Sentiment analysis on twitter's american users during the 2020 USA presidential election

***Abstract*—The dynamics of the 2024 U.S. presidential election have been notably altered by the unexpected withdrawal of President Joe Biden, resulting in Vice President Kamala Harris stepping in as the Democratic nominee. This shift has intensified public and media engagement, particularly on social media platforms like Twitter. This research explores the evolving electoral landscape by analyzing tweets directed at Donald Trump and Joe Biden during the 2020 election cycle, using Kamala Harris's recent entry as a contemporary lens for understanding political sentiment. The study employs Natural Language Processing (NLP) techniques for data preprocessing and applies sentiment analysis via the VADER tool to gauge emotional tones in user-generated content. Additionally, Latent Dirichlet Allocation (LDA) is utilized to identify prevalent topics and themes. The findings aim to elucidate how social media has reshaped public sentiment and discourse, revealing insights into voter polarization and candidate influence.**

***Index Terms*—Sentiment analysis , Topic modelling, Election, Twitter, NLTK, VADER, LDA.**

I. Introduction

The current electoral race underway in the United States is becoming truly unique. What initially appeared to be another contest between Republican Donald Trump and Democrat Joe Biden, the current president, took an unexpected turn when Biden stepped aside due to health problems, allowing Vice President Kamala Harris to take his place. Social media has been in fermet lately due to Biden's numerous gaffes, significantly increasing interactions, especially on platforms like Twitter. However, with Biden's withdrawal, the situation has changed. While Trump's victory once seemed certain, the entry of Kamala Harris has brought the race for the White House back into balance.

In recent years, the advent of social media has revolutionized how we evaluate public sentiment and opinion on a multitude of topics, particularly in situations as critical as political elections. Platforms like Twitter, Facebook, and Instagram have become virtual town squares where millions of users freely express their views, reactions, and preferences in real-time.

This digital landscape offers unprecedented access to a wealth of public debate that was previously inaccessible or constrained to traditional polling methods. For political elections, social media serves as a dynamic and influential arena where candidates, authorities, and voters converge to discuss policies, candidates' platforms, and the latest developments. Unlike traditional surveys that sample a limited number of respondents, social media aggregates opinions from a various and global audience, providing a comprehensive snapshot of public sentiment across demographics, geographies, and ideological spectrums.

The aim of my research is to understand the sentiment and mood of users posting tweets directed at either Biden or Trump. I seek to analyze the language used to determine whether these political candidates tend to polarize the political thoughts of these users or if their opinions are more ambiguous. Given that the election is set for November 5, 2024, and the situation within the Democratic Party is unstable, I have decided to analyze the 2020 election, which also featured Biden and Trump as direct opponents (although, as mentioned earlier, Kamala Harris has now stepped in).

To achieve this goal, the study will follow several methodological steps:

- Data Collection
- Data Preprocessing
- Sentiment Analysis
- Topic Modeling
- Interpretation of the results

This introduction sets the stage for examining how social media has transformed the way we measure public sentiment and opinion, particularly during significant events such as political elections.

II. Literature review

The use of social media as a tool for analyzing public sentiment has gathered significant academic attention in recent years. Studies have shown that platforms like Twitter, Facebook, and Instagram offer a rich source of real-time data that can be leveraged to understand public opinion on a wide array of topics, including political events. According to Tumasjan et al. (2010) [7], Twitter can serve as a valid indicator of political sentiment and election outcomes, providing insights that traditional polls may miss due to their limited reach and delayed reporting.

The 2020 US Presidential Election between Donald Trump and Joe Biden further underscored the importance of social media in political communication. Research by Barber´a (2020) [1] demonstrated that social media engagement could influence public opinion and voter behavior, particularly during critical periods of the election cycle. Additionally, studies by Bovet and Makse (2019) [3] highlighted how misinformation and echo chambers on social media could distort public perception and polarize voters.

Political polarization is on the rise not only in the United States, but also across the world. Today political elites, elected officials and everyday people are polarized. There are two

distinct forms of political polarization. The first is ideological polarization, which is the divergence of political opinions, beliefs, attitudes, and stances of political adversaries [4].The second is affective polarization, which is based on work considering the role of identity in politics and how identity salience within groups (e.g. political parties) can exacerbate out-group animosity. Affective polarization assesses the extent to which people like (or feel warmth towards) their political allies and dislike (or feel lack of warmth towards) their political opponents.

Higher levels of polarization can be beneficial for society – predicting higher levels of political participation, and perceptions of electoral choice . However, political polarization can also be bad for democracy, increasing the centralization of power, congressional gridlock, and making citizens less satisfied (Wagner, Citation2021). Previous work has also highlighted interpersonal implications of polarization, including an unwillingness to interact with people with opposite political opinions, and dehumanization towards (Mason, Citation2018) political adversaries.

Given that people are unwilling to engage in day-to-day interactions with their political adversaries, many build their impressions of opponents via the media – meaning (social) media is increasingly shaping how we perceive the political environment. As media has become more fragmentated, people have become more polarized both ideologically and affectively. However, media may not always have a polarizing effect on viewers. Some suggest social media and traditional media have no effect on political polarization.

Sentiment analysis has emerged as a powerful method for interpreting the emotional tone of social media posts. The VADER tool, developed by Hutto and Gilbert (2014) [5], is particularly effective for analyzing sentiments expressed in short, informal texts such as tweets. VADER's ability to capture both polarity (positive, negative, neutral) and intensity of sentiment makes it a valuable tool for political sentiment analysis.

Since the themes of debate can significantly affect the level of polarization among the population, Topic modeling, particularly Latent Dirichlet Allocation, is another essential technique used to identify the main topics within large datasets of text. Blei, Ng, and Jordan (2003) [2]introduced LDA as a method for discovering the underlying structure in text corpora by classifying words into topics. This approach has been applied in various studies to analyze political discourse on social media. For instance, Roberts et al. (2014) [6] utilized LDA to uncover topics related to policy issues and political ideologies discussed on Twitter during election campaigns.

In the context of the 2020 presidential elections in the USA, this study aims, using the techniques mentioned above, to classify the language of American users' tweets in order to understand whether one candidate rather than the other (in this case Biden or Trump) tends to polarize public opinion or, on the contrary, do not clearly reveal the individual thoughts of users. Furthermore, this research may be particularly useful in the future to study how user sentiment has changed, through tweets, after Joe Biden's resignation and Kamala Harris's entry into the scene.

III. Research Methodology

A. Data collection

For this study, a dataset sourced from Kaggle containing a total of over 900,000 tweets discussing Donald Trump and approximately 800,000 tweets discussing Joe Biden was selected. These tweets were collected over nearly a month, spanning from October 15, 2020, to November 9, 2020. The dataset includes variables such as timestamp, tweet content, likes, city, country, and state. Due to the presence of numerous NaN (missing values) values, extensive data cleaning was performed, resulting in a refined dataset comprising 389,106 tweets in total—175,797 about Joe Biden and 213,309 about Donald Trump.

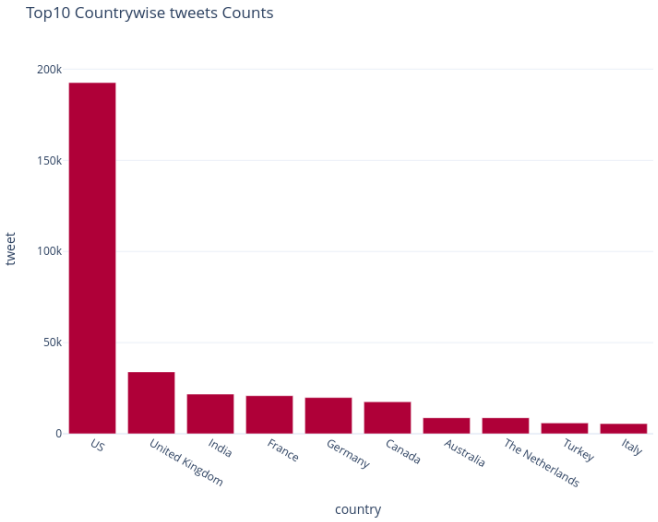

Fig. 1: Top 10 countries with the highest number of tweets.

During the initial exploratory analysis phase, the focus was on deriving meaningful insights from the available data. For example in 2 illustrates the distribution of tweets by country, highlighting the top ten countries with the highest tweet volumes. Unsurprisingly, the USA stands out with the highest number of tweets, followed by the UK, India, France, and Germany, with similar tweet frequencies among the latter three countries.

Delving deeper into the analysis, choropleth maps focused on the USA showcase which states generated the highest numbers of tweets regarding Trump (depicted in red) and Biden (in blue). Additionally, a comparative view displays the total tweet counts per state. These visualizations provide a geographical perspective on the intensity and distribution of Twitter discussions surrounding the two candidates during the specified timeframe.

B. Data preprocessing

Text processing is a crucial step in Natural Language Processing that prepares raw text data for analysis by transforming

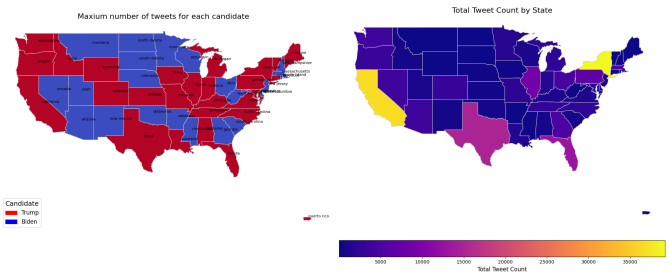

Fig. 2: left map: cities with the higher number of tweets related to Trump (red) and Biden (blue); right map: total number of tweets

it into a clean and structured format. These steps are essential for ensuring the data is accurate, relevant, and ready for subsequent analyses such as sentiment analysis and topic modeling. Basically text processing consists into two main steps:

- Data cleaning
- Tokenization

Data cleaning is a critical step to ensure the data is accurate and relevant for analysis by removing URLs, lowercasing text, stripping special characters, tokenizing text, lemmatizing words, and excluding common and election-specific stopwords to highlight meaningful content. The purpose of these steps is to refine the dataset, ensuring that the text data is clean, consistent, and ready for further analysis.

Tokenization is the process of splitting the cleaned text into individual tokens or words. This step is essential for analyzing word frequency and for performing further NLP tasks such as sentiment analysis and topic modeling.

These preprocessing steps are foundational for any NLP task, ensuring that the input data is of high quality and suitable for extracting meaningful insights.

C. Sentiment Analysis

After the data cleaning and tokenization steps, the next crucial phase is sentiment analysis which involves determining the emotional tone behind a body of text. This is a valuable tool for understanding public sentiment towards specific topics, such as political candidates during an election. In this study, we focus on the 2020 US Presidential Election between Donald Trump and Joe Biden, utilizing VADER to perform sentiment analysis on tweets related to these candidates.

VADER is a sentiment analysis tool that is particularly well-suited for analyzing social media text, especially short texts like tweets. It uses a combination of a pre-defined lexicon and grammatical rules to evaluate the sentiment conveyed in the text (Hutto and Gilbert, 2014). This makes VADER an excellent choice for this study, given the nature of Twitter data. VADER assigns sentiment intensity scores to words in a text and accounts for the context provided by degree modifiers and negations to determine the overall sentiment intensity and polarity of the text. It generates a compound sentiment score, a normalized, weighted composite score ranging from -1 (most negative) to +1 (most positive), which provides a summary of the overall sentiment of the text.

For this analysis, we applied VADER to the cleaned and preprocessed dataset. Each tweet is processed by VADER to obtain a sentiment score, with the compound score being used as the primary indicator of sentiment. Based on these compound scores, tweets are categorized into positive, negative, and neutral sentiment groups. This classification helped us understand the general sentiment trends related to each candidate. By leveraging VADER for sentiment analysis, it is possible to quantify the emotional tone of tweets related to the 2020 US Presidential Election, providing insights into public sentiment dynamics and how different segments of the population perceived the candidates. Understanding these sentiment trends is crucial for examining how social media reflects and potentially influences public opinion during significant political events.

D. Topic modelling

Following sentiment analysis, the next step in the methodology is topic modeling to uncover the main themes discussed in the tweets. Topic modeling is a statistical model used to discover abstract topics within a collection of documents. In this study, as alreasy mentioned, it is employed Latent Dirichlet Allocation , a widely used topic modeling technique, to identify key topics related to the 2020 US Presidential Election.

LDA is a generative probabilistic model that assumes documents (in this case, tweets) are composed of a mixture of topics, and each topic is characterized by a distribution of words (Blei, Ng, and Jordan, 2003). The goal of LDA is to infer these topics from the text data, providing a high-level overview of the main themes being discussed. Concretely The tweets are prepared for LDA by creating a document-term matrix, where each tweet is represented as a vector of word counts. LDA model is then trained on the document-term matrix, with the number of topics specified to adjust the desired granularity of the analysis. The trained LDA model identifies the dominant topics within the tweet corpus, assigning each tweet a distribution over these topics and characterizing each topic by a distribution of words.

Using LDA for topic modeling allows to uncover the main themes discussed in tweets about the 2020 US Presidential Election. This analysis provides a deeper understanding of the issues that dominated public debate and how different topics are associated with each candidate. By combining sentiment analysis and topic modeling, we might gain a significant and interesting view of public sentiment and the thematic landscape of social media discussions during the election period.

E. Interpretation of the results

The sentiment analysis results reveal the distribution of positive, neutral, and negative tweets for each candidate based on their polarity scores. For tweets about Donald Trump, 36.4% are classified as negative, 32.5% as positive, and 31.1% as neutral. These findings suggest that a significant portion of

users tweeting about Trump either have a negative view of him or tend to express their opinions in a more aggressive manner.

Conversely, the results show that 38.8% of the tweets regarding Joe Biden are positive, 36.7% are neutral, and only 24.5% are negative. Overall, Biden has a higher proportion of positive and neutral tweets, while Trump stands out for the higher percentage of negative sentiment among Twitter users.

TABLE I: Sentiment Analysis of Tweets for Joe Biden and Donald Trump

Candidate	Positive	Negative	Neutral
Joe Biden	38.8%	24.4%	36.8%
Donald Trump	32.6%	32.4%	31.0%

These results might suggest a preference for Joe Biden. However, the substantial number of neutral tweets makes it challenging to fully understand the true opinions of those users about him. Additionally, the sentiment analysis for Donald Trump reveals a clearer polarity of political thought, given the lower percentage of neutral tweets. This indicates that opinions about Trump are more decisively polarized, whereas Biden's sentiment distribution reflects a more nuanced public perception.

For what concern the topic modelling step, the results obtained are not very clear or exhaustive. Attempts were made to identify major topics for each candidate and correlate these with the sentiment analysis results. However, the diversity of terms used by users when posting tweets has likely contributed to the lack of clear, dominant topics. In contrast, tweets posted directly by the candidates themselves might have been more easily detectable, as politicians often use consistent slogans and messaging to persuade the public to support their ideas. This consistency in language could lead to more distinct and identifiable topics. Below it can be seen an example of the keywords found.

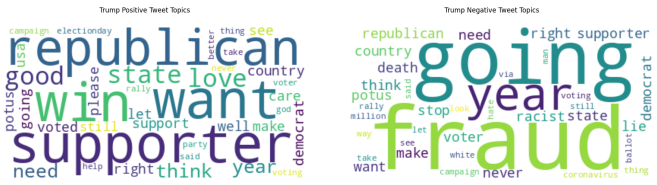

Fig. 3: Left: the figure shows the keywords about positive tweets towards Trump; Right:the figure shows the negative keywords. However, the topic associated with each group of keywords is unclear.

IV. Conclusion and final analysis

This study analyzed public sentiment and thematic content related to Donald Trump and Joe Biden during the 2020 U.S. presidential election, using Twitter data as the primary source. By employing Natural Language Processing (NLP) techniques, including sentiment analysis with VADER and topic modeling with Latent Dirichlet Allocation (LDA), the research aimed to provide insights into voter polarization and the influence of social media on political discourse.

The sentiment analysis revealed a significant disparity between the perceptions of Trump and Biden on Twitter. Tweets about Trump displayed a higher proportion of negative sentiment compared to those about Biden, who garnered more positive and neutral reactions. This finding suggests a polarized public perception of Trump, contrasted with a more balanced or favorable view of Biden.

Topic modeling, however, encountered challenges in delineating clear thematic trends. The diverse and informal nature of Twitter discourse, characterized by varied language use and rapid content generation, complicated the identification of dominant topics. This limitation underscores the difficulty in capturing coherent narratives from social media data, where informal language and brevity often obscure broader thematic patterns.

Several issues and limitations influenced the outcomes of this study:

- **Dataset Limitations**: The dataset, though extensive, was not exhaustive. The presence of missing values and the inherent biases in social media data—such as the overrepresentation of certain demographics or the influence of bots—may skew the results. The sample might not fully represent the overall sentiment of the electorate.
- **Methodological Challenges**: Sentiment analysis using VADER, while effective for short texts, may not capture the full nuance of sentiment, especially given the context-dependent nature of tweets. Similarly, LDA's ability to identify topics is limited by the variability in user language and the noise inherent in social media data.
- **Polarization and Echo Chambers** : The study did not account for the potential effects of echo chambers and polarization within Twitter, where users are often exposed to information that reinforces their existing beliefs. This factor could influence the sentiment and topics identified in the analysis.

In summary, while this study provides valuable insights into the sentiment and thematic content of tweets related to Trump and Biden, it also highlights the complexities and limitations inherent in social media analysis. Future research could benefit from incorporating more sophisticated methodologies and broader datasets to better capture the nuances of public sentiment and its impact on political discourse. By addressing these limitations, researchers can enhance the accuracy and relevance of sentiment and topic modeling in the context of evolving political environments.

REFERENCES

[1] Pablo Barberá. How social media reduces mass political polarization: Evidence from germany, spain, and the u.s. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM 2014)*, Newport Beach, CA, 2014. The AAAI Press.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003.

[3] Alexandre Bovet and Hernán Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, 10, 01 2019.

[4] Russell J. Dalton. Generational change in elite political beliefs: The growth of ideological polarization. *The Journal of Politics*, 49(4):976–997, 1987.

[5] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[6] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm: An r package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019.

[7] Andranik Tumasjan, Timm O. Sprenger, Philipp Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM 2010)*, Washington, D.C., 2010. The AAAI Press.