Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach

Rincy Jose
Department of Computer Science and Engineering
Rajagiri School of Engineering and technology
Ernakulam, India
rinujk@gmail.com

Varghese S Chooralil
Department of Computer Science and Engineering
Rajagiri School of Engineering and technology
Ernakulam, India
varghesesc@gmail.com

***Abstract*—Sentiment analysis is the computational study of opinions, sentiments, evaluations, attitudes, views and emotions expressed in text. It refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Sentiment analysis over Twitter offers people a fast and effective way to measure the public's feelings towards their party and politicians. The primary issue in previous sentiment analysis techniques is the determination of the most appropriate classifier for a given classification problem. If one classifier is chosen from the available classifiers, then there is no surety in the best performance on unseen data. So to reduce the risk of selecting an inappropriate classifier, we are combining the outputs of a set of classifiers. Thus in this paper, we use an approach that automatically classifies the sentiment of tweets by combining machine learning classifiers with lexicon based classifier. The new combination of classifiers are SentiWordNet classifier, naive bayes classifier and hidden markov model classifier. Here positivity or negativity of each tweet is determined by using the majority voting principle on the result of these three classifiers. Thus we were used this sentiment classifier for finding political sentiment from real time tweets. Thus we have got an improved accuracy in sentiment analysis using classifier ensemble Approach. Our method also uses negation handling and word sense disambiguation to achieve high accuracy.**

***Index Terms*—Sentiment Analysis; WordNet; SentiWordNet; Word Sense Disambiguation; Machine Learning Methods; ensemble Approach.**

I. Introduction

Currently, twitter is becoming one of the most popular micro-blogging platforms. Millions of users can share their thoughts and opinions about different events and people on the micro-blogging platform. Therefore, Twitter is considered as a rich source of information for sentiment analysis.

Sentiment analysis can be considered as the use of natural language processing, text analysis and computational linguistics to identify and extract sentiment information in source materials. Generally, sentiment analysis aims to find the attitude of a writer with respect to some relevant topic or the overall contextual polarity of a document.The main task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature level[1] whether the expressed opinion in a document, a sentence or an feature is positive, negative, or neutral.The accuracy of a sentiment analysis is based on how well it agrees with human judgements. This can be measured by using precision and recall[2].

In this paper, we introduces an accurate sentiment classifier by combining machine learning classifiers with lexicon based classifier for finding political sentiment and sentiment towards new released movies from real time tweets. Section 2 contains detailed study of the method. Section 3 includes implementation details and results. Section 4 is the conclusion.

II. Methodology

The system mainly deals with the tweets extraction and sentiment classification. Here we use a sentiment classifier by combining machine learning classifiers with lexicon based classifier. The classifiers using are SentiWordNet classifier, naive bayes classifier and hidden markov model classifier.Our proposed system consists of mainly six modules. They are

1) Data acquisition
2) Pre-processing
3) Sentiment Classification using SentiWord-Net
4) Sentiment Classification using Naive Bayes
5) Sentiment Classification using HMM
6) Sentiment Classification using ensemble Approach.

Block diagram of proposed architecture for sentiment classification is shown in Figure 1. This classifier is used to find political sentiment from extracted tweets in real time manner.

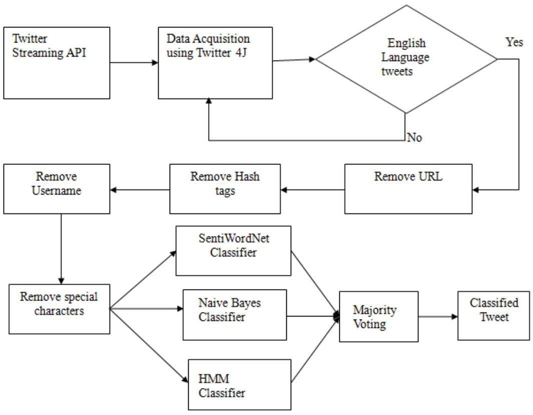

Fig. 1. Proposed Architecture

Tweets extracted in real time manner is serve as input to pre-processing module and then they are further classified as positive, negative or neutral.

A. Data acquisition

To obtain the Twitter feeds in continuous fashion Twitter streaming API tool is used[3]. It allows real time access to publicly available data on Twitter. These tweets serve as input to pre-processing module and then they are further classified as positive, negative or neutral.

B. Pre-processing

It first identifies the presence of URL using a regular expression and removes all the URLs from the extracted tweet. Then it removes all the private usernames identified by @ user. Then it removes all the Hash tags identified by the # symbol and all the special characters. Refined tweets are then classified using classification scheme.

Negation handling is one of the factors that contributed significantly to the accuracy of our classifiers. A major problem occurring during the sentiment classification is in the negation handling. Since here we use each word as feature, the word “win” in the phrase “not win” will be contributing to positive sentiment rather than negative sentiment .This will leads to the errors in classification. This type of error is due to the presence of “not” and this is not taken into account. To solve this problem we applied a simple algorithm for handling negations using state variables and bootstrapping. We built on the idea of using an alternate representation of negated forms[4]. This algorithm stores the negation state using a state variable. It transforms a word followed by a nt or not into “not_ ”+ word form. Whenever the negation state variable is set, the words read are treated as “not_ ”+ word. When a punctuation mark is encountered or when there is double negation, the state variable will reset.We have applied negation handling to our three classifiers separately for accurate classification.

C. Sentiment Classification using SentiWordNet

Sentiment classification is done on twitter data using SentiWordNet(SWN) and WordNet. Concept of word sense disambiguation is used for accurate classification[5]. WordNet is lexical database for the English language that groups English word into set of synonyms called synset. SentiWordNet is an extension of WordNet that assigns to each synset of WordNet three sentiment numerical scores, positivity, negativity and objectivity.

WordNet lexical relations are not always a good indicator of polarity detection. Synonyms may have different polarity based on the part of speech of the word in that sentence. This can be solved by using sense-tagged word lists.For that here we introduce Sentiment Classifier using Word Sense Disambiguation which is based WordNet. SWN classifier assigns different sentiment weights to different words. It also depends on the how the word is being used in the sentence i.e. identification of “part of speech” for the word is necessary to be classified by SWN classifier.

D. Sentiment Classification using Naive Bayes

The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. It relies on very simple representation of document as Bag of words. The model works with the bag of words feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label. For twitter sentiment analysis bigrams from the twitter data are used as features on Naive Bayes .It Classifies tweets into positive and negative labels.

E. Sentiment Classification using HMM

Our sentiment tagging system makes use of the Viterbi forward backward algorithm to traverse the states: where each state represents a possible prediction of the sentiment over the context traversed by the algorithm. In the continuity sentiment tags are predicted as we move forward with the algorithm. The Viterbi algorithm is a search algorithm that avoids the polynomial expansion of a breadth first search by trimming the search tree at each level using the best ”m” Maximum Likelihood Estimates (MLE) where ”m” represents the number of tags of the following word. The HMM models make use of two kinds of probabilities to keep account of the state of the sentence sentiment for the current: an emission probability keeps track of the sentiment tag given the word and its frequency of occurrence in the training data. The second probability is known as the transition probability accounts for the current state of the system given the state that the system was in previously. The advantages of this is that if there exists enough evidence to incite belief that the system is no longer analysing a positive, negative or neutral sentence, a necessary transition may be made to a better or more probable state. The decision for this transition is decided taking into context the previous state of the system, the current probability and the probability given the next word to be introduced will cause a transition. The Markov assumption takes into consideration the states preceding and succeeding the current states.

F. Sentiment Classification using Ensemble Approach

we introduce an approach that automatically classifies the sentiment of tweets by combining machine learning classifiers with lexicon based classifier[6]. Thus we are taking advantages of these three classifiers(SentiWordNet classifier, naive bayes classifier and hidden markov model classifier) for accurate classification of political data. Here positivity or negativity of each tweet is determined by using the majority voting principle on the result of these three classifiers. SentiWordNet classifier uses opinion lexical resource SentiWordNet and WordNet along with Word Sense Disambiguation for accurate classification of tweets which is extracted in real time manner. Thus this classifier considers the most suitable sense of word in its context. Other two classifiers are working based on the training data. Thus we have developed an accurate sentiment classifier for finding political sentiment from real time tweets.

III. Implementation Details

We extracted tweets about arvind kejriwal and kiran bedi for 3 weeks during the Delhi election days. Then we applied sentiment analysis on these tweets using classifier ensemble Approach.Tool used for Data Acquisition is Twitter Streaming API. Using Twitter Streaming API, it retrieved related tweets for a given query string in real time manner and it had been added to .csv file. Then pre-process the extracted tweets and refined tweets are classified using classification scheme.

We extracted 12000 tweets about Arvind Kejriwal for 3 weeks during the Delhi election days. Here we uses three classifiers for accurate classification of political data. Here positivity/negativity of each tweet is determined by using the majority voting principle on the result of these three classifiers.

The figure 2 comparing the probability of positivity of real time extracted tweets about Arvind Kejriwal and Kiran Bedi for 3 weeks during the Delhi election days.

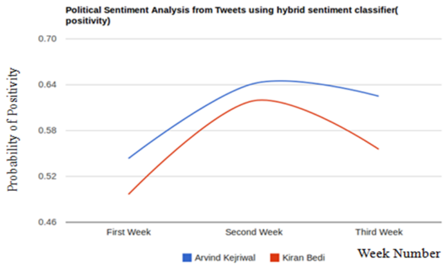

Fig. 2. political sentiment analysis from tweets using classifier ensemble approach

From this graph, it is clear that public sentiment towards Arvind Kejriwal is higher than that of Kiran Bedi.

We extracted tweets about two new released movies,namely Baahubali and Bajrangi Bhaijaan for 3 weeks. Then we applied sentiment analysis on these tweets using classifier ensemble approach.Then we got a graph in the following figure 3.

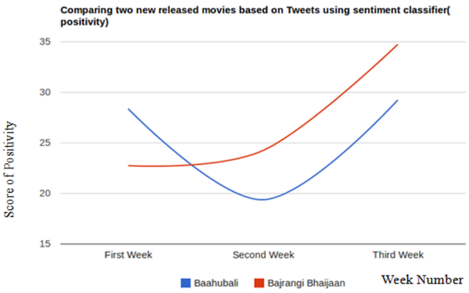

Fig. 3. sentiment analysis on new released movies using classifier ensemble approach

From this graph,it is clear that public sentiment towards film ’Baahubali’ is higher than that of ’Bajrangi Bhaijaan’ upto the end of first week. During the second week and third week ’Bajrangi Bhaijaan’ has higher public sentiment than ’Baahubali’.The accuracy improvement in this sentiment analysis method is shown in Table1.

While considering accuracy, it is observed that classification using ensemble Approach has higher accuracy compared to the individual classifiers.

TABLE I
ACCURACY OF DIFFERENT CLASSIFIERS

Classifier	Accuracy (in %)
SentiWordNet	21.05
Naive Bayes	69.92
HMM	64.06
classifier using ensemble approach	71.48

IV. CONCLUSION

In this work, we have Implemented a real time twitter sentiment analyser using classifier ensemble approach. Here it combines machine learning classifiers with lexicon based classifier. Thus we are taking advantages of these three classifiers such as SentiWordNet classifier, naive bayes classifier and hidden markov model classifier for accurate classification of political data. Thus we were developed a novel accurate sentiment classifier for finding political sentiment from real time tweets. Then Compared political sentiment towards two politicians by plotting graphs using results of sentiment analysis on extracted twitter data. Again this classifier was used to Compare two new released movies by the sentiment analysis on twitter data. Finally,we were analysed improvement in the accuracy of our classifier compared to the individual ones.

REFERENCES

[1] N. D. Valakunde, Dr. M. S. Patwardhan, *Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process* IEEE, International Conference on Cloud and Ubiquitous Computing and Emerging Technologies ,2013.

[2] Bing Liu, *Sentiment Analysis and Opinion Mining* Morgan and Claypool Publishers, May 2012.

[3] A. Bifet, G. Holmes, B. Pfahringer, *MOA-TweetReader: real-time analysis in twitter streaming data* LNCS 6926,Springer-Verlag, Berlin Heidelberg, 2011, pp. 4660.

[4] Vivek Narayanan1, Ishan Arora2, Arjun Bhatia3, *Fast and accurate sentiment classification using an enhanced Naive Bayes model* 2012.

[5] Pulkit Kathuria ,Kiyoaki Shirai, *Example Based Word Sense Disambiguation towards Reading Assistant System* The Association for Natural Language Processing,2012.

[6] Farhan Hassan Khan, Saba Bashir, Usman Qamar, *TOM: Twitter opinion mining framework using hybrid classification scheme* Decision Support Systems 57 (2014) 245257, 2013 Elsevier B.V.