been seen by many as an effective tool to monitor user choice and inclination. Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to achieveSentiment analysis. The efficiency of these algorithms is contingent upon the quantity as well as the quality of the labeled training data.
Keywords— sentiment analysis, Naive Bayes, SVM, training data, labeling, Twitter

I. INTRODUCTION

Social media are usually used to show judgement and perception about companies, products, services, hobbies, politics, etc. Therefore, enterprises, organizations, governments, and different groups in general have shown concern in the opinions of users for their activities. They are also interested to known the way users use these media, the communication style and some users trais such as gender, age, geographical location, political orientation, etc. In general, the main aim is to provide personalized services, particularized offers, or simply to know what people think about something in order to improve their activities.

Data is gathered from Twitter API. Search is given by using #Hashtag followed by the name like #FAN, #BajarangiBhaijaan, #assemblyelection etc. Roughly 17000 tweets have been collected from the various movie tweets. Reviews can also be explored by #Hash tags followed by respective movie stars, directors, and production house and music record companies. In twitter hash tags turn into the necessary symbol to discover about something and it provides user limit of 140 words to show their views and attitude.

Suppose the information contain word which has been appeared more than two times continuously then it has to be changed. For example great great great great party can be convert to great party. Break words like “a”,” is”,” the”, “etc” etc; These words has nothing to do with the emotion, so has to be removed from the message. Now next step is to train the data using supervised classifier.

Naive Bayes classifier is a prominent method for text classification problems were given a document or article the classifier has to decide the category of the document. Naive Bayes classifier is based on probabilistic technique of classification which derives its roots from Bayes Theorem. It is based on the belief of independence between the various components. It is an extensible classifier and can run accurately with large data sets. Naive Bayes classifier is fast as compared to other classifiers and thus is used as a baseline for text classification problems.

SVM (Support Vector Machine) works on the basis of Supervised Learning. SVM needs a training set and labels combined with it. After training, if a test data is fed in, the model allocate it to one category or the other. It performs well with linear classification. It can even work efficiently on a nonlinear classification using a kernel trick by mapping the inputs into high dimensional feature space. It builds a hyper-plane for classification. The hyper- plane is chosen such that the distance between the nearest data point on either side is maximized.

II. LITERATURE SURVEY

Sentiment Analysis is the comprehensive research of how opinions and perspectives can be relate to ones emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached upto great achievement which can surpass the positive vs negative and deal with whole arena of behavior and emotions for different communities and topics.

Pang and lee [1] designed the system where opinion can be positive or negative was found out by the ratio of positive word to total words. Later in 2008 the authors developed the procedure where tweet outcome can be divided by term in the tweet. Compare to baselines that are generated by humans, the results are pretty good when machine learning techniques are used. Regardless of using different types of features the authors did not obtain desired accuracies over topic based categorization.

Jiang et al. [2] focuses on target dependent sentiment classification. Here target dependent means whether the sentiment is positive, negative or neutral depends on nature of the question that is asked. The authors proposed to make better target dependent sentiment classification by joining features of target dependent and considering related tweets. The authors also discussed that there is a need of consideration of current tweets to the related tweets by employing graph based optimization. As believed by authors experimental results, the graph based optimization increases the performance.

Chen et al.[3] designed the feed forward BPN network and uses sentiment orientation to calculate the results at each neuron. The authors developed the procedure based on neural network which is a combination of machine learning classifiers and semantic orientation indexes. In order to obtain accuracy in methodology, semantic orientation indexes used as inputs for neural network. The proposed methodology outperforms other neural networks and traditional approaches by increasing efficiency in both training as well as classification time.

Malhar and Ram [4] selected supervised machine learning techniques and artificial neural networks to segregate twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers. The authors proposed a methodology to predict the outcome of election results by utilizing the user influence factor. To carry out contraction in dimension the authors combined the Principle Component Analysis with SVM.

Anton and Andrey [5] analysed the existing techniques and developed a model for automatic sentiment analysis of twitter messages using unigram, bigram and jointly i.e. hybrid feature. The purpose of the authors is to analyse and produce approaches for analyzing the accent of the messages in social media. The authors reviewed existing automatic sentiment analysis approaches and in order to maintain the context of growing methods the character feature of social media statements were studied.

III. METHODOLOGY

The tweets that were posted by users in the form of hashtags to show their opinions about present political trends were considered. We then keep the retrieved tweets in the database. After pre-processing, the remaining data was divided into the training set tweets and the test set tweets. Polarity and subjectivity were calculated using three different libraries, SentiWordNet,W-WSD and TextBlob. We applied the Naive Bayes on training set and built a classification model. The model was tested onthe training dataset to obtain the accuracy result of each classifier. The framework of sentiment analysis is as shown in Fig 3.

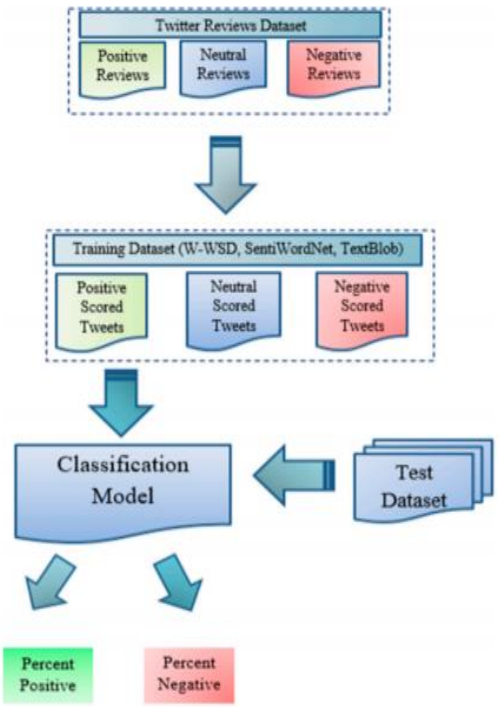

Fig. 3 Sentiment Analysis Framework

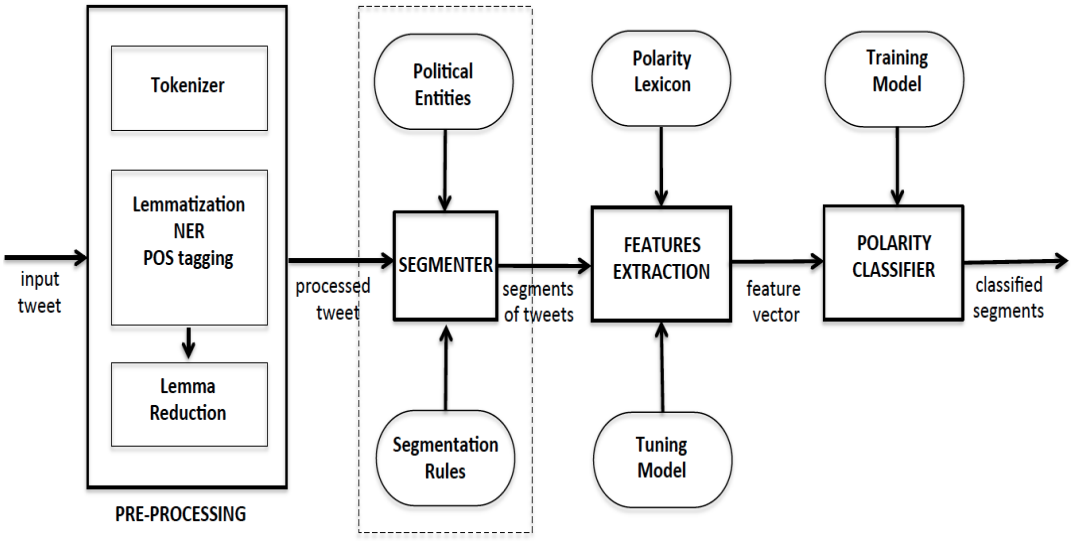

Fig 3.1. System Architecture

A. Data Collection

To collect public opinion based on collected hashtags related to views about political parties including Twitter top trends, Tweepy is used. Tweepy has the cursor object, managing and passing data that spans multiple pages. An account on Tweepy API is linked to Twitter account. Data collection includes Tweets, Followers, Following, Todays status (the followers and following count that day). To retrieve the tweets, Tweepy API accepts parameters and provides the Twitter account's data in return. Body of the tweet will be extracted and stored in a CSV file.

B. Data Pre-Processing

Data pre-processing is an important tool. An unstructured Tweeter data set it is a collection of information where people entered his/her feelings, opinions etc. To improve the accuracy of the Tweeter data, the tweets were divesting special characters like '@' and URLs. Then the pre-processed data is subjected to classification. It includes data cleaning, normalization, feature extraction.

Data cleaning: It removes a bad errors data and reduces unwanted information in data. It also removes the stop words like I, me my, myself etc.

Normalization: To get the accurate results from classifier tweets must be processed properly. Since tweets are in user language, we have to clean the data which are irrelevant to the data. Examples of irrelevant data are URLs, repeated words. Normalization is also called as "scaling down" transformation of features.

Feature selection: It is the process of identifying and removing irrelevant and redundant features thereby it reduces the dimensionality of the data. Features are selected from the pre-processed tweets to obtain the feature vector.

C. Classification

The data mining techniques are used for classification. Naive Bayes algorithm is a powerful algorithm for classification. It is used for classifying the data on basis of probabilities. It is also known as Maximum a Posterior Naive Bayes. Bayes theorem relates the conditional and marginal probabilities of two random events. It calculates the probability each individual attribute and finally find out the yes or no probability.

IV. COMPARISON

In existing system, ILDA Technique uses Probabilistic graphical model at Document level. Each review is considered as a mixture of latent aspects and ratings. Accuracy in rating is not specified.

In proposed method, Machine learning technique is used at document level. Approximately 75% of accuracy in rating is achieved. Text document get classified into sentiments of different categories like positive, negative and neutral.

V. CONCLUSION

The political tendency, that takes into account the polarity of entities related to the political parties that appear in tweets of user. The sentiment analysis system developed in order to obtain the polarity of these entities was also presented. However it is possible that in other political context different to Spanish, the left, center and right tendencies also need to be adapted on this point we are working on a system in which political tendency, as defined in this paper, will be a future within a wider classification system. We have described our approach for political tendency identification of Twitter users. We have defined a metric, called Political Tendency that takes into account the polarity of entities related to political parties that appear in the tweets of the user. The Sentiment Analysis system developed in order to obtain the polarity of these entities was also presented.

ACKNOWLEDGEMENT

With great pleasure, we would like to express our sincere thanks to our guide, Mrs. Shruthi T R, Assistant Professor, Dept. of Computer Science and Engineering, Malnad College of Engineering, Hassan, whose constant guidance and encouragement at every stage of this work was very much valuable. We would like to express our sincere thanks to our HOD, Dr. Geetha Kiran A, Dept. of Computer Science and Engineering, Malnad College of Engineering, Hassan, for providing sufficient facility to carry out this project

REFERENCES

[1] Vignesh Rao, JayantSachdev,*"A Machine Learning Approach to classify News Articles based on Location"*, International Conference on Intelligent Sustainable Systems,2017.
[2] B. Pendharkar, P. Ambekar, P. Godbole, S. Joshi, and S. Abhyankar, *"Topic categorization of rss news feeds*", Group vol. 4, p. 1, 2007.
[3] Leo Breiman, Random forests, *Machine Learning*. vol.45, no.1, pp.532, 2001.
[4] J.K.M. Han,*" Data Mining: Concepts and Techniques"*, 2nd ed. 2006.
[5] H. a. K. S. Yu, *"SVM tutorial: Classification, regression, and ranking*", Handbook of Natural Computing 2009.

