ISSN 2395-1621

Empirical Study of Sentiment Analysis of Multi-Lingual Real Time Tweets for Psephology

Prof. Rachana Sable, Ayush Chhajer, Sudhanshu Ranjan, Ajam Bhandari, Patrick D'souza

rachana.sable@raisoni.net
ayush.chhajer@gmail.com
santirudra1994@gmail.com
aazambhandari1998@gmail.com
patrickdsouza07@gmail.com

Department of Information Technology Engineering
G.H. Raisoni Institute Of Engineering And Technology
Wagholi,Pune,India

ABSTRACT

Due to the properties of social media being interactive in nature and real time, it has grabbed much attention which makes it best suited for political analyzing related work. Recent research have probed that social media platforms can record the mindsets of netizens which in turn are the valuable assets to be considered for elections. A large amount of raw data is generated by these social media platforms that can be used for making decisions which can be turning point in election.

Keywords — sentiment analysis; prediction; elections.

ARTICLE INFO

Article History

Received: 14th March 2020

Received in revised form : 14th March 2020

Accepted: 21st March 2020

Published online : 21st March 2020

I. INTRODUCTION

Social media platforms like Twitter, Facebook, etc. are increasingly being used to express their opinions and interests by people. This has piqued the interest of businesses and other social entities, which are employing very sophisticated tools to gather and make use of the large amounts of data available on these social media platforms. "Big data" is referred to as the huge amount of data generated by these platforms. After scrutinizing a lot of research related to election prediction, a survey paper is created in which every work related to psephology using social media is incorporated.

Sentiment analysis of the data obtained from social media platforms that has proven to be an effective way of capturing the opinion of the people and predict trends, thereby improving the decision making the process. One such application of this is the Prediction of Election Results.

Our prediction for the maximum livelihood of a party winning will be done by our proposed machine learning model. We mined data from Twitter. Mined data was cleaned and relevant data was fed into Text Blob for sentiment analysis. The polarity of tweets obtained will be used in many models. Real time data pertaining to the Elections was used for training the model, which was then used to predict the results of the 2019 elections.

Prediction is done and accuracy of the model is obtained by comparing the results with the values of the votes obtained by the two parties in polls.

1.1 Motivation

The election process is an integral part of the fabric of a country. It is a complex process and many socio-political institutions, including political parties, attempt to predict the trends and mood of the people to gain an insight into the process and make use of the gathered information. Machine Learning tools are perfect for analysis of such a process and far surpass the traditional means.
Due to rapid increase in the number of users in social media, a powerful platform has been provided to netizens to voice their opinions. Social media Platforms like Twitter and Facebook is being actively used to share reviews, ratings, and recommendations. This vast array of information can be actively used for marketing and social studies. Political Campaigns have used this information available on the above platforms to design their marketing campaigns. Huge monetary investments by politicians in digital marketing campaigns right before an election along with arguments

and debates between their opponents and supporters only enhance the claim that opinions and views posted by netizens may predict the results of an election.

The practice which is widely gaining momentum throughout the world is the ability to extract insights from social media data and forms a fascinating area of study with the ability to provide a wider view on how public opinion is shaped. Therefore, to get a clear idea of what deed is done by a Politician our system will act as a feedback.

1.2 Contribution

The work provides a brand new approach to tackle the task of prediction of polls using data gathered from social media site Twitter. In contrast to the existing work where sentiment analysis is performed on gathered data and classification methods like Naive Bayes or SVM are applied to categorize the data into one or more classes, the problem will be tackled as one of Regression where a Machine Learning has been used to compare the correlation between sentiment expressed on Twitter and the number of votes a party receives. Moreover tweets, hashtags are the important features which are mined from Twitter for that specific application. Hence, it is mandatory to implement a labeling technique for the mined Twitter data which can make a balance between accuracy and speed.

The remainder of this paper has been organized into 4 sections. 'Related Work' discusses the existing literature and previous work done in the domain. Further, it is discussed how our work builds on top of the current work done. 'Material and Methods' section describes the approach used and provides a step by step account of the process followed as well as details about the machine learning methods used. 'Conclusion' section summarizes the work done.

II. RELATED WORK

The major portion of related corpus focuses on the process of Data Gathering and subsequent Sentiment Analysis of preprocessed data.

In cross-domain dataset has been used to train the sentiment analysis model, to make up for the lack of availability of contextually relevant data. In these product reviews (15166), How Net and NTUSD, Chinese Original Basic Lexicon (OBL) datasets are used. It focuses on Features classification into four categories:

Sentiment words (Number of positive words, negative words in document, Sum of P and N, Difference of P and N, Appearance number of sentiment words), n-Gram(Number of 2-Grams in document),

Statistical information (Number of question marks, exclamation marks, negations in document, Length of document,Number of adversative conjunctions,adjectives in document) and Results (Score, Number of sentences judged to be positive, negative & neutral) of lexicon-based method Features. It's accuracy for Hotels: H (0.858), Books: B (0.929), Electronic devices: E (0.816). It can be made not inferior to the domain of electronics. [1]

The authors discuss how gathered data can be effectively labeled and prepared for classification. Further comparison of different classification models has been provided. In this Twitter dataset (30,000 tweets for the training data set) is used. It focuses on Hashtags features. It's accuracy for MNB (nltk)-0.54 MNB (Scikit-learn)-0.97 SVM (nltk)-0.58 SVM (Scikit-learn)-0.99. We can use such algorithm packages which provide more accuracy for classification. [2]

In data in Bengali corpus has been classified in six feeling classes and to annotate the sentences. It uses Twitter Data as dataset. It focuses on Annotators from blog as Features. In future, it can adapt a corpus-driven method for building a lexicon of emotion words and phrases and extend the emotion analysis tasks in Bengali language. [3]

Authors have not considered the emojis which are a relevant aspect when explaining the polarity of a tweet. Since data was labeled manually the size i.e. 36,465 was not big enough to provide more accuracy, so we can fetch more tweets and label them. It uses Twitter data as Dataset. It focuses on words and pharses which have been taken out from the tweets are consider as Features. It's accuracy generated were Naïve Bayes (62.11%), SVM (78.42%), Dictionary based (34.1%). [4]

Authors uses dictionaries of good and bad words to determine the polarity. It uses Twitter dataset (9021 tweets from 21 different domains) and Twitter dataset (16,454 tweets).

Generic lexicon in the lexeme space can be use in future for classification, where it cannot distinguish between the various senses of a word. A concept expansion approach, to expand the feature vectors, may prove to be useful. This is because of the extensive world knowledge embedded in the tweets. [5]

Authors Taboada, Brooke, and Stede explain the model that integrate pack of-words in talk markers with the slant demand by 5% exactness. It focuses on weighting, multiple cut-offs as Features. It uses Polarity Dataset. In future it is not enhanced which could have been done. With multiple sources of knowledge (Taboada, Brooke, and Stede 2009). [6]

Authors recommended portraying Hindi reviews as positive, neutral, negative. They crack another score smashing point and used it for two different methods. Moreover, fusion takes place between the POS Tagged Ngram and central N-gram. Twitter data is used as dataset. [7]

Authors try out two non-identical strategies, Multinomial Naïve Bayes (MNB) and SVM. They found that MNB methodology surpasses SVM on scaled scale areas with short meterial. Twitter data is used as dataset. Its accuracy is 74.85 %. Disadvantage is we can take only 1000 documents per class in line with the movie review corpus. This let us to remove any primary sentiment angle which may be present. [8]

To recognize the sentiment from Hindi content, Authors give rise to an well organised approach. By adding more opinion words, they developed Hindi language entity and improve the present Hindi Senti Word Net (HSWN). 80% precision has been measured by them. It uses annotated dataset. It focuses on words in the document as features. Its accuracy is 80.21%. Disadvantage is that the dataset is not extended for the better and generalized results. [9]

Authors equate quantities of public opinion estimated from survey with sentiment computed from text. They scrutinize some studies on public faith and ministerial ideas from year 2008 to 2009, and discover that they collectively agreed to sentiment idiom frequencies in coetaneous Twitter messages. While our outcome differs over datasets, in individual instances the mutuality are high about 80%, and seize noteworthy extensive drift. The outcome focuses the prospects of text streams as a replacement and extension for traditional ballot. [10]

Authors come up with an interpreted record on the subject of electoral prediction using Twitter data. The order is sequential, In the middle of papers there are cross-references, and few of them are just assigned for the advantage of comprehensiveness. In addition to that, seminal papers not entirely related to the topic are included plus papers on associated sectors such as integrity, gossips, and anthropology of twitter users. [11]

Authors inspected over 100,000 tweets, bring up political organizations in advance to the German federal election 2009. All-inclusive, we observe that Twitter is as a matter of fact used as a podium for state meditation. The bare number of tweets comes adjacent to traditional opinion polls and revolves voter fondness, while the emotion of tweets intimately equivalent to political agendas, applicants portrait, and affirmation from the media covering scope of the offensive series. [12]

Authors coated some a hundred and fifteen studies on the employment of Twitter in politics. For this discussion, studies area unit sorted in 3 topical categories: studies addressing the employment of Twitter by politicians and campaigns; studies addressing the employment of Twitter by varied publics throughout election and issue campaigns; and comments on Twitter throughout campaign events–such as televised debates, party conventions, and polling day coverage. [13]

Authors, during this work, we've studied the employment of twitter by house senate and politician candidates throughout the midterm (2010) elections within the U.S. our information is comprehensive of virtually 700 candidates and over 690k documents that they made and cited within the three.5 years resulting in the elections. we've used graph and text mining techniques to research variations between Democrats, Republicans and party candidates, and recommend a completely unique use of language modeling for estimating content cohesiveness. Our findings have shown vital variations within the usage patterns of social media, and recommend conservative candidates used this medium additional effectively, conveyancing a coherent message and maintaining a dense graph of connections.

Despite the shortage of party leadership, we discover party members show each structural and language-based cohesiveness. Finally, we tend to investigate the relation between network structure, content Associate in Nursingd election results by making a proof-of-concept model that predicts candidate conclusion with an accuracy of eighty eight.0%.[14]

Authors operate the current Irish General Election as a chronolgy for exploring the capability to represent menesterial emotions by way of excavatation of social media. Our perspective unite emotion studies using supervised learning and volume-based initiatives. We appraise in opposition to the stereotypical public opinion polls and the final election outcome. [15]

Existing work is limited due to the very fact that it focuses solely on the sentiment analysis phase and simply classifies the gathered data into classes according to specific lingual.

In this text - sentiment analysis, although a very important phase, has been looked at like the first step in preparing a machine learning model. Gathering of tweets happened.

Tweets were used so that sentiment analysis could be performed, the overall positive,negative,neutral sentiment towards the two major parties, BJP and Congress, is determined by assigning a score to the three polarities. Then a Regressi on model will be established how the positive, neutral and negative sentiment as expressed by netizens on Twitter relates to the two parties obtain. This is in contrast to existing classification centric models used as it goes beyond simply analyzing the data and predicts how the opinion of public affects the number of votes obtained.

Authors	Algorithms Used	Techniques	Dataset	Accuracy Measure	Accuracy
K. Mao, J. Niu, X. Wang, L. Wang and M. Qiu 2015	lexiconbased algorithms,IG algorithm	Combines Lexicon-based and Learn-based techniques (CLL)	product reviews(15 166), HowNet and NTUSD, Chinese Original Basic Lexicon (OBL)	F-Measure,A ccuracy for classificatio n	H(0.858),B(0.929), E(0.816)
Jyoti Ramteke, Darshan Godhia, Samarth Shah and Aadil Shaikh	Multinomial Naive Bayes and Support Vector machines	two stage framework which can be used to create a training data from the mined Twitter data	Twitter data(30 thousand tweets for the training data set).	F-1 score	MNB(nltk)-0.54 MNB(Scikit-learn)-0.97 SVM(nltk)-0.58 SVM(Scikit-learn)-0.99
D. Das and S. Bandyopad -hyay 2010	Conditional Random Field (CRF) based word level emotion classifier,SVM	Ekman's six emotion classes.	Twitter Data	kappa metric.	Classes (# Words),CRF,SVM Happy (106) 67.67 80.55, Sad (143) 63.12 78.34, Anger (70) 51.00 66.15 Disgust (65) 49.75 53.35 Fear (37) 52.46 64.78 Surprise (204) 68.23 79.37
Parul Sharma, Teng-Sheng Moh 2016	Naive Bayes, Support Vector Machine, Dictionary based	POS Tagged Ngram and central N-gram fusion takes place.	Twitter data.	Precision And Recall	Naïve Bayes(62.1%), SVM(78.4%), Dictionary based(34%)
Subhabrata Mukherjee, Pushpak Bhattachar yya 2012	Support Vector Machine	Lexicon based classification and supervised classification	Twitter dataset(850 7 tweets from 20 different domains) and Twitter dataset(15, 214 tweets based on hashtags.)	bag-of-words model	76.5%
Maite Taboada,Ju lian Brooke,Ki mberly Voll,Manfr ed Stede 2011	Support Vector Machine	Semantic orientation (SO), SO-CAL, the Semantic Orientation CALculator	Polarity Dataset	positive and negative reviews	75.85%
Adam Bermingha m & Alan Smeaton 2010	Support Vector Machine (SVM)and Multinomial Naive Bayes (MNB) classifiers	templating approach to extract positive, negative and neutral blog post content and comments using the TREC relevance judgments as labels	Twitter data	10 fold cross-validation	74.85%
N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek 2005	HindiSentiWor dNet (HSWN)	a method is proposed to increase the coverage of the Hindi Senti WordNet for better classification results.	annotated dataset	Positive and negative reviews	80.21%

Table 1: Reference Work

Authors	Algorithms Used	Techniques	Dataset	Accuracy Measure	Accuracy
O'Connor et al. 2010	Lexicon-based sentiment analysis.	Lexicon-based sentiment analysis. Aggregated results at national level. No prediction attempted.	February to November 2008. Candidate names used as keywords.	Correlation against pre-electoral polls.	No significant correlation found.
Gayo-Avello 2011	Lexicon-based sentiment analysis.	Lexicon-based sentiment analysis. Individual votes. Aggregated results at state level. Vote rates	June 1 to November 3, 2008. Presidential and vice-presidential candidate name	MAE against actual electoral results.	MAE 13.10% (uncompetitive with traditional polls).
Tumasjan et al. 2010	MAE	Number of tweets. Aggregated results at national level. Vote rates.	August 13 to September 19, 2009. Parties present in the Bundestag and politicians from those parties.	MAE against actual electoral results.	MAE 1.65% (comparable with traditional polls, although larger).
Jungherr et al. 2014	MAE	Number of tweets. Aggregated results at national level. Vote rates.	Different time windows from August 13 to September 27, 2009. Parties running for election	MAE against actual electoral results.	Unstable MAE depending on time window but larger than MAE reported by Tumasjan et al. 2010. Incorrect prediction when taking into account all parties running for election
Livne et al. 2011	Regression models	Regression models for binary results of races which included external data.	March 25, 2007 to November 1, 2010. Tweets and social graph for 700 candidates.	Winner prediction	81.5% accuracy when using external data alone. 83.8% accuracy when incorporating tweets (but not graph data). Not noticeable improvement.
Bermingha m & Smeaton, 2011	ML-based sentiment analysis.	ML-based sentiment analysis. Aggregated results at national level. Vote rates.	February 8 to 25, 2011. Major parties	MAE against actual electoral results	MAE 5.58% (uncompetitive with traditional polls).
Skoric et al., 2018	MAE	Number of tweets. Aggregated results at national level. Vote rates.	Tweets by 13,000 Singaporean political engaged netizens.	MAE against actual electoral results	MAE 5.23% Inconclusive since pre-electoral polls are banned in Singapore.

Table 2: Reference Work

III. OBJECTIVES

- Collecting group opinions
- Social media is also considered as an important part as a toolbox for political campaign.

IV. PROBLEM STATEMENT

- Not restricting it to a particular lingual.
- Emojis are taken into consideration.
- Concept expansion approach needs to be used.
- Dataset can be enhanced for generalization output.
- There are many discriminative features but Punctuation is not considered which plays a significant role, therefore we will be incorporating it.
- Multiple sources of information can be used to enhance the system.
- To perform better in every domain like electronics , books ,hotels etc.
- To overcome noise, resolving inconsistencies in data by removing '@' & URL's.

V. ACKNOWLEDGEMENT

This research was guided by Prof. Rachana Sable who provided their experience and their knowledge, that was a great resource of knowledge. We are also thankful to Prof. Jyoti Chauhan and Prof. Ganesh Kadam for their insights that guided us to improve the quality of our work.

VI. CONCLUSION

For the prediction of election results, we use social media which gives us difficulty at different stages of development of the system . Our paper, handles the issue of less data is tackled. Finally, a model is proposed which uses TextBlob .This model may not be itself suffice to predict the results ,therefore it can be combined with other tools like mathematical models ,surveying and various offline techniques.

Our proposed model's dataset will be formed by collecting the real time tweets .

Our proposed system can be automated to automatically collect data from social media sites periodically since to predict the elections we need a collection of data from which we can get access to the past history data ,present data and with the help of which we can predict the future. Therefore the collected dataset will be used for analysis purposes and can be stored for other purposes too.

The attributes of the dataset extracted from new data should be used for the analysis purposes.

Therefore we suggest to create such a model that will be able to be a participant of reinforcement learning which will be able to automate the process of mining the data and fetch only the relevant data that can play a crucial role to affect the output of the system.

REFERENCES

[1].Cross-Domain Sentiment Analysis of Product Reviews by Combining Lexicon-Based and Learn-Based Techniques, by K. Mao, J. Niu, X. Wang, L. Wang and M. Qiu (2015)

[2]. Election Result Prediction victimization Twitter sentiment Analysis by Jyoti Ramteke, Darshan Godhia, Samarth sovereign and Aadil Shaikh.(2016)

[3].Labeling feeling in Bengali journal corpus - a fine-grained tagging at the sentence Level by D. Das and S. Bandyopadhyay(2011)

[4] Prediction of Indian Election victimization Sentiment Analysis on Hindi Twitter by Parul Sharma, Teng-Sheng Moh(2010)

[5] Sentiment Analysis in Twitter with light-weight Discourse Analysis by Subhabrata Mukherjee, Pushpak Bhattacharyya(2009)

[6] Lexicon-Based ways for Sentiment Analysis by Maite Taboada, solon Rupert Brooke, Kimberly Voll, Manfred Stede(2012)

[7] Classifying Sentiment in Microblogs: Is Brevity a plus By Adam Bermingham & Alan Smeaton(2010)

[8] Sentiment Analysis of Hindi Review supported Negation and Discourse Relation by N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek(2005)

[9] From Tweets to Polls: Linking Text Sentiment to opinion statistic by Brendan Flannery O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge and Noah A. Smith(2010)

[10] "I needed to Predict Elections with Twitter and every one I got was this Lousy Paper". A Balanced Survey on Election Prediction victimization Twitter information by Daniel Gayo-Avello.(2011)

[11] Predicting Elections with Twitter: What a hundred and forty Characters Reveal regarding Political Sentiment by Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe(2011)

[12] Twitter in Politics: A Comprehensive Literature Review by Andreas Jungherr (2014)

[13] The Party is Over Here: Structure and Content within the 2010 Election by Avishay Livne , Matthew P. Simmons , Eytan Adar and Lada A. Adamic (2011).

[14] On victimization Twitter to watch Political Sentiment and Predict Election Results by Adam Bermingham and Alan F. Smeaton (2012).

[15] Predicting elections from social media: a threecountry, three-method comparative study by Kokil Jaidka, Saifuddin Ahmed, Marko Skoric & Martin David Hilbert (2018).

