Hackathon 2024 Project Report
Avengers

1. Introduction

Objective: Our project uses social media data to analyze and predict how political trends will develop by using large language models (LLMs) to identify and forecast emerging patterns.
Scope: Our analysis concentrated on short-term trends, utilizing a dataset comprising 14 days of tweets to generate predictive insights for the following weeks.
Relevance: By offering insights into the changing attitudes and worries of the public, trend prediction is essential to comprehending changes in public opinion. This is particularly useful in political contexts where prompt awareness of these changes can impact public relations campaigns, policy decisions, and campaign tactics.

2. Background and Literature Review

Previous Research: Gujral et al. [1] examined social media's political impact using LLMs for nuanced election predictions based on Twitter data. Williams et al. [2] found that LLMs can generate convincing election disinformation, raising concerns about their misuse in political contexts. More recently, Yu et al. [3] introduced a framework utilizing LLMs to predict election outcomes, validated with real-world and synthetic voter data.

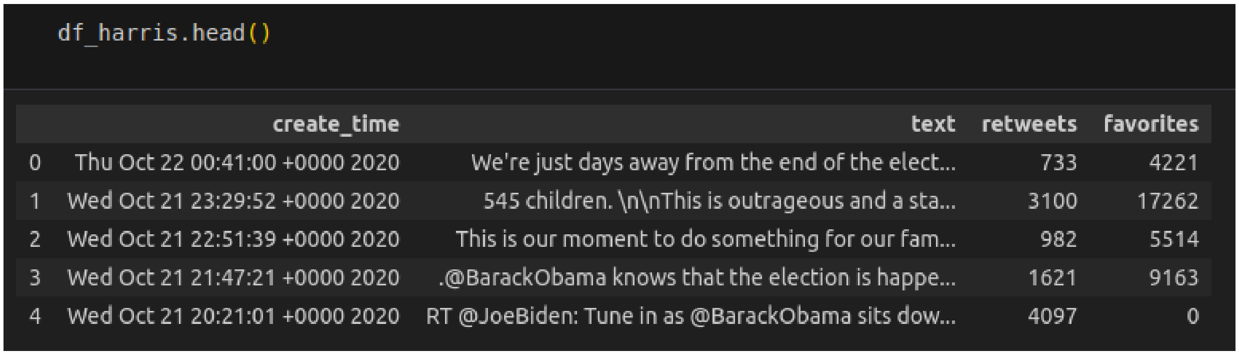
LLMs for Social Media Analysis: By processing enormous volumes of textual data from sources like social media, news articles, and forums, large language models (LLMs), like GPT, are used to assess and forecast trends. Accurate forecasting of trends in public opinion, consumer behavior, and political landscapes is made possible by their use of sophisticated natural language processing to recognize patterns, attitudes, and emerging subjects. LLMs also assist in giving organizations and brands a more thorough understanding of the factors that may affect their business.

Challenges: Inherent biases in user demographics, the dissemination of false information, and the massive amount of data created are some of the difficulties in predicting social media trends. Additionally, rapid shifts in trends, contextual nuances in language, noise from irrelevant content, and evolving user engagement dynamics complicate accurate analysis and forecasting, necessitating continuous adaptation of predictive models.

3. Methodology

Data Collection: Given our primary focus on social media platforms, we selected X, formerly known as Twitter, due to its prominence during the 2020 U.S. Presidential elections, where both candidates actively engaged with the electorate. The increasing popularity of social media as a news source further underscored its relevance. However, due to challenges encountered in collecting tweets directly from X, we utilized a pre-existing dataset available on Kaggle [4] to support our project.

Data Preprocessing: The dataset contained 4 JSON files namely biden_timeline.json, harris_timeline.json, pence_timeline.json, and trump_timeline.json. Each of these files in JSON format was loaded into a Python object. This object was then used to create a Pandas DataFrame, and several new columns (text, created_at, retweet_count, favorite_count) were added in the process. The resulting dataset looks like below (data from harris_timeline.json)



```
df_harris.head()
```

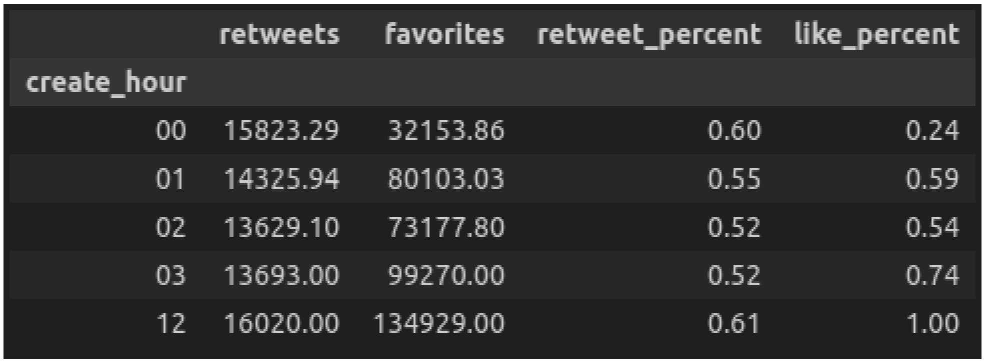
	create_time	text	retweets	favorites
0	Thu Oct 22 00:41:00 +0000 2020	We're just days away from the end of the elect...	733	4221
1	Wed Oct 21 23:29:52 +0000 2020	545 children. \n\nThis is outrageous and a sta...	3100	17262
2	Wed Oct 21 22:51:39 +0000 2020	This is our moment to do something for our fam...	982	5514
3	Wed Oct 21 21:47:21 +0000 2020	.@BarackObama knows that the election is happe...	1621	9163
4	Wed Oct 21 20:21:01 +0000 2020	RT @JoeBiden: Tune in as @BarackObama sits dow...	4097	0

By importing the NLTK library in Python for Natural Language Processing, a set of predefined stopwords from the NLTK data repository was downloaded. The main reason was to eliminate common words that do not contribute significant meaning to the text. The Punkt tokenizer model was used for tokenizing text into sentences and words, which aided in breaking down the text into manageable pieces for analysis.

In addition to stopwords, newline characters, URLs, and non-ASCII characters were also removed from the text. The resulting lines were converted to lowercase and tokenized into words.

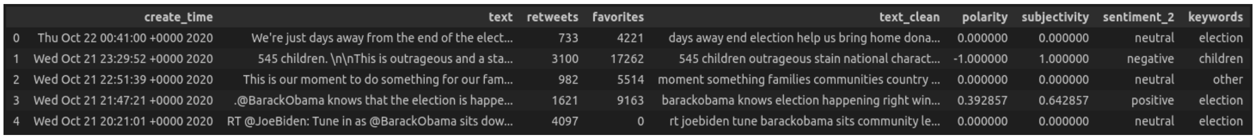
We used the TextBlob library to find the sentiment polarity with a floating range between -1.0 and 1 where -1.0 indicates a negative sentiment, 1 indicates a positive sentiment and 0 indicates a neutral sentiment. We also found the sentiment subjectivity ranging from 0 (objective statements) to 1 (subjective statements).

We then proceeded to analyze the average retweets and the favorites by hour of the day. Using this, we calculated the maximum retweets and favorites for the average retweets, and favorites which were grouped by the hour. The percentage of each hour's retweets and favorites relative to the maximum values was computed.



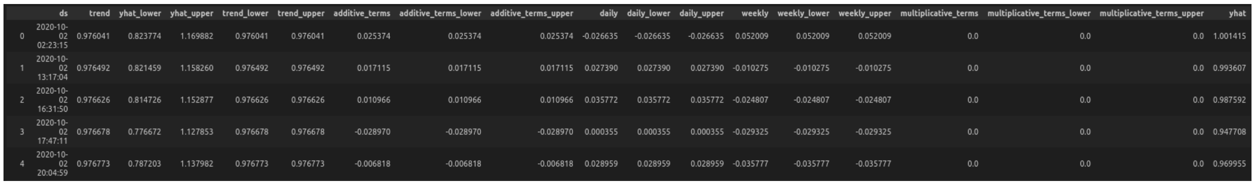
	retweets	favorites	retweet_percent	like_percent
create_hour				
00	15823.29	32153.86	0.60	0.24
01	14325.94	80103.03	0.55	0.59
02	13629.10	73177.80	0.52	0.54
03	13693.00	99270.00	0.52	0.74
12	16020.00	134929.00	0.61	1.00

Each cleaned tweet was categorized based on the presence of keywords such as "election" and "children".



	create_time	text	retweets	favorites	text_clean	polarity	subjectivity	sentiment_2	keywords
0	Thu Oct 22 00:41:00 +0000 2020	We're just days away from the end of the elect...	733	4221	days away end election help us bring home dona...	0.000000	0.000000	neutral	election
1	Wed Oct 21 23:29:52 +0000 2020	545 children. \n\nThis is outrageous and a sta...	3100	17262	545 children outrageous stain national charact...	-1.000000	1.000000	negative	children
2	Wed Oct 21 22:51:39 +0000 2020	This is our moment to do something for our fam...	982	5514	moment something families communities country ...	0.000000	0.000000	neutral	other
3	Wed Oct 21 21:47:21 +0000 2020	.@BarackObama knows that the election is happe...	1621	9163	barackobama knows election happening right win...	0.392857	0.642857	positive	election
4	Wed Oct 21 20:21:01 +0000 2020	RT @JoeBiden: Tune in as @BarackObama sits dow...	4097	0	rt joebiden tune barackobama sits community le...	0.000000	0.000000	neutral	election

We used the Prophet library designed for forecasting time series data, which created a predictive model based on the data. We used this data to create the required political trend forecast. A sample of the populated data is given below,



	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_terms	additive_terms_lower	additive_terms_upper	daily	daily_lower	daily_upper	weekly	weekly_lower	weekly_upper	multiplicative_terms	multiplicative_terms_lower	multiplicative_terms_upper	yhat
0	2020-10-02 02:23:15	0.976041	0.823774	1.169882	0.976041	0.976041	0.025374	0.025374	0.025374	-0.026635	-0.026635	-0.026635	0.052009	0.052009	0.052009	0.0	0.0	0.0	1.001415
1	2020-10-02 13:17:04	0.976492	0.821459	1.158260	0.976492	0.976492	0.017115	0.017115	0.017115	0.027390	0.027390	0.027390	-0.010275	-0.010275	-0.010275	0.0	0.0	0.0	0.993607
2	2020-10-02 16:31:50	0.976626	0.814726	1.152877	0.976626	0.976626	0.010966	0.010966	0.010966	0.035772	0.035772	0.035772	-0.024807	-0.024807	-0.024807	0.0	0.0	0.0	0.987592
3	2020-10-02 17:47:11	0.976678	0.776672	1.127853	0.976678	0.976678	-0.028970	-0.028970	-0.028970	0.000355	0.000355	0.000355	-0.029325	-0.029325	-0.029325	0.0	0.0	0.0	0.947708
4	2020-10-02 20:04:59	0.976773	0.787203	1.137982	0.976773	0.976773	-0.006818	-0.006818	-0.006818	0.028959	0.028959	0.028959	-0.035777	-0.035777	-0.035777	0.0	0.0	0.0	0.969955

4. Implementation

Tools and Technologies: The tools we used to clean our dataset and predict the trend for the 2020 US Presidential elections are listed below,

- Python
- Visual Studio Code

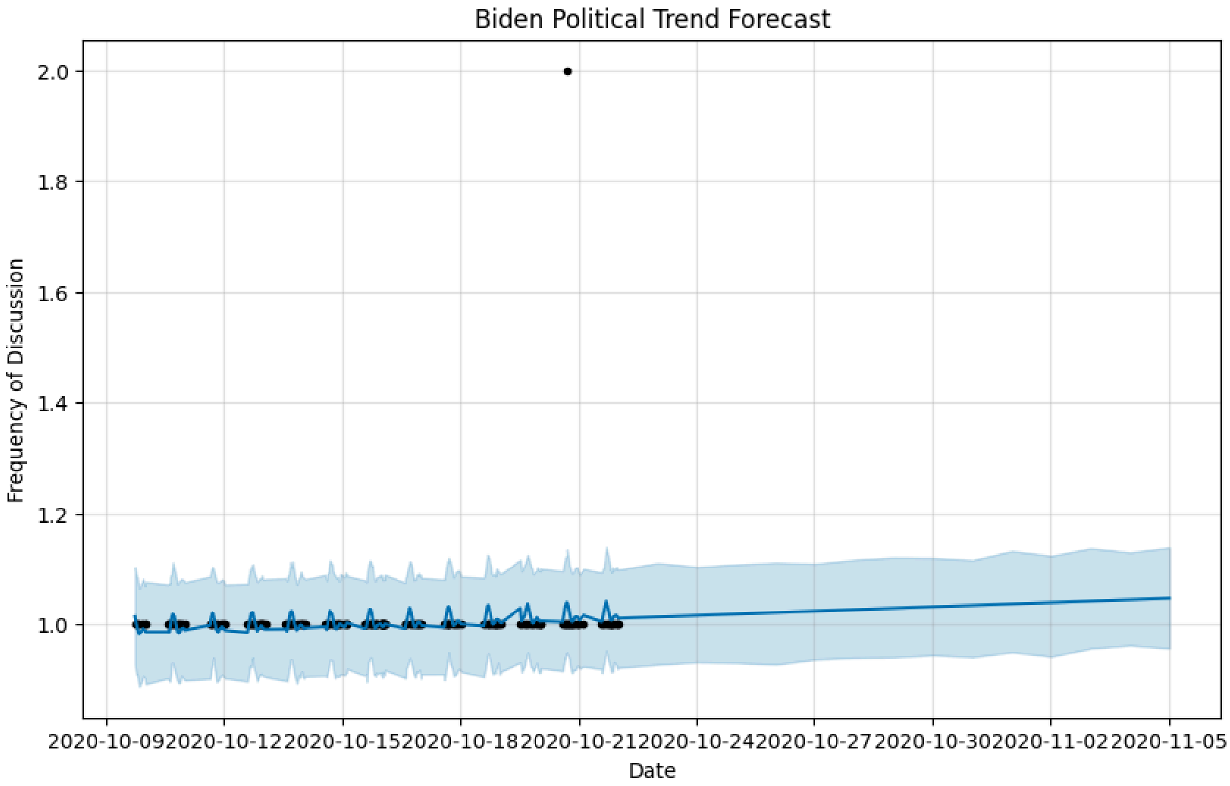
A few libraries we used are as follows,

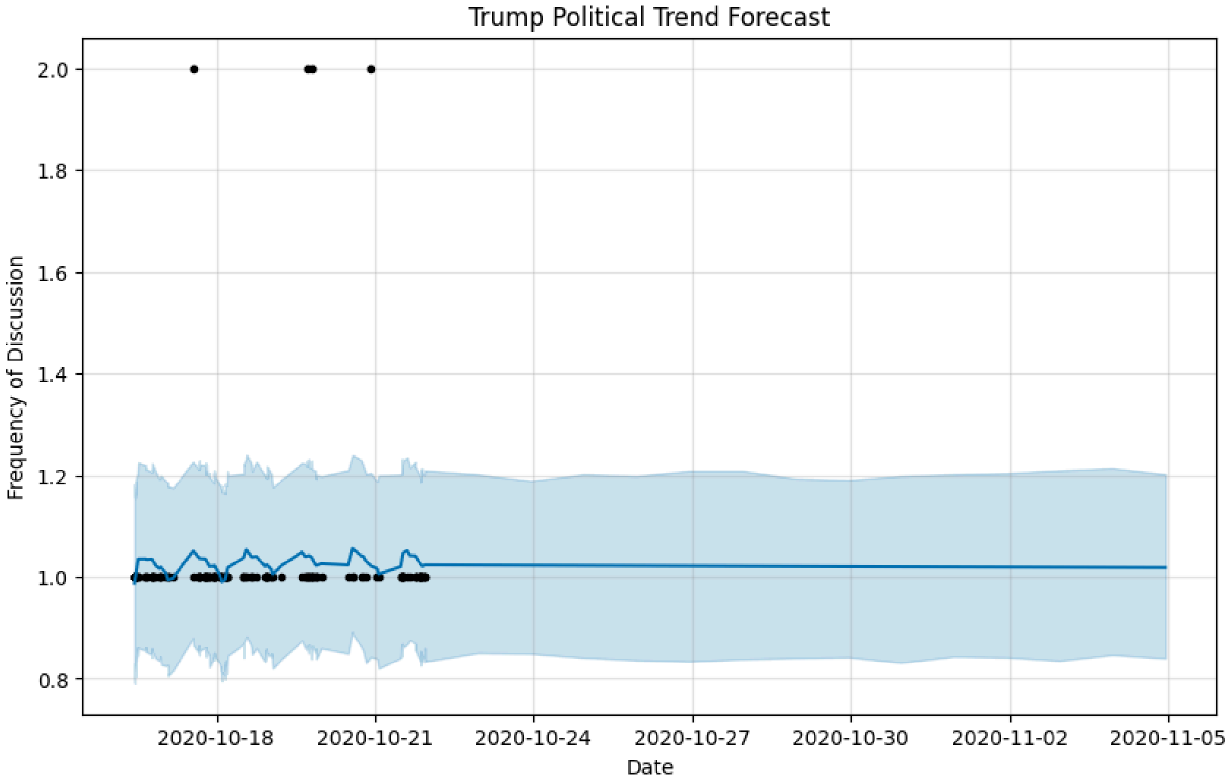
- NLTK
- Textblob
- Matplotlib
- Pandas

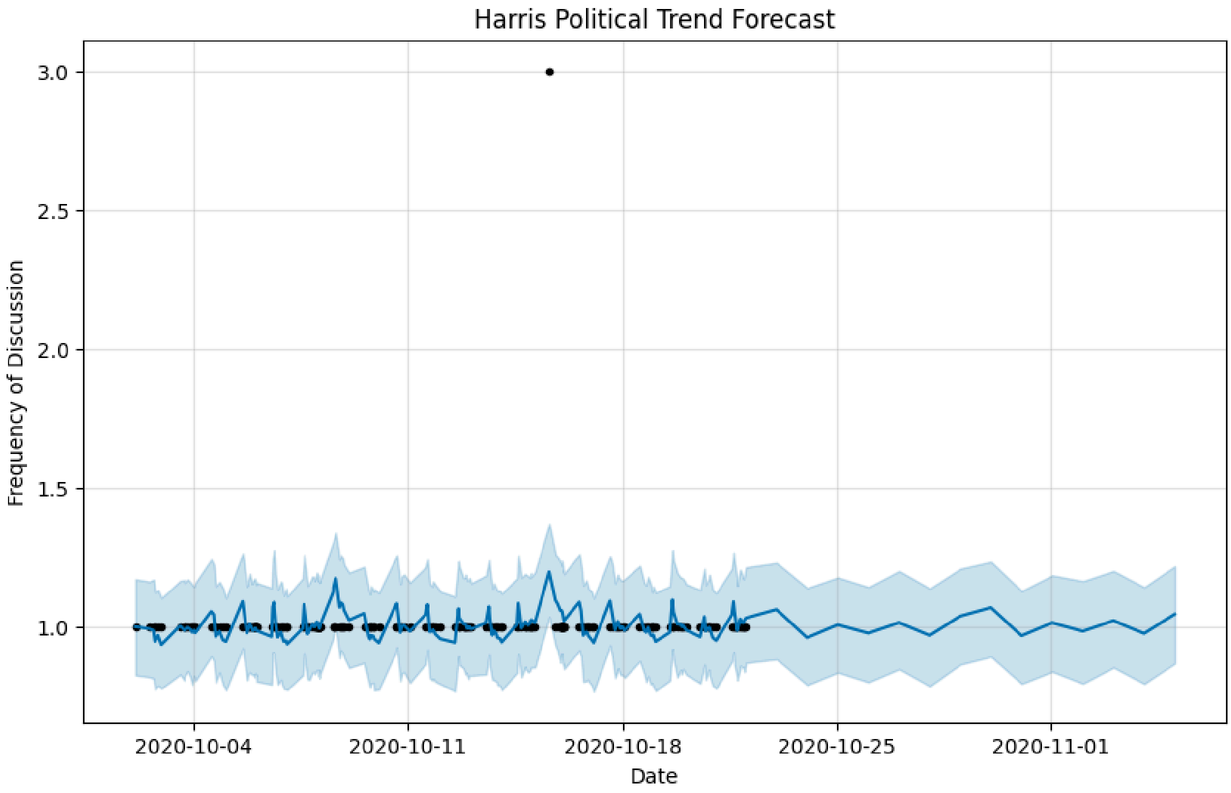
The most prominent tool we used for prediction is Prophet. Prophet is an open-source forecasting tool developed by Facebook specifically for time series data that may contain seasonality, trend shifts, or missing data. It is highly effective for quickly generating accurate predictions, such as identifying and projecting social media trends over specified periods.

5. Results

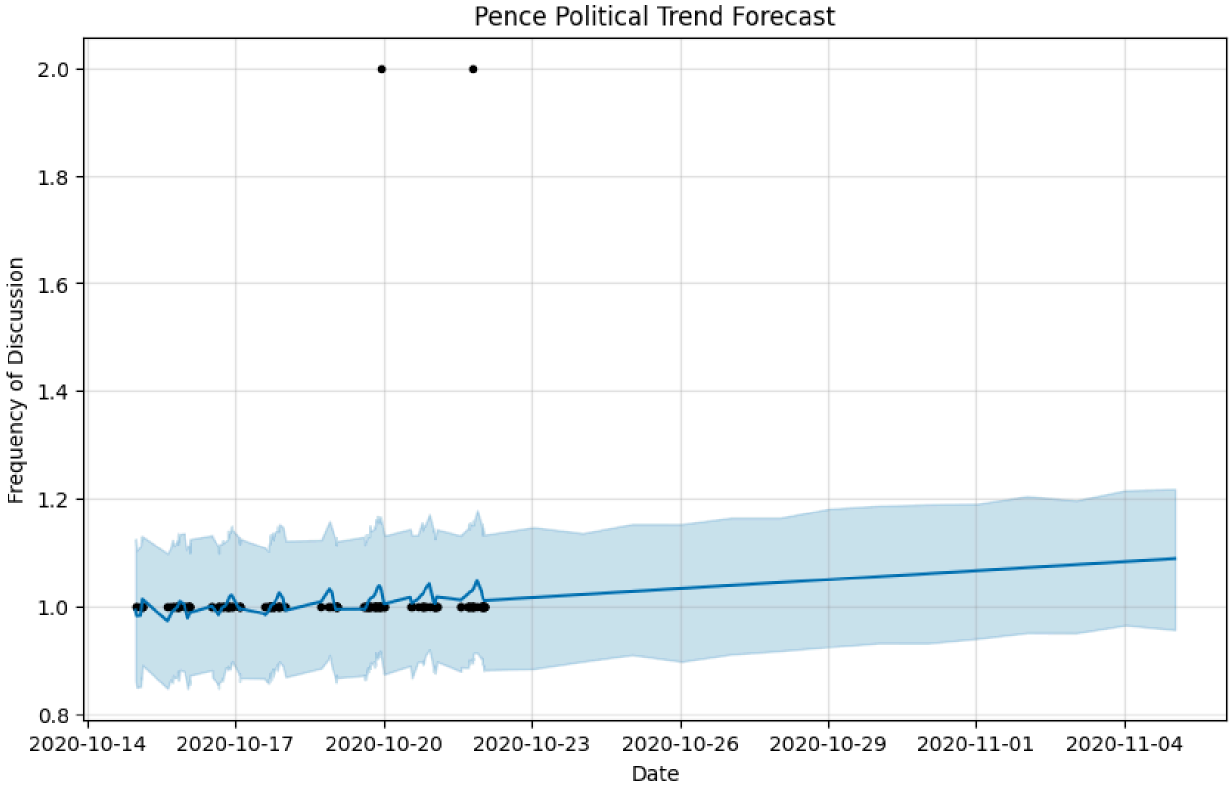
Trend Analysis: We plotted the trends using the tweets by both the presidential and vice-presidential candidates.


Biden Political Trend Forecast
Frequency of Discussion
2.0
1.8
1.6
1.4
1.2
1.0
2020-10-09
2020-10-12
2020-10-15
2020-10-18
2020-10-21
2020-10-24
2020-10-27
2020-10-30
2020-11-02
2020-11-05
Date


Trump Political Trend Forecast
Frequency of Discussion
2.0
1.8
1.6
1.4
1.2
1.0
0.8
2020-10-18
2020-10-21
2020-10-24
2020-10-27
2020-10-30
2020-11-02
2020-11-05
Date



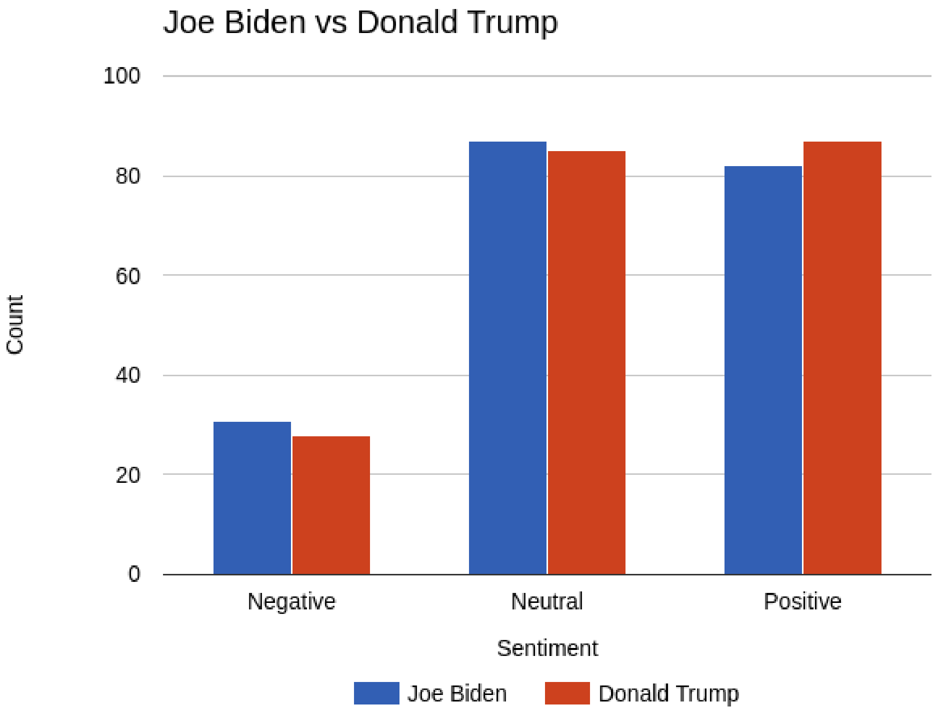
Harris Political Trend Forecast
Frequency of Discussion
3.0
2.5
2.0
1.5
1.0
2020-10-04
2020-10-11
2020-10-18
2020-10-25
2020-11-01
Date

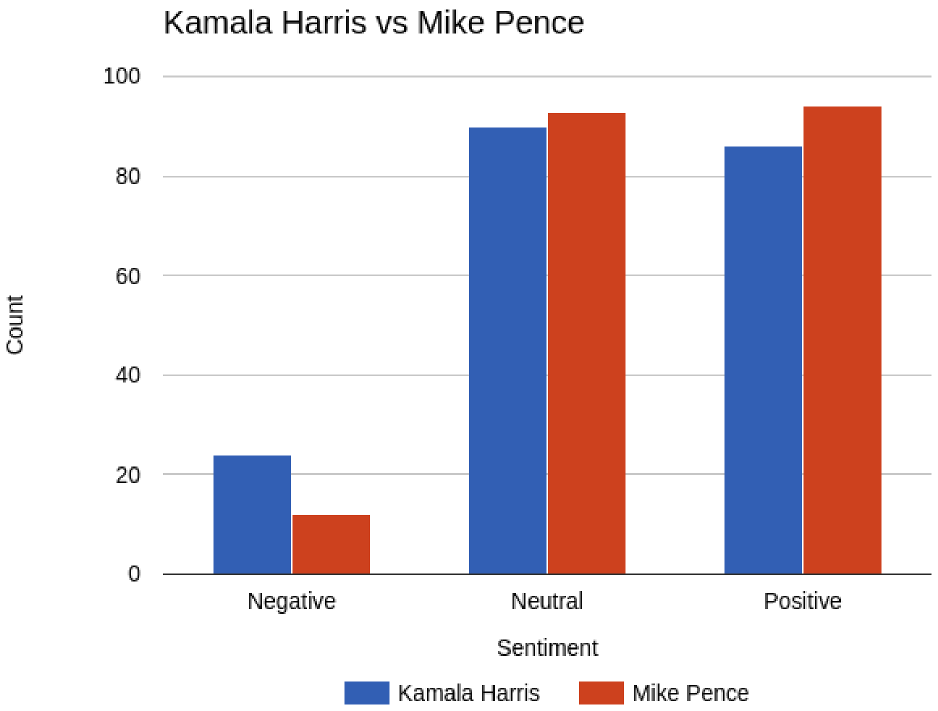


Pence Political Trend Forecast
Frequency of Discussion
2.0
1.8
1.6
1.4
1.2
1.0
0.8
2020-10-14
2020-10-17
2020-10-20
2020-10-23
2020-10-26
2020-10-29
2020-11-01
2020-11-04
Date

Prediction of Trends: The model predicts that Biden's trends will continue to rise steadily, while Pence's trends show an even more pronounced upward tendency. While Harris's trajectory shows erratic swings, Trump's track stays steady with no notable shifts. The model has moderate confidence in the stability and variability of Trump's and Harris's trajectories, respectively, whereas it has high confidence in Biden and Pence's upward tendencies.

Sentiment Count Comparison: We have also plotted the number of sentiment counts against both the presidential and vice-presidential candidates as shown below.





6. Discussion

Insights: Both Biden and Pence show significant political momentum, according to the trend study, with Pence's engagement rising more sharply. In the meantime, the variations in Harris's trend and the consistency of Trump's point to varying degrees of public interest or sentiment stability, providing information about contemporary political dynamics.

Limitations: Since X has placed limitations on the use of their API, gathering data from them was the most difficult task. Although we used pre-existing datasets, we are unable to confirm their legitimacy.

7. Conclusion and Future Work

Summary: The main conclusions show that social media trends for Biden and Pence have been steadily rising, with Pence's gain being considerably more pronounced, indicating that public interest in these leaders is increasing. While Harris's trend fluctuates, reflecting shifting public mood, Trump's trend stays steady, demonstrating sustained interest. All things considered, the model successfully reflected these dynamics, offering trustworthy forecasts of short-term political trends as well as practical insights into shifting public interest levels.

Future Directions: To improve trend robustness and expand insights, future developments might involve gathering data from other social media networks. Additionally, we can gather data over a longer period of time, which will improve the predicted accuracy of the model. Sentiment analysis and real-time data updates are examples of more sophisticated methods that potentially increase responsiveness to abrupt changes in public interest.

8. References

[1] Gujral, P., Awaldhi, K., Jain, N., Bhandula, B., & Chakraborty, A. (2024). Can LLMs Help Predict Elections? (Counter) Evidence from the World's Largest Democracy. arXiv preprint arXiv:2405.07828.
[2] Williams, A.R., Burke-Moore, L., Chan, R.S.Y., Enock, F.E., Nanni, F., Sippy, T., Chung, Y.L., Gabasova, E., Hackenburg, K. and Bright, J. (2024). Large language models can consistently generate high-quality content for election disinformation operations. arXiv preprint arXiv:2408.06731.
[3] Yu, C., Weng, Z., Li, Z., Hu, X., & Zhao, Y. (2024). Will Trump Win in 2024? Predicting the US Presidential Election via Multi-step Reasoning with Large Language Models. Predicting the US Presidential Election via Multi-step Reasoning with Large Language Models (October 22, 2024).
[4] Kaggle, United States Political Tweets. Twitter timeline of most prominent American politicians:
https://www.kaggle.com/datasets/tunguz/united-states-political-tweets/data?select=pence_timeline.json