Election Result Prediction Using Sentiment Analysis

Abdulrahman Ahmed
Electrical Engineering
University of Ottawa
Cairo, Egypt
aahme275@uottawa.ca

Abdulrahman Elmashtoly
Electrical Engineering
University of Ottawa
Cairo, Egypt
aelma045@uOttawa.ca

Mahmoud AboEl-wfa
Electrical Engineering
University of Ottawa
Cairo, Egypt
maboe017@uottawa.ca

Nada Aboelfetouh
Electrical Engineering
University of Ottawa
Cairo, Egypt
naboe016@uottawa.ca

Nada Abd-Elmageed
Electrical Engineering
University of Ottawa
Cairo, Egypt
nabde013@uottawa.ca

***Abstract*—**

A world without problems is useless, people's passion to have more in our world, makes us use technology in each step. Sentiment analysis new field of artificial intelligence that depends on analyzing public opinion due to huge amounts of opinion on the web, social media, and blogs. We used social media especially Twitter to apply automatic means of polarity(positive, negative, and neutral). The election in the United States is a hot issue that will affect various aspects of the World and the election took a huge amount of tweets throughout the election duration.

Classification algorithms such as Logistic Regression, support vector machine, Decision tree, Random Forest, K-Nearest Neighbor, and Gaussian Naïve Bias in the other hand we used Deep Learning techniques like Neural network to train those models we need training data to perform sentiment analysis. Quality and quantity of data will be important factor that will affect the accuracy. Our goal in this paper we collect some data and apply cleaning then classify every opinion of the mentioned candidate then train our models based on this data to predict the winner.

Keywords*— *Supervised Models, Machine Learning, Sentiment analysis, Artificial intelligence, Deep Learning.

Introduction

Social media are interactive technologies that facilitate the creation and sharing of information, ideas, interests, and other forms of expression through virtual communities and networks. While challenges to the definition of social media arise due to the variety of stand-alone and built-in social media services currently available, there are some common features[1]. That could use by a number of data analytics companies on a variety of subjects.

Machine Learning models will learn some patterns and relations from the features that extracted from the data set. Twitter is an online social service that allows users to interact with other's tweets or replay their opinion. It is interesting and could grab attention to change a lot of matters. A huge number of tweets makes the problem of classification so, we could know how things go out there without being there [2].

Using social media platforms such as Twitter, Instagram and Facebook are a common way of expressing yourself. It is common for people to share news, discuss political events, and comment on certain global events. Thus, social media is used in political campaigns, social and development projects, and to express opinions about elections. It was during the U.S. presidential election that social media was first used for a political campaign.

Election sentiment analysis has its drawbacks as well, for instance, it might be challenging to identify sarcasm in casual contexts. Because of their writing styles, sometimes people classify negative feelings as positive. Comparing Facebook to Twitter, sentiment analysis is generally significantly more effective on the latter platform. This suggests that some social media users may not be sincere when presenting their true sentiments. Therefore, their opinions do not accurately represent the situation. Furthermore, because not all voters are present, social media may not accurately reflect voter sentiment during elections. While social media may not fully represent everyone, it does offer a representative sampling of people's views. Additionally, some people might not want to share their opinions because of privacy concerns, so even if they are on social media, they might not express their genuine viewpoints [3].

I. System Architecture

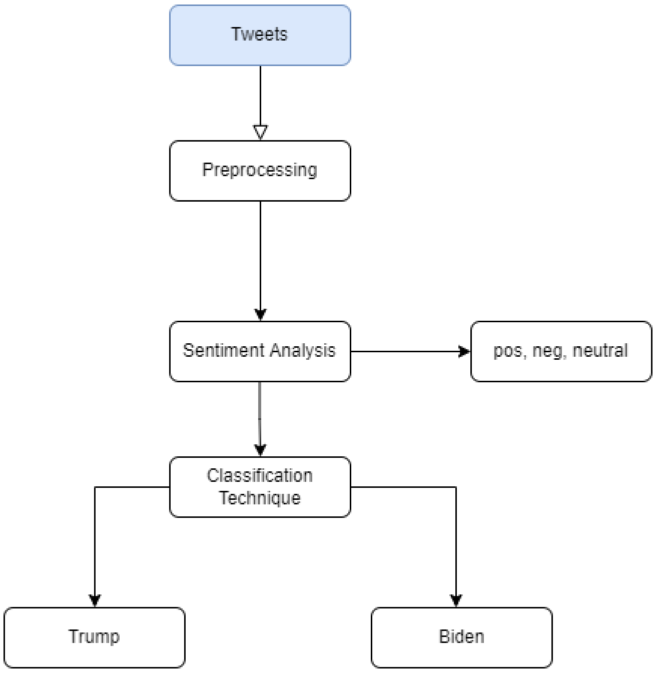

Fig.1: SYSTEM ARCHITECTURE

II. MODEL

Our proposed method consists of four main steps, which include: data collection / cleaning and pre-processing, sentiment analysis and classification.

A. *Data Collection*

Gathering data from twitter is a crucial part and effective to predict future scenarios or actions. Collecting tweets made before the election duration using Tweepy python library is easy for accessing the Twitter API.

B. *Data Cleaning*

Data Quality is the main issue of data preprocessing. Raw data contain errors and unwanted data. In data cleaning we Prepared a data set for analysis by removing irrelevant data, removing punctuations, URLS and digits to get better performance and powerful insights from our data.

Tweets contain many unwanted characters for Sentiment analysis we used the Regular Expression Library: to perform pattern matching within Target Sequence.

C. *Data Mapping*

Collecting data from Various resources so we map each sample to its main location for analysis. the data must be assimilated in a way that makes it accessible to decision-making. We manually mapped it by creating a list of US states and another list of US state codes, then extracting tweets from the location data that contained these. as in Massachusetts Mapped to Massachusetts.

D. *Sentiment analysis*

Sentiment refers to the positivity or negativity expressed in Writings Sentiment analysis provides an effective way to evaluate written or spoken language to discover if the expression is favorable, unfavorable, or neutral, and to what degree. To determine the emotion of each tweet, we apply the polarity scores () method to the text. The result of this method call is a dictionary that displays the intensity of negative, neutral and positive sentiment in the tweet. All these three values are the overall compound sentiment of the tweet. we can categorize every tweet data as either positive, negative, or neutral, we can aggregate it by the state to get the general perception of the public of that state regarding [4].

- Vader

  one of the most popular NLP sentiment analysis Packages. Based on the return the probability of a given sentence. It is designed for social media data and can produce effective results when combined with data from Facebook, Twitter, and other sources. The primary drawback of the rule-based approach for sentiment analysis is that the method only cares about individual words and completely ignores the context in which it is used. After we apply it on tweets there are the results contain five different decision to determine which of the following states a state belongs to, we employ the methodology mentioned above. Decisions

  [Trump, Biden, distract neg votes, distract pos votes, Insufficient Data]

  distract votes: The margin between positive tweets for both candidates is more than the margin for negative tweets.

  The other distract votes: The margin between positive tweets for both candidates is less than the margin for negative tweets.

  Insufficient Data: States in which the number of tweets made on either competitor is less than 15

  the final result from Vader analysis Strongly Biden was the winner [8].

- Text Blob

  Text blob produces polarity and subjectivity as its two outputs after receiving a sentence. The output that falls between [-1,1], where -1 denotes a negative emotion and +1 denotes a pleasant emotion, is polarity. Subjectivity is the output that falls within the range [0,1] and pertains to subjective assessments [8].

  The results are the same as Vader

E. *Classification Models*

A machine learning model is trained to recognize certain types of patterns. By using algorithm to analyze and learn from a set of data, you may train it to perform calculations on those data.

Supervised Machine Learning Models

Supervised learning: Supervised ML for labeled data for regression and classification problems (binary or multiclassification). For regression problems predicted value is continuous but for classification problems y predicted must be discrete values, 1 or 0 for binary classification and many classes for multi-classification. Based on the advantage of each algorithm, the shape of the data, and the type of problem we choose the suitable algorithm and tuning hyperparameters to improve accuracy [7].

- SVM

Used for supervised ML for regression and classification problems but usually used for classification problems. For linear cases, it is based on support vector, hyperplane, margin.

Support vector: the closest data points that exactly lie on margin.

Hyperplane: the line that divides training data points into classes.

Margin: distance between the hyperplane and the support vector.

- Decision tree

Tree-based methods create a series of decision rules splitting the predictor space to predict a target. Each branch of the tree represents a predictor.

If the problem is regression so, without determining hyperparameters' number of leaves, it would take the mean of the training dataset and that result would be the predicted value. If the problem is classification, then the predicted point would be classified to the group that contains the highest amount of data [12].

- Random Forest

Algorithm that constructed from many Decision Trees used to solve classification and regression problem. Is Trained through bagging is Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The prediction depends on taking the average or mean of the output from various trees [5].

- **Gaussian NB**

  It is a Generative Model in that each class follows a Gaussian distribution. assume that each feature makes an independent and equal contribution to the outcome in details -We presume that there are no dependent feature pairs. For instance, the fact that it's "Hot" outside has nothing to do with the humidity, and the fact that it's "Rainy" outside has nothing to do with the winds. Therefore, it is presumed that the traits are independent Second, each Feature is given equal weight (or importance). For instance, it is impossible to precisely anticipate the outcome based solely on the temperature and humidity. All of the characteristics are relevant and are seen as having an equal impact on the result [6].

- **K -Nearest Neighbors**

  The K-Nearest-Neighbors is a non-parametric classification method, which is simple but effective in many cases in order to classify a data record, its k closest neighbors are obtained, creating a neighborhood of t. The classification with or without consideration of distance-based weighting is often decided by a majority vote among the data records in the neighborhood [11].

F. *Deep Learning Models*

- **Neural Network**

  Neural network technologies are based on deep learning, sometimes referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), a subset of machine learning. Its structure and terminology are inspired by the human brain, reflecting the communication between biological neurons. a node layer of an artificial neural network (ANN) consists of an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, is connected to others and has a weight and threshold that go along with it. Any node with output exceeds the defined threshold value is activated and begins providing data to the network's highest layer. Otherwise, no information is moved to the network's next layer [10].

III. Perfomance Evaluation

This dataset was divided into a training dataset (80%) and a testing dataset (20%). The models were evaluated using the 20% testing dataset after being trained. Table.1 with Fig.2 displays the results and for the champion model Fig.3, Fig.4 display the confusion matrix and classification report. The accuracy of the champion model is 67.9 %. And the confusion matrix for that model shown in Fig.2.

Model Trained	Accuracy
Logistic Regression	**55%**
Support vector Machine	**51.6%**
Decision Tree	**67.9%**
Random Forest	**67.7%**
Gaussian Naïve Bayes	**52.6%**
K-nearest neighbor	**66.9%**
Neural Network	**57.2%**

Table.1

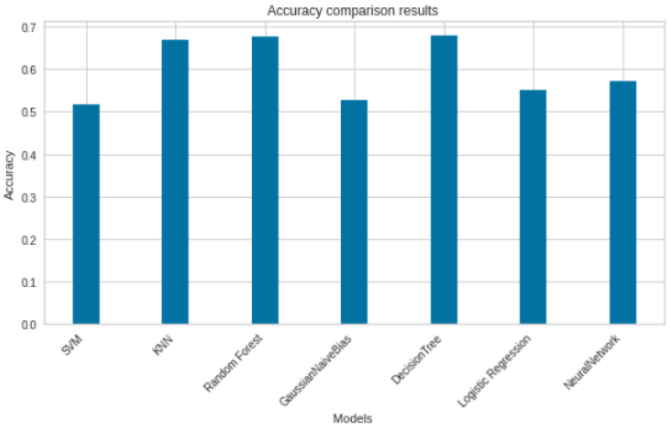

Fig.2: Comparison of Accuracies

A. *Champion Model(Decision Tree)*

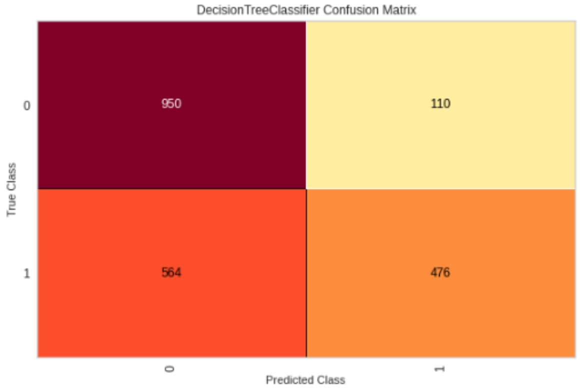

Fig.3: Confusion Matrix for Decision Tree

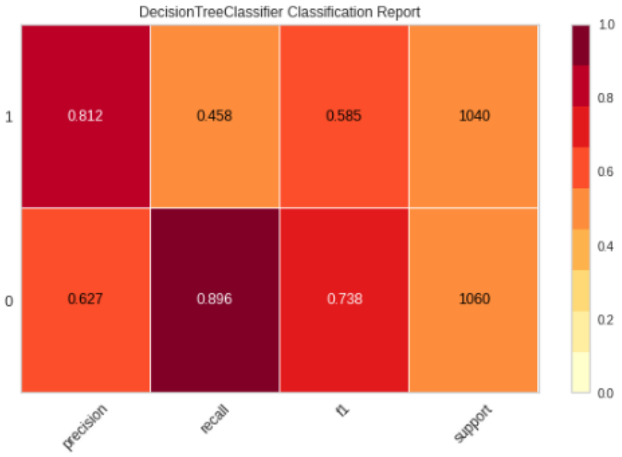

Fig.4: Classification Report for Decision Tree

IV. Summary and Conculsion

The Twitter dataset was collected for the analysis and prediction of elections by scrutinizing and researching public opinions on social media during the U.S presidential election between Trump and Biden. In this project, VADER and Textblob were used to extract features from the tweet, and many machine learning algorithms were applied to determine whether the given candidate has a positive or negative sentiment and make a decision could be right or not. For future work, we can also take form the official Twitter API to get more data such as the user's longitude and latitude to confirm their location. By using these parameters so, we can get better and accurate results also we can extract emojis and understand emotions and avoid fake tweets. The data are not distributed well. This was brought up earlier in my analysis, and I believe that this may have introduced some bias. The ideal situation would be to have the same number of tweets from both candidates in each state. But in the real world, it is challenging, if not impossible, to accomplish this. To get my dataset's distribution of states closer to uniformity, though, I might consider adding some distribution models. It is possible to enhance the judgment's ability to forecast a state. The reasoning I used to determine whether a state would lean Republican or Democratic in the long run should be improved. I created a rationale for this computation using what little technological expertise and subject knowledge I had, but it is still available to improvement.

References

[1] Jyoti Ramteke, Darshan Godhia, Samarth Shah, and Aadil Shaikh. Election Result Prediction Using Twitter sentiment Analysis, 2018Hutto, C.J. & Gilbert, E.E. (2014).

[2] VADER: A Parsimonious Rule based Model for Sentiment Analysis of Social Me ia Text. Eighth International Conference on Weblogs and Social Meia (ICWSM-14).Ann Arbor, MI, June 2014.

[3] Budiharto, Widodo, and Meiliana Meiliana. “Prediction and Analysis of Indonesia Presidential Election from Twitter Using Sentiment Analysis.” Journal of Big Data, vol. 5, no. 1, Dec. 2018, p. 51. DOI.org (Crossref), https://doi.org/10.1186/s40537-018-0164-1.

[4] Maurya, Ankit, et al. “Election Result Prediction Using Sentiment Analysis.” International Journal of Advanced Research in Science, Communication and Technology, Apr. 2021, pp. 118–22. DOI.org (Crossref), https://doi.org/10.48175/IJARSCT-V4-I3-018.

[5] “Introduction to Random Forest in Machine Learning.” Engineering Education (EngEd) Program Section, https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/. Accessed 2 Aug. 2022.

[6] Hrouda-Rasmussen, Stefan. “(Gaussian) Naive Bayes.” Medium, 7 May 2021, https://towardsdatascience.com/gaussian-naive-bayes-4d2895d139a.

[7] QuinnRadich. What Is a Machine Learning Model? https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model. Accessed 2 Aug. 2022.

[8] ES, Shahul. “Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch.” Neptune.Ai, 9 Oct. 2020, https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair.

[9] Chaudhry, Hassan Nazeer, et al. “Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020.” Electronics, vol. 10, no. 17, Aug. 2021, p. 2082. DOI.org (Crossref), https://doi.org/10.3390/electronics10172082.

[10] What Are Neural Networks? https://www.ibm.com/cloud/learn/neural-networks. Accessed 2 Aug. 2022.

[11] Joby, Amal. *What Is K-Nearest Neighbor? An ML Algorithm to Classify Data*. https://learn.g2.com/k-nearest-neighbor. Accessed 2 Aug. 2022.

[12] z_ai. “Decision Trees Explained.” *Medium*, 26 Sept. 2021, https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6.