Report On

Election Result Prediction using Sentiment Analysis

Submitted in partial fulfillment of the requirements of the Course project in Semester VIII of Final Year Computer Engineering

by
Salman Ansari (Roll No. 49)
Parth Desai (Roll No. 53)
Sanket Suhagiya (Roll No. 76)

Mentor
Dr. Tatwadarshi Nagarhalli

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(A.Y. 2021-22)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the Mini Project entitled **"Election Result Prediction using Sentiment Analysis"** is a bonafide work of **Salman Ansari (Roll no. 49), Parth Desai (Roll no. 53), Sanket Suhagiya (Roll no. 76)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in Semester VIII of Final Year **"Computer Engineering" .**

_____
Dr. Tatwadarshi Nagarhalli
Mentor

_____
Dr. Megha Trivedi
Head of Department

_____
Dr. H.V. Vankudre
Principal

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Course Project Approval

This Mini Project entitled **"Election Result Prediction using Sentiment Analysis"** by **Salman Ansari (Roll no. 49), Parth Desai (Roll no. 53), Sanket Suhagiya (Roll no. 76)** is approved for the degree of **Bachelor of Engineering** in in Semester VIII of Final Year **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Date:

Place:

Contents

ABSTRACT

In the recent year, social media has provided end users a powerful platform to voice their opinions. Opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Twitter.

Businesses need to identify the polarity of these opinions to understand user orientation and thereby make smarter decisions. One such application is in the field of politics, where political entities need to understand public opinion to determine their campaigning strategy. Twitter is indeed used extensively for political deliberation. We find that the mere number of messages mentioning a party reflects the election result. This data is used to predict outcome of election by using sentiment analysis.

Sentiment analysis of the tweets determine the polarity and inclination of vast population towards specific topic, item, or entity. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis. These algorithms are utilized to build classifier and classified the test data as positive, negative, and neutral. A two-stage framework can be formed to create a training data from the mined Twitter data and to propose a scalable machine learning model to predict the election result.

1.1 INTRODUCTION

An election is a most important part in the democracy. It is the instrument of democracy where the voters communicate with the representatives. Due to their important role in politics, there has been a big interest in predicting an election outcome. Traditional polls are too costly and still accuracy won't be achieved. To overcome this problem social media has been used as it is easy and freely available.

Social media sites have become valuable sources for opinion mining because people post everything right from the details of their life to the products and services they use, to give opinions about the current issues such as political and religious views. Millions of messages are being posted every day on popular social media sites such as Twitter, Instagram, and Facebook. Twitter is an online social networking service that enables users to send and read short 240-character messages called "tweets". Along with the short messages, users can use the hashtags before a relevant keyword or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter Search. The use of hash tags makes the problem of text classification relatively easier since the hash tag itself can convey an emotion or opinion. Currently around 6500 tweets are published per second, which results in approximately 561.6 million tweets per day. Facebook is the most popular and well-known Social Networking Service (SNS) all over the world. According to reports 86.3 percent people reported using the Internet in the previous six months. So, Facebook is also one of the good sources among various social media's available to be considered as a data source. In June 2016, the number of active Instagram users, those who use the application on a weekly basis, surpassed the number of active users on Twitter reaching a total of more than 500 million, of these some 300 million use their accounts at least once every day. This total has reached even greater heights in 2017 with the social media company boasting a total of 700 million active users, making Instagram the second most widely used social media platform after Facebook. Since Instagram attracts in surplus of 700 million users worldwide, allowing them to share and promote material at little cost, it would seem rational for political parties to want to take advantage of this communication tool.

This is an interesting research area that combines politics and social media which both concern today's society.

1.2 PROBLEM STATEMENT

The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Businesses (or similar entities) need to identify the polarity of these opinions to understand user orientation and thereby make smarter decisions.

One such application is in the field of politics, where political entities need to understand public opinion and thus determine their campaigning strategy. Sentiment analysis on social media data has been seen by many as an effective tool to monitor user preferences and inclination. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis. The accuracy of these algorithms is contingent upon the quantity as well as the quality (features and contextual relevance) of the labeled training data.

Since most applications suffer from lack of training data, they resort to cross domain sentiment analysis which misses out on features relevant to the target data. This, in turn, takes a toll on the overall accuracy of text classification. In this paper, we propose 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. Finally, we propose a scalable machine learning model to predict the election results using our two-stage framework.

OBJECTIVES

1. To predict the outcome of elections.
2. To use publicly available information from social media platforms (e.g., Twitter).
3. To apply natural language processing.
4. To apply Sentiment Analysis and predict outcome.

1.3 SCOPE

For our proposed model, we perform multistage classification and identify whether the sentiment of a tweet is positive or negative w.r.t. one of the election candidates. In this regard, we first classify the tweet based on the candidate that it is addressing or is relevant to. The first classifier is an 'entity classifier' which classifies a general stream of data into the respective entities. In the next stage, the classification is performed based on the sentiment of the text w.r.t. that candidate. Thus, each candidate has a classifier associated with him/her. The entity classifier is trained with the entire data set labeled by the entities. The sentiment classifier is trained with data set pertaining to only its candidate.

2.1 Introduction

This section summarizes some of the scholarly articles and research works in the field of Machine Learning and data mining to analyze sentiments on Twitter and preparing prediction model for various applications. The use of social media in last few decades have been helpful in determining people's attitude with respect to specific topics or events, a wide research interest in natural language processing and determining the sentiments based on it.

In [1], the collected tweets are analyzed using lexicon-based approach to determine the sentiments of public and a comparison is made among the candidates over the type of sentiment. Also, a word cloud is plotted representing most frequently appearing words in the tweets. Sentiment analysis on social media data has been done by authors of [2] as it is an effective tool to monitor user preferences and inclination. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis. Corpus collection, linguistic analysis and training a classifier was performed step by step in [3].

Corpus is a collection of written texts. They collected a corpus of 300,000 text posts from Twitter then evenly split into three sets of texts: Positive, Negative and Neutral. It is based on the Naive Bayes classifier that uses N-gram and POS-tags as features. With respect to [4], the mere number of messages reflects the election result and even comes close to traditional election polls. Using LIWC text analysis software, they conducted analysis of over 100,000 messages containing a reference to either a political party or a politician. Their result shows that Twitter is indeed used extensively for political deliberation and mere number of messages mentioning a party reflects the election result.

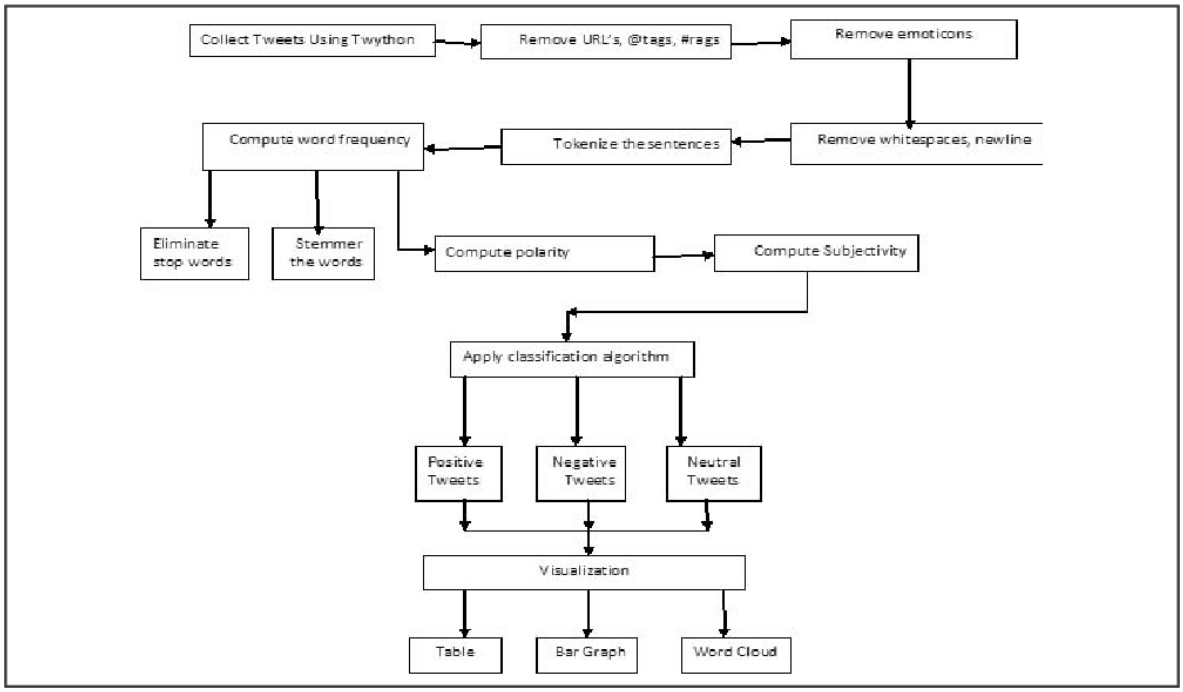
Utilization of Dictionary Based, Naive Bayes and SVM algorithm to build the classifier was done by [5] and classified the test data as positive, negative, and neutral. They identified the sentiment of Twitter users towards each of the considered Indian political parties. The results of the analysis for Naive Bayes were the BJP (Bhartiya Janta Party), for SVM it was the BJP (Bhartiya Janta Party) and for the Dictionary Approach it was the Indian National Congress.

In [6], authors have used Lexicon based approach with machine learning to find emotions in tweets and predict sentiment score. The research also showed that lexicon-based sentiment analysis improves the prediction result, but the improvements also vary in different states. The authors of [7], used fast and in memory computation framework 'Apache Spark' to extract live tweets and perform sentiment analysis. This paper provides a method for analyzing sentiment score in noisy twitter streams and reports on the design of a sentiment analysis by

classifying user's perception via tweets into positive and negative. The result of Indonesian Election was predicted by authors of paper [8].

Indonesian people had not elected a president until 2004, so Presidential candidate in Indonesia become a hot and interesting conversation among Indonesian citizen, and many of them expressed it through social media. In this research, the authors focused on tweets related to 2019 Presidential election with top keywords. Among Jokowi and Probowo result is produced by using R language showed that Jokowi leads the election.

2.2 Architecture / Framework / Block diagram



2.3Algorithm and Process Design

TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks. Here, we use this library to perform text classification in either positive or negative based on sentiment analysis.

This library is just like a Python string with the functionality of that you can easily use its functions. It provides a cool functionality that can easily summarize the text, provide you with sentiments of the text, spelling correction, translation, and language detection.

After using TextBlob, we get, Polarity ranging from -1 to +1(negative to positive) and tells whether the text has negative sentiments or positive sentiments. Polarity tells about information. Subjectivity also ranging from -1 to +1(negative to positive). So more +ve subjectivity means less factual data and mostly public opinion.

There are many cases where polarity is zero because there is some data which either doesn't contain any text or simply have links or hashtags only. So, we will drop such data in further steps.

2.4 Details of Hardware & Software

Hardware

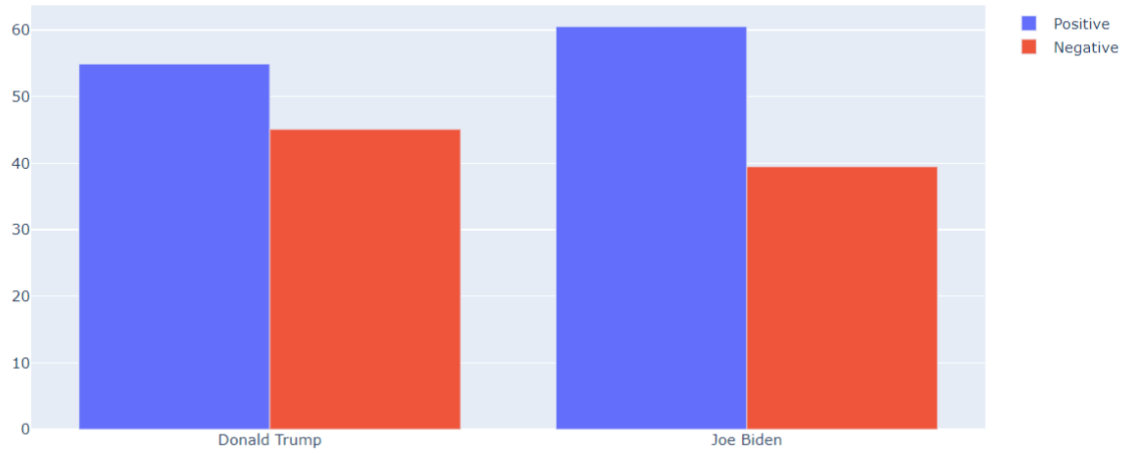
- Intel i5 processor
- RAM – 8GB
- Hard disk – 10GB
- Web browser
- Internet Connection

Software

- Jupyter Notebook
- Python — a programming language
- Tweepy — a type of RESTful API specifically for Twitter
- Textblob — processed textual data library tool (already trained on numerous textual data.)
- Pandas — data manipulation and analysis library
- NumPy — scientific computing library
- Matplotlib — plotting library
- Plotly — plotting library

2.5 Experiment and Results for Validation and Verification

Visualizing data gives a clearer picture of what we are doing. Here, we compare Negative and Positive tweets on Trump's tweets with that of Joe Biden to get a better understanding through visualization that who will be going to win this election.



From the above figure, it is very evident that Joe Biden is getting more positive replies as compared to negative reviews whereas Trump is getting both types of reviews in approx. equal ratio.

There are more chances for Joe Biden to win the Election.

2.6 Conclusion and Future work.

The use of social media for prediction of election results poses challenges at different stages. In this paper, we first tackle the scarcity of training data for text classification by providing a two-stage framework. Finally, we propose our model for election result prediction which uses the labeled data created using our framework. While our model alone may not be sufficient to predict the results, however it becomes a crucial component when combined with other statistical models and offline techniques (like exit polls). We implemented the proposed model on a dataset which was created by mining Twitter for 3 days. However, this model can be extended in the future to create an automated framework which mines data for months since election result prediction is a continuous process and requires analysis over long periods of time. Features should be extracted from newly mined data and compared with existing set of features. Some similarity metrics can be used to compare the new and old features. Only in cases where the metric value crosses a threshold, the newly mined data should be labeled using the two-stage framework. Thus, we recommend creating an Active learning model wherein the model itself recommends what data should be labeled. This would minimize the efforts for labeling while making sure that there is no compromise on contextual relevance.

References

[1] Ms. Farha Nausheen et al., “Sentiment Analysis to Predict Election Results Using Python”, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018).
[2] Jyoti Ramteke et al., “Election Result Prediction Using Twitter Sentiment Analysis”, 2016 International Conference on Inventive Computation Technologies(ICICT).
[3] Alexander Pak, et al., “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the Universit e de Paris-Sud, Laboratoire LIMSI-CNRS.K. Elissa, “Title of paper if known,” unpublished.
[4] Andranik Tumasjan et al., “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.
[5] Parul Sharma et al., “Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter”, 2016 IEEE International Conference on Big Data (Big Data).