A Comprehensive Examination of Regional Polling Result Forecasting Using Data from Social Media

Uman Sohail[1] Dr. Shaik Shavali Tailor Kanekal[2]
[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana
[2]Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

***Abstract*—The introduction and widespread use of modern social media (SM), such as Facebook, Twitter, and Instagram social networks (SNs), have changed how politicians interact with the public and conduct political campaigns. Due to the inherent strengths of SM, namely the vast amount of data that can be accessed in real time, a new area of study has formed that focuses on using SM data to forecast election results. Despite the fact that numerous research have been done in the last ten years, the findings are frequently disputed. In this context, this article attempts to analyse and summarise how research on election prediction using SM data has developed from its inception, to lay out the state of the art and the state of the practise, and to pinpoint future research directions in this area. In terms of methodology, we conducted a methodical review of the literature, evaluating the quantity and calibre of publications, the electoral context of studies, the main approaches and traits of successful studies, as well as their main advantages and disadvantages, and compared our findings with those of previous reviews.**

The primary goal of the research is to forecast election outcomes using information from Twitter. We used the Kaggle open source dataset for the 2019 Indian General Election Tweets. All of Rahul Gandhi's and Narendra Modi's tweets make up this corpus, which was fed to the suggested model using cutting-edge machine learning methods. The project's scope is restricted to determining who has an advantage over others and calculating the correctness of the suggested model. The system administrator uses training data to develop the suggested model. To determine whether a tweet has a favourable or negative sentiment, test data can be provided to the model. The project's scope does not include user account maintenance.

The most important findings are the poor performance of the most popular method, namely volume and sentiment analysis on Twitter, and the better outcomes with novel methods, including regression techniques trained with conventional polls. The implementation of cutting-edge machine learning techniques has to be properly researched, among other things, and is mentioned in a vision of future research on integrating advancements in process definitions, modelling, and evaluation.

Keywords—Elections, Social Media, Social Networks, Machine Learning, Systematic Review.

I. INTRODUCTION

Social media (SM) has played a central role in politics and elections throughout this

decade. We have entereda new era mediated by SM in which politicians conduct permanent campaigns without geographic or time constraints,and additional information about them can be obtained notonly by the press but also directly from their profiles on social networks (SNs) and through other people sharing and amplifying their voices on SM. In this new scenario, SM isused extensively in electoral campaigns [1], and an online campaign's success can even decide elections. In practice,recent examples of SM engagement and electoral success include the 2019 Indian General election, when Narendra Modi focused his campaign on free-media marketing [2], and the 2018 Brazilian presidential election, when the candidate with more SM engagement but little exposition on traditional media was elected [3].

Moreover, in some way, it is possible to measure how apolitician's message is spreading over SM and try to estimate how much attention a candidate is receiving or how many people are talking about a candidate. Thus, considering a largeamount of data available in real time and the low cost of their acquisition, combined with the advances of techniques for processing them, a new research subject has emerged, focusing on using the SM data to predict election outcomes. Many studies claimed very positive results, others challenged the predictive power of SM, and even the same study may achieve positive results in one context and negative results in another context [18].Thus, there is not yet a common perspective on the literature or well-established methods, processes, and tools for predicting election results based on the SM data. Moreover, even the SM context has changed over the years. For example,Facebook surpassed the number of active users of Twitter, andnew SN has emerged, such as Instagram.

In this context, this article aims to give a thorough review and investigation of the state of both the art and practice of predicting election outcomes based on the SM data and identify key research challenges and opportunities in this field. We systematically reviewed 83 studies from 2008 to 2019,identify the context of studies, main models, strengths, and challenges of this new area, as well as the main characteristics present on successful studies, and deeply discuss future directions

II. RESEARCH BACKGROUND

Research Gap:

The way politicians communicate with the electorate and run electoral campaigns was reshaped by the emergence and popularization of contemporary social media (SM), such as Facebook, Twitter, and Instagram social networks (SNs). Due to the inherent capabilities of SM, such as the large amount of available data accessed in real time, a new research subject has emerged, focusing on using the SM data to predict election outcomes. Despite many studies conducted in the last decade, results are very controversial and many times challenged.

Problem Statement:

The existing methodologies to predict the election results based on volume and sentiment of tweets is found to be ineffective. the number of publications in this areais increasing and research is spread across 28 countries fromall continents. Nevertheless, there cannot yet be found anyprominent researchers, research groups, or clusters performingsustainable research in the area. In addition, there was noidentification of a common well-known forum for publicationon this subject, and results are spread across many forums.By combining studies' characteristics and success we foundthat, despite being the most used approach, volume/sentimentdoes not present high success rates, which is consistent withthe conclusions of previous surveys.

Aim Of The Project

The main of the project is to predict election results based on twitter data. We have used the 2019 Indian General Elections Tweets dataset which is an open source dataset on Kaggle. This corpus consists of all tweets for Rahul Gandhi and Narendra Modi and fed it to the proposed model that is built using advanced machine learning techniques.

Scope Of The Project

The scope of the project is limited to the compute the accuracy of the proposed model and compute who has a leading edge over other. The admin of the system trains the proposed model with training data. The test data can be given to the model to find if the sentiment of tweet is positive or not. Maintenance of user accounts does not fall under the scope of the project.

Proposed System

This article aims to investigate and summarize how research on predicting elections based on the SM data has evolved since its beginning, to outline the state of both the art and the practice, and to identify research opportunities within this field. In terms of method, we performed a systematic literature review analyzing the quantity and quality of publications, the electoral context of studies, the main approaches to and characteristics of the successful studies, as well as their main strengths and challenges and compared our results with previous reviews. We identified and analyzed 83 relevant studies, and the challenges were identified in many areas such as process, sampling, modeling, performance evaluation, and scientific rigor. Main findings include the low success of the most-used approach, namely volume and sentiment analysis on Twitter, and the better results with new approaches, such as regression methods trained with traditional polls.

Technical process involved in proposed model:

1. Identification of Dataset
2. Data preprocessing
3. Tweet analysis
4. Feature extraction
5. Creation of Model
6. Testing the model.

Advantages:

- High accuracy
- Can be extended to real time environments.

III. ALGORITHMIC PROCESS

Technical Approach

Below is the technical approach to address the problem:

1.Identification of dataset

2.Explorative Data Analysis

3.Cleaning the dataset and applying NLP techniques

4.Feeding the dataset to multiple algorithms and finding the best algorithm that suits the scenario

5.Training the final classifier and creating a model for the final classifier

6.Testing the final classifier and saving the results.

Data Preparation:

1.Filtering out tweets of Narendra Modi and Rahul Gandhi with some keywords and hashtags in it.

2.Remove stop words

3.Remove punctuation marks.

Algorithm for Sentiment Analysis

4.Calculate the polarity score using vadersentiment analyzer

5.Traverse through the polarity scores for each tweet and assign the Final Emotion as per the highest score among positive, negative, neutral.

6.Repeat Step 4 and Step 5 using Flair Classifier

7.Classify the total number of tweets based on Final Emotion for Narendra Modi and Rahul Gandhi.

8.Below are the results:

Model creation:

1.Use label encoding for categorizing the sentiment

2.Use tokenization to convert textual tweet data into numerical formats.

3.Split the dataset for training and testing

4.Feed the dataset to various machine learning algorithms.

5.Below are the accuracy results:

Algorithm	Accuracy in %
Decision Trees	80
Random Forest	85
Naive Bayes	75
Logistic Regression	87
SVM	79

Data Preparation for final classifier:

1.Read the dataset

2.Drop duplicates

3.Clean tweets

4.Perform stemming

5.Apply Tf-Idf vectorization

6.Find polarity

7.Perform one hot encoding for training set and testing

8.Split the dataset for training and testing

9.Feed the dataset to final classifer

Model :

Create the final classifier using bidirectional LSTM

We observe that the accuracy of the final classifier using Bidirectiona LSTM is 94.45%

Proposed Modular Implementation

Creating Model:

We focused on the task of predicting election results based on social media posts and tweets. After analyzing the tweets in the datasets, we propose a novel model that can generate the sentiment of the tweet. We have applied NLP techniques on the electoral tweets dataset, did the required exploratory data analysis and fed the processed dataset to multiple machine learning algorithms to test their performance. We finally built a machine learning model using Machine Learning and deep learning techniques that can get the sentiment of each tweet and computed the overall sentiment for Trump and Biden and its accuracy is about 94.45%.

This notebook includes the following:

1. Loading the dataset:
2. We will use 80% of data as training, 20% as validation data
3. Bag of words
4. Sanity check
5. Removing duplicates
6. Cleaning tweets
7. Removing punctuatuons
8. Cleaning tweets
9. Removing punctuatuons
10. Transforming the vectorizer
11. Converting to DataFrame
12. Final Classifier Creation

Below is the proposed modular implementation of the project. It consists of modules:

1. Admin

Admin Module:
The admin of the system is responsible for the activities like:

1. Uploading the dataset
2. Analysis of Electoral Tweets Dataset.
3. Comparison of various machine learning algorithms on the Tweets Dataset.
4. Review the performance of the algorithms on the given dataset
5. Build model for prediction of election results.
6. Test the model for to detect sentiment of tweet using test data and compute overall sentiment based on the entire dataset

A. SYSTEM DESIGN

1. Data Flow Diagram: Admin

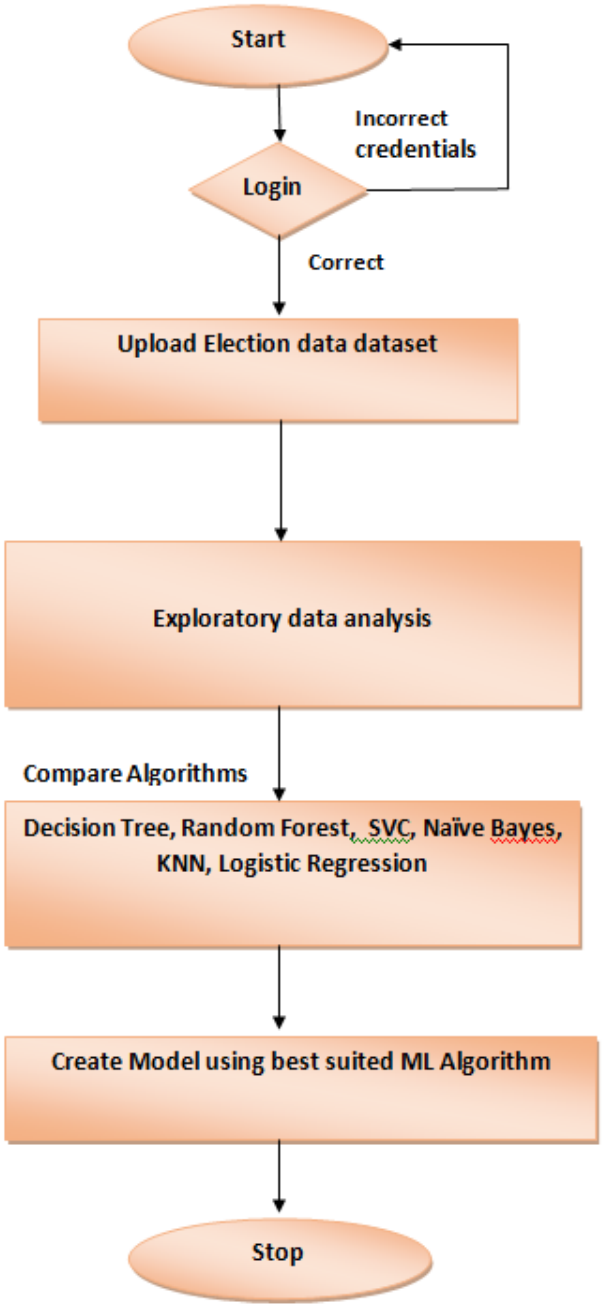


Figure 1: A Data Flow Diagram for Admin

IV. IMPLEMENTATION AND RESULT ANALYSIS

Home page:

This is the starting page of the application when the application is executed on Pycharm, the application is hosted on a web server and URL is generated to access the application once the user clicks on the URL the below page is opened on the browser.

Fig: Home Page

Admin Login:

This is the login page for the admin module. The admin need to login into the system with his credentials in order to perform operations like uploading the dataset, Training the dataset, Exploratory data Analysis of the dataset, Feeding the dataset to different Machine learning Algorithms to find the Algorithm that can meet the best accuracy and Create a model that can be hosted on the Flask Application to be used by the users.

Fig: Admin Login

Upload Dataset:

On this page, the administrator of the system can upload datasets that are used for training the machine learning models. The admin has to select the file by clicking on the Choose file button and click on the upload button to upload the file to the server. Once the upload is complete, a success message would be displayed that the file is successfully uploaded. For this project we are using IndianElection19TwitterData_full.csv as a dataset.

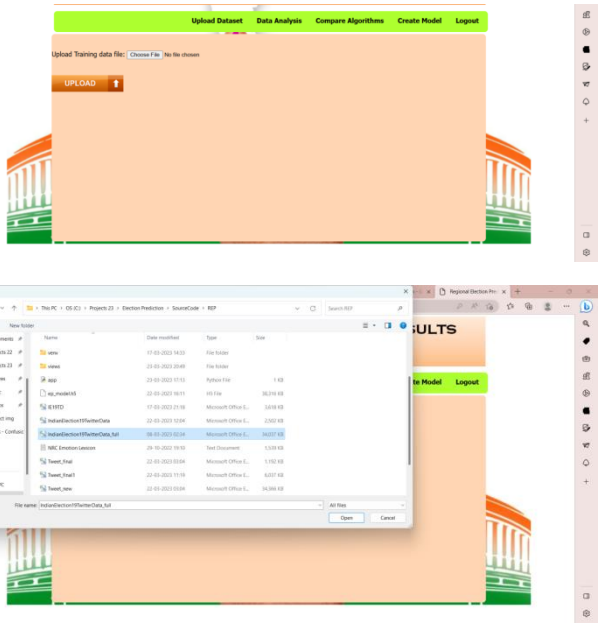
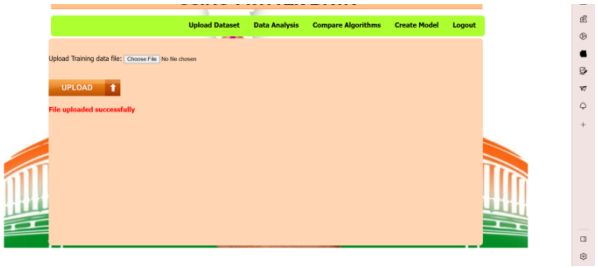


Fig: Upload Dataset & File Uploaded Successfully.



Data Analysis:

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, identify the relationships of various parameters of the outputs with the help of graphs, statistics etc.

Fig: Data Analysis

Public Tweet Analysis:

The below graph shows the Public Tweet Analysis over data present in the dataset.

Fig: Public Tweet Analysis

Narendra Modi Tweet Analysis:

The below graph shows the Narendra Modi Tweet Analysis over data present in the dataset.

Fig: Narendra Modi Tweet Analysis

Rahul Gandhi Tweet Analysis:

The below graph shows the Rahul Gandhi Tweet Analysis over data present in the dataset.

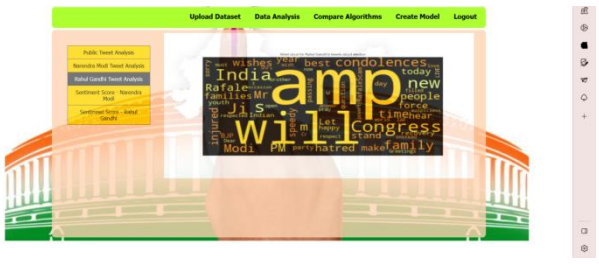


Fig: Rahul Gandhi Tweet Analysis

Sentiment Score - Narendra Modi:

The below graph shows the Sentiment Score - Narendra Modi over data present in the dataset.

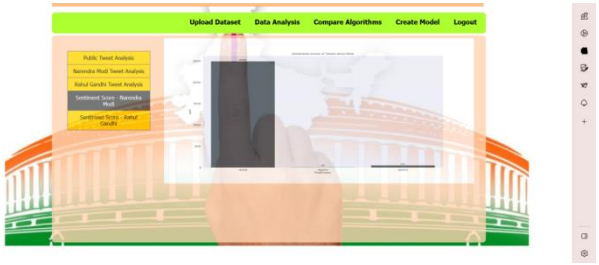


Fig Sentiment Score - Narendra Modi

Sentiment Score - Rahul Gandhi:

The below graph shows the Sentiment Score - Rahul Gandhi over data present in the dataset.

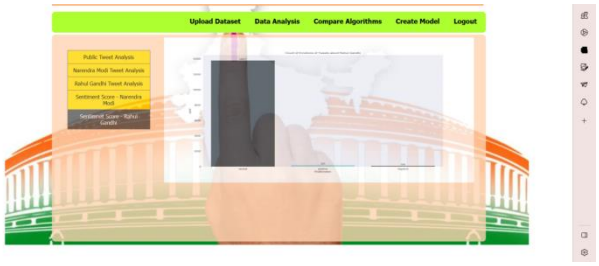


Fig Sentiment Score - Rahul Gandhi

Compare Algorithms:

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm. When the dataset is feed to various algorithms to evaluate the situation with some parameters like Accuracy, F1-Score , Recall…

Logistic Regression Classifier:

When the dataset is feed to Logistic Regression algorithm we observe that the test accuracy is 43.4%.

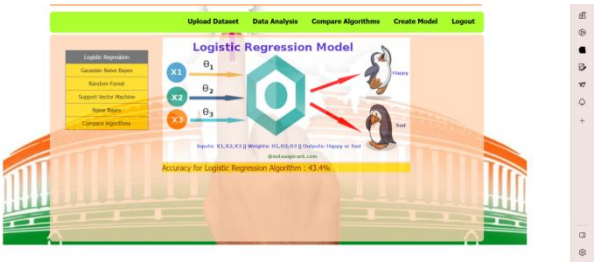


Fig: Logistic Regression Classifier

Gaussian Naive Bayes Classifier:

When the dataset is feed to Gaussian Naive Bayes algorithm we observe that the test accuracy is 44.54%.

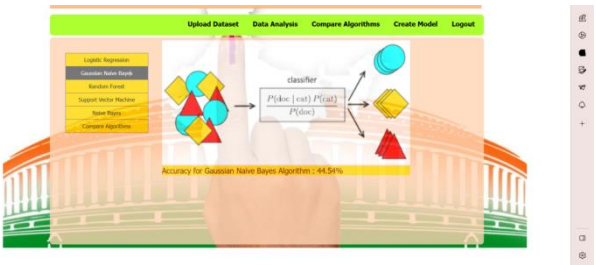


Fig: Gaussian Naive Bayes Classifier

Random Forest Classifier:

When the dataset is feed to Random Forest Classifier algorithm we observe that the test accuracy is 39.79%.

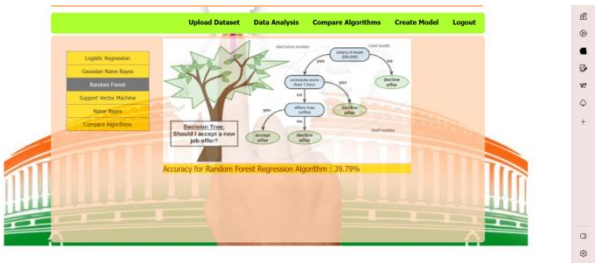


Fig: Random Forest Classifier

Support Vector Machine Classifier:

When the dataset is feed to Support Vector Machine Classifier algorithm we observe that the test accuracy is 67.62%.

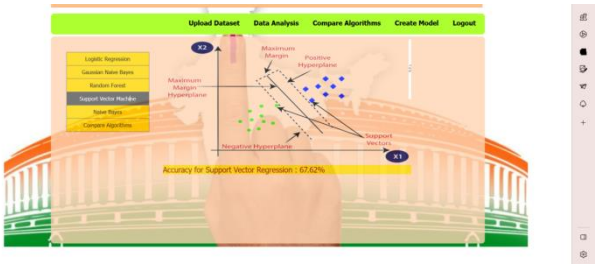


Fig: Support Vector Machine Classifier

Multinomial Naive Bayes Classifier:

When the dataset is feed to Multinomial Naive Bayes Classifier algorithm we observe that the test accuracy is 44.44%.

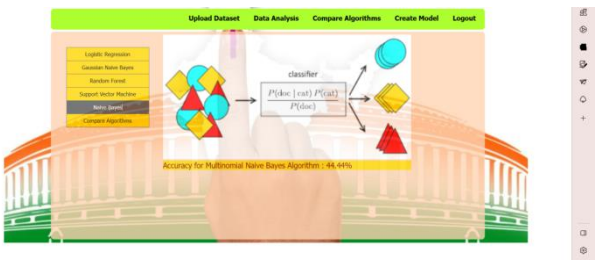


Fig: Multinomial Naive Bayes Classifier

Compare Algorithm Summary:

On this page, the admin can feed the dataset to various Algorithms to train them, get the test accuracy for each algorithm and their accuracies are summarized here.

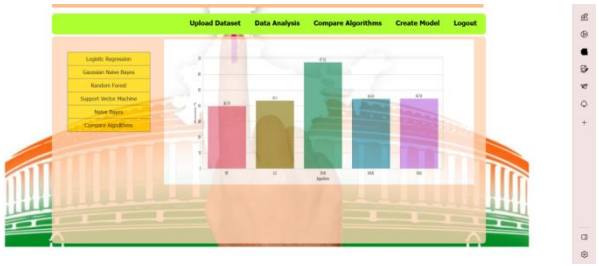


Fig: Compare Algorithm Summary

Create Model:

This screen shows the Accuracy of the Final Classifier Model is 94.45%.

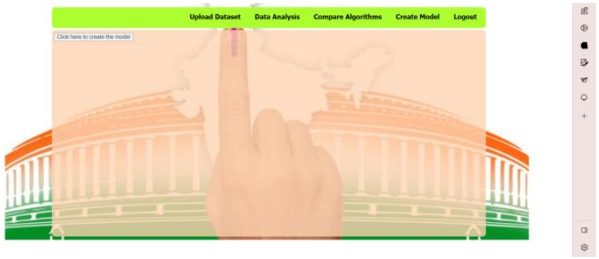
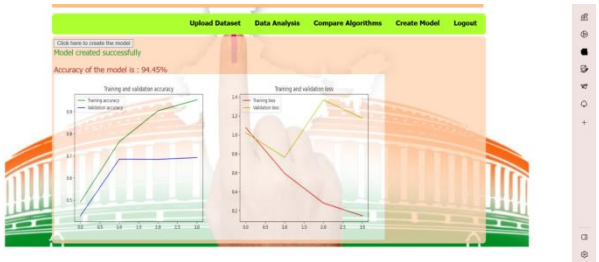


Fig: Create Model



V. CONCLUSION AND FUTURE SCOPE

In this work, we concentrated on the task of forecasting election outcomes based on tweets and posts on social media. We suggest a novel algorithm that can determine the sentiment of the tweet after analysing the tweets in the datasets. On the electoral tweets dataset, we used NLP techniques, performed the necessary exploratory data analysis, and then fed the processed information to various machine learning algorithms to test their performance. Using machine learning and deep learning techniques, we were able to create a machine learning model that can accurately predict the sentiment of each tweet and compute the aggregate sentiment for Rahul Gandhi and Narendra Modi.

Future Scope:

We used Twitter data in our study to forecast the outcome of the election. We intend to analyse data from Facebook and other social media platforms in the future and combine them to predict election outcomes generally. We may also use conversational images, audio, and video.

REFERENCES

[1] A. Jungherr, “Twitter use in election campaigns: A systematic literature

review,” J. Inf. Technol. Politics.

[2] P. L. Francia, “Free media and Twitter in the presidential election:

The unconventional campaign of Donald Trump,” Social Sci. Comput.

Rev.

[3] K. Brito, N. Paula, M. Fernandes, and S. Meira, “Social media and

presidential campaigns–preliminary results of the Brazilian presidential election,” in Proc. 20th Annu. Int. Conf. Digit. Government Res.

[4] S. Tilton, “Virtual polling data: A social network analysis on a student

government election,” Webology, vol. 5, no. 4, pp. 1–8.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting

elections with Twitter: What 140 characters reveal about political

sentiment,” in Proc. 4th Int. AAAI Conf. Weblogs Social Media, pp. 1–8.

[6] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith,

"From tweets to polls: Linking text sentiment to public opinion time

series," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, pp. 1–8.

[7] E. Sang and J. Bos, "Predicting the Dutch senate election results

with Twitter," in Proc. Workshop Semantic Anal. Social Media, pp. 53–60.

[8] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts?

How sentiment analysis of social media can improve our knowledge of

citizens' political preferences with an application to Italy and France,"

New Media Soc., vol. 16, no. 2, pp. 340–358.

[9] K. Singhal, B. Agrawal, and N. Mittal, "Modeling Indian general

elections: Sentiment analysis of political Twitter data," in Information

Systems Design and Intelligent Applications (Advances in Intelligent

Systems and Computing). New Delhi, India: Springer.

[10] N. Dwi Prasetyo and C. Hauff, "Twitter-based election prediction in the

developing world," in Proc. 26th ACM Conf. Hypertext Social Media

(HT), pp. 149–158.
.