Opinion Mining and Sentiment Analysis: Tamilnadu Pre-Election: A Survey

L.Hari Uthra, R.darshini, K.Megha and S.Prince Sahaya Brighty

Department of computer science and engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India
hariuthralokendran@gmail.com
darshinilatha95@gmail.com
kuzhupillymegha@gmail.com
brighty.s@srec.ac.in

ABSTRACT

Nowadays Opinion Mining has become a budding topic of research due to lot of blinkered data available on Blogs & social networking sites. Tracking different types of opinions & summarizing them can present valuable insight to different types of opinions to users who use Social networking sites to get reviews about any product, service or election polls. This Paper presents approach for analyzing the sentiments of users using data mining classifiers. We have used natural language processing to routinely read reviews and used Naive Bayes classification to determine the polarity of reviews. We have also extracted the reviews of pre-election poll and the polarity of their reviews. One such application in the field of politics, where political entities need to understand public opinion and thus determine their campaigning strategy. Finally, we calculated the number of positive and negative sub-tweets for each candidate and made a graph based analysis.

Keywords: Opinion mining, Sentiment analysis, Polarity, Machine learning, Postagging, Stemming.

1. INTRODUCTION

The domain is on Data mining, which is *the extraction of hidden predictive information from large databases,* is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Opinion mining [1] aims at automatically identify opinionated content, and determining people's sentiment, perception or attitude towards an entity or topic. Using opinion mining, it is possible to automatically analyze this vast and rich user-generated content, and develop less expensive and lower latency solutions for public opinion elicitation, with a reasonable degree of accuracy.

Opinion mining tools allow businesses to understand new product opinion, product sentiments, brand view and reputation management [7]. These tools help users to perceive product opinions or sentiments on a global scale. Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a given domain. However, sentiment is expressed differently in different domains, and it is expensive to interpret data for each novel domain.

Textual information can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are subjective expressions that describe an individual's sentiments, appraisals or feelings toward entities, events and their properties [4]. People express their opinions not only about products and services, but also about various topics and issues especially from social domains [3].

Sentiment analysis is a method of computing and stratifying a view of a person given in a piece of a text, especially in order to identify persons thinking towards a specific topic product etc is positive or negative or neutral. Social media portals have been globally used for expressing a sentiments publicly through text based message and images [13]. Currently, Twitter, facebook linkedIn, flickr etc enables users to post their view publicly. Among all social media.Twitter widely get used for expressing view on certain topic. Twitter allow tweeple(i.e users of twitter) to post their opinion on political environment, entertainment, industry, stock market etc.

During election each authoritative organization publishes about 12 polls every year, and most of them are concentrated during the month that precedes the election. The time elapsed between any two published polls varies
Enormously (days to months), One may wish to look into whether public indicators about health, education or security may be predicted based on the public opinion expressed about these services. These indicators are infrequent, and usually are produced according to specific conditions (e.g. census). In this paper, we develop a case study to investigate whether it is possible to predict variations in vote intention polls, based on the sentiment articulated on user-generated comments in internet. More specifically, we consider the Tamilnadu political circumstances in which public election-related polls represent sparse data, i.e., there are few data points and the time elapsed between two polling execution varies. Public polls are very important, as political parties discover their results as a major part of their campaigning strategy. Political parties can conduct their own polls, but results cannot be published. In addition, this possibility is subject to budgetary constraints, which may influence their precision.

2. RELATED WORKS

The studies regarding predicting election outcomes in several countries using Twitter data have been conducted in the past few years and the various prediction results of the product.
In this paper, they have described the approach in predicting the results of election with respect to Indonesian Presidential Election in 2014[2].It performed an automatic buzzer detection on Twitter dataset to remove buzzers, such as computer bot, paid users, and fanatic users that usually produce noise in the data distribution. Then, it used fine-grained political sentiment analysis to partition each tweet into several subtweets and subsequently assigned each sub-tweet with one of the candidates and its sentiment polarity. Finally, to predict the election results, we leveraged the number of positive subtweets for each candidate. The mean absolute error (MAE) of our prediction is comparable to the results obtained by the other Twitter-based prediction approach conducted by Prasety[2]. Moreover, the most interesting part is that Twitterbased prediction successfully outperforms all prediction results published by several survey institutions. This would suggest that Twitter is a useful resource for predicting the outcomes of Presidential Election, at least in Indonesia.

It has a dataset, and the dataset is labeled in two stages, it can be used to train a supervised machine learning model to perform public sentiment analysis and predict election outcome. We split the dataset in 80:20 ratios to prepare the training and testing sets [12].In this work, relatively recent in this area is especially content analysis of social networks, through the analysis of polarity and intensity of opinion. The first objective of our work consisted of studying the different issues of social network analysis and text mining for politics purposes. The second objective of their work consisted of searching and locating the content corresponding to the French Presidential Election 2012 in posted tweets. It is common that the *tweets* contain a link to a web page, which means to have multiple levels of processing of the message content. We will therefore add to the analysis of *tweet*, analysis of the Web page to which it refers [8]. In this paper, they have examined whether the sentiment extracted from user-generated content with regard to political news could be used to forecast variations in vote intention polls. They also developed an approach for Extracting sentiment of user-generated comments in Portuguese and examined two methods for opinion mining and proposed two types of features to represent the sentiment, summarization and bursts.

With the increased use of social media the current paper focused mainly on use of social media as a tool for election campaign. India which is known to be one of the wired countries in the world with having more than 65 % of its youth below age-group of 35; The proposed systems will try to analyze the Maharashtra state assembly election; to study the impact of social media on Maharashtra politics system found people can express their views in 140 characters more efficiently and openly[15].In this research, they proposed a framework to mine opinions in microblog domain and then build a real-time analysis system to monitor the sentiment propagation, sentiment fluctuation and hot opinions in Weibo. Opinions are abstracted as triples, based on which we then design an opinion mining method. We build a microblog-oriented sentiment lexicon and propose a lexiconbased sentiment analysis algorithm to classify sentiments[11].

This paper focuses on the frame work on opinion mining to perform the analysis on the enterprises which have been done in each phases. Opinion Mining allows to quickly identify tonality of the key conversations, trends and issues surrounding business in real-time. It also allows businesses to track positive and negative spikes in the reviews of the customers[6].

In this paper we have presented a relatively novel domain of calamitous situations for crowd sentiment detection. develop a baseline model using the Naïve Bayes Classifier trained on features like unigram and Part of Speech (POS) tagging and tested it on Kashmir floods data set. An overall accuracy of 66.88% is achieved, further to improve the performance of classifier the training set needs to be enhanced[3].In this paper was to capture polarity of the sentiments captured from twitter data. We have used Case Study of Digital India mission to achieve our goals. The results are encouraging as we are able to segregate sentiments as 250 positive, 150 neutral and 100 negative sentiments. We got these results on a small data set of 500 tweets, which is quite small for this case study but we would try to implement the same on larger data set of twitter corpus[4]. It can be observed from the experimental results that data mining classifiers is a good choice for sentiments prediction using tweeter data. In a experimentation, knearest neighbour(IBK) outperforms over all three classifier namely RandomForest, baysNet, Naive Baysein. RandomForest also gives good prediction accuracy. There is a no need to use of ensemble of classifier for sentiments predictions of tweets as single classifier (i.e knearest neighbour) gives a better accuracy over all combinations of ensemble of classifier .

3. PROPOSED SYSTEM

The architectural view of the proposed system (fig 3.1). In this paper, we study the various methods and techniques related to different domains like opinion summarization, sentiment analysis, etc. This framework collects the different set of opinions and applies all techniques which we already explained for getting summary. This summary is shown in the form of graphs and also in text which are easily understandable. The proposed system involves pre-processing of the textual data which is entered by the users about the elections which is conducted by the election commission.

The users will be provided with the separate user login to register their reviews about election with their city. The user can also view other reviews from his account. The admin also have an account so he can view the user reviews and he can calculate the polarity for each party with city.

The following segment gives a complete view of the proposed work. The proposed system uses customer reviews to take out aspect and mine whether given is positive or negative opinion. Each review is split into individual sentences.

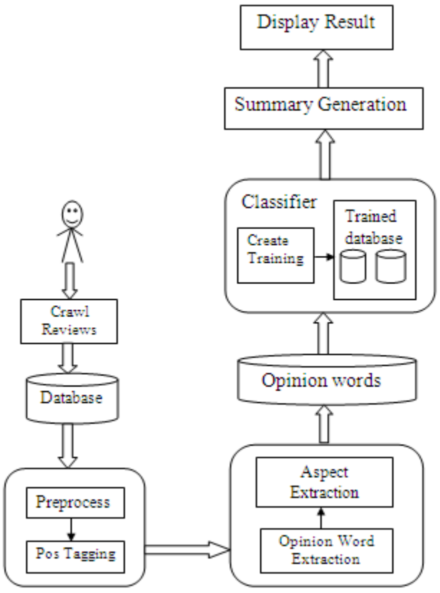

Fig.3.1 Architectural diagram

A review sentence is given as input to data preprocessing. Next, it extracts aspect in each review sentence. Stop word removal, stemming and pos tagging are data preprocessing. Sentiment orientation is used to identify whether it is positive or negative opinion sentence. Then it identifies the number of positive and negative opinions of each aspect.

3.1. Stop Word Removal

Most frequently used words in English are not useful in text mining. Such words are called stop words. Stop words are language specific functional words which carry no information. It may be of types such as pronouns, prepositions, conjunctions. Stop word removal is used to remove unwanted words in each review sentence. Words like is, are, was etc. Reviews are stored in text file which is given as input to stop word removal. Stop words are collected and stored in a text file. Stop word is removed by checking against stop words list.

3.2. Stemming

Stemming is used to form root word of a word. A stemming algorithm reduces the words "longing", "longed", and "longer" to the root word, "long". It consist many algorithms like n-gram analysis, Affix stemmers and Lemmatization algorithms. Porter stemmer algorithm is used to form root word for given input reviews and store it in text file.

3.3. POS Tagging

The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Common POS categories in English grammar are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the task of labeling (or tagging) each word in a sentence with its appropriate part of speech. POS tagging is an important phase of opinion mining, it is essential to determine the features and opinion words from the reviews. POS tagging can be done either manually or with the help of POS tagger tool. POS tagging of the reviews by human is time consuming. POS tagger is used to tag all the words of reviews. Stanford tagger is used to tag each word in an online review sentences. Every one sentence in customer reviews are tagged and stored in text file.

3.4. Aspect Extraction

Frequent item set mining is used to find all frequent item sets using minimum support count. Here, every sentence is assigned as single transaction. Noun Words in each sentence is assigned as item sets for single transactions. Aspect extraction is implemented using figure 2. This algorithm first extracts noun and noun phrases in each review sentence and store it in a text file. Minimum support threshold is used to find all frequent aspects for a given review sentences. Aspects like pictures, battery, resolution, memory, lens etc. Then, the frequent aspects are extracted and stored in text file.

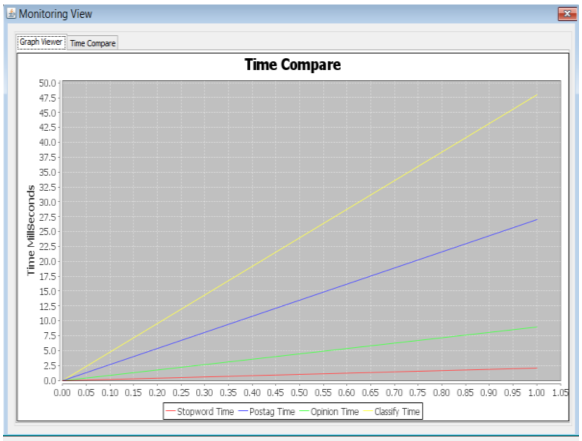

Fig 3.3 Timeline graph of the process

3.5. Sentence and Aspect Orientation

The proposed system first determines the number of positive and negative opinion sentence in reviews using opinion words. The positive and negative labels are collected labels in opinion words. Examples of positive opinion words are long, excellent and good and the negative opinion words are like poor, bad etc. And the next step is to identify the number of positive and negative opinions of each extracted aspect. Both sentence and aspect orientations are implemented using Naïve Bayesian algorithm using supervised term counting based approach. The probabilities of the positive and negative count are found according to the words using Naïve Bayesian classifier.

3.6 .Results

Our application allows calculating the current tendency for the set of selected candidates. To classify the tweets as positive or as negative naive bayes algorithm is used. We save all the results in the database with all the previous results. We use this data for the different statistical analyses done in different periods of time. Such a temporal analysis allows observing and comparing in easy way the changing of trends. The very interesting example is observation of trends changing during the last period of election. Fig 3.2 represents the dmk party's support at Coimbatore and the result shows the positive outlook based on the opinions given by the users in that city. Fig 3.3 shows the timeline graph of the sentiment analysis process.

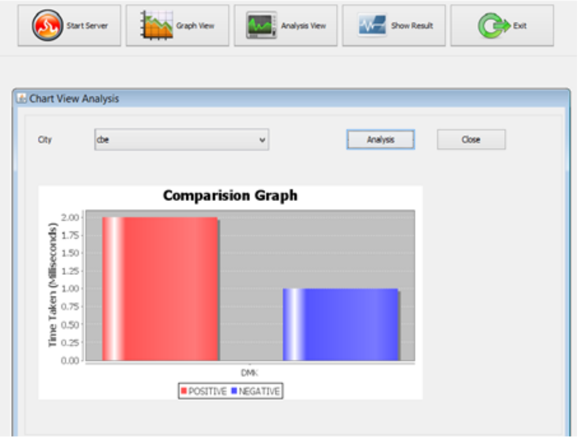

Fig 3.2 chart of Dmk support in coimbatore

Naïve Bayesian algorithm

Steps are as follows:

1. The positive labels, negative labels and review sentences are stored in separate text file.2. Split the sentence into the combination of words. It means first combination of two words and then single words.

3. First compare the combination of two words, if it matched then delete that combination from the opinion. Again start comparing of single word.

4. Initially, the probabilities of positive and negative count to zero [positive=0, negative=0].

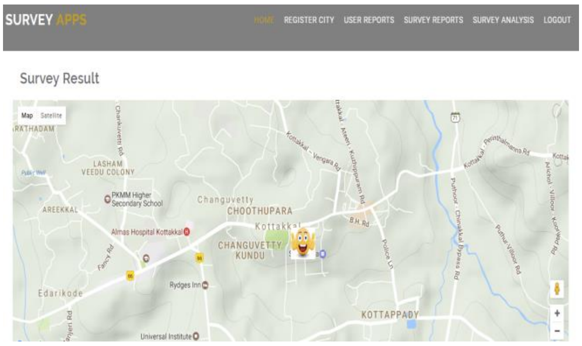

Fig 3.4 Expression(happy) of the people in Coimbatore city

Opinion word rule gives that, if word is

Matched with positive opinion words then positive count get increment, or it is negative opinion word then negative count get increment.

Two rules must be applied:

1. Negation Negative->Positive. This will increment positive count.

2. Negation Positive ->Negative. This will increment negative count.

After comparing all the words of the sentence, the found probabilities of the positive and negative counts are compared in the following manners.

a) If the probability of positive count is greater than the negative count, then the sentence or opinion is positive.

b) If the probability of negative count is greater than the positive count, then the sentence or opinion is negative.

c) If the probability of positive count minus probability of negative count is zero, then it is neutral. Finally system identifies the number of positive and negative opinion of each extracted aspect in customer reviews.

4. CONCLUSION AND FUTURE WORK

It this paper, we examined whether the sentiment extracted from user-generated content with regard to political news could be used to forecast variations in vote intention polls. The use of social media for prediction of election results poses challenges at different stages. The analysis of social networks is a topic very interesting, full of prospects, with possible applications in various fields of science such as sociology, politics, medicine, communication, marketing, etc. Our work consisted of studying the different issues of social network analysis and text mining for politics purposes, we have presented the detailed procedure to carryout sentiment analysis process to classify highly unstructured data of Twitter into positive or negative categories using naive byes algorithm.

The public opinion surveys are analyzed by admin. Finally, the results are visualized with the help of map and emoticons based on the reviews by the user. We reached an accuracy of 70% of prediction for the binary class problem, mainly based on negative sentiment, which we are able to detect with significant confidence. Unlike other works, mentions to candidates revealed very poor predictive power, compared to sentiment-based features.

We are presently developing more experiments with new testing data, among them the second round of the 2020 elections. In addition, we could experiment the proposed approach for other meager governmental indicators, such as popularity or

Government approval, for census data per critical area .As future work, we want to develop mechanisms to incorporate sentiment articulated in various medias, each one with its own form of expression and media representativeness. Some of the issues that need to be solved are: what techniques are suitable to each media, and its underlying expression behavior; how to discover the representativeness of the population interacting through the media, and in which section the opinion should account for prediction; in the midst of others. Another important line of work is to address reactions to news beyond direct comments on the newspapers, such as repercussion on Facebook or Twitter.

REFERENCES

[1] Diego Tumitan and Karin Becker, (2014) ”Sentiment-based Features for Predicting Election Polls: a Case Study on the Brazilian Scenario”International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) PP: 127-133 IEEE.

[2] Mochamad Ibrahim, Omar Abdillah, Alfan F. Wicaksono, Mirna Adriani, (2015) “Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in A Twitter

Nation” 15th International Conference on Data Mining Workshops PP:1348-1353 IEEE.

[3] Harvinder Jeet Kaur, Rajiv Kumar, (2015) “Sentiment Analysis from Social Media in Crisis Situations” International Conference on Computing, Communication and Automation (ICCCA2015) pp: 251-256 IEEE.

[4] Prerna Mishra , Dr. Ranjana Rajnish , Dr.Pankaj Kumar, (2016) “Sentiment Analysis of Twitter Data:Case Study on Digital India” International Conference On Information Technology pp:148-153 IEEE.

[5] ISonali J. Bagul , Prof. Rakhi D. Wajgi, (2016) “Design Feedback Analysis System for E-Commerce Organization “ Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR’16) IEEE.

[6] Pankaj Kumar, Kashika Manocha . Harshita Gupta, (2016) “Enterprise Analysis Through Opinion Mining” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) PP:3318-3323 IEEE.

[7] A.Jeyapriya ,C.S.Kanimozhi Selvi , (2015) “Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm” IEEE sponsored 2nd international conference on electronics and communication systems(icecs) PP: 548-552.

[8] Katarzyna Wegrzyn-Wolska, Lamine Bougueroua, (2012) “Tweets mining for French Presidential Election” Fourth International Conference on Computational Aspects of Social Networks (CASoN) PP: 138- 143 IEEE.

[9] Vaanchitha, Kalyanaraman, Sarah Kazi, Sangeeta Oswal , (2014) “Sentiment Analysis on News Articles for Stocks” PP: 10- 15 IEEE.

[10] Venkata Rajeev P, Smrithi Rekha V, (2015) “Recommending Products to Customers using

Opinion Mining of Online Product Reviews and Features” International Conference on Circuit, Power and Computing Technologies [ICCPCT] IEEE.

[11] Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun, (2014) “Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach” 7th International Conference on Utility and Cloud Computing PP: 890-895 IEEE.

[12] Jyoti Ramteke, Darshan Godhia, (2015) “Election Result Prediction Using Twitter sentiment Analysis” IEEE.

[13] Anurag P. Jain, Mr. Vijay D. Katkar, (2015) “Sentiments Analysis Of Twitter Data Using Data Mining” International Conference on Information Processing (ICIP) PP: 807-810 IEEE.

[14] Delenn Chin, Anna Zappone, Jessica Zhao “Analyzing Twitter Sentiment of the 2016 Presidential Candidates “ American Journal Of Science and Research.

[15] Nilesh V. Alone , Gayatri P. Wani, (2015) “Analysis of Indian Election using Twitter” International Journal of Computer Applications pp: 37-41 .