Real Time Sentiment Analysis of Tweets Using Naive Bayes

Ankur Goel
JSS Academy of Technical Education,
Noida, INDIA
goel_ankur@hotmail.com

Jyoti Gautam
JSS Academy of Technical Education,
Noida, INDIA
jyotig@jssaten.ac.in

Sitesh Kumar
JSS Academy of Technical Education,
Noida, INDIA
sitesh226@gmail.com

Abstract— **Twitter[1] is a micro-blogging website which provides platform for people to share and express their views about topics, happenings, products and other services. Tweets can be classified into different classes based on their relevance with the topic searched. Various Machine Learning algorithms are currently employed in classification of tweets into positive and negative classes based on their sentiments, such as Baseline, Naive Bayes Classifier, Support Vector Machine etc. This paper contains implementation of Naive Bayes using sentiment140 training data using twitter database and propose a method to improve classification . Use of SentiWordNet along with Naive Bayes can improve accuracy of classification of tweets, by providing positivity, negativity and objectivity score of words present in tweets. For actual implementation of this system python with NLTK and python-twitter APIs are used.**

Key Words: Twitter, SentiWordNet, Machine Learning, NLTK, Python, Sentiment Analysis.

I. INTRODUCTION

Now a day's social media became so popular that it has a significance impact on trade and marketing. People share their precious views, opinion, experiences on social sites so that others can get benefit from these. Twitter is one of such platform where common people shares their reviews in short blogs i.e. in 140 characters. These reviews can be for anything like any product or service such as movie, stock market, schools, colleges, politics and much more. Here, people shares the unbiased opinions about anything they wanted, that's why one can consider these reviews as more generalized and real one. There are five basic steps involved to implement this whole system- first step is selection of training data which is selected on the basis of type of problem, second step is preprocessing of training data which means removing irrelevant information like URLs, user names, slang words, symbols etc., third step is to establish connection with twitter database using twitter API from where recent tweets can be extracted for analysis purpose, in fourth step various machine learning algorithms like Naive Bayes, Support Vector Machine are used for the classification of tweets into different classes and in the final step results are displayed on the basis of polarities of tweets after their classification. SentiWordNet[2] is lexical resource which is broadly used for opinion mining and it provides sentiment score to each synset of WordNet. These score are of three types; positivity score, negativity score and objectivity score. So instead of using Naive Bayes in traditional way, these scores can be taken into account so that more accuracy can be attained.

In third section, all theoretical work done is explained like how to gather resources, their integration and their simulation, in fourth section traditional method of classification is compared with proposed method of classification using the help of SentiWordNet, after that future scope, appendix and references are provided.

II. LITERATURE SURVEY

Numerous research work has been already done in field of sentiment analysis. But the informal tone of tweets has always been a challenge for the analysis. Because of the unbiased nature of tweets , twitter database has been in lime light for sentiment analysis of movies , products , popularity or anything of that sort. Sentiment analysis has given way to a wide range of researches ranging from document level classification (Tirney , 2002 ; Pang and Lee , 2004) to sentence level Hu and Liu ,2004 ; Kim and Hovy ,2004) leading to phrases (Esuli and Sebastini ,2006) . Classifying tweets into positive and negative classesusing distant supervision was presented by Alec Go,Richa Bhayani and Lei Huang , 2009 .They presented an approach for automatically classifying the sentiment of twitter messages with respect to a query term . They presented results of machine learning algorithms (Naive Bayes , maximaum Entropy , and SVM) for classification . Use of linguistic features for detecting sentiment of twitter messages was investigated by Efthymios Kouloumis , Theresa Wilson and Johanna Moore (2011). they used hashtagged dataset(HASH) for development and training .The main aim of this paper was

[1] https://about.twitter.com/
[2] http://sentiwordnet.isti.cnr.it/

to take a supervised approach using the existing hashtags in twitter data for building training data . Hassan Saif , Yulan He and Harith Alani , 2012 explained an approach of adding semantics as additional features into training set for sentiment analysis. For each extracted entity its semantic concepts were added as an additional feature and the correlation of the representative concepts with positive / negative sentiments were calculated . Use of sentiwordnet as a lexical resource for sentiment analysis (Stefano Baccianella, Andrea Esuli , and Fabrizio Sebastiani ,) added to the efficiency of previous traditional methods of classification . This paper explained how use of SENTIWORDNET 3.0 improves efficiency of classifications with respect to previous versions of WORDNET and SENTIWORDNET.

III. SYSTEM MODEL

Basic architecture of the sentiment model is explained using the block diagram in fig.1, which shows various phases for the sentiment analysis of real time tweets. Various steps involved in this system are as explained as follows.

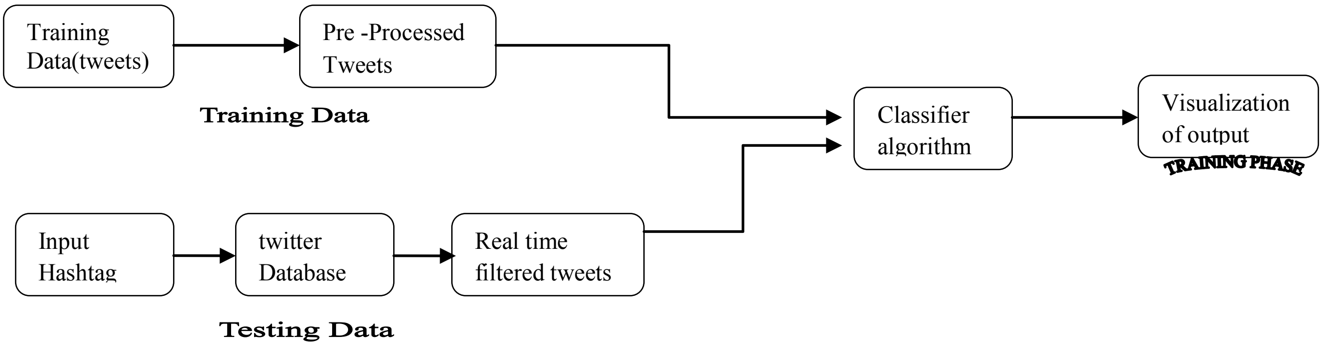

FIGURE 1: FLOW DIAGRAM FOR SYSTEM

TRAINING DATA

Training data is most important part of the whole system as training of the system is wholly depends on it and classification of testing data is done on the basis of this result only. While choosing training data, type of problem should be taken into consideration as similar type of training data should be taken so that it can provide more efficient and accurate results like if problem is related to movie review then training data can be taken from IMDb or if problem is related to food review then it is better to use review of zomato or if problem is related to product review then data can be taken from amzon.com because this type of training data is more associated to the type of problem statement. There are some frequently used training sets are given below with their description.

Training corpus of Sentiment140[3] is used as training data for the classification of tweets as positive and negative classes. It contains labeled tweets as positive=4 , neutral=2 and negative=0, id, text date etc. It has around 16 millions tweets in the training corpus which makes it more relevant to use as training Data.

Movie review dataset provided by Large Movie Review Dataset(LMRD)[4] for binary sentiment classification. 25K highly popular movie reviews are given for training purpose and 25K for the testing purpose.

PRE PROCESSING

Huge amount of data is already present related to different problems, But using this given data as it is, may not produce desired output as data contains many irrelevant things which makes it tough to handle. So it is always suggested to remove irrelevant portion of data or make it relevant which can further increase efficiency of the system

Taking example of twitter data, tweets contains so many data types such as User ID (staring with @), URLs, text, date and time, location, multimedia files (images, videos etc), emoticons, hashtags (staring with #). Each of these have their own significance during the sentiment analysis and some are irrelevant which do not have any significant effect so it is suggested to omit these data while sentiment analysis.

User name in tweets always starts with @ symbol, this is to tell who tweeted this particular tweet , while doing sentiment analysis there is no significant effect of user name so by applying filters, user name is excluded from training as well from testing data.

[3] http://help.sentiment140.com/for-students

[4] http://ai.stanford.edu/~amaas/data/sentiment/

There is character limitation in tweets so users includes some URL links to explain it better, these URLs (generally start with http://) are not needed during training and testing of data as they do not contain any useful information in it which can be used during sentiment analysis.

Text is actual body of tweet which can be maximum 140 character longer and contains everything that any user want to tweet. This text is mainly used for sentiment analysis but it should be free of irrelevant data.

Date and time stamp is attached with every tweet which tells that about when a particular tweet is tweeted. This feature is very useful to find how frequently any user tweets and to find most buzzing word of twitter for any particular time interval.

Location can also be traced in tweets but this is optional feature of user whether he/she wants to share location or not, this feature helps in to find the trends in any particular demographical area. Here in above mentioned system this feature is not used and omitted.

Emoticons are very useful symbols present in the text of tweets, they emphasize the sentiments of tweets and also help to find out the true sentiment of tweets.

Table 1 : Types of Emoticons

Type	***Examples***
Positive	:-j =p :] :-P ;) :p :3 =] :b :-) 8) _/ :') ;-) :-p :S
Negative	:-/ : :'(:[= :/ :@ :'-(:c ;(=/ v.v

Words starting with # symbols are termed as 'Hashtags', they are advanced kind of tags which are associated with the tweets and describe tweets in more detailed manner so that one can easily understand in which context this tweet is tweeted and these hastags forms a cluster of tweets which are greatly associated with the hashtag and it is easy to understand whole event at once. Generally tweets are the subpart of hashtags.

TWITTER APP MANAGEMENT

To access tweet data, Twitter provides twitter api to developers, which lets you register your application or project for accessing data from twitter database. For registering any new application a form including details of the application , purpose, website for application, URL etc needs to be provided. After registration an access token is generated which provides means to connect to twitter database. Generation of access token provides complete OAuth settings which include details like consumer key , consumer secret , request token URL ,authorize URL ,access token URL etc.

CLASSIFICATION

Classification basically means categorizing data into different classes based on some computation which determines the sentiment behind the data. Number of classes depends on the type of problem. For example, for movie data the classes could be good , bad and average or simply positive or negative . Many classifiers can be used for classification process like Naive Bayes classifier , Support vector machine , Baseline etc. Here Naive Bayes classifier is being used for the classification process. Naive Bayes is mostly preferred for classification due to its speed and simplicity.

Naive Bayes classifier assumes that the presence of a particular feature in a class is not related to the presence of any other feature . For example , a fruit may be considered to be an apple if it is yellow , round and about 3 inches in diameter. Even if these features depend on each other , all of these properties contribute to the property that this fruit is orange and thus the name "Naive". Mathematically , for a word w and class c , by Bayes Theorem

$$\mathbf{P(c/w)=[P(w/c)P(c)]/P(w)} \qquad \textbf{.....(1)}$$

where P(c/w) is probability of class c given word is w.
P(c) is probability of class c and P(w) is probability of word w .

Naive Bayes classifier will be

$$\mathbf{c^*=arg\ maxc\ P(c/w)} \qquad \textbf{.....(2)}$$

IV. RESULTS

The first phase in sentiment analysis has been performed using training dataset consisting of 1.6 million tweets of sentiment 140 dataset . After training system, when tested with a small testing dataset of 100 tweets the system gives an efficiency of 58.40%. Figure 2 shows the dynamically changing movie ratings at different times for different movie name like batman , deadpool , superman , civilwar etc. Rating is calculated using the positive percentage of tweets which are extracted from twitter database .

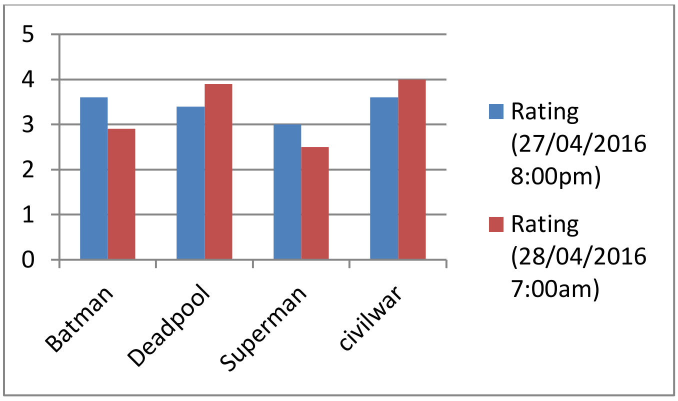

Figure 2 : Movie Ratings

V. PROPOSED SYSTEM MODEL

For sentiment analysis of tweets training , preprocessing and classification are the core steps . We are keeping the first two steps i.e., training and preprocessing as mentioned in previous section as it is. But to improve the efficiency we are modifying the classification part . Instead of using just Naive Bayes , we will be using combined results of Naive Bayes with SentiWordNet . For this we are proposing a new classification technique which is a bit different from the previous work but helps in increasing the efficiency of the overall system.

SentiWordNet

WordNet is lexical database for English language. It contains definitions of words according to their part of speech. Additionally, it provides synset(cluster) , synonyms of words and phrases. There are many research labs which provides their own version of WordNet. SentiWordNet is one of such type of WordNet which provides positivity, negativity and objectivity scores of the word in addition to the definition of the word.

These scores are quite useful while classifying tweets as these scores are set after deep observation and analysis of semantics of that word in the real world.

Each word can be classified according to its part of speech in English language. That's why, its positivity, negativity and objectivity score can differ for the different parts of speech.

For Example: lexicon '(dog.n.01)' has PosScore=0.0, NegScore=0.0 and ObjScore=1.0, lexicon '(sad.a.01) has PosScore=0.125, NegScore=0.75 and ObjScore=0.125, lexicon '(beat.v.01) has PosScore=0.125, NegScore=0.0 and ObjScore=0.875.

Now, this score can be integrated to posterior probabilities while calculating Naive Bayes classifier so that these words have greater role while determining the overall sentiment of the tweet.

Proposed Work Flow

1.First train the system using a good and relevant corpus of tweets.

2.Now, take the testing corpus of tweets and take one tweet at a time.

3. Tokenize the tweet and preprocess it.

4. Taking all tokenize word find the posterior probability according to the training dataset.

5.Now, find the SentiWordNet score of all tokenize words in particular tweet and add it with its posterior probability.

6. Compute the class conditional probability using Naive Bayes classifier .

Mathematically , for each word w in given tweet , class c (positive / negative class) and sentiwordnet score s (positivity score / negativity score) for word w.

$$\mathbf{P(c/w)=[[P(w/c)+s]P(c)]/[P(w)+s] \qquad ....(3)}$$

Naive Bayes classifier will be

$$\mathbf{c^{*}=arg\ maxc\ P(c/w) \qquad ....(4)}$$

VI. FUTURE SCOPE AND APPLICATIONS

Sentiment analysis of large datasets has always been a hot shot. Although many methods are there for analysis but all are somewhat lacking when it comes to accuracy . To improve the accuracy one way is to train your system in a way such that it gets the sentiment of word based on the entire tweet i.e if a word in the tweet has more than one meaning then it compares all the meanings of the word and takes the one which best suits the sentiment of entire tweet . This approach needs development of advanced highly intelligent algorithms . But this approach is surely going to add to the present accuracy of sentiment analysis systems.

The above explained sentiment analysis model has a flaw that it takes a lot of time in fetching data from twitter and in data management. This flaw can be overcome by use of Hadoop ecosystem instead of python as Hadoop ecosystem provides high speed processing of large data .

Application

Sentiment analysis of tweets can be extended to any review related website for example product review to understand products popularity , movie review etc . It can also be highly useful in sub component technology such as detecting antagonistic ,heated language in mails ,context sensitive information detection ,spam detection etc . Determining consumer attitudes and trends is one of the major applications of sentiment analysis of data.

VII. CONCLUSION

Sentiment analysis of tweets for movies provides ratings based on unbiased user tweets . Although many classifiers are available but Naive Bayes have been used because of its speed. When trained 1.6 million tweets of sentiment140 dataset and tested with most recent 100 tweets the system shows an accuracy of 58.40%. By use of sentiwordnet along with Naive Bayes for classification this accuracy can improved to a considerable extent as proposed .

APPENDIX

Python:

Python is very famous open source high level and dynamic programming language. There are many versions of it are already present to use. Version 2.7 is very famous and still many are stick on this version. For analytics version 3.0 or higher are suggested to use so that all types of APIs which may be used while classifying or performing any analytics can be easily combine and work.

NLTK:

Natural Language Toolkit (NLTK) is popular platform to interpret human language with the help of python. Now a days, there is tremendous amount of data and it is not possible for human being to interpret all data into some sensible information. For this purpose, NLTK libraries provides functionalities for tokenization, tagging, stemming etc. NLTK provides over 50 corpora and lexical resources along with the WordNet integration.

REFERENCES

[1]. *Efthymios Koulompis, TheresaWilson, Johanna Moore (2011),* " Twitter Sentiment Alaysis: The Good tha Bad and the OMG!," in: The fifth International AAAI Conference on Weblogs and Social Media.

[2]. *G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K.Miller*. 1990. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244

[3]. *Saif, Hassan; He, Yulan and Alani, Harith* (2012), "Semantic sentiment analysis of twitter," in: The 11th International Semantic Web Conference (ISWC 2012), 11-15 November 2012, Boston, MA, USA.

[4]. *Alec Go, Richa Bhayani and Lei Huaug, Stanford university(2009),* "Twitter Sentiment Classification using Distant Supervision ," in:The Third International Conference on Data Analytics.

[5]. *A. Kumar qnd T.M. Sebastian,* "Machine Learning assistedSentiment Analysis". Proceedings of International Conference on computer science and engineering (ICCSE'2012), 2012.

[6]. Bifet and E. Frank, "Sentiment Knowledge Discovery In Twitter Streaming Data", In proceedings of 13th International Conference of Discovery Science, Berlin, Germany : Springer, 2010

[7]. *Bo Pang and Lillian Lee*, Opinion mining and sentiment Analysis

[8]. *Tom M. Mitchell*, generative and discriminative classi_ers: Naive Bayes and Logistic Regression

[9]. *Christopher M. Bishop*, Pattern Recognition and Machine Learning

[10]. *Mark Lutz* , Programming python 4th Edition Trainingdata.zip Available at : http://help. sentiment140. com /for-students