International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

A Survey on Political Bias Recognition using Tweets

Nayana A D[1*], Dr. Ananda Babu J[2], Ananya Anand[3], Kruthika M V[4], Raksha R[5]
BE student, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, India[1]
Associate Professor, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, India[2]
BE student, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, India[3]
BE student, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, India[4]
BE student, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, India[5]

Abstract
As, each country's augmentation banks on political leaders, it is incumbent to acknowledge each individual regarding politics. Every political enthusiast intends to share their opinions on a particular political party or a politician. In regard to this, Twitter plays a favourable platform. Hinging on American Trends Panel survey, it is said that politically inclined people are more active on twitter. It allows to personalize of the opinions on numerous issues related to politics. For each user, their tweets are collected and entities related to politics are extracted. This paper focuses on political bias recognition of tweets. Here, the main concentration is on NLP, Decision tree and KNN methods.

Keywords
Political party, political enthusiast, tweets, NLP, Decision tree, KNN.

1.Introduction
In this divided world, social media plays a crucial role in connecting people since everyone is hooked towards it. Social media facilitates delivering of information within no time and also conveying ideas and thoughts among people.

People feel more convenient to express their opinions and views on different political leaders or political parties in whole through social media. Here, the main focus is on twitter since it has become a popular platform for micro-blogging in which users express their opinions on political subjects in real-time. Twitter has also got a better searching algorithm compared to any other social media, which helps to retrieve data [1].

Coming up with one more advantage, tweets once posted cannot be edited, which does not affect the polarity extracted from the data.

During elections, it also helps the different political parties to prepare the intelligence report and to determine the deserving candidate based on the tweets collected.

Political bias recognition through twitter replaces the conventional method in which the news reporters had to visit the public to get their reviews on political issues during elections. Coming up with the better way, in which the news broadcasters can give the probability of different political parties using sentiment analysis technique. Methodologies that could be used to identify political tendency are Natural Language Processing (NLP), Decision Tree, KNN, Support Vector Machine (SVM), Naïve Bayes etc.

The tweets of the sole users are collected and entities are extracted related to political issues. Data are pre-processed where unwanted data are discarded through normalization. In the next step the feature extraction is done and then the data is classified into positive, negative and neutral. On this basis the probability table is built. Therefore, Sentiment Analysis (SA) technique plays a vital role as it deals with classification and interpretation of emotions.

2.Related works
There will be always handful of approaches for accomplishing a task. Sentiment diffusion models were used in Pakistani and Indian elections to forecast the eventual winner of these elections [2]. The Spanish election from 2011 to 2012 is used to portray that the experiment can be benefited in campaigns to track candidate allegiance and predict voting results [3]. Barack Obama was predicted as the winner in 2012 Presedential election, US through census twitter mode [4]. In relate to this, we have come across copious methodologies for identifying political tendency. Here, NLP, Decision Tree and KNN methods are conferred.

3.Methodologies
3.1Natural Language Processing:
Natural Language Processing (NLP) involves intelligent analysis of written language.

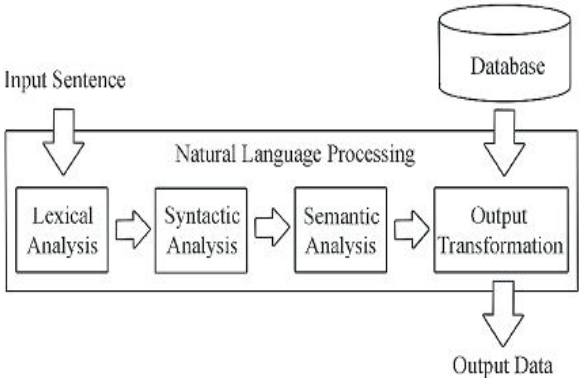

Figure 1 Stages in NLP

Figure 1 shows different stages that are comprised in NLP:

- Lexical Analysis involves splitting up input sentence into tokens.
- For the purpose of validity checking, tokens are parsed with pre-defined syntax in Syntactic Analysis process.
- Semantic Analysis is concerned with ensuring sentence exactness.
- The entailed result can be retrieved from the previous output in the Output Transformation process.

The factors that hinders the efficiency of NLP include,

- A tweet can accommodate only 140 characters [5] with no restriction on the form and content thus, the researches of NLP have to face an arduous task.
- Not all tweets will have an aptness in vocabulary. In such cases NLP will decline or breakdown.
- Human/system errors will also interrupt in getting the requisite result [6].

3.2Decision Tree Analysis:

Decision tree is an archetypal classification and disclosure algorithm in supervised machine learning which can be practised to categorise or segregate the given data, standing on the preceding knowledge of the training data.

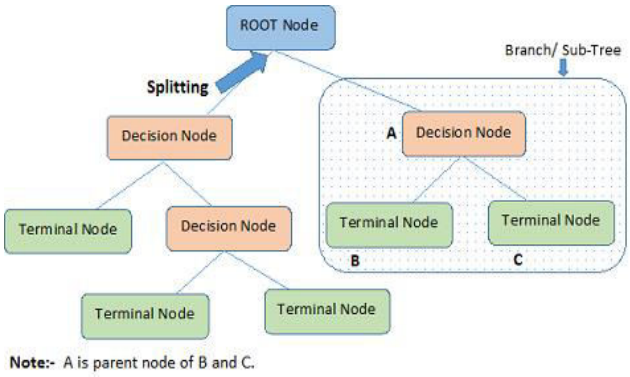

Figure 2 Decision tree working

Decision tree dawns from the root node. Further, the root node branches into possible fallouts called decision nodes. Each of those fallouts points to further nodes, which diverges into alternative possibilities. The nodes which do not have the further possibilities are called terminal nodes. The conclusion is made based on the final outcome of the tree.

Evaluating the deliberate effects of the decision tree, we have come up with few of the cons:

- There will be colossal change in the structure of decision tree even if there is a small change in the data. So, it can be said that decision trees are unstable.
- If there are many unsettled values, calculations will become compounded.
- While modelling decision trees, one of the major difficulties faced is over-fitting. It performs well with training data set, but fails to be accurate while predicting samples that are not part of the training set [7].

3.3K-Nearest Neighbours:

KNN is a supervised learning algorithm used for regression and classification of the data. To classify new information, it uses previously known data set.

To ameliorate this algorithm, weights are allotted to each of the k-points according to their distance from the test point. Another noticeable point here is that the dataset size influences the value of k [8].

Working of KNN algorithm:
Feature similarity is used by k-nearest neighbours algorithm for predicting values of new data points [9]. Here are few steps to give brief explanation regarding the working of KNN algorithm.

- Both training data and test data are loaded.
- Choosing nearest data points (value of k) is essential.

- Euclidean, Manhattan or Hamming distance between test data and each row of training data, which are then organized in sorted format.
- Top k rows from the sorted area will be chosen and then test point is assigned with a class.

Figure 3 KNN Classifiers

There are few merits and demerits of KNN, but the reason why we are only concentrating on demerits here is, the accuracy factor is more dependent on or is affected by the disadvantages of the methods that are being used. So, choosing the method which performs better to be more accurate is important.

- The main snag of KNN algorithm is that it is slack learner, as it simply uses training data itself for sorting rather than learning from it [11].
- After each prediction it must quantify the distance and sort all the training data. This process may lag if there are a large number of training examples.
- Storage of memory for this algorithm is required more than other supervised learning algorithm [12].
- KNN algorithm doesn't work well with high dimensional data with vast number of dimensions, it becomes arduous for the algorithm to calculate the distance in each dimension.
- Dataset must be standardized and normalized before applying KNN algorithm. If this feature scaling is not done then the algorithm may result with wrong predictions [13].

4. Real life circumstances

People or more specifically, the political enthusiasts always prefer social media for expressing their feelings rather than a poll conducted through any other mass media. They feel that they will be identified uniquely by their social media profiles and so is the reason analysts use social media as a tool for their works. Accordingly, Twitter is more preferable for analysts to get accurate prediction conclusions about election or any other issues.

Table 1 US Presidential Election prediction using tweets, 2017

Accuracy	Place	Candidates		
		Donald Trump	Hillary Clinton	Neutral Tweets
Analysis Prediction	Florida	58.26%	35.4%	6.44%
Actual Result		49.1%	47.8%	3.1%
Analysis Prediction	Ohio	56.2%	35.75%	8.05%
Actual Result		52.1%	43.5%	4.6%
Analysis Prediction	N.Carolina	53.26%	38.05%	8.69%
Actual Result		50.5%	46.7%	2.8%

Figure 4 CNBC report regarding US Presidential election, 2017

As it is evident in table 1, the sentiment analysis predictions were so close to the actual results. Though mass media reported that Hillary Clinton would win the elections (figure 4), Donald Trump has won the elections.

In relate to this, Sentiment Analysis results were so accurate because, users of twitter used to mention Donald Trump whenever they posted a tweet regarding the elections. According to the survey [10], lots of tweets were retrieved about Donald Trump than Hillary Clinton, while request was made with the Search API using certain hashtags to retrieve tweets.

Twitter has variety of opinions to download data which makes it easy for analysts to convene information for their works.

4.1Comparison of Methodologies:

Table 2 Implementation of methodologies on political issues.

Methodologies	Year	Implementation	Accuracy	Dataset
NLP	2011 2012 2015	Predicting percentage of voters who will vote for certain candidate	95% 80% 98%	7 million tweets
	2017	Political Bias Prediction	70%	1) 12 online publishers 2) 103k articles
Decision tree	2019	Twitter based outcome prediction of 2019 Indian General Elections using Decision tree	97%	5000 tweets
	2014	India Lok Sabha elections	77.14%	Attributes such as name of the state, total voters, total percentage of turnout etc
KNN	2014	Indian Lok Sabha elections	48.57%	Attributes such as name of the state, total voters, total percentage of turnout etc

Table 2 shows the accuracy of methodologies that are discussed in this paper, when implemented on political issues. It is noticeable that NLP does not maintain constant accuracy.
Though decision tree gives best accuracy when compared to other two methods (NLP and KNN), it is not user friendly, as we have discussed about it while mentioning its demerits.

It is evident that accuracy obtained will be poor if KNN method is implemented. So, choosing the best among the rest is important.

5. Conclusion

After surveying on various methodologies like Natural Language processing, Decision Tree Analysis and K-Nearest Neighbor, it can be concluded that Machine Learning is the powerful technique to identify the political tendency. Besides,
According to our knowledge the most appropriate method on recognizing the political bias on twitter can be done using Naïve Bayes Classifier.

Finally, our heed is on using Naïve bayes method. On this point, we are going to implement a system that recognizes the political bias of the various tweets fetched by the twitter users.

Acknowledgment

None.

References

[1] "Is Twitter the best social media in the world?" . https://www.quora.com/Is-Twitter-the-best-social-media-in-the-world. Accessed 09 March 2020.

[2] Vadim Kagan and Andrew Stevens, Sentimetrix V.S. Subrahmanian, University of Maryland, "Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election", Published by the IEEE Computer Society, 2015.

[3] A. Hernandez-Suarez, G. Sanchez-Perez, V. Martinez-Hernandez, H. Perez-Meana, K. Toscano-Medina, M. Nakano and V. Sanchez, "Predicting Political Mood Tendencies based on Twitter Data", Published by IEEE in 2017 5th International Workshop on Biometrics and Forensics(IWBF).

[4] Murphy Choy, "US Presidential Election 2012 Prediction using Census Corrected Twitter Model", Submitted on 5 Nov 2012.

[5] Ferran Pla and Lluis-F. Hurtado, "Political Tendency Identification in Twitter using Sentiment Analysis Techniques", Proceedings of COLING 2014, the 25th International
Conference on Computational Linguistics: Technical Papers, pages 183–192, Dublin, Ireland, August 23-29 2014.

[6] Chalermpol Tapsai, Phayung Meesad and Choochart Haruechaiyasak, "Thai Language Segmention by Automatic Ranking Trie", The 9th International Conference Autonomous Systems, At Cala Millor, Spain, October 2016.

[7] "What is a Decision Tree? How does it Work?". https://clearpredictions.com/Home/DecisionTree. Accessed 13 March 2020.

[8] Andre Faria ,"Predict political orientation of Twitter users", Instituto Superior Tecnico .

[9] Onur Varol, Emilio Ferrara, Filippo Menczer, Alessandro Flammini, "Early Detection of Promoted Campaigns on Social Media", Published on 05 July 2017.

[10] Lazaros Oikonomou, Christos Tjortjis, "A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter", International Hellenic University, Thermi, Greece.

[11] Naresh E, Vijaya Kumar B. P and Niranjanamurthy M, "Challenges and Issues in Test Process Management", Journal of Computational and Theoretical Nanoscience, Volume-16(9), PP-3744–3747, September 2019.

[12] Naresh E, Vijaya Kumar B. PandMadhuri D Naik "Survey on test generation using machine learning technique", International Journal of Recent Technology and Engineering, Volume-7(6S), March 2019.

[13] Naresh E, Vijaya Kumar B. P and Sahana P Shankar "Comparative Analysis of the Various Data Mining Techniques for Defect Prediction using the NASA MDP Datasets for Better Quality of the Software Product" Advances in Computational Sciences and Technology, Volume 10, Number 7, Research India Publications, June-2017.