Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation

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 issues in previous sentiment analysis techniques are classification accuracy, as they incorrectly classify most of the tweets with the biasing towards the training data. In opinion texts, lexical content alone also can be misleading.Therefore, here we adopt a lexicon based sentiment analysis method, which will exploit the sense definitions, as semantic indicators of sentiment. Here we propose a novel approach for accurate sentiment classification of twitter messages using lexical resources SentiWordNet and WordNet along with Word Sense Disambiguation.Thus we applied the SentiWordNet lexical resource and Word Sense Disambiguation for finding political sentiment from real time tweets. Our method also uses a negation handling as a pre-processing step in order to achieve high accuracy.**

***Index Terms*—Negation Handling; Sentiment Analysis; WordNet; SentiWordNet; Word Sense Disambiguation.**

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 the source materials. Generally, sentiment analysis aims to find the attitude of a writer with respect to any 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 a feature is positive, negative, or neutral.Document level sentiment analysis is the classification of the overall sentiments mentioned by the reviewer in the whole document text in positive, negative or neutral classes.

Sentiment Classification techniques can be roughly divided into the machine learning approach, lexicon based approach and hybrid approach [2]. The machine learning approach applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches.

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 [3].

In this paper, we introduce a Sentiment Classifier using Word Sense Disambiguation (WSD)

978-1-4673-7349-4/15/$31.00 ©2015 IEEE

which is based on lexical resources WordNet and SentiWordNet for finding the political sentiment from real time tweets. Section 2 contains a detailed study of the method. Section 3 describes the implementation details and results. Section 4 is the conclusion.

II. Literature Survey

Bifet A., Holmes G., and Pfahringer B. [4] discussed the handling of tweets in real-time. The research paper introduced a system, MOATweetReader, which processes the tweets in real-time despite their dynamic nature. The system performs two functions: First, it detects the changes in term frequencies and second, it performs sentiment analysis in real-time. The authors only use positive or negative classes for sentiment classification. There is a possibility that a tweet may be neutral, which is not considered in this paper.Montejo-Raez A. [5] proposed an unsupervised approach for sentiment polarity detection from twitter tweets. The polarity scores are calculated from SentiWordNet and the random walk algorithm is used to calculate the weights from the tweet. The proposed algorithm has comparable performance with SVM algorithm. The benefit of the proposed technique is that there is no need of training corpus as required in supervised learning techniques and there is no dependency on the model domain. The limitations include handling of negation, manual labelling process for certain tweets and facing flaws in the calculation of the final polarity score.Farhan Hassan Khan [6] proposed a twitter opinion mining framework using hybrid classification scheme.This paper presents an algorithm for twitter feeds classification based on a hybrid approach. This involves pre-processing steps and a hybrid scheme of classification algorithms.The proposed classification algorithm incorporates a hybrid scheme using an enhanced form of emotion analysis , SentiWordNet analysis and an improved polarity classifier using list of positive or negative words. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy when compared to similar techniques.

III. Methodology

The system mainly deals with the tweets extraction and sentiment classification. To incorporate semantics in sentiment analysis, we use a database of sentiment organized words, SentiWordNet along with WSD concept. Then this system used to measure political sentiment. The system consists of mainly three modules. They are

A) Data acquisition
B) Pre-processing
C) Sentiment classification.

Block diagram of the 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.

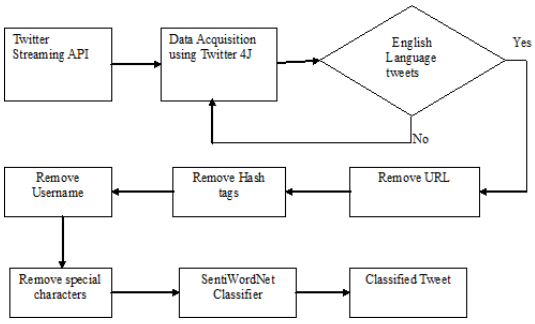

Fig. 1. Proposed Architecture

A. Data acquisition

To obtain the Twitter feeds in continuous fashion Twitter streaming API tool is used [4]. It allows real time access to publicly available data on Twitter. These tweets serve as input to preprocessing 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 by identifying "@username". Then it removes all the hashtags 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 classifier. The major problem occurs during the

sentiment classification is in the negation handling. When we use each word as a feature, the word "win" in the phrase "not win" will be contributing to positive sentiment rather than negative sentiment .This will lead 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[7]. This algorithm stores the negation state using a state variable. It transforms a word followed by a n't 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.

Many words with strong sentiment occur only in their normal forms in their training set. But their negated forms would be of strong polarity. We solved this problem by adding negated forms to the opposite class along with normal forms during the training phase. It means that if we encounter the word "fail" in a negative document during the training phase, we increment the count of "fail" in the negative class and also increment the count of "not_ fail" for the positive class. This is to ensure that the number of "not_ " forms are sufficient for classification. This modification resulted in a significant improvement in classification accuracy due to bootstrapping of negated forms during training.

C. Sentiment Classification

Sentiment classification is done on twitter data using SentiWordNet(SWN) and WordNet.Concept of word sense disambiguation is used for accurate classification[8]. The input tweets are retrieved by using Twitter streaming API.The Twitter streaming API allows real time access to publicly available data on Twitter. The tweets serve as input to the pre-processing module and then they are further classified as positive, negative or neutral. WordNet is a lexical database for the English language that groups English word into set of synonyms called synset.WordNet distinguishes between nouns, verbs, adjectives, adverbs. SentiWordNet is an extension of WordNet that assigns three sentiment numerical scores to each synset.The scores are:

- PosScore [0,1]: positivity measure
- NegScore [0,1]: negativity measure
- ObjScore [0,1]: objective measure.

Sample of SentiWordNet fragment is shown in Figure 2.

Category	WNT Number	pos	neg	Synonyms
A	01123148	0.875	0	good#1
A	00106020	0	0	good#2 full#6
A	01125429	0	0.625	bad#1
A	01510444	0.25	0.25	big#3 bad#2
N	03076708	0	0	trade_good#1 good#4 commodity#1
N	05144079	0	0.875	badness#1 bad#1

Fig. 2. SentiWordNet Fragment

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.

Here we introduce the Sentiment Classifier using Word Sense Disambiguation which is based on WordNet and SentiWordNet. First Tokenization and Speech Tagging was done on refined tweet. Then Word Sense Disambiguation was applied in classification. The techniques of WSD are aimed at the determination of the meaning of every word in its context. For each word in a tweet, disambiguation selects the synset in WordNet that best represents this word in its context.For that it selects the synset with the highest similarity score from the WordNet.

SWN classifier assigns different sentiment weights to different words. It also depends on how the word is being used in the sentence.It means that identification of "part of speech" for the word is necessary to be classified by SWN classifier. The sentiment value of each word is calculated by calling the SentiWordNet. First,the Sentiment weight of each word is found and then Sentiment weight of each sentence is calculated by adding sentiment weight of each word in that sentence.

IV. 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 sentiment lexicons SentiWordNet and WordNet.

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.For that we had to create an app with twitter after logging in through our account. Thus we will get authentication fields such as consumer_ key, consumer_ secret, access_ token and access_ token_ secret.The keyword to be searched, count of tweets to be extracted and filename to which tweets to be written must be specified in config.json file. Then pre-processed the extracted tweets and refined tweets are classified using classification scheme. Thus we applied the SentiWordNet lexical resource and Word Sense Disambiguation for finding political sentiment from real time tweets.

We extracted 12000 tweets about Arvind Kejriwal for 3 weeks during the Delhi election days. Then we calculated total positive sentiment score and total negative score by applying sentiment analysis on these tweets using sentiment lexicons SentiWordNet, WordNet.

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

From this graph, it is clear that public sentiment towards Arvind Kejriwal is higher than that of Kiran Bedi.This sentiment analysis technique shows an accuracy of 78.6% for finding the political sentiment from real time extracted tweets.In the exsisting method[6] the accuracy achieved was 75% .In our method accuracy improvement is due to the use of word sense disambiguation and negation handling in the sentiment analysis.

V. Conclusion

In this paper, we have implemented a real time, domain independent twitter sentiment analyser using sentiment lexicons such as SentiWordNet and WordNet.It compared political sentiment towards two politicians by plotting graphs using results of sentiment analysis on real-time extracted twitter data.This was done by applying WSD and negation handling in order to increase accuracy

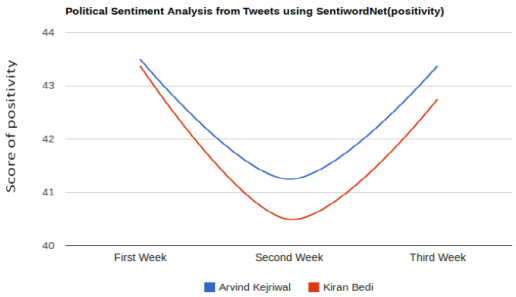

Fig. 3. political sentiment analysis from tweets using SentiWordNet and WSD

of sentiment analysis. Negation handling results in 1% improvement in classification accuracy and WSD results in 2.6% improvement in classification accuracy. This twitter sentiment analyzer can also be used to compare two new released movies or two products in real time manner.

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] Walaa Medhat a, Ahmed Hassan b, Hoda Korashy b, *Sentiment analysis algorithms and applications: A survey* Ain Shams Engineering Journal science direct 2014.

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

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

[5] A. Montejo-Raez, E. Martnez-Camara, M.T. Martn-Valdivia, L.A. Urena-Lopez, *Random Walk weighting over SentiWordNet for sentiment polarity detection on Twitter* Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2012, pp. 310.

[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.

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

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

[9] Alaa Hamouda , Mohamed Rohaim , *Reviews Classification Using SentiWordNet Lexicon* The Online Journal on Computer Science and Information Technology (OJCSIT), Vol. (2) No. (1), 2012.