A Comprehensive survey of Machine Learning Algorithms for Sentiment Analysis of Political Tweets

DNSB.Kavitha[1], Prasad Reddy P.V.G.D[2], K. Venkata Rao[3]
Department of Computer Science & Systems Engineering
Andhra University College of Engineering(A)
Andhra University, Visakhapatnam

Abstract- There is a rapid growth in the usage of social networks, such as Facebook and Twitter, Many people post their tweets related to particular topic with different opinions. Especially marketing and political campaigns leverage from opinions expressed on twitter or other social communication platforms. There is a need for processing the computations of sentiment analysis in documents. To do this many machine learning algorithms were proposed earlier like Navie Bayes classifier, Sentoword Net algorithm, Support Vector Machine to classify opinions into 'Positive',' Negative',' Neutral'. The present research work is aimed at analyzing data from social networks such as Twitter and estimate the mood of public towards various parties contesting for elections, and try to make some estimates of winning chances by taking various parameters into consideration.This paper provides a survey and a comparative study of existing methods for sentiment analysis.Limitations of Sentiment Analysis on Twitter are also discussed in this paper.

Keywords Twitter, Tweetes, Sentiment analysis, Political data

1. INTRODUCTION

There has been a rapid change in web services, internet technology,various types of social media sites such as discussion forums, micro blogs, and system to system networks provides posting online opinions/reviews about a political party or politician or product has tremendously became a popular way for sharing their opinions or thoughts about particular political party/politician or services. Twitter one of the social networking site will be the targeted social media site for this paper.

Nowadays, billions of users are using the twitter can be used as rich source of data for mining information. There is an interface for retrieving the twitter data known as twitter corpus which provides free information in the form of a stream.Analysis of this information has led to a variety of research. Examples include prediction of elections, estimation of public sentiment during elections . To assess the general public's opinions,sentiment towards politician, political parties, products and services such a user generated data can be very useful . Recently conducted surveys have been revealed that such online reviews from public has played very important role in strategies of political parties ,e-commerce companies. Lots of new technologies are coming out in market for automatic sentiment analysis and to dig out hidden knowledge from such a huge amount of user-generated data on social media. In this paper, we are dealing with sentiment analysis of political tweets of user corresponding to their favorite political party or candidate.

In this research paper,we will try to find answer for two research questions namely RQ1 and RQ2.

RQ1.Can Twitter able to find whether a person's opinion towards political party/candidate remain the same for a long time?

RQ2.If changes, what % it may change?

We will discuss these two problems in this paper. First we will present earlier research in this field, and then the techniques and models used in this area. Second , we will compare different models used so far. Finally, conclude with the findings from previous research in this area.

2. RELATED WORK

Tarek Elghazaly et al.[1] measured sentiment of people from the tweets collected during presidential elections in Egypt 2012.They used two techniques Support Vector Machine (SVM) and Naïve Bayesian (NB) to measure the accuracy and time.They observed that the Naïve Bayesian method is the highest accuracy and gives the lowest error rate with the most popular text evaluation measures (F-measure, Recall, and Precision).

Kodmelwar M.K et al.[2] developed a system called automatic sentiment analysis based on machine learning technique to integrate both large-scale data mining techniques as well as qualitative analysis. They focused on public's Twitter posts where public post their opinions/reviews about political leaders/parties to understand issues and problems that they have with them (political parties/leaders).They used multinomial naïve Bayes to train detector which will automatically detect sentiments (happy, sad, disgusting, and angry) from tweets and micro blogs.

Delenn Chin et al.[3] collected tweets related to 2016 presidential elections of presidential candidates Donald Trump, Ben Carson, Hillary Clinton, Bernie Sanders.In this project they concentrated on Emojis and classified into happy, sad, angry, laughter ,scared. They implemented three algorithms Naive Bayes, Support Vector Machine, Nearest Neighbors on each of the presidential candidate tweets and observed the accuracy levels of each algorithm i.e, SVM 49.22% accuracy , Naive Bayes 49.02% accuracy and Nearest Neighbors 45.99% .

Martin Ringsquandl et al.[4] identified that NLTK's built-in pre-processing functionalities were found to be not completely sufficient for informal text corpora. they extracted relevant aspects with a newly introduced combination of the PMI measure and phrase frequency as constraint. Aspect extraction and pruning methods presented in this paper can be applied in any domain where a meronymy relationship of opinion targets and aspects holds true. They verified this relationship for political candidates and their campaign topics. The evaluation of the PMI adjusted measure as constraint on aspect extraction shows that the meronymy relationship between politicians and their campaign holds.

Kartik Singhal et al.[5] manually downloaded tweets related to AAP and BJP of general elections from a time period of 28 February 2014 to 28 March 2014. proposed an unsupervised hybrid approach of Lexicon Based and Rule Based Sentiment Analysis which will analyze words related to other words, thus giving overall sentiment of the sentence. For lexicon, SentiWordNet is used which can give us the sentiment scores of a word. A negative score signifies negative connotation and a positive score signifies positive connotation of the word. They followed 4 steps Dependency extraction, Set Distribution, Context Rules Formation, Determining Sentiment Scores.Their results shows that 34.91% of users were positive towards AAP party in NCT Delhi region.

Bholane Savita Dattu et al.[6] compared different sentiment classification techniques on different datasets and concluded that SVM on Amazon product review data and ChnSentiCorp dataset gave 89.8% accuracy, NB Amazon product review data and ChnSentiCorp dataset gave 89.4% accuracy, NBSVM on SEMEVAL 2013 twitter dataset gave 79.4% accuracy, MNB on SEMEVAL 2013 twitter dataset gave 71.14% accuracy, SentiStrength+ TwitterSentiment model on SNAP twitter dataset gave 69.7%, TwitterSentiment model on SNAP twitter dataset gave 57.2%.

Nicholas Beauchamp[7] used a new linear regularization feature-selection method on 1,200 state-level polls during the 2012 presidential campaign with over 100 million state-located political tweets and showed that out-of-sample testing when properly modeled, the Twitter-based measures track and to some degree predict opinion polls, and can be extended to unpolled states and potentially substate regions and subday timescales.

Vishal A. Kharde and S.S. Sonawane[8] had studied various papers and identified that sentiment analysis can done at various levels of granularity like word level,sentence level,document level and feature based sentiment and concluded that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document and said that more the cleaner data, more accurate results can be obtained and Use of bigram model provides better sentiment.

Daniel Gayo-Avello[9] identified many flaws in Current Research regarding Electoral Predictions using Twitter Data like 1.There is not a commonly accepted way of "counting votes" in Twitter,2. There is not a commonly accepted way of interpreting reality. 3. All the tweets are assumed to be trustworthy,4. Demographics are neglected and recommended for future research regarding Electoral Predictions using Twitter Data like 1.There are elections virtually all the time, thus, if you are claiming you have a prediction method you should predict an election in the future,2. Clearly define which is a "vote" and provide sound and compelling arguments supporting your definition. For instance, why are you using all of the users even if they have just a few tweets on the topic? Or, why are you dropping users because they have few tweets on the topic? I know, it is hard and probably not even fair but we need to know the way votes are to be counted... 3.

International Journal of Research in Advent Technology, Vol.7, No.5, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Clearly define the golden truth you are using,4. Sentiment analysis is a core task and concluded that the predictive power of Twitter regarding elections has been greatly exaggerated.

Brendan O'Connor et al.[10] retrieved messages with the keywords economy,job,jobs for consumer confidence,obama for presidential approval,obama and mccain for elections in the year 2008 and implemented correlation analysis to check whether text sentiment is a leading indicator of polls and concluded that time-intensive polling can be supplemented with the text data that is generated from online social networking sites like twitter.

Andranik Tumasjan et al.[11] have examined 104,003 political tweets, which were published on Twitter's public message board between 13th August and 19th September 2009, prior to the German national election. They collected all tweets that contained the names of the 6 parties represented in the German parliament (CDU/CSU, SPD, FDP, B90/Die Grünen, and Die Linke) and tried to answer three questions 1)Does Twitter provide a platform for political deliberation online? 2) How accurately can Twitter inform us about the electorate's political sentiment? 3) Can Twitter serve as a predictor of the election result ?and concluded that Twitter can be considered as a valid indicator of political opinion .

Daniel Gayo-Avello[12] has identified that there are number of characteristics and sub-characteristics which defines method to predict electoral results from Twitter, namely: Period and method of collection, Data cleansing measures, Prediction method and its nature, Performance evaluation .

Paltoglou et al.[13] had proposed an unsupervised, lexicon-based algorithm which estimates the level of emotional strength in text in order to make a final prediction on three data sets namely Digg,Myspace and Twitter and tested on two complimentary tasks: subjectivity detection and polarity classification.The advantages of the approach is that it requires no training and thus can be readily applied into a wide selection of environments.

Paltoglou et al.[14] proposed an algorithm, SentiStrength, to extract sentiment strength from informal English text, using new methods to exploit the de-facto grammars and spelling styles of cyberspace. They applied this algorithm on MySpace comments and identified that SentiStrength is able to predict positive emotion with 60.6% accuracy and

negative emotion with 72.8% accuracy, both based upon strength scales of 1-5.

Agus Zainal Arifin et al.[15] proposed a method NonNegative Matrix Factorization(NMF) for analyzing the link between features.They applied this on dataset of 764 tweets having 5 emotions i.e. happy (senang), angry (marah), fear (takut), sad (sedih), and surprise(terkejut) collected from twitter using Indonesian language and computed users emotion by kNearest Neighbor(kNN) approach.

Bo Pang et al.[16] employed three machine learning algorithms namely Naive Bayes, maximum entropy classification, and support vector machines on movie dataset to classify the document not by topic but by overall sentiment and concluded that factors that make the sentiment classification problem are more challenging.

Adam Bermingham et al.[17] had proposed an approach which combines sentiment analysis using supervised learning and volume-based measures and found that social analytics using both volume-based measures and sentiment analysis are predictive.They have implemented this method on Irish General Election, 2011 dataset.

Ramteke, J. et al.[18] had proposed a two stage framework which can be used to create a training data from the mined Twitter data without compromising on features and contextual relevance and then proposed a scalable machine learning model to predict the election results. They collected Twitter data for two candidates – namely Donald Trump and Hillary Clinton for the dates March 16th, 2016 and March 17th, 2016.

Manish Gaurav et al.[19] had proposed methods to augment the counts by counting not only the presence of candidate's official names but also their aliases and commonly appearing names. In addition, they devised a technique to include relevant and filter irrelevant tweets based on predefined set of keywords and then compute the probability of a candidate to win in two different ways 1)Moving Average Aggregate Probability (MAPP) 2) Moving Average Aggregate Probability Using Counts (MAPC) .This approach is successful in predicting the winner of all three presidential elections held in Latin America during the months of February through April, 2013.

Akshat Bakliwal et al.[20] had performed a series of 3-class sentiment classification experiments on Irish General Elections in February 2011.They omitted

International Journal of Research in Advent Technology, Vol.7, No.5, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

sarcastic tweets and achieved an accuracy of 61.6% using supervised learning and a feature set consisting of subjectivity-lexicon-based scores, Twitterspecific features and the top 1,000 most discriminative words.

Shadi Abudalfa and Moataz Ahmed[21] had presented the comparative study between target-dependent and open domain targeted sentiment classification on same dataset with different supervised learning algorithms and concluded that the accuracy for target-dependent sentiment classification is 71.5% and open domain targeted sentiment classification is 44.13%.

Xiaodong Zhang[22] had implemented model related to lpa(likes for pages) and lpo(likes for posts) on Taiwan's 2016 general elections and achieved an accuracy of 80% in prediction of election results.

Campanale, M et al.[23] had proposed Navie Bayes Multinominal Text classification algorithm to conform that the twitter is massively used for sentiment analysis of predicting the election results.They collected and processed 1,200,000 tweets in the context of the 2016 Italian Constitutional Referendum.

Abdul Fatir Ansari et al.[24] had developed a project on twitter sentiment analysis with ensemble model classifier which is a meta learning algorithm that combines different classifiers to improve the prediction accuracy.

3. FINDINGS OF THE LITERATURE SURVEY

From the above literature review,it is concluded that most researchers used supervised learning algorithms like SVM,Navie Bayes,Maximum Entropy.With respect to RQ1 and RQ2 we found that analyzing the political tweets in a short period may not answer these questions but not only considering the text tweets we also need to observe how many twitter accountants are following the video links of campaigning of the politicians which can change the opinion.The analysis concludes the following points:

- The researchers have applied mainly Navie Bayes and Support Vector Machine(SVM) on political tweets collected during presidential election 2016 and found that Navie Bayes gives more accuracy than SVM.
- For enhancing the accuracy of sentiment classification text evaluation measures F-measure, Recall, and Precision are to be used.
- Not only the tweets in English but other language tweets sentiment can be analyzed with two techniques Dictionary based Lexicon approach and Corpus based Lexicon approach.
- Majority of the techniques used political datasets such as German election 2009 US presidential election 2016.
- Recent researchers have applied ensemble classifiers method to achieve the sentiment classification accuracy of 83.58% on Kaggle public learderboard.
- The sentiment analysis with unigram gives less accuracy.so we can improve the accuracy by considering bigrams,trigrams.
- Credibility is the major concern to be considered.

Limitations

As data were limited to the tweets which contain the names of parties and politicians that are considered as search terms, they may have missed some replies belonging to a discussion thread because those who respond do not always repeat these names in every message. So for the analysis many tweets may be missing.

As many of the techniques use only labeled data they suffer from scarcity of data.

4. CONCLUSION

We have studied the Machine learning sentiment classification algorithms.This study included single,ensemble classifiers.As there are huge number of features in a dataset which are redundant and irrelevant it essential to employ feature reduction technique to improve the training dataset.It is there by recommended to use Extreme Learning Machine based Classifier for an improvement of sentiment analysis.

REFERENCES

[1] Tarek Elghazaly, Amal Mahmoud, Hesham A. Hefny(2016) "Political Sentiment Analysis Using Twitter Data." ACM. ISBN978

[2] AmrutaTarlekar, Kodmelwar M.K "Sentiment Analysis of Twitter Data from PoliticalDomainUsingMachineLearningTechniques. "IJIRCCE Vol. 3, Issue 6, June 2015, ISSN : 2320-9798

[3] Delenn Chin, Anna Zappone, Jessica Zhao,"Analyzing Twitter Sentiment of the 2016 Presidential Candidates."

[4] Martin Ringsquandl and Dušan Petković "Analyzing Political Sentiment on Twitter."2013 AAAI Spring Symposium

[5] Kartik Singhal , Basant Agrawal and Namita Mittal "Modeling Indian general elections: Sentiment Analysis of Political twitter data." Springer India

2015 J.K. Mandal et al. (eds.), Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing 339

[6] Bholane Savita Dattu, Prof.Deipali V. Gore "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques."(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 5358-5362

[7] Nicholas Beauchamp, "Predicting and Interpolating State-Level Polls Using Twitter Textual Data."American Journal of Political Science, Vol. 00, No. 0, xxxx 2016, Pp. 1–14

[8] Vishal A. Kharde and S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques." International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[9] Daniel Gayo-Avello, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" journal arXiv preprint arXiv:1204.644,2012.

[10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series."proceedings of the fourth international AAAI conference on web logs and social media 2012.

[11] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe,"Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment ."proceedings of the fourth international AAAI conference on web logs and social media 2010.

[12] Daniel Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from Twitter data." article in Social Science Computer Review , Vol 31, Issue 6, 2013.

[13] Paltoglou, G. and Thelwall, M. 2012."Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media."ACM Trans. Intell. Syst. Technol. 3, 4, Article 66 (September 2012), 19 pages

[14] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). "Sentiment strength detection in short informal text."Journal of the American Society for Information Science and Technology, 61(12), 2544–2558

[15] Agus Zainal Arifin, Yuita Arum Sari,Evy Kamilah Ratnasari".Emotion Detection of Tweets in Indonesian Language using Non-Negative Matrix Factorizatio." I.J. Intelligent Systems and Applications, 2014, 09, 54-61 Published Online August 2014 in MECS (http://www.mecs-press.org/)

[16] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan,"Thumbs up? Sentiment Classification using Machine Learning Techniques" Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[17] Adam Bermingham and Alan F. Smeaton, "On Using Twitter to Monitor Political Sentiment and Predict Election Results."

[18] Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). "Election result prediction using Twitter sentimentanalysis."2016 International Conference on Inventive Computation Technologies (ICICT).

[19] Manish Gaurav, Anoop Kumar, Amit Srivastava, Scott Miller, "Leveraging Candidate Popularity On Twitter To Predict Election Outcome." 2013 ACM 978-1-4503-2330-7 .

[20] Akshat Bakliwal , Jennifer Foster, Jennifer van der Puil , Ron O'Brien , Lamia Tounsi and Mark Hughes,"Sentiment Analysis of Political Tweets: Towards an Accurate Classifier." Proceedings of the Workshop on Language in Social Media (LASM 2013), pages 49–58, Atlanta, Georgia, June 13 2013. c 2013 Association for Computational Linguistics.

[21] Shadi Abudalfa and Moataz Ahmed,"Comparitive Study on Efficiency of Using Supervised Learning Techniques for Target-Dependent Sentiment Polarity Classification in Social Media.:International Journal of Computing and Digital Systems ISSN(2210-142X)

[22] Xiaodong Zhang ,"Social media popularity and election results: A study of the 2016 Taiwanese general Election." PLOS ONE | https://doi.org/10.1371/journal.pone.0208190 November 28, 2018

[23] Campanale, M., & Caldarola, E. G.,Revealing," Political Sentiment with Twitter: The Case Study of the 2016 Italian Constitutional Referendum." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[24] Abdul Fatir Ansari,Abinava Seenivasan,Anusha Anandan, "Twitter Sentiment Analysis."a course project developed in GitHub.