Journal of Intelligent Systems and Internet of Things
Vol. 15, No. 01, PP. 185-197, 2025
DOI: https://doi.org/10.54216/JISIoT.150116

AI-Powered Election Insights: Predicting the 2024 Trump vs. Kamala Election Showdown with Machine Learning

Yazan Alnsour[1,*], Mohammad Alsharo[2], Malik AL-Essa[3], Aseel Smerat[4]

[1]Prince Mohammad Bin Fahd University, Kingdom of Saudi Arabia

[2]AL AL-Bayt University, Jordan

[3]AL-Ahliyyah Amman University, Jordan

[4]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India Applied science research center, Applied science private university, Amman 11931, Jordan

Emails: Yalnsour@pmu.edu.sa; Mohammad.alsharo@aabu.edu.jo; M.alessa@ammanu.edu.jo; smerat.2020@gmail.com

Abstract

The United States presidential elections receive a substantial attention not only from American voters, but also from news agencies, politicians, and international governments due to the local and global impact of the outcome. Therefore, different parties strive to predict the election's results ahead of time, and opinion polls remain the predominant prediction method despite their bias and flaws. Online political communication has immensely evolved in recent years, especially on social media websites like Reddit, which has become a key platform in political discourse offering a valuable resource for studying public opinions on key issues. This study aims to utilize advanced machine learning methods to predict the outcome of the upcoming 2024 U.S. presidential election with a focus on the two primary candidates, former President Trump and Vice President Harris. Employing deep learning techniques to analyze more than 25 thousand online posts on Reddit, the results indicate that on the national level, Harris has more favorable sentiment in comparison to Trump among online users. However, analyzing the data associated with the battleground states, our model predicts that Trump has an edge over Harris, which may result in Trump winning the majority of the electoral votes in these states. This study underscores the importance of integrating social media data with machine learning capabilities for enhanced data-driven forecasts in upcoming elections and major public events.

Received: July 08, 2024 Revised: October 10, 2024 Accepted: December 30, 2024

Keywords: Elections; Deep Learning; Political forecast; Trump vs Harris; Public opinion Analysis; Reddit

1. Introduction

Predicting the outcome of the United States presidential election generate interest in both practice and academia, as the electoral outcomes of the country with the largest economy and military power has a substantial impact on social, economic, and world political issues [1, 2]. Election forecasting has a history as an "art, not a science," but it has become ever more sophisticated in recent years as it combines statistics, social, and political sciences [3]. For instance, Professor Alan Lichtman, a distinguished historian in American University in Washington D.C., has successfully predicted the outcome of nine out of the last ten presidential races using his thirteen keys prediction system [4]. In academia, several political science scholars have introduced statistical models they claim to predict the elections, some of these models take into consideration the economic factors and household income [5, 6] while other models utilize public opinions and approval rates of former presidents [7]. Moreover, due to the nature of the American electoral vote system, which designates a specific number of electoral votes for each state, the winner prediction process is complex and could have several dimensions. Therefore, for every past presidential election, there were nationwide and statewide prediction models using either just a few snapshots of data or a handful of methods aiming to conclude the election's outcomes [8, 9].

In recent years, social networks have become a popular outlet for political and social opinions, especially among younger generations [10, 11]. Social Media platforms such as Twitter, Facebook, Reddit and others have emerged as important sources of real-time data to depict public sentiment, offering a continuous stream of data that reflects the dynamic and rapidly changing landscape of users' opinions [12-14]. Not only that but they are also valuable since they provide spontaneous and unfiltered thoughts of large number of users which depict the picture of people's sentiment in comparison to traditional polling methods [8, 15, 16]. For example, a recent study by Boulianne et al. [10] argued that social media outlets like Twitter (now referred to as X), Reddit, and others have become venues for political and social discussions over the past few years. In a previous study, Beauchamps [15] stated that online users discussions and interactions and their sentiments reflected in their posts can provide important insights about new social trends and possible future electoral outcomes. In addition, Rajadesingan et al. [17] highlighted that an online platform Reddit can serve as a valuable data source where conversations and discussions that take places in different online communities also known as subreddits would reveal trends and provide valuable insights into the community members feelings, opinions, and sentiment which could be used to predict their future behavior. A major challenge with using unstructured data in online posts to obtain meaningful information and insights comes from the large volume of available data and unstructured way it is presented in. Bovet and Maksa [18] indicated that the utilization of Machine Learning (ML) techniques to analyze a large volume of unstructured data can offer new opportunities for both academia and industry. Natural Language Processing (NLP) models which are a type of ML can analyze large volumes of unstructured data from a variety of resources to extract meaningful information and insights that can be leveraged to predict future trends [19]. NLP models can be developed to understand and interpret human like text [20].

By processing and extracting relevant information from unstructured text data, NLP models can reveal hidden patterns, trends, and sentiments that might otherwise go unnoticed. In social media where users freely share their thoughts, opinions, and reactions NLP becomes a powerful tool can sift through vast amounts of user's generated text to identify key themes, emotional tones, and sentiment [21]. In this paper, we aim to analyse Reddit's posts that share users' opinions regarding former President Donald Trump and Vice President Kamala Harris. Using more than twenty-five thousand posts our goal is to provide detailed and precise projection of how the election might unfold. As discussed by Bovet and Makse [18] analysing social media data can be valuable for decision makers and different stakeholders that might be impacted by the outcomes of an electoral vote. With communication evolving in the digital era, the findings and strategies outlined in this study will be of importance to both researchers and practitioners who seek to better understand what affects democratic procedures and their outcomes. The remainder of this paper is structured as follows: the next section reviews the relevant literature and presents related work. This is followed by a detailed description of the predictive model, research methodology, and results. Finally, the paper concludes with a discussion of the findings, the study's limitations, and directions for future research.

2. Related Work

Predicting the outcome of elections with a degree of certainty is crucial for stakeholders like decision makers and political groups alike. Candidates and parties may attempt to influence projected results by utilizing means to sway voters and alter their views, on Election Day. In addition, they can assess voter responses to social and political issues then tailor their messaging to improve how potential voters perceive their candidates. Furthermore, certain financial entities and investors might show interest in forecasting election outcomes due to the influence they wield over financial market swings. Preparing accurately predicting market instability can reduce the risks that some investors face or present lucrative investment prospects [22]. In the following sections, we will explore existing studies. Pinpoint where the research falls short at present.

A. Traditional Methods of Election Prediction

Traditional election predictions rely on surveying individuals for insights regarding voters' preferences at a certain point of time prior to elections. Pasek [23] argues that traditional methods suffer deficiencies due to their small sample sizes and potential biases. Opinions polls typically draw from a small subset of potential voters which may result in many times inaccurate predictions. For example, traditional polling did not accurately forecast the results of the Brexit referendum in the UK nor the 2016 U.S presidential election [24]. This was due to their failure to capture changing opinions and their oversight of specific voter groups along with outdated sampling methods that led to forecasting inaccuracies. Also offering a glimpse into voter preferences at a certain period that only capture a snapshot of the voters' sentiment. Such would potentially miss voters' opinion if changes due to various factors, like debates or scandals that can influence opinions significantly during an election cycle. Moreover, there is the concern of nonresponse bias when people who opt out of surveys may hold opinions compared to those who take part in them leading to distortion of the findings [15].

B. The Rise of Social Media as a Predictive Tool

The rise of media has brought about a change in the way public sentiment is assessed and understood in today's world. Platforms such as X, Reddit, and Facebook are now channels for capturing real time feedback from the public and provide a flow of information that mirrors the evolving views of users [25]. Researchers can use different social media platforms to gather data in real time and across different time points that can provide a better picture of the public opinion and possible changes across time and due to certain events in comparison to the conventional techniques that are used in. The usage of such data to predict future behaviour and possible outcomes will be more precise polling [15, 26]. Social media platforms can allow researchers to gather a large amount of data with a cost significantly lower than conventional polls. Such allow including different viewpoints from different users regarding different topics and discussions. In addition, social platforms provide flexibility to gather data before and after main events or incidents and allow comparing how such affected the public [27].

Many social media platforms such as Facebook, X, Reddit and others are becoming a venue for different types of discussions due to their accessibility, east of use, and how it can offer users to share their ideas in anonymous way [28]. Previous research has shown that the analysis of social media posts can reveal different types of trends among the public regarding different social and political events such as local and national elections [29]. The different opinions, emotions, and sentiment that online users may portray in their posts can show where they stand from different causes, events, or issues [30]. The community based social news website and forum have become very popular in the recent few years especially in the US. Reddit users engage with different communities and events in what is known as subreddits that focuses on certain topics and subjects. The platform became a valuable source for social media data where researchers can use to take a close look at the public opinion in certain topics, reveal different opinions, and highlight certain trends [17].

C. Challenges of Using Social Media Data

Contrary to the classical methods of opinion polling, data from social media is free of charge and only requires technical knowledge to extract needed information. Both researchers and practitioners can benefit from such features, especially when there are limited resources. Not only that but regarding data volume most polling would contain few hundreds and might get up too few thousand but when it comes to social media researchers can collect data from thousands or even millions of individuals that are willingly sharing their opinions, thoughts, and feelings regarding different issues, incidents, and events. Nonetheless, in some cases users tend to use short comments, use slang, emojis, or even sarcasm, which make it challenging to extract viewpoints about complex issues and topics [31, 32]. In addition, as discussed by Bessi al. [33] the existence of fake profiles and social media bots that can create fake comments and posts to distort the public sentiment. Bots and fake accounts can be used to magnify certain messages and sway some trends with or against. In their recent paper Bovet and Makse [18] mentioned that in the 2016 elections there were fake social media accounts that were used to provide support to certain political figures to enhance their visibility not only that but also spread incorrect information about rivals. Lastly as noted by Wilson et al. [34] social media platforms are mainly utilized by younger voters in comparison to older ones, thus analyzing results from such platforms might be biased towards the opinions of younger demographics.

D. Machine Learning and Sentiment Analysis in Election Prediction

To overcome analyzing a large amount of missy and unstructured data researchers have been leveraging machine learning algorithms that can digest and analyze a large amount of messy data and transform it into a structured format to analyze the extracted future to reveal hidden trends and highlight common themes [18]. In addition, Chauhan et al. [35] argue that machine learning models and techniques can be used to leverage social media data and predict the outcomes of future events such as electoral activities and the odds of certain candidates and their campaign to attract voters during elections. Using social media data that is dynamic and continuously updating will give a better picture of what is happening in comparison to traditional polling techniques that are limited to certain time slot and a relatively speaking smaller sample size that may not capture the diverse population opinion. Based on that we can conclude that employing machine learning models to analyze individuals' sentiment to better understand the public opinion and inclination can assist in better anticipating voters' behavior and choice on the ballots [15]. There are different types of ML models ranging from classical ones like Naïve Bayes (NB), Decision Trees (DTs), and Support Vector Machines (SVM) to more sophisticated models that consist of multiple functions. NB for example has been used in classification to analyze the sentiment of a given text. SVM on the other hand has been utilized to classify data sets that are high in dimensionality where there are many features (words) for consideration. Decision Trees on the other had have been commonly used where the researcher would like a clear reason for how a given text is being classified [35]. In addition, researchers utilized ensemble techniques that combine multiple models together to enhance their outcomes as if Random Forests are ensembles of DTs [19].

Advanced techniques known as Deep Learning (DL) that use multiple layers of Neural Networks (NNs) that mimic human intelligence in learning and analyzing data have been utilized by researchers. Deep learning models are particularly well suited for handling large-scale, high-dimensional data and have achieved significant breakthroughs in a variety of fields, including computer vision, Cybersecurity, and NLP [36]. Unlike traditional machine learning algorithms, which often require handcrafted features and specific domain expertise to perform well, deep learning models can automatically learn relevant features from raw data. Deep learning models excel at capturing patterns and relationships in data. In the context of NLP, deep learning has enabled significant advances in sentiment analysis [37]. An advanced deep learning model known as Long Short-Term Memory (LSTM) networks uses layers of recurrent neural networks (RNN). Those networks are ideal for analyzing temporal sequences of data [16, 38]. Recently, transformers and Large Language Models (LLMs) are considered cutting-edge in the field of NLP, particularly in tasks like language translations, text generation and text summarizations [39]. NLP have seen the rise of pre-trained transformer models, which are trained on vast amounts of text and then fine-tuned for specific tasks, e.g. BERT [40] and Generative Pre-trained Transformers (GPT) [41].

3. Methodology

A. Data Collection

We used Python, a high-level programming language, to collect large data sets pertaining to the 2024 US presidential election. In our data gathering, we focused on the main nominee of the Republican Party the former President Donald Trump and Vice President Kamala Harris that accepted the Democratic Party nomination after the withdrawal of President Joe Biden. As discussed above and due to the popularity of Reddit in political discussion especially in the United States we targeted different discussion forums or what is known as "subreddit" to collect data. To that extent, we leveraged via Python an Application Programming Interface (API) to gather online posts and comments from thousands of users that discussed Trump and Harris. We started the data gathering when Harris accepted the Democratic Party nomination after the withdrawal of Biden. To guarantee a wide spectrum of opinions and discussions we gathered data from several subreddits that would represent different ideologies to make sure that the analysis results won't be biased to one candidate over the other as voters are becoming more ideologically polarized in recent years.

B. Data Preprocessing

We applied systematic preprocessing on the data collected in the previous section. Data preprocessing is an important part of the NLP that transformed unstructured textual data into a structure format that can be used as an input for different mathematical models, which are the essence of machine learning models [42]. Preprocessing also involves cleansing the data to remove noise and ensure the consistency of input data that allows algorithms to focus on relevant features and patterns within the data [35]. We categorize posts as the ones that have discussed Trump vs the ones discussed with Harris. In addition, we have also labeled the battle ground states as swing states if the users that made the post indicated that they are a resident of one of those states. We made sure that in the post text we remove unrelated items such as web links, tags, and other unrelated items [43]. By eliminating those unrelated elements from the text data, we managed to simplify the data for the machine learning models. The next stage focused on excluding words and punctuation marks known as stop words from the text data. Stop words such as "the", "s", and "and" are frequently found in writing but do not provide valuable information or convey the emotions in the text [19], similarly punctuation marks [44]. In addition to the previous steps, we choose to uniform the text by changing it to lowercase for example "Trump” and "trump".

After the previous steps applied tokenization that breaks down each text into smaller components known as tokens that represent a word or phrase. This step is crucial as it enables machine-learning models to analyze word frequencies and their co-occurrences. This step is important for sentiment analysis and for models to effectively understand patterns [19, 32]. After that, we applied stemming and lemmatization methods to simplify words and reduce them to the root form. This method combines forms of words (like "vote", "votes," and "voting" becoming vote") to enhance model precision by grouping them as a single concept for analysis purposes. There was attention paid to emojis, and special characters commonly found in social media content due to their significance, in expressing emotions and sentiments; for instance, a smiling emoji typically conveys positivity while a sad face denotes negativity. Depending on what the model needs emojis were either taken out or turned into text labels to keep their meaning. Special characters, without information were deleted to make the cleaner [16]. In the processing stage, they considered negations like "not good". Doesn't support," which could greatly change the tone of a sentence. Addressing negations is crucial to understanding sentiments, avoiding situations where "not happy" is mistakenly labeled as positive. After these initial steps were carried out, the data set was converted into an organized layout suitable for more sophisticated sentiment analysis and machine learning purposes. Ensuring the accuracy of models. Deriving predictions from top quality data hinges upon the meticulous preprocessing stage [42].

C. Feature Extraction

Feature extraction was crucial in getting the text ready for sentiment analysis by changing text into numbers that machine-learning models could work with and understand effectively. In this research project, various feature extraction methods were used. Focused notably on Term Frequency Inverse Document Frequency (TF IDF). This technique assessed the significance of words in the data by studying how often they appear within documents, versus their occurrence across the entire collection of documents. The characteristics that were obtained were used as data, for intelligence algorithms to forecast the results of the 2024 United States presidential election by analyzing the attitudes expressed towards the prominent candidates Donald Trump and Kamala Harris. To boost the models capacity for understanding sentiment n-grams, like uni-grams,bi-grams, and Tri-grams were integrated. Sequences of words known as n-grams are used to capture more, than terms in text analysis tasks like sentiment detection; they help identify significant phrases, like "not supporting Trump" or "backing Harris for president." This method maintains the order of words to better understand the subtleties and nuances of the text’s meaning.

Furthermore, then ngrams part of speech (POS) tagging was utilized to categorize words into classes (for example nouns or verbs). This provided another level of depth to the examination by recognizing the functions that words fulfilled in sentences. For example, descriptive words signified sentiments strongly while action words, like "back" or "resist" disclosed user behaviors or inclinations. When we added POS tags to the features used by the model and analyzed comments, with them in mind it helped the model understand the comments better and led to accuracy in classifying sentiments on a sentiment analysis task. We then improved our sentiment analysis further by calculating sentiment scores using tools, like VADER or SentiWordNet on each post or comment. These tools gave scores to words based on their meanings. Positive words got higher scores, and negative words got lower ones. The combined sentiment ratings for every comment played a role in helping the model understand the general sentiment conveyed in the text content. When dealing with datasets techniques like Word2Vec or GloVe were used to grasp the connections, among words. Unlike TF IDF that views words as entities word embeddings, placed words in a vector space bringing similar meaning words closer together. For instance, "vote" and "support" could have vector representations that allow the model to grasp the connections between words and their contexts effectively. Word embeddings offer an understanding of word meanings and improve the model’s capacity to grasp intricate sentiment patterns.

Once the important features were identified and extracted from the data gathered in the study process, various techniques for selecting these features were employed to ensure that only the model used the pertinent details. Approaches, like Chi Square and Principal Component Analysis (PCA) were utilized to streamline the feature set by eliminating duplicate elements. This did not boost the model’s effectiveness and efficiency. Also safeguarded against overfitting. By homing in on the features, the model could concentrate on the crucial factors that shape voter opinions thereby enhancing its ability to predict accurately. Ultimately. In summary – the process of extracting features played a role in converting unprocessed text data into organized inputs that machine-learning models could leverage to forecast the outcome of the 2024 U.S presidential election effectively.

D. Model Building and Training

Using prelabeled Reddit data that contains more than fifty-eight thousand Reddit post various machine-learning models were built and adjusted carefully to guarantee that the final predictive model could accurately grasp and understand the sentiment expressed in the voters’ data. We start our experimentation with classical models such as Logistic Regression (LR), Decision Tree (DC), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF) to more sophisticated and advanced Deep Learning models that use different types of Neural Networks such the long short-term memory (LSTM). We have used Python to import prelabeled data and split it to 70% for training, 15% for validation, and 15% for testing. We have extracted different features from the training data using different techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, word embeddings, and Generative Pre-Training Transformer (GPT). Some models have been tuned with different hyperparameters to enhance their performance. For example, when it came down to Logistic Regression and SVM models’ adjustments were made on parameters such as regularization strength and kernel type while for Random Forests the focus was on optimizing the number of trees and their depth. On the other hand, for LSTM networks hyperparameters like the number of layers learning rate and batch size were adjusted for results. Techniques such as Grid Search and Random Search were employed to determine the combination of hyperparameters for each model in order to ensure the best performance.

Following the training phase was the evaluation of the models using both validation and test sets to gauge their predictive performance levels. The effectiveness of the models, in categorizing the online posts to negative sentiment towards the candidates was assessed using metrics like accuracy, precision, recall and F1 score (see Figure 1). Regarding LSTM networks, the temporal nature of user discussions was captured, enabling the analysis of how sentiment evolved over time. This approach helped the model to not only classify sentiment at a given moment but also detect trends leading up to Election Day, offering predictions that were dynamic and contextually aware. To avoid overfitting. A situation where a model excels on the training data but struggles with data. Methods like cross validation were utilized. Cross validation was employed to evaluate the performance of models across data subsets and guarantee their ability to adapt well with inputs. Moreover, regularization methods like L1 and L2 regularization were implemented in models such as Logistic Regression and SVM to discourage models and promote simpler and broadly applicable solutions.

E. Model Evaluation

We compared different models such as LSTM (Long Short-Term Memory), XGBoost (Extreme Gradient Boost) SVM (Support Vector Machine), CatBoost (Categorical Boost), Random Forest, and Logistic Regression (LR) using different metrics (as indicated in Table 1 and Figure 1). These metrics offer an insight into how each model can categorize sentiment into positive or negative categories while also considering neutrality, within the data set. Accuracy serves as a performance indicator by calculating the percentage of predictions made by each model; however, its interpretation may be deceiving when dealing with class distribution disparities. To offer a rounded viewpoint the F1 score combines precision and recall as a mean to assess the models especially in situations where both precision and recall are equally significant.

Table 1: Performance of Models on Validation Data

Model	Accuracy	Weighted F1-Score
LSTM	0.881	0.892
XGBoost	0.852	0.850
SVM	0.844	0.831
CatBoost	0.833	0.824
Random Forest	0.823	0.813
Logistic Regression	0.814	0.804
Decision Tree	0.801	0.792
KNN	0.762	0.753

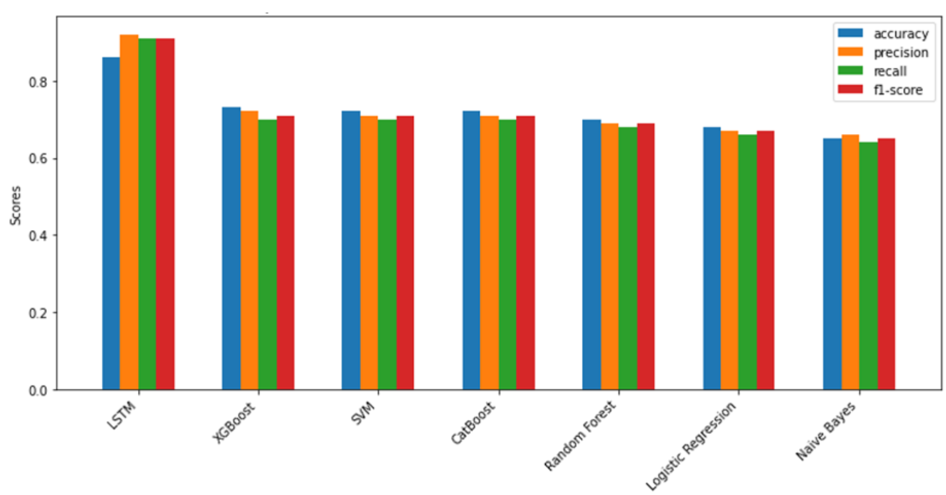

Figure 1. Comparison of Developed Models Performance

In sentiment classification tasks, across metrics like accuracy and precision among others LSTM emerges as the top performer outclass them all by capturing sequential patterns in text. A key factor in comprehending the context and flow of information is sentiment analysis. While XGBoost and SVM also show performance with rounded results across key evaluation criteria, like accuracy, precision, recall and F1 score (see Figure 1). These models show their strength in handling sentiment analysis tasks when dealing with uneven class distributions while also providing quick training times and consistent classification results. While CatBoost and Logistic Regression show results, in performance metrics compared to Random Forests in text analysis tasks and sentiment interpretation tasks; however, LSTM and XGBoost outperform them in handling changes in sentiment present within the dataset with greater accuracy and nuance perception capability. On the contrary, the KNN model exhibits the performance among all models due to its struggle in capturing details and contextual nuances inherent, in textual data resulting in decreased classification accuracy levels. In these assessments, it is evident that LSTM stands out as the choice, for sentiment analysis because of its adeptness in handling data and grasping contextual nuances effectively. However, different deep learning techniques were proposed in the literature to predict the results of the elections, for example, Brito [40] used MLP-BP and GRNN models in his experiments. These findings underline the effectiveness of deep learning methodologies in tasks requiring data analysis like text processing where interpreting context and word sequencing plays a role, in precise sentiment inference. As shown in Figure 2 below, the confusion matrix for the LSTM model shows strong performance in classifying positive, neutral, and negative sentiment accurately.

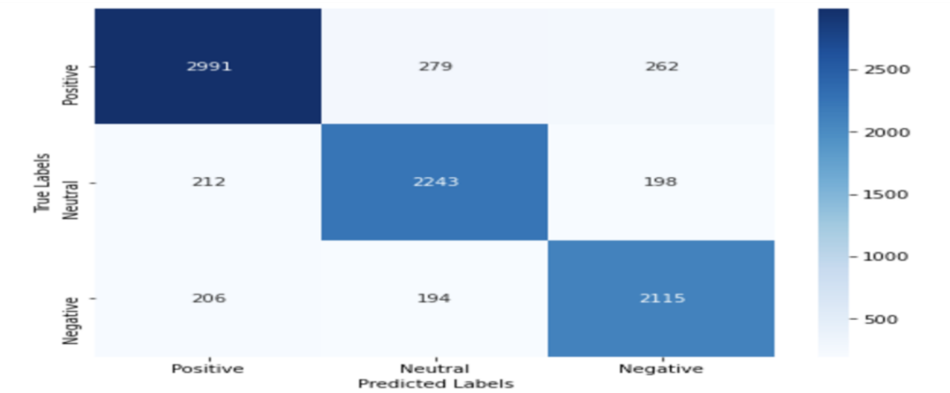

Figure 2. LSTM Confusion Matrix

4. Results and Discussion

After selecting the LSTM model, we applied the mode to analyze 28,268 Reddit posts that discussed the 2024 presidential candidates, particularly focusing on the public perception of Donald Trump and Kamala Harris and were downloaded using an API using Python script. A notable observation from the collected data is the preference for shorter comments, with a significant number of posts being under 100 characters and peaking around 50 characters. This indicates a tendency for Reddit users to engage in concise discussions when sharing their views on political matters (see Figure 3). Using the LSTM model classification model developed in this study, we evaluated the tone of these discussions and summarized the results (see Figures 4 and 5).

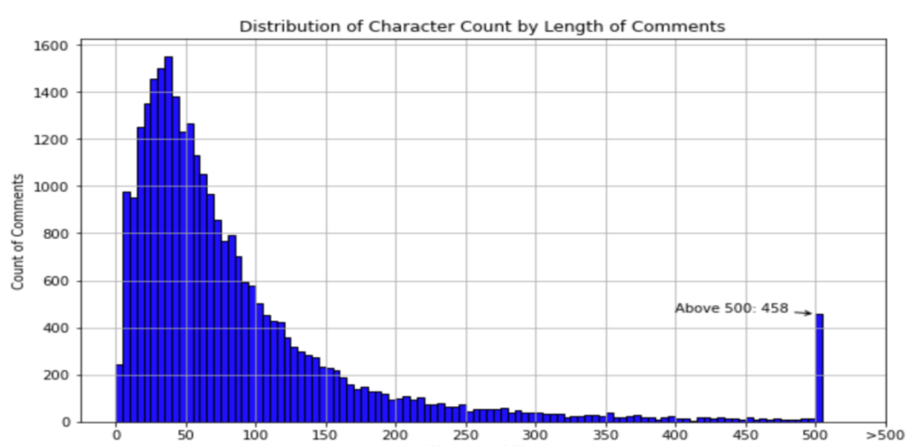

Figure 3. Distribution of Posts' Characters

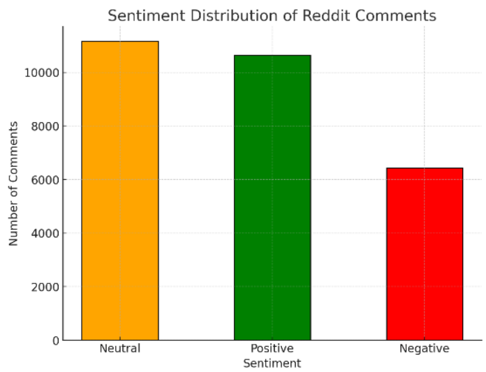

Figure 4. Sentiment Distribution

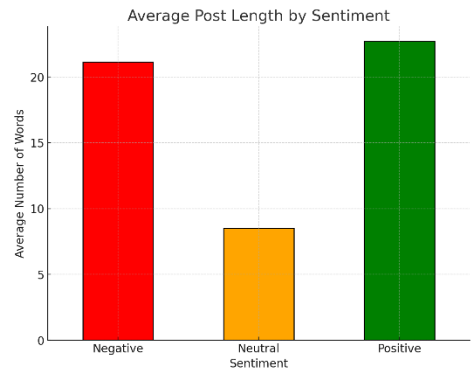

Figure 5. Post Length vs Sentiment

As seen in Figure 6, the findings provide a comparative look at the public sentiment towards both candidates on a national level. Kamala Harris was found to have a significantly higher proportion of positive sentiment, with approximately 40% of the comments expressing favorable views towards her. Donald Trump, in comparison, positive sentiment was slightly above 30%. This implies that Harris enjoys broader support in terms of positive discussions, reflecting stronger approval from the public regarding her candidacy. In terms of neutral sentiment, Harris and Trump were very close. On the other hand, Trump faced a larger proportion of negative sentiment, with over 40% of comments being critical of him compared to 30% for Harris. This indicates that public discourse around Trump tends to be more polarized.

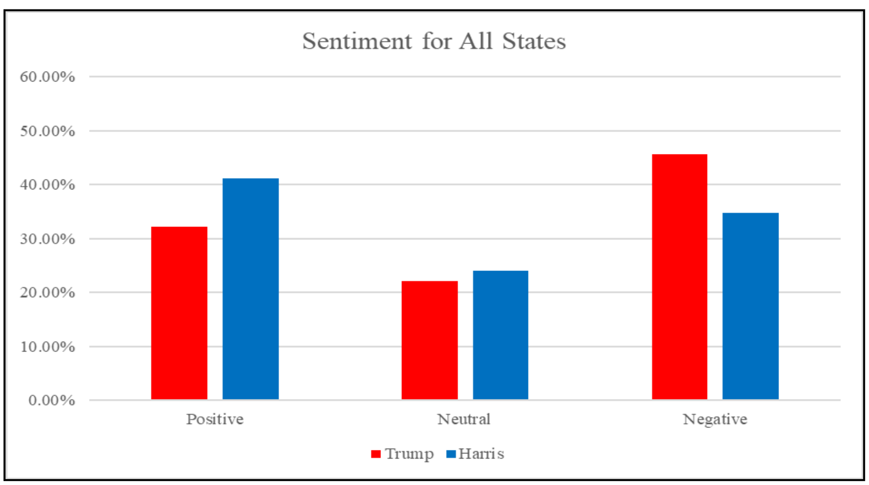

Figure 6. Sentiment Analysis

When trying to zoom the analysis of posts sentiment in key swing states, including Arizona, Georgia, Michigan, North Carolina, Nevada, Pennsylvania, and Wisconsin, we see a different picture. In these states, Trump garnered a higher proportion of positive sentiment at 40%, while Harris saw 35%. This suggests stronger support for Trump in these battleground areas, where voter sentiment is crucial. Harris, however, maintained a narrow edge in neutral sentiment, reflecting more balanced discussions about her candidacy. When it comes to negative sentiment in swing states Harris surpassed Trump, with 36% of the comments being critical of her compared to 32% for Trump. This contrasts with the national trend, where Trump had a higher proportion of negative sentiment overall. In these politically pivotal states, the sentiment towards Trump seems to be more favorable, reflecting a more competitive electoral climate.

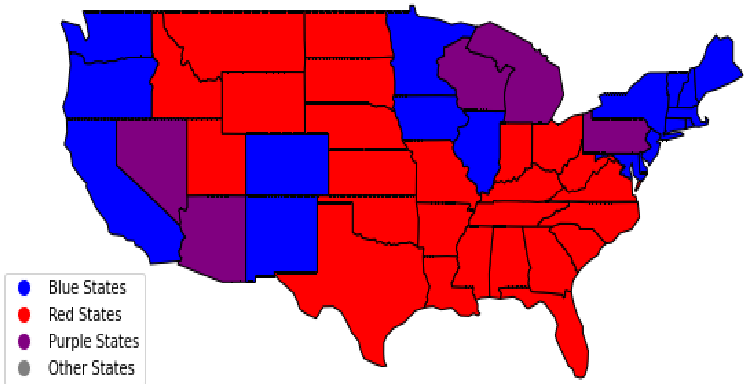

Figure 7. US states (projected Blue, Red, and Purple)

Our sentiment analysis revealed a contrast in how potential voters discuss their ideas, promises, and possible future policies. We found that the public sentiment as reflected in the Reddit posts since Harris became officially the main contender for the Democratic Party favors her over Trump with stronger positive and neutral overall sentiment. In addition, the posts show higher negative overall sentiment regarding Trump in comparison to Harris (as shown in Figure 8). In contrast and interestingly speaking when zooming to the swing states or what is also known as purple states Georgia, Nevada, Wisconsin, Michigan, Arizona and Pennsylvania our results suggests that although Harris may be slightly leading in the neutral posts, Trump is leading in positive sentiment posts in addition to that Harris has more negative sentiment than Trump. Although the national trend may favor Harris over Trump or discuss her more favorably than Trump the trends in the swing states which may be the ones determined the outcomes of the elections are favoring Trump over Harris. This brings back to our mind those 2016 elections where Hillary Clinton won the majority vote but lost the electoral vote to Trump.

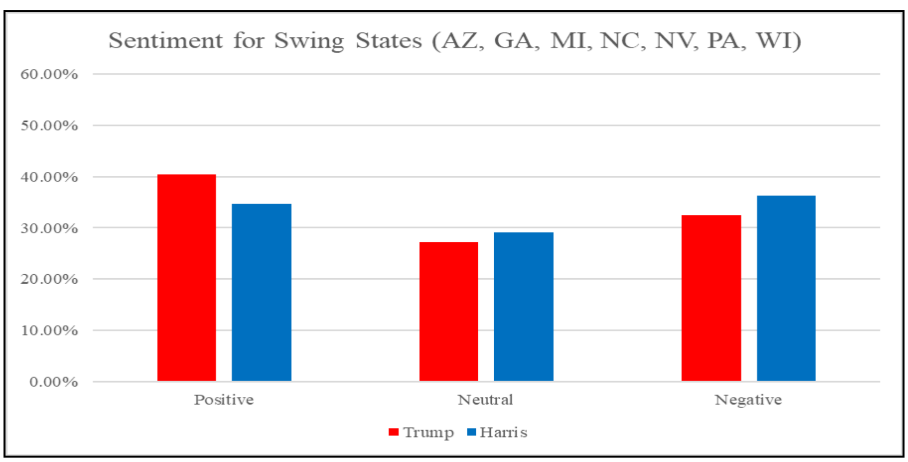

Figure 8. Sentiment Analysis for Swing States

5. Conclusion

In this paper, we used machine-learning models to analyze the sentiment of online Reddit posts that discussed the US 2024 presidential elections, mainly sentiment regarding the two main contenders President Trump and Vice President Harris. Using LTSM deep learning model, we successfully classified Reddit posts with an accuracy that exceeded 85%. The LTSM model was used to classify the posts for each contender into three categories (positive, neutral, and negative) then we use the aggregate number for each category to compare the candidates to predict the outcomes of the upcoming elections. Our study highlights the importance of social media analytics, particularly when related to political events and discussions. The developed predictive model has unveiled interesting trends in public sentiment at a macro and micro levels. As though the overall sentiment may favor Harris and sentiment in the swing states seem to be in Trump's favor, which may result in latter winning the upcoming elections. The findings of our study highlight the importance of taking a continual data set in comparison to a snapshot of data as more traditional polls tend to use. The tools and techniques employed in our study can be leveraged to other political events, as social media has become a powerful tool in shaping public opinion, especially among young voters.

6. Limitations and Future Work

In this article, we have focused on data that was extracted from Reddit, a community driven social media platform that has been used as a venue for financial, social, and political discussions and information sharing. Although the platform is a rich source for unstructured data, in particular to US politics future studies can also incorporate data from other platforms such X and Facebook. In addition, future studies can incorporate demographic data that can show more insights about how age, gender, and race could influence electoral outcomes. In addition, we have collected posts that are written in the English language. It is well known that today in the US elections the Hispanic vote is a significant factor that affects the outcome of the electoral votes in many states. The Hispanic/Latino community in the US is the biggest in terms of number and Future studies can collect and analyze posts written in Spanish.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] N. J. Spykman, *America's strategy in world politics: the United States and the balance of power*, Routledge, 2017.

[2] B. Heredia, J. D. Prusa, and T. M. Khoshgoftaar, "Location-based twitter sentiment analysis for predicting the US 2016 presidential election," in *Proc. The Thirty-First International Flairs Conference*, 2018.

[3] D. Ebanks, J. N. Katz, and G. King, "If a statistical model predicts that common events should occur only once in 10,000 elections, maybe it's the wrong model," presented at *39th Annual Meeting of the Society for Political Methodology*, 2022.

[4] A. J. Lichtman and A. Lichtman, *Predicting the next president: The Keys to the White House, 2024*, Rowman & Littlefield, 2024.

[5] R. C. Fair, "The effect of economic events on votes for president: 1984 update," *Political Behavior*, vol. 10, pp. 168-179, 1988.

[6] M. S. Lewis-Beck and T. W. Rice, "Forecasting presidential elections: A comparison of naive models," *Political Behavior*, vol. 6, pp. 9-21, 1984.

[7] R. Brody and L. Sigelman, "Presidential popularity and presidential elections: An update and extension," *Public Opinion Quarterly*, vol. 47, no. 3, pp. 325-328, 1983.

[8] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *Journal of Information Technology & Politics*, vol. 13, no. 1, pp. 72-91, 2016.

[9] C. Kennedy, et al., "An evaluation of the 2016 election polls in the United States," *Public Opinion Quarterly*, vol. 82, no. 1, pp. 1-33, 2018.

[10] S. Boulianne, C. P. Hoffmann, and M. Bossetta, "Social media platforms for politics: A comparison of Facebook, Instagram, Twitter, YouTube, Reddit, Snapchat, and WhatsApp," *New Media & Society*, 2024, doi: 14614448241262415.

[11] S. Lee and M. Xenos, "Social distraction? Social media use and political knowledge in two US presidential elections," *Computers in Human Behavior*, vol. 90, pp. 18-25, 2019.

[12] J. Massachs, et al., "Roots of Trumpism: Homophily and social feedback in Donald Trump support on Reddit," in *Proc. 12th ACM Conference on Web Science*, 2020.

[13] S. Stier, et al., "Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter," in *Studying Politics Across Media*, Routledge, 2020, pp. 50-74.

[14] A. H. Juma'h and Y. Alnsour, "Using social media analytics: The effect of President Trump's tweets on companies' performance," *Journal of Accounting and Management Information Systems*, vol. 17, pp. 100-121, 2018.

[15] N. Beauchamp, "Predicting and interpolating state-level polls using Twitter textual data," *American Journal of Political Science*, vol. 61, no. 2, pp. 490-503, 2017.

[16] H.-H. Nguyen, "Enhancing sentiment analysis on social media data with advanced deep learning techniques," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 5, 2024.

[17] A. Rajadesingan, P. Resnick, and C. Budak, "Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits," in *Proc. International AAAI Conference on Web and Social Media*, 2020.

[18] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature Communications*, vol. 10, no. 1, p. 7, 2019.

[19] K. Naithani and Y. P. Raiwani, "Realization of natural language processing and machine learning approaches for text-based sentiment analysis," *Expert Systems*, vol. 40, no. 5, p. e13114, 2023.

[20] A. Rajput, "Natural language processing, sentiment analysis, and clinical analytics," in *Innovation in Health Informatics*, Elsevier, 2020, pp. 79-97.

[21] H. Hapke, C. Howard, and H. Lane, *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*, Simon and Schuster, 2019.

[22] C. Redl, "Uncertainty matters: Evidence from close elections," *Journal of International Economics*, vol. 124, p. 103296, 2020.

[23] J. Pasek, "Predicting elections: Considering tools to pool the polls," *Public Opinion Quarterly*, vol. 79, no. 2, pp. 594-619, 2015.

[24] A. N. Philippou, "Why do polls fail? The case of four US presidential elections, Brexit, and two India general elections," arXiv preprint arXiv:2107.14166, 2021.

[25] A. Mitchell, et al., "In Western Europe, public attitudes toward news media more divided by populist views than left-right ideology," *Pew Research Center*, 2018, vol. 14.

[26] D. J. S. Oliveira, P. H. d. S. Bermejo, and P. A. dos Santos, "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls," *Journal of Information Technology & Politics*, vol. 14, no. 1, pp. 34-45, 2017.

[27] M. Felt, "Social media and the social sciences: How researchers employ Big Data analytics," *Big Data & Society*, vol. 3, no. 1, p. 2053951716645828, 2016.

[28] T. W. Smith, "Making the most of online discussion: A retrospective analysis," *International Journal of Teaching and Learning in Higher Education*, vol. 31, no. 1, pp. 21-31, 2019.

[29] M. Brown and T. Johnson, "Secular change in metamorphism and the onset of global plate tectonics," *American Mineralogist*, vol. 103, no. 2, pp. 181-196, 2018.

[30] A. Tumasjan, et al., "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. International AAAI Conference on Web and Social Media*, 2010.

[31] S. H. D. Kolagani, A. Negahban, and C. Witt, "Identifying trending sentiments in the 2016 US presidential election: A case study of Twitter analytics," *Issues in Information Systems*, vol. 18, no. 2, pp. 80-86, 2017.

[32] U. Yaqub, et al., "Analysis of political discourse on Twitter in the context of the 2016 US presidential elections," *Government Information Quarterly*, vol. 34, no. 4, pp. 613-626, 2017.

[33] A. Bessi and E. Ferrara, "Social bots distort the 2016 US presidential election online discussion," *First Monday*, vol. 21, no. 11-7, 2016.

[34] G. Wilson, et al., "Understanding older adults' use of social technology and the factors influencing use," *Ageing & Society*, vol. 43, no. 1, pp. 222-245, 2023.

[35] P. Chauhan, N. Sharma, and G. Sikka, "The emergence of social media data and sentiment analysis in election prediction," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 2601-2627, 2021.

[36] I. H. Sarker, "Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, p. 154, 2021.

[37] G. Gurung, R. Shah, and D. P. Jaiswal, "Recent challenges and advancements in natural language processing," in *Federated Learning for Internet of Vehicles: IoV Image Processing, Vision and Intelligent Systems*, Bentham Science Publishers, 2024, pp. 350-369.

[38] P. K. Yechuri and S. Ramadass, "Classification of image and text data using deep learning-based LSTM model," *Traitement du Signal*, vol. 38, no. 6, 2021.

[39] D. Luitse and W. Denkena, "The great transformer: Examining the role of large language models in the political economy of AI," *Big Data & Society*, vol. 8, no. 2, p. 20539517211047734, 2021.

[40] E. Brito and H. Iser, "MaxSimE: Explaining Transformer-based semantic similarity via contextualized best matching token pairs," in *Proc. 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[41] G. Yenduri, et al., "GPT (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, 2024.

[42] M. Abbas, *Analysis and Application of Natural Language and Speech Processing*, Springer, 2023.

[43] S. Kumari and Z. Ali, "Extracting feature requests from online reviews of travel industry," *Acta Scientiarum: Technology*, vol. 44, 2022.

[44] S. Chirgaiya, et al., "Analysis of sentiment-based movie reviews using machine learning techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 5, pp. 5449-5456, 2021.