

**MIT WORLD PEACE UNIVERSITY**

**School of Computer Engineering and Technology**

B.Tech. Capstone Project - End Term T12 Project Phase II

Group No.: B-81

**Predicting Election Outcomes Using Social Media Data**

Presented By

**Shreyas Khairnar - 1032180112**

**Gopal Joshi - 1032180211**

**Dhruv Kaliraman - 1032180760**

**Abhishek Raj - 1032180748**

**Under the Internal Guidance of**

Prof. Chitrakant Banchhor

**School of Computer Engineering and Technology**

**MIT World Peace University, Kothrud,**

**Pune 411038, Maharashtra,India,**

**Academic year:2021-22**



**SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY**

C E R T I F I C A T E

This is to certify that,

**Shreyas Khairnar - 1032180112**

**Gopal Joshi - 1032180211**

**Dhruv Kaliraman - 1032180760**

**Abhishek Raj - 1032180748**

of BTech. (Computer Science & Engineering) have completed their project titled **Predicting Election Outcomes Using Social Media Data** and have submitted this Capstone Project Report towards the fulfillment of the requirement for the Degree-Bachelor of Computer Science & Engineering (BTech-CSE) for the academic year 2021-2022.

**Chitrakant Banchhor Dr. Vrushali Kulkarni**

Project Guide Head of School

School of CET School of CET

MIT World Peace University, Pune MIT World Peace University, Pune

Internal Examiner: Prof. Chitrakant Banchhor

**Date: June 3, 2022**

**Acknowledgments**

This paper and the research behind it would not have been possible without the exceptional support of our guide, Prof. Chitrakant Banchorr Sir as well as external examiners Prof. Siddhivinayak Kulkarni, Prof. Pradnya Kulkarni, Prof. Sajida Shikalgar and Prof. Renuka Suryavanshi. Their enthusiasm, knowledge, and exacting attention to detail have been an inspiration and kept my work on track from my first encounter with the abstract writing to the final draft of this paper. Chirantan Joshi and Karan Patil, our colleagues at University, have also looked over our transcriptions and answered numerous questions about the project with unfailing patience. The generosity and expertise of one and all have improved this study in innumerable ways, those mistakes that inevitably remain are entirely our own responsibility.

Shreyas R. Khairnar

Gopal Joshi

Dhruv Kaliraman

Abhishek Raj

**Abstract**

After the Inception of Social Media Platforms in the early 2000s, they saw an exponential rise in popularity due to the ease of sharing opinions and connecting with close ones and networking with people all over the globe. As a result of a huge user base, Social Media Platforms generate a large amount of raw data which can be used to analyze the mood of people. With the increase in popularity of Social Media Platforms in the late 2000s, there’s been a massive transformation in the ways the electoral campaigns are managed as well as the way in which politicians use to connect with their voter base. A huge amount of Real-Time Data is been generated through platforms like Facebook, Reddit, Instagram, and Twitter which can be processed and fed to Natural Language Processing(NLP) algorithms which can generate interesting and useful patterns which can be used by election strategists as well as to predict the outcomes of elections with help of various Machine Learning techniques. Though biased tweeting in favor of certain political parties for incentives results in sample error in collected data and may skew the models. In our research, the main aim is to perform Bias Detection and Sentiment Analysis over real-time data (including regional language data) to predict which party will win the state or national election held in India. In our work we get the data from Twitter where the citizens give their opinion about the political parties and the analysis of these sentiments is done after bias detection to conclude the result.

**List of Figures**

| 5.1 | High-Level Diagram …………………………………………. 16 |
| --- | --- |
| 5.2 | Block Diagram ……………………………………………….. 16 |
| 5.3 | Project Flow ………………………………………………….. 17 |
| 5.4 | Data Cleaning Flow ………………………………………….. 18 |
| 6.1 | Project Plan(Timeline Chart for entire SDLC) ………………. 19 |
| 7.1 | Data Preprocessing …………………………………………… 21 |
| 7.2 | Bias Detection ………………………………………………... 22 |

**List of Tables**

| 41. | Requirement Rationale | 14 |
| --- | --- | --- |
| 4.2 | Risk factors | 15 |
| 8.1 | Volume Share(Punjab Assembly Elections 2022) | 24 |
| 8.2 | Percentage Share per party(Punjab Assembly Elections 2022) | 24 |
| 8.3 | Percentage share per Volume(Punjab Assembly Elections 2022) | 25 |
| 8.4 | Results(Punjab Assembly Elections 2022) | 25 |
| 8.5 | Volume Share(Delhi Assembly Elections 2020) | 25 |
| 8.6 | Volume Share(Delhi Assembly Elections 2020)(Similarity Threshold=0.65) | 25 |
| 8.7 | Percentage Share per party(Delhi Assembly Elections 2020) | 26 |
| 8.8 | Percentage share per Total Volume(Delhi Assembly Elections 2020) | 26 |
| 8.9 | Results(Delhi Assembly Elections 2020) | 26 |

**Contents**

Abstract ……………………………………………………...……………4 List of Figures …………………………………………………………… 5 List of Tables …………………………………………………………….. 6

| 1 | Introduction ……………………………………………………………. | 8 |
| --- | --- | --- |
| 2 | Literature Survey ……………………………………………………… | 9 |
| 3 | Problem Statement ……………………………………………………. | 12 |
| 4 | Project Requirements ………………………………………………….. | 14 |
| 5 | System Analysis/Proposed Architecture ………………………………. | 16 |
| 6 | Project Plan ……………………………………………………………. | 19 |
| 7 | Implementation ……………………………………………………….. | 20 |
| 8 | Result and Analysis ………………………………………………….. | 23 |
| 9 | Applications ………………………………………………………….. | 26 |
| 10 | Conclusion and Future Scope ………………………………………… | 27 |

**1. Introduction**

**Project Statement:** Predicting Election Results with the use of Social Media Data and Natural Language Processing Techniques

**Area:**

● Natural Language Processing

● Social Media Analytics

**Project Introduction and Aim:**

For a long time, Traditional methodologies have been carried out for analyzing the mood of the people and predicting Elections based on the mood. These traditional methodologies consist of mailing people for input, person-to-person interviews, or telephonic interviews. Despite all the work put in, much of the time these predictions are mostly inaccurate and no better than random guesses due to sampling errors. Also, these traditional methods prove to be very expensive and time-consuming. The past decade observed a widespread acceptance of Social Media Platforms like Twitter, Instagram, Facebook, Telegram, Reddit, etc which provide users with the ability to freely voice their opinions, share their feelings with other people, and in the process create huge amounts of raw data. Twitter is one of the most used platforms where people can freely share their opinions. It is possible to extract this data using Twitter API and many researchers in this space take advantage of this easy-to-acquire data. In Twitter, it is important to extract keywords and useful features and give them appropriate weights as well as identify biases. Using Natural Language Processing techniques and Data Mining, it is possible to extract the data shared by politicians or normal users and process it in such a way that it is possible to measure the extent of popularity of the candidate/ political party in real-time as well as in excessively cheaper ways as compared to traditional methodologies.

Here, we aim to develop a robust and generalizable model which will take Twitter data as input, perform NLP tasks with the latest state-of-the-art techniques like BERT, introduce a bias detection function to get a good sample, and will showcase the popularity/vote share of a particular party as a result. This model can be used by researchers for further research in the space which will act as a base for comparing performance as well as any normal person to analyze the trends not only in elections but also in other domains like stock markets, etc. just by tweaking the data scraping strategy.

**2. Literature Survey**

During the presidential election in the United States, a paper was published and the purpose of the paper is to predict Accurate election results and interesting data trends. The paper said that the analysis of social networks and sentiment Analytics can be used to predict political as well as economic situations and future changes. It succesfully predicted 2016 presidential elections when all exit pools were unable to do so.

Another analysis in 2020, the author categorized the tweets into positive and negative tweets that supported the sentiment values. Author additionally premeditated to plot the foremost frequent word to match the recognition of every of the candidates. Pie-chart was made to support the geo-locations of the users by considering the share of tweets that belong to a number of the famous states of Bharat. It absolutely was found that a large share of tweets were from Bombay, the national capital and Bangalore. The author additionally calculated the retweet frequency distribution and found out that BJP tweets are far more frequent as compared to Congress. The number of BJP tweets were somewhere between 100–250 in contrast to Congress whose average retweet frequency was 20–30.

In 2012 before the French presidential election, the authors published a paper entitled "For Tweet Mining". French presidential election ". Its purpose was not only to predict election results, but also analyzing how they differ trends and voting scenarios are affecting the masses of various social media survey methods. In addition, the importance of the intensity of opinion introduced to provide an opportunity to draw results shows how honest and dedicated voters are. As an author of political claims, social network analysis and text mining the purpose is an area that presents many challenges, which may be a convenient and accurate way to predict both future political and economic trends.

In 2013, Mahmood et al., published an article on “Mining Twitter big data to predict the winner of Pakistani elections 2013”. Their purpose was to fetch tweets, preprocess them, store them in the database and try to draw conclusions about the winners of the elections. In the end, PMLN is the winner. The authors used three different approaches to complete the implementation task. The three models are decision tree CHAID, Naïve Bayes model and support vector machine (SVM). Their analysis shows that PTI will win, but the winner is different.

In this paper, the authors used the multi-cognitive layer model of the neural network to predict election results. They created a dataset based on the questions that are then used to train the model. Though the conclusions of the models are not promising and share a little doubt about predictions of results.

The authors explore the role of twitter in the political domain. For this, they analyze tweets using LIWC text analysis software. They analyze over a million tweets using political parties or the names of politicians as keys. Findings show a lot of political use of Twitter

Domain and had over-fitting in data. Paper indicates that the information flow can be arranged in the original two types are offline information flow and the second is online information flow and speciality types of algorithms used in information fetching and add their poor services, for example: Regular examples, Naive Bayes, Decision Trees, etc. The creator tries to give the data that this calculation used for offline mining is not appropriate to exploit information online on the basis that in the online world, information changes rapidly as indicated by time to explore online information flow is a test mission these days.

In this paper, unlike earlier works six different models with different analytical approaches are be done. Furthermore, the Harvard psychologist LoughranMcDonald Dictionary of Financial Emotions used to construct the space of emotions. The results show that (1) models with better sentiment analysis than bag of words model in

independent validators and testers;

(2) model which used the poles of sensation cannot provide useful prediction; (3) There is a slight difference between models used for two different sentiment dictionaries.

Authors in this paper studied whether an outcome similar to the results of traditional research methods can be achieved by using different kinds of methods or tools, such as word processing. For this purpose, they developed a system to predict the general election of Ireland 2011. For a review of these, MEA methods were selected. Furthermore, for sentiment analysis, tweets were collected and Manually annotated by 9 annotators. Result found that the MNB classifier obtained the best classification accuracy.

In this paper authors analyzed the predictive power of Twitter data and used multiple races for US Congress in the 2010 election. In this study, two different methods were used to predict elections for multiple Senate races. The first of these is called Twitter Volume. In this method, Tweets posted by candidates were counted and the final tally for each candidate was divided by the total. Another method was implemented called Sentiment Analysis. Based on this method, each tweet was predicted to be positive, negative and neutral by recognition of words. Therefore, according to this paper, the results of prediction with Twitter data usage have no better chance in this electoral estimate.

In this paper, the authors attempted to take an analysis of friends and supporters of each candidate of the constituency in social networks. However, most of the candidate profiles on social networks are incomplete networks. To analyze the relationship between the number of people supporting each candidate on social networks network for a specific date and the result of the election,so majorly two regression model had been proposed: a linear ordinary least squares model vote division model and logistic regression model with election the result is the dependent variable. However, their results were not accurately predicted on social media Data.

**3. Problem Statement**

**Project Scope:** We aim to create a generalized and robust model for predicting election outcomes. For this, we use Natural Language Processing techniques like Weighed Sentiment with volume as weight, Similarity Detection & Machine Translation. We also aim to take the help of the latest state-of-the-art models like BERT for Sentiment Analysis and techniques like Levenshtein Distance for bias detection.

**Project Assumptions:**

● There is a research gap in terms of the accuracy of elections being predicted with the current tools.

● Taking into account regional language data will help in getting more accurate results. ● Removal of bias will increase the model's robustness.

● Weighing the data with help of volume will help in getting better results.

● Using the latest state-of-the-art techniques like the BERT model will identify sentiments more accurately than the traditional popular model like the sentiment.Vader, etc. due to the ability to identify the context.

**Project Limitations:**

● Data typed in English but semantically non-English cannot be dealt with efficiently.

● Elections don’t happen frequently and thus it is hard to take into account as many elections as possible to validate the model.

● There needs to be an approach to identify data from news channels which adds a lot of noise to data that skews the model and is not useful in analysis.

**Project Objectives:**

● To develop a robust and generalizable, end-to-end model from scraping data to translation, bias detection, sentiment analysis and finally predicting the election results.

● Regional Language Data: Data in regional languages was not considered by existing

studies. Thus, it may create a Sampling bias by only considering the proportion of the English-speaking population. In multilingual countries like India, it is very important to

consider Regional Language Data. Hence we use regional tweets as well.

● These days probably every major Political party has what is called ‘IT Cell’, which is a massive group of people getting paid for Biased Posts in favor of a particular party. There are no studies where any technique is used to handle or not consider posts from IT Cell accounts which may cause huge deviations in predictions. We try to get past this by removing such biased posts.

● Almost all the studies validated models only on Single Election which is not enough to reach any conclusion and the model can’t be generalized to other elections. We aim to generalize this by creating a robust model.

**4. Project Requirements**

**Reusable Components:**

● Data Collection Pipeline

● Data Translation Pipeline

● Data Cleaning Pipeline

● Similarity Detection Pipeline

● Sentiment Analysis Pipeline

**Requirement Rationale:**

| **Requirements** | **Rationale** |
| --- | --- |
| Sentiment Analysis | Sentiment extracted from tweets can be a good measure for predicting the election outcome. |
| Sentiment Magnitude | Sentiment magnitude can be a good indicator in terms of how strongly the public feels about the party. |
| Volume Analysis | Volume signifies how popular a political party is on that current timeline. |

Table 4.1 Requirement Rationale

**S/W Requirements:**

● Python Libraries like Pandas, and Numpy for Data Preprocessing.

● NLTK Library for Natural Language Processing tasks & Sentiment Analysis. ● Visualization libraries for analyzing data

● APIs/Libraries/Tools for Data Collection and Data Translation.

**H/W Requirements:**

● No Special H/W Requirements

○ For better performance minimum of 8GB RAM is necessary

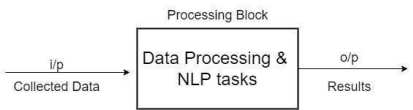
**Risk Factors:**

| **High**  ● Changes in Data Collection API limitations  ● Changes in Twitter data collection  rules |
| --- |
| **Medium**  ● Changes in Data Translation API limitations |
| **Low**  ● Very less data availability for some elections which can only be found out  after scraping the data which results in  time wastage |

Table 4.2 Risk Factors

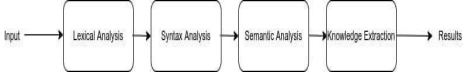
**5. System Analysis/ Proposed Architecture**

**High-Level Design**

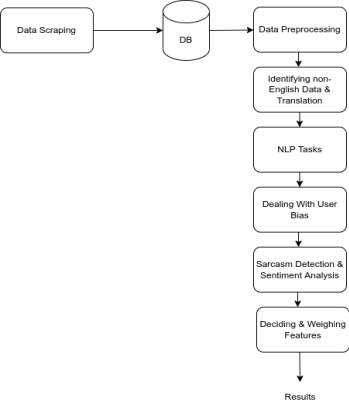
****Fig.5.1 High-Level Design

**Block Diagram**

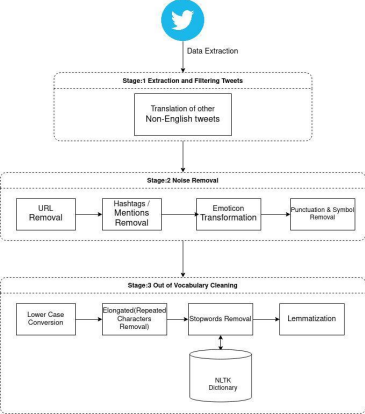
Fig.5.2 Block Diagram 



**Project Flow**

****Fig.5.3 Project Flow

**Data Cleaning Flow**

****Fig.5.4 Data Cleaning FLow

**6. Project Plan: (Timeline Chart for entire SDLC)**

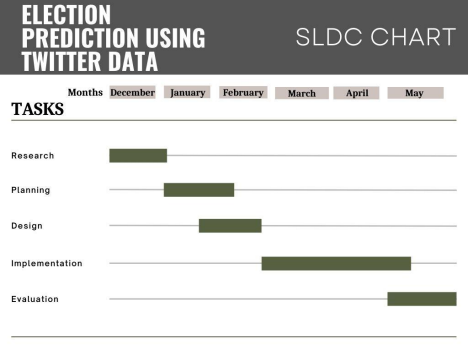


Fig.6.1 Timeline chart for entire SDLC

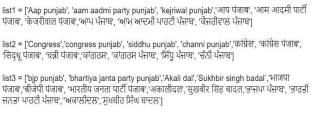
**7. Implementation**

The proposed methodology consists of 5 phases:

***A. Data Scraping:*** We considered Punjab Assembly 2022 elections for predicting results. Data is collected from Twitter as a data source using Python script. Script to collect data was developed on the Jupyter notebook platform with the use of the Snscrape library which takes the help of Twitter API to collect tweets. Keywords were decided for the data scraping using domain knowledge.

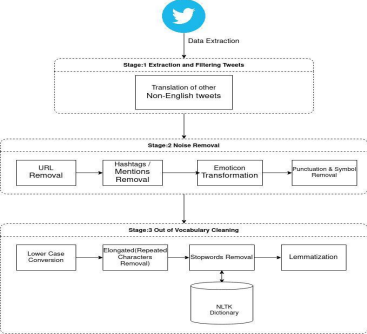
The timeframe for collecting tweets was considered to be 20 days prior to the election. The regional language tweets were also extracted as it reduces the sampling error that may occur when considering only English language tweets, All the tweets containing the decided querying keywords were extracted. The data is collected in JSON format and later converted to pandas dataframe. The dataframe is later exported to a *.csv* file. The attributes extracted for each tweet were: *'url', 'id', 'datetime', 'location', 'username', 'displayname', 'verified', 'followerCount', 'friendsCount', 'language', 'content', 'bio', 'likes', 'retweets', 'quotes', 'comments', 'hashtags'*.

The following keywords are selected for data scraping:



***B. Translation of regional language data:*** As a result of diverse Indian culture, many different languages are spoken throughout the country. Thus, it becomes important to collect and process the tweets posted in regional languages. Otherwise, considering only English data might result in a sample error as it represents a very small class of the whole population. Regional languages were identified and translated with the help of the Google Translation API based on the Neural Machine Translation approach which helps in context identification. Identifying context becomes very important in translation when translating multi-sentence tweets. Google Translate was selected over other popular translators like Yandex Translate, Bing Translate, and LibreTrans due to its wide language support, ability to translate according to context, and easy to use API.

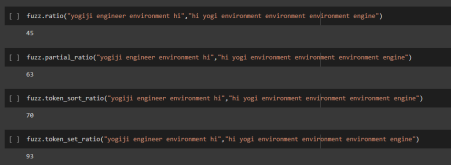
***C. Data Preprocessing:*** Preprocessing is very important to smooth the data as well as to remove noise i.e. words/symbols which are not interpretable by language models and can’t be interpreted for any kind of sentiment analysis. Regular expressions are used to identify patterns for such words using the *re* library in python and replace them using the required approach. Stopword removal is also an essential part of data preprocessing as stopwords generally don’t add much meaning to the data as well cannot interpret sentiment. Additionally, emojis were replaced by the respective interpretations using the custom emoji to meaning dictionary, all the data was converted to lowercase for uniformity, URLs were replaced with the keyword URL, and hashtags/ mentions were removed along with punctuation. Also, redundant characters for e.g. *i* in *hiii* are removed. Finally, the data was lemmatized i.e. converting the word to its base form e.g. *played/playing/plays* when lemmatized is transformed to *play* which helps in similarity matching of tweets for our bias detection algorithm.

Fig.7.1 Data Preprocessing

***D. Deciding and Calculating Weights:***

***Bias Detection:*** For bias detection, we make use of a fuzzy-wuzzy library that takes into account the Levenshtein distance. However, Levenshtein distance on its own is not an accurate metric to detect similarity. The fuzzy-wuzzy package cuts through that as it allows to calculate scores based on order and frequency of words. The way we identify biased users or bots is by seeing how similar their tweets are. We scrape 5 tweets from every Twitter user and compare them with each other to produce scores. Every Twitter user is then given a summed-up score which impacts how much importance we give to that user to train our model and use it for prediction. If the similarity score is high, it points to the fact that the user may be a bot or that the user is highly biased. If the score is low, it points to the fact the user is not a bot.

A few examples of the fuzzy-wuzzy library are as follows:

Fig.7.2 Bias Detection

***Sentiment Analysis:***

After preprocessing the data and detecting bias the next step is to capture the sentiments of users’ tweets. In this section, we are going to build a Sentiment Classifier to analyze our collected dataset for sentiment analysis using the BERT model and Hugging Face library. We are using the pre-trained Roberta model in order to tap into sentiments.

The roBERTa model is an extension of Bidirectional Encoder Representation from Transformers (BERT). The BERT and roBERTa fall under the Transformers family that was developed for sequence-to-sequence modeling to address the long-range dependencies problem which is essential in text-processing as context is important. This is a roBERTa-base model trained on 124M tweets and fine-tuned for sentiment analysis with the TweetEval benchmark. The model is trained on 124M tweets.

**8. Results and Analysis**

In order to calculate the percentage share per volume of each party, we used the following formula:

***% Share per volume = (1\* Positive value) + ((Negative values)\*(-1 )) / sum of all the values of each party***

We come to a conclusion using volume sentiment analysis and bias detection. We predict tweets into 2 categories, namely positive and negative using the roBERTa-base for Sentiment Analysis. We initially analyze the results using weighted sentiment analysis. Then we consider the threshold to be 0.65 for bias detection. The threshold conveys that 5 tweets per user were scraped and cleaned after our analysis concluded whether that user was biased or not. A score of 0.65 portends the fact that the similarity between those 5 tweets was 65%. Considering that we have used techniques like removing stop words, and lemmatization, it is safe to say that 65% is a good threshold to keep.

The features used to calculate results were sentiment polarity, volume share of the party, and similarity score obtained from the bias detection algorithm. We assign a weight of +1 for positive tweets and -1 for negative tweets.

Finally, we calculate each party’s vote share with the following formula:

***Vote Share(%) = Weighed Volume Share of Party / Total Weighted Volume Share (All Parties)* Punjab Election Results 2022**

**Volume Share(Punjab Assembly Elections 2022)**

|  | Positive | Negative | Total |
| --- | --- | --- | --- |
| AAP | 47758 | 15912 | 63670 |
| CONGRESS | 39201 | 12606 | 51807 |
| OTHERS | 20550 | 5816 | 26366 |

Table 8.1 Volume Share(Punjab Assembly Elections 2022)

**Percentage Share per party(Punjab Assembly Elections 2022)**

|  | Positive | Negative |
| --- | --- | --- |
| AAP | 75.0086 | 24.9914 |
| CONGRESS | 77.4877 | 24.3326 |
| OTHERS | 77.9412 | 22.0587 |

Table 8.2 Percentage Share per party(Punjab Assembly Elections 2022)

**Percentage share per Volume(Punjab Assembly Elections 2022)**

|  | Positive | Negative | Total |
| --- | --- | --- | --- |
| AAP | 0.336684 | 0.112176 | 0.44886 |
| CONGRESS | 0.276359 | 0.08887 | 0.365229 |
| OTHERS | 0.144873 | 0.041002 | 0.185875 |
| **Total** | 0.757916 | 0.242048 | 0.999964 |

Table 8.3 Percentage share per Volume(Punjab Assembly Elections 2022)

**Results(Punjab Assembly Elections 2022)**

| Vote  Share(%) | Weighed Volume Share of Party / Total  Weighed Volume Share of Party / Total  **Predicted** | Weighted Volume Share (All Parties)  Weighted Volume Share (All Parties)  **Actual** |
| --- | --- | --- |
| AAP | 43.9% | 42.01% |
| Congress | 36.07% | 22.98% |
| Others | 19.98% | 24.98% |

Table 8.4 Results(Punjab Assembly Elections 2022)

We were able to predict the winning party’s vote share with negligible error. The vote share of others was also decently predicted but an error in the vote share of Congress is noticed which might be explained by the skewness in the volume of scraped dataset due to the large no. of tweets made by News Agencies in midst of the notorious tussle between major nominees for CM candidate.

**Delhi Election Results 2020:**

**Volume Share(Delhi Assembly Elections 2020)**

|  | Positive | Negative | Total |
| --- | --- | --- | --- |
| AAP | 60169 | 23668 | 83837 |
| CONGRESS | 1539 | 409 | 1948 |
| BJP | 31851 | 8304 | 40155 |

Table 8.5 Volume Share(Delhi Assembly Elections 2020)

**Volume Share(Delhi Assembly Elections 2020) (Similarity Threshold=0.65)**

|  | Positive | Negative | Total |
| --- | --- | --- | --- |
| AAP | 54033 | 21560 | 75593 |
| CONGRESS | 1406 | 381 | 1787 |
| BJP | 28875 | 7591 | 36466 |

Table 8.6 Volume Share(Delhi Assembly Elections 2020)(Similarity Threshold=0.65)

**Percentage Share per party(Delhi Assembly Elections 2020)**

|  | Positive | Negative |
| --- | --- | --- |
| AAP | 71.47% | 28.52% |
| CONGRESS | 78.98% | 21.40% |
| BJP | 79.18% | 20.81% |

Table 8.7 Percentage Share per party(Delhi Assembly Elections 2020)

**Percentage share per Total Volume(Delhi Assembly Elections 2020)**

|  | Positive | Negative | Total |
| --- | --- | --- | --- |
| AAP | 0.474552 | 0.189354 | 0.663906 |
| CONGRESS | 0.012374 | 0.003346 | 0.01572 |
| BJP | 0.253598 | 0.066669 | 0.320267 |
| **Total** | 0.74052 | 0.242048 | 0.999893 |

Table 8.8 Percentage share per Total Volume(Delhi Assembly Elections 2020)

**Results(Delhi Assembly Elections 2020)**

| Vote  Share(%) | Weighed Volume Share of Party / Total  Weighed Volume Share of Party / Total  **Predicted** | Weighted Volume Share (All Parties)  Weighted Volume Share (All Parties)  **Actual** |
| --- | --- | --- |
| AAP | 59.2% | 53.57% |
| Congress | 1.87% | 4.26% |
| BJP | 38.85% | 38.51% |

Table 8.9 Results(Delhi Assembly Elections 2020)

With a relatively tiny margin of error, we were able to accurately estimate the winning party's vote share. Other party's vote percentages were similarly fairly predicted.

**9. Applications**

● The model created can be used not only for election predictions but in various other industries such as finance, sports, etc. For instance, the sentiment that people hold for a company can be represented in their tweets and that can work as a valuable signal for investors and further act as a signal of how the company is perceived around the world.

● Getting accurate results from election polls is not possible and assuming so would be a fallacy. What our model aims to do is provide accurate insights and statistics about various parties competing in elections.

● It is entirely possible that an individual may look at erroneous poll data and assume that the party that he/she supports is leading by a comfortable margin and would most definitely win the election. This would lead to the individual not voting without knowing the actual consequences of not voting. Hence it is extremely important for the general public to know accurate poll data and our model does just that.

**10. Conclusion & Future Scope**

To generalize the model more elections need to be taken into account so that the model gets completely robust. Furthermore, more research can be done on the effects of how text cleaning affects the outcome of the BERT sentiment predictor. It is also evident that the model gives a lot of responses as neutral which can be enhanced and fine-tuned to get either positive or negative outcomes. Finally, the bias detection techniques can be improved and more comprehensive techniques can be used for detecting bots. It is also recommended to keep a check on how Twitter evolves. With the addition of Elon Musk as CEO, his new vision is for removing bots from Twitter altogether. Therefore, researchers must monitor the changes in technology on Twitter.

**References:**

[1] K. d. S. Brito and P. J. L. Adeodato, "Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207147.

[2] A. Lopardo and M. Brambilla, "Analyzing and Predicting the US Midterm Elections on Twitter with Recurrent Neural Networks," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 5389-5391, doi: 10.1109/BigData.2018.8622441.

[3] F. J. J. Joseph, "Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree," 2019 4th International Conference on Information Technology (InCIT), 2019, pp. 50-53, doi: 10.1109/INCIT.2019.8911975.

[4] M. Bilal, S. Asif, S. Yousuf and U. Afzal, "2018 Pakistan General Election: Understanding the Predictive Power of Social Media," 2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), 2018, pp. 1-6, doi: 10.1109/MACS.2018.8628445.

[5] B. C. Das, M. M. Anwar and I. H. Sarker, "Reducing Social Media Users’ Biases to Predict the Outcome of Australian Federal Election 2019," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411633.

[6] B. C. Das, M. M. Anwar and I. H. Sarker, "Reducing Social Media Users’ Biases to Predict the Outcome of Australian Federal Election 2019," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411633.

[7] K. dos Santos Brito, R. L. C. S. Filho and P. J. L. Adeodato, "A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2021.3063660.

[8] M. Tsai, Y. Wang, M. Kwak and N. Rigole, "A Machine Learning Based Strategy for 28

Election Result Prediction," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 1408-1410, doi: 10.1109/CSCI49370.2019.00263.

[9] E. Soylu and S. Baday, "Predicting the June 2019 Istanbul Mayoral Election with Twitter," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1-6, doi: 10.1109/ASYU50717.2020.9259800.

[10] S. M. Srinivasan, R. S. Sangwan, C. J. Neill and T. Zu, "Twitter Data for Predicting Election Results: Insights from Emotion Classification," in IEEE Technology and Society Magazine, vol. 38, no. 1, pp. 58-63, March 2019, doi: 10.1109/MTS.2019.2894472.

[11] B. Bansal and S. Srivastava, “On predicting elections with hybrid topic based sentiment analysis of tweets,” Procedia Comput. Sci., vol. 135, no. 2018, pp. 346–353, 2018.

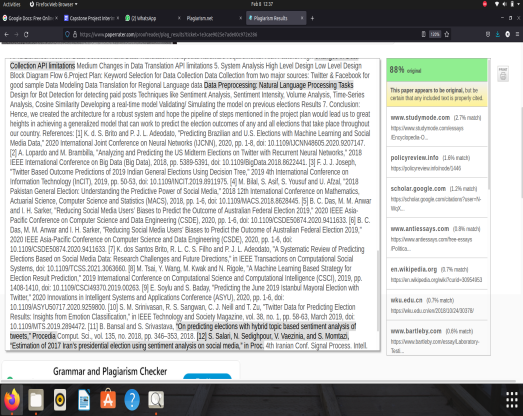
[12] S. Salari, N. Sedighpour, V. Vaezinia, and S. Momtazi, “Estimation of 2017 Iran’s presidential election using sentiment analysis on social media,” in Proc. 4th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS), Dec. 2018, pp. 77–82.

[13] D. Kreiss, R. G. Lawrence, and S. C. McGregor, “In their own words: Political practitioner accounts of candidates, audiences, affordances, genres, and timing in strategic social media use,” Political Commun., vol. 35, no. 1, pp. 8–31, Jan. 2018.

[14] M. Bilal, A. Gani, M. Marjani, and N. Malik, “Predicting elections: Social media data and techniques,” in Proc. Int. Conf. Eng. Emerg. Technol. (ICEET), Feb. 2019, pp. 1–6

[15] J. Kwak and S. K. Cho, “Analyzing public opinion with social media data during election periods: A selective literature review,” Asian J. Public Opin. Res., vol. 5, no. 4, pp. 285–301, 2018.

**Plagiarism Report:**

****