Background: Duke Kunshan University (DKU) is an interdisciplinary institution that grants dual undergraduate degrees, an MOE Chinese degree and a degree from Duke University in Durham, United States. The principal structure of DKU majors is robustly interdisciplinary. No student confines their study to a single discipline (for example, biology or economics). Instead, all students engage in broad inquiry related to a subject or question (for example, political economy or global health) and take a wide variety of courses related to that area (for example, in public policy, history, ethics, or economics). As a result, our graduates are prepared to engage in a wide variety of inquiries using multiple methodologies to address complex issues that require interdisciplinary approaches.

This has implications for our vision and expectations of undergraduate theses and design projects, which reflect this broad interdisciplinary training. At DKU, every student completes a two-year project known as signature work which consists of multiple interconnected parts including thematic courses, experiential learning, capstones, and a final product. It seeks to integrate students' interdisciplinary educational experience and culminates in the creation of a product in a scholarly, creative, or applied nature in leu of an undergraduate thesis or design required by JED. Because DKU encourages students to cultivate their independence and creativity as one of its institutional student learning outcomes, the student-led signature work projects often reflect students' own particular interdisciplinary interests and training. In addition, signature work has an intensive emphasis on problem-solving and skill-development which is much needed for any interdisciplinary inquiry; thus, students' final products are evidence of transferrable skills that students have acquired and demonstrated through the 2-year program, rather than content knowledge narrowly defined by disciplinary training.

In sum, while the Chinese major declared with any given student might be construed narrowly, the experience of our students is much broader—and intentionally so. This is a distinctive feature of our curriculum, and this distinctiveness results in broadly interdisciplinary submissions from our graduates' submitting theses or design projects. We have designed this to prepare our students for a wide variety of graduate programs in China and the West, where interdisciplinary training is a competitive advantage.

SENTIMENT ANALYSIS ON INSTAGRAM AND X FOR THE 2024 U.S. PRESIDENTIAL ELECTION

by

Forrest Leung

Signature Work Product, in partial fulfillment of the
Duke Kunshan University Undergraduate Degree Program

March 9, 2025

Signature Work Program
Duke Kunshan University

APPROVALS

Mentor: Feng Tian, Division of Natural and Applied Sciences

CONTENTS

ACKNOWLEDGMENTS

I am deeply grateful to my capstone advisor, Feng Tian, for his guidance throughout this project. His invaluable insights, feedback, and support have been instrumental in bringing my Signature Work to fruition by challenging me to be more critical at each step.

I would also like to thank Christopher Van Velzer for his steadfast encouragement in shaping my academic journey. His faith in me gave me the confidence to navigate challenges beyond the classroom and pursue my future studies with renewed purpose.

A special note of appreciation goes to the Division of Natural and Applied Sciences faculty for fostering an interdisciplinary learning environment. Through my thematic courses, they have been catalysts for my intellectual growth by broadening my academic horizons and engaging me with ideas beyond the boundaries of a single discipline.

To my friends both near and far, thank you for the memories, camaraderie, and support. Your compassion has been a beacon of strength and joy in recent years, and I know our friendships will outlast the undergraduate years we shared.

And above all, my deepest gratitude goes to my family. Their sacrifices have given me opportunities I could have only dreamed of. My achievements reflect their love, and I carry their encouragement into the next chapter of my life.

ABSTRACT

Social media's rise as an easily accessible medium for political discourse has reshaped voter engagement and ideological divides, especially during electoral cycles. This paper examines sentiment trends surrounding the 2024 United States presidential election by analyzing social media posts from Instagram and X, formerly known as Twitter. Utilizing natural language processing (NLP) techniques like sentiment analysis, topic modeling, and geographic sentiment mapping, we assess how political discourse on these two platforms evolved before and after the election.

A dataset of over 50,000 posts was collected using automated web scraping, and was analyzed using *Valence Aware Dictionary and sEntiment Reasoner (VADER)* and *TextBlob*. The results indicate that Instagram serves as a platform for voter mobilization and engagement, whereas X exhibits moderate polarization and reactionary discourse. Pre-election sentiment was largely neutral or positive, reflecting civic enthusiasm, whereas post-election discussions saw sharper sentiment shifts across partisan powerhouses and swing states. Geospatial analysis further revealed that sentiment trends were more stable in states aligned with election results.

This research showcases the role of social media in shaping public perception, demonstrating platform-specific differences in user engagement and distribution. Findings suggest that algorithmic amplification and digital environments influence voter sentiments in ways that extend past traditional polling.

Keywords: *sentiment analysis, natural language processing (NLP), data visualization, 2024 U.S. presidential election, social media*

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Context

The rise of social media in the 21st century has transformed political communication, providing an instantaneous medium for political parties, candidates, and voters to engage with one another. Since President Barack Obama's Facebook campaign in 2008 aimed at mobilizing young voters [29], social media platforms like Facebook, Instagram, X, Reddit, and TikTok have become key environments for shaping public discourse, electoral behavior, and democratic engagement [28, 36]. While these platforms have effectively mobilized voters, they have simultaneously introduced new challenges, such as increased misinformation and disinformation [1]. These problems have actively compounded ideological divides and polarized partisan perceptions, which is commonly seen during the American electoral cycle [38]. Likewise, research indicates the global scale of this issue, as disinformation campaigns have been identified in the electoral processes of over 80 countries [8].

For the United States, the 2016 presidential election between Democratic candidate Hillary Clinton and Republican candidate Donald J. Trump marked a turning point where social media became a significant concern for candidates and constituents. Popularized during Trump's campaign, the term 'fake news' has since been synonymous with attempts to delegitimize credible news sources and describe fabricated stories [2]. X emerged as a key component of the political strategies for both parties, as exemplified by President Donald Trump's deliberate efforts to engage with American voters through strategic posting to generate viral content [9].

Social media as a political tool has reshaped how voters engage with information and each other. As a result, voters have gradually drifted away from civil discourse as

politically charged disinformation has become more frequently disseminated on social media platforms, which have become a primary news source for most Americans [17]. Coupled with growing political tensions, recent research has shown that algorithmic biases in platforms and artificially generated material, when combined with rising political tensions, compound the problem by magnifying content that supports users' preexisting opinions and fostering political polarization [32, 37].

This study is timely in the context of the 2024 U.S. presidential election as it explores how social media has contributed to shaping public sentiment as a primary medium of political discourse. Using sentiment analysis, topic modeling, and geographic mapping, we seek to unveil whether Instagram and X have broadened the electorate's perspective or have entrenched existing ideological divides.

1.2 Scope and Significance

This paper examines the political landscape regarding the 2024 U.S. presidential election, which took place on November 5, 2024. This electoral cycle was characterized by significant technological advancements, namely in artificial intelligence, with candidates employing artificially generated content to effectively target voters [6]. By analyzing data from Instagram and X, two of the most influential social media platforms [24], this study aims to provide an updated view of how social media has influenced voter perception and political discourse in an ever-polarizing society [7].

This paper highlights the impact of social media on voter turnout and electoral outcomes. While previous studies focused on disinformation and algorithms, few have examined their intersection in the context of the 2024 election [31, 41].

1.3 Related Work and Continuity

Election prediction using sentiment analysis is an evolving field that leverages NLP and machine learning techniques to forecast electoral outcomes by analyzing the sentiments expressed in polls, news articles, and social media discourse [4]. Existing literature, such as those that have extensively examined the role of social media sentiment in political campaigns by analyzing millions of X posts before and after presidential elections [10], demonstrate that sentiment analysis can serve as a valuable tool for assessing public opinion trends and, in many cases, align closely with election outcomes [21, 30].

Machine learning models, such as Naive Bayes and Term Frequency-Inverse Docu-

ment Frequency (TF-IDF) feature extraction, have traditionally identified correlations between social media sentiment and election outcomes, capturing shifts in public opinion, especially in swing states. Unlike human analysis, these models can uncover hidden patterns within large datasets, revealing trends that might go unnoticed [34]. However, while sentiment analysis provides a useful predictive tool, it remains an indicator rather than a definitive measure of electoral outcomes.

Most studies have primarily focused their analysis on a singular channel. During the 2020 election, sentiment analysis of 127.3 million geotagged tweets revealed stark ideological divides in voter priorities, yet it lacked cross-platform comparisons [30]. This constraint highlights the necessity for a broader strategy in analyzing digital discourse. Unlike previous research, this study adopts a broader perspective by incorporating a cross-platform comparison from Instagram to better comprehend how digital conversations shape public opinion and influence electoral dynamics.

1.4 Research Objectives

This study aims to:

1. Analyze sentiment trends the week before and after the 2024 U.S. presidential election to assess possible shifts in voter sentiment.
2. Identify key discussion topics and interactions using topic modeling and co-occurrence networks to map dominant narratives in political discourse.
3. Compare sentiment variations across geographic locations to visualize differences in political discourse in partisan strongholds and swing states.

Chapter 2

METHODOLOGY

The workflow of this study was adapted from Chaudhury et al. [10], featuring three main steps: data retrieval and pre-processing, feature extraction, and sentiment analysis. This paper incorporates modifications tailored to the unique characteristics of social media discourse and Instagram data. Unlike Chaudhury et al. [10], this paper's approach prioritizes sentiment polarity over topic modeling by reordering the workflow, placing sentiment analysis before feature extraction. By ensuring topic categorization is informed by sentiments, this approach can lead to more contextually aware political discourse.

2.1 Data Collection

The data for this study was gathered over three weeks, beginning on October 29, 2024 and ending on November 16, 2024. Initially, direct data acquisition was planned through the X API; however, restrictions introduced after Elon Musk's takeover and Twitter's rebranding to X in 2022 necessitated a shift to an alternative. Following Musk's appointment as CEO, X imposed strict limits on new unverified accounts, capping them at 500 daily posts to curb "extreme levels of data scraping & system manipulation" [12]. Additionally, Musk significantly increased API pricing to monetize the platform under new ownership and to price out smaller developers, reinforcing the need for alternative data collection methods.

As a result, given the real-time nature of political discourse on social media, an automated scraping approach using PhantomBuster was implemented to systematically collect cloud-based data. Scraping was preferred over traditional crawling methods due to its ability to efficiently extract targeted, high-relevance data within a con-

strained time frame, ensuring that all collected content was directly related to the 2024 presidential election [23]. This method retrieved structured data, including timestamps, user metadata, engagement metrics, and text content, providing a comprehensive foundation for sentiment and topic analysis.

2.2 Data Processing

Prior to conducting sentiment and topic analysis, preprocessing is necessary to ensure accuracy and consistency by filtering misspellings, slang, abbreviations, emojis, and extraneous metadata from raw posts. Using Python's *regex* package, non-textual components such as special characters, hyperlinks, user mentions, and emojis are removed to prevent bias or distortion in sentiment analysis and topic clustering. Text cleaning refines the dataset by preserving relevant textual information.

Given the raw dataset of 50,000+ posts, a stratified random selection strategy was applied to ensure balanced temporal representation across key election periods. A capped subset of 10,000 posts, at most 5,000 before and after Election Day, was extracted for analysis. This selection was made to control for bias introduced by temporal surges in political discourse, ensuring that sentiment and topic modeling accurately reflect sustained trends rather than momentary spikes. The Instagram dataset *result.csv* was then supplemented with an additional dataset derived from itself by re-scraping 4,356 posts for detailed metadata. This new dataset was saved as *posts.csv* (see Appendix B). Posts from *posts.csv* were matched with their counterparts in *result.csv* using post URLs, and the extra metadata was integrated.

Considering the multi-source nature of the dataset, timestamps from both platforms were standardized using the *dateutil* module in Python to align time markers across all posts. This step was essential for accurate time-series analysis, ensuring that sentiment trends were comparable regardless of the original format in which the data was collected.

Following date standardization, each post was tokenized using *Natural Language Toolkit (NLTK)* to facilitate text analysis. By omitting common stopwords using *NLTK stopwords*, which includes 179 words like "the," "is," and "on," etc., we were able to isolate core phrases to retain higher informational value while improving computational efficiency and thematic classification.

To maintain dataset integrity, missing values, particularly within timestamps and text fields, were systematically dropped using *pandas* functions in Python. To further

enhance linguistic consistency, text was converted to lowercase, and words were lemmatized using *NLTK*. This process reduced words to their base forms, ensuring that different inflected versions of the same word were treated uniformly. Standardization prevented redundant variations in the modeling, ultimately improving classification accuracy.

To validate the data processing, Instagram data before and after the election was briefly visualized in Figure 2.1 using Python's *WordCloud* package to ensure no erroneous data remained in the new datasets.

Figure 2.1: Instagram Word Cloud Before (left) and After (right) Election Day.

Following this, the Instagram dataset was split into *final_before.csv* and *final_after.csv*, and X data was saved to *tweets_before.csv* and *tweets_after.csv* (see Appendix B).

2.3 Sentiment Analysis

Sentiment analysis is a NLP technique that determines a piece of text's emotional tone. This method is widely used in social media monitoring to evaluate public opinion on specific topics. A positive, negative, or neutral sentiment score is assigned to each text entry based on its content. In this study, we used a rule-based technique to conduct sentiment analysis. Due to its effectiveness in handling short, informal text common on platforms like Instagram and X, *VADER* was selected as this study's primary sentiment analysis tool.

2.3.1 Rule-Based Sentiment Analysis

VADER is a tool and lexicon designed to parse social media content. Unlike other models, which rely on supervised machine learning and require large labeled datasets, *VADER* is pre-trained on a dictionary of words with a predetermined sentiment score [20]. Each word is mapped to a corresponding valence score and a compound score is

then computed by summing the valence score of each word in the lexicon, adjusting to account for parameters, and then normalized between -1 and +1.

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

where x is the sum of valence scores and α is the normalization constant, which defaults to 15. Negative values indicate negative sentiment, and positive values indicate positive sentiment [20].

One of *VADER*'s advantages is its ability to account for contextual sentiment intensity using heuristics [42]. For example, capitalization (e.g., "VOTE") increases the emphasis on a sentiment, punctuation (e.g., "Yes!!!") amplifies sentiments, and degree modifiers (e.g., "very good" versus "somewhat good") adjust the sentiment score proportionally. Figure 2.2 showcases the VADER Compound score distribution of the sentiments before (left) and after (right) Election Day after preliminary data transformation was done.

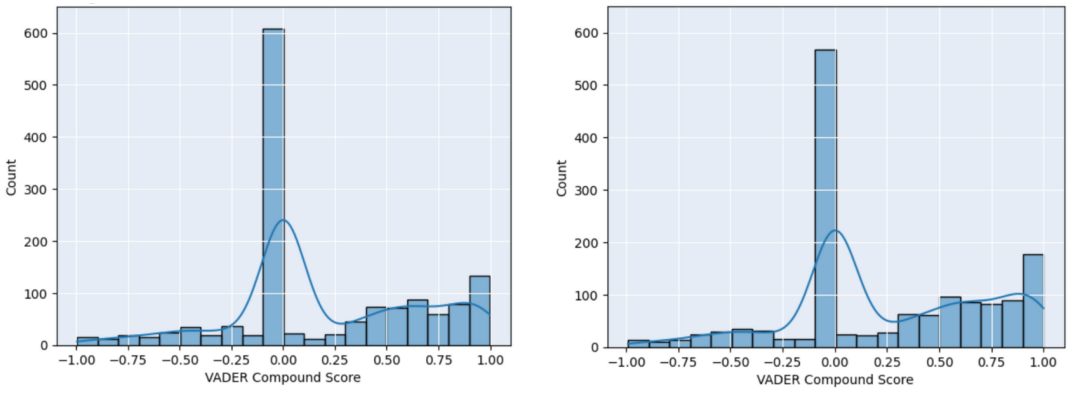

Figure 2.2: Compound Score Distribution Before (left) and After (right) Election Day

Since *VADER*'s primary use-case is for sentiment analysis on social media posts, *TextBlob* was incorporated as a secondary sentiment verification method to enhance classification accuracy on plain text. *TextBlob*, another lexicon-based sentiment analysis tool, measures subjectivity in addition to polarity, helping to distinguish opinion and fact-based statements. This makes it particularly useful for detecting nuanced shifts within discourse.

TextBlob determines sentiment using a pre-trained lexicon and rule-based approach, assigning two key scores: a polarity score ranging from -1 (negative) to +1 (positive),

which indicates the overall tone of a post, and a subjectivity score ranging from 0 (objective) to 1 (subjective), which reflects how opinionated or fact-based content is.

When calculating sentiment for a single word, *TextBlob* takes the average of different entries for the same word and finds words and phrases to which it can assign polarity and subjectivity, averaging them together when working with longer text.

2.3.2 Sentiment Score Aggregation and Trend Analysis

Once individual sentiment scores were assigned, they were aggregated to analyze sentiment trends before and after Election Day. This helped us identify any shifts in public opinion and understand how major political events influenced online discourse.

A comparative analysis was conducted between Instagram and X to examine platform-specific differences in political engagement. Sentiment trends were analyzed separately for each platform, revealing potential disparities in user attitudes, discussion intensity, and overall discourse patterns. The differences in sentiment trends may reflect platform-specific biases, variations in user demographics, or differences in content moderation policies that influence the tone and nature of discussions.

2.4 Topic Modeling

Topic modeling is a natural language processing technique used to uncover themes within large text corpora. This study utilized topic modeling to analyze social media captions through Latent Dirichlet Allocation (LDA), a three-level hierarchical Bayesian inference model used to estimate distributions. Although the LDA model is over 20 years old, it remains accessible and performs strongly compared to modern NLP models according to recent studies [18, 19]. LDA assumes text is a blend of multiple topics, each modeled as an infinite mixture over an underlying set of topics [5]. The process involves preprocessing, vectorization, and training the model to identify latent topics.

To ensure high-quality input, the text was first preprocessed using *regex* to remove URLs, hashtags, mentions, special characters, and two-letter words. Additionally, non-English posts were filtered using the *langdetect* Python library to ensure linguistic consistency. The cleaned text was further processed by removing any words within *NLTK stopwords* to retain only meaningful content.

Once preprocessed, the text data was vectorized using TF-IDF with the *TfidfVectorizer* package in Python. TF-IDF assigns weights to words based on their importance

in distinguishing between documents to ensure that commonly used but uninformative words do not dominate the topic modeling process. The vectorization process was parameterized by setting $min\_df = 10$ and $max\_df = 0.6$. min_df ignored words that appeared in fewer than 10 posts to prevent noise from rare words and max_df excluded words that appear in more than 60% of the dataset as they are too frequent to contribute meaningfully to topic differentiation.

The LDA model was trained with 10 topics, using word co-occurrence to identify hidden themes from the dataset. The trained model generated topic distributions for each document, and each post was assigned the most probable topic. To ensure topic interpretability, the top 15 words per topic were extracted, with priority given to political keywords such as "Trump," "Kamala," "election," "vote," "poll," and "ballot."

The resulting 10 topics were mapped and labeled based on thematic relevance, each featuring eight words. These topics were then tracked over time and visualized in 3.10.

2.5 Geographic Analysis

By mapping sentiment scores to geo-located posts, we can explore how political discourse evolved across states before and after the election. This process involves isolating posts containing location metadata, standardizing geographic references, and filtering out non-U.S. locations to ensure the dataset focuses exclusively on domestic political discussions. This step only used Instagram data as the X data was limited in size and did not collect location metadata.

To achieve that, we implemented a function to standardize location data and extract and normalize U.S. state names from the location field, ensuring consistency across datasets. Posts without location data or those linked to non-U.S. regions were excluded to maintain dataset integrity, though this also reduced the datasets to a few hundred available posts. Once location data was cleaned, sentiment scores were calculated using *VADER* and aggregated at the state level. This aggregation assessed sentiment differentiation across states, providing insights into the political climate in various regions before and after the election.

Given the substantial variation in the volume of posts across different states, raw sentiment scores alone were insufficient for direct comparisons. States with higher engagement could disproportionately influence overall sentiment trends, making it necessary to implement normalization techniques to ensure a balanced analysis.

To account for these differences, Z-score normalization was applied, producing a

standardized sentiment measure. This transformation adjusted for disparities in posting volume, allowing sentiment scores to be interpreted relative to each state's overall engagement levels rather than absolute sentiment values. Additionally, a normalization factor for post count was introduced to control for biases stemming from variations in engagement intensity.

To further refine the analysis, an adjusted sentiment score was calculated by dividing the Z-score of sentiment by the normalized post count plus one. This final adjustment prevented states with significantly lower posting volumes from disproportionately affecting sentiment rankings. By implementing these normalization techniques, the analysis provided a more equitable comparison of sentiment trends across states, ensuring that the influence of high-traffic regions did not overshadow insights from lower-volume areas.

With the sentiment scores standardized, geographic trends were visualized by generating a choropleth map using Python's *Folium* library. A GeoJSON file containing U.S. state boundaries was loaded to provide the necessary spatial context, and a dictionary mapping state sentiment scores was built to map each state to a color scale, with red representing negative sentiment and blue representing positive sentiment. Missing data was handled by shading states gray.

Additionally, we examined correlations between state-level sentiment trends and the 2024 election outcomes. A scatterplot with variable bubble sizes was generated to compare sentiment scores against election winners and identify differences in online engagement between Democratic and Republican-leaning states. By mapping sentiment trends alongside electoral results, this analysis provides insight into whether states with different political leanings exhibited distinct sentiment patterns before and after the election.

2.6 Data Visualization

Given the complexity of analyzing large-scale textual data, visualization techniques are crucial in revealing patterns, identifying key discussion themes, and tracking sentiment variations over time. By leveraging a combination of static and interactive visualizations, this study aimed to present findings in an intuitive and analytically rigorous manner.

A range of Python-based libraries (see Appendix A) was employed to generate visual representations, including word clouds, scatter plots, network graphs, stacked

bar graphs, violin plots, and overlaying time-series. These tools facilitated a structured analysis of sentiment dynamics, allowing for both macro-level trends and granular insights into platform-specific engagement.

To highlight broader themes and validate consistency, word clouds were created separately for Instagram data before and after Election Day. A shift in narrative focus was observed by comparing the word clouds, and data preprocessing could be depicted for interoperability.

For trend analysis, *pyplot* produced static line graphs that tracked daily sentiment fluctuations over time. To further analyze sentiments over time, *mdates* was used to format labels on the x-axis for readability. Bar charts were used for comparative sentiment analysis across different political topics and platforms, highlighting variations in discourse engagement. Furthermore, a stacked bar graph was generated to compare positive, neutral, and negative sentiment proportions for X hashtags. *Matplotlib Patches* was integrated into sentiment trend graphs to distinguish factions and topic clusters, allowing for clear demarcation of political affiliations and discussion groupings.

In addition to sentiment analysis, topic and hashtag co-occurrence network graphs for the X dataset were employed to quantitatively illustrate the connectivity and frequency of hashtags, identifying central political figures, misinformation narratives, and other narratives.

To contextualize statistical visualizations, *Seaborn* generated a choropleth map to depict geospatial sentiment distributions and a violin plot that analyzed dataset variability across different states and political alignments.

Chapter 3

RESULTS

The findings reveal platform-specific differences in sentiment trends, temporal shifts in discourse, and geographic variations in sentiment distribution before and after Election Day.

3.1 Instagram Data Figures

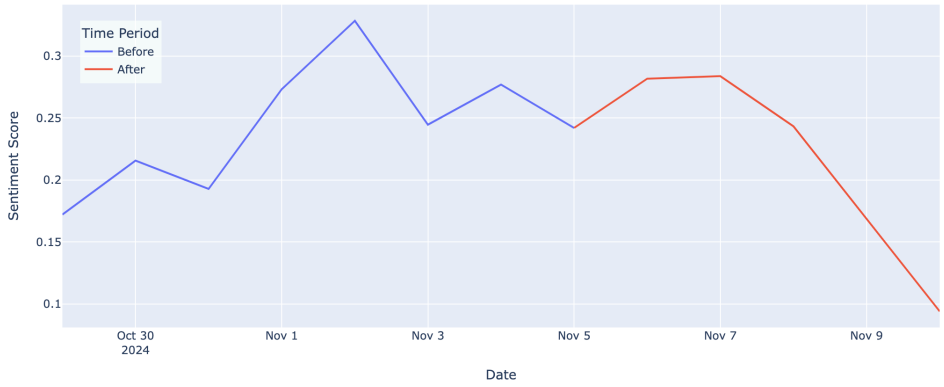

Figure 3.1: Daily Sentiment from October 29 to November 10, 2024.

Figure 3.1 depicts a line graph comparing the daily average sentiment scores on Instagram before and after the election. The y-axis represents the range of sentiment scores while the x-axis represents dates leading up to and following the election.

The increase in sentiment before the election suggests rising enthusiasm and engagement in political discussions. On the other hand, the post-election decline could signal a large volume of disappointment, polarization, or discourse as results were processed [13]. The abrupt drop-off in sentiment score after Nov 7 may correlate with external political events, challenges to election results, and policy reactions.

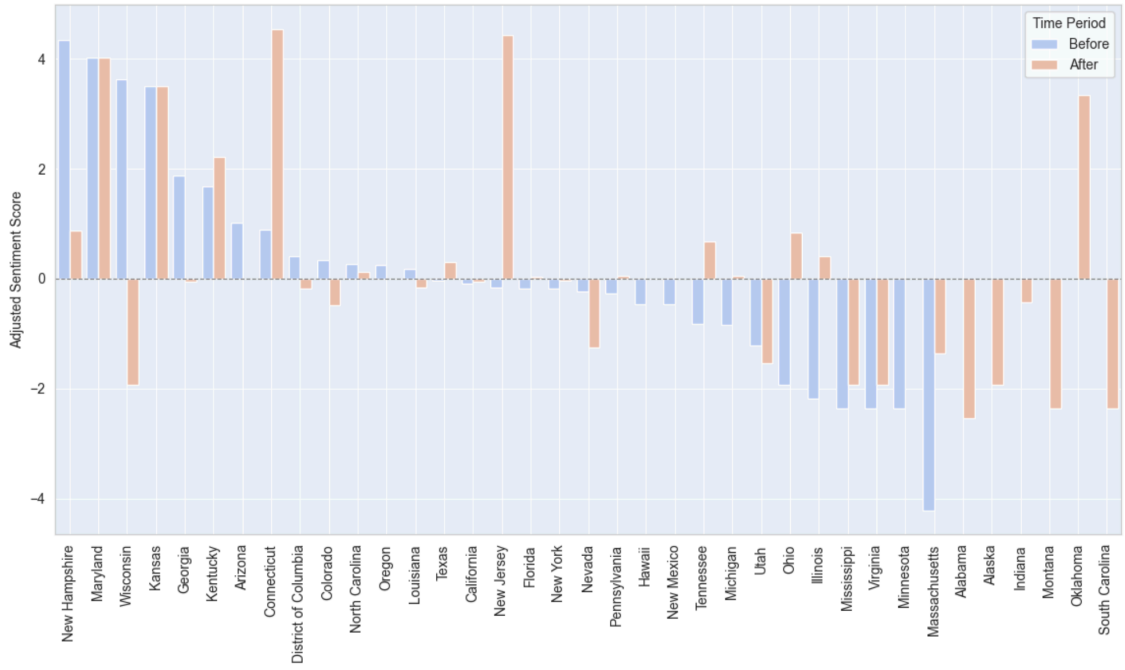

Figure 3.2: Standardized State Sentiment Rankings.

Figure 3.2, a comparative bar chart, shows the standardized sentiment scores (see Section 2.5) across the U.S. on Instagram before and after the election, seen in blue and orange, respectively. The y-axis represents the adjusted sentiment score, which measures how sentiment in each state deviated from the national average.

Across the board, declines in traditionally Democratic states suggest disappointment or unrest, whereas increases in sentiment in Republican states may indicate approval of the results. Many swing states from the Rust and Sun Belts have [14], with states like Wisconsin and Georgia dropping sentiment drastically even though Trump won both states in the election. This indicates that the users most active on Instagram were not aligned politically with the election results.

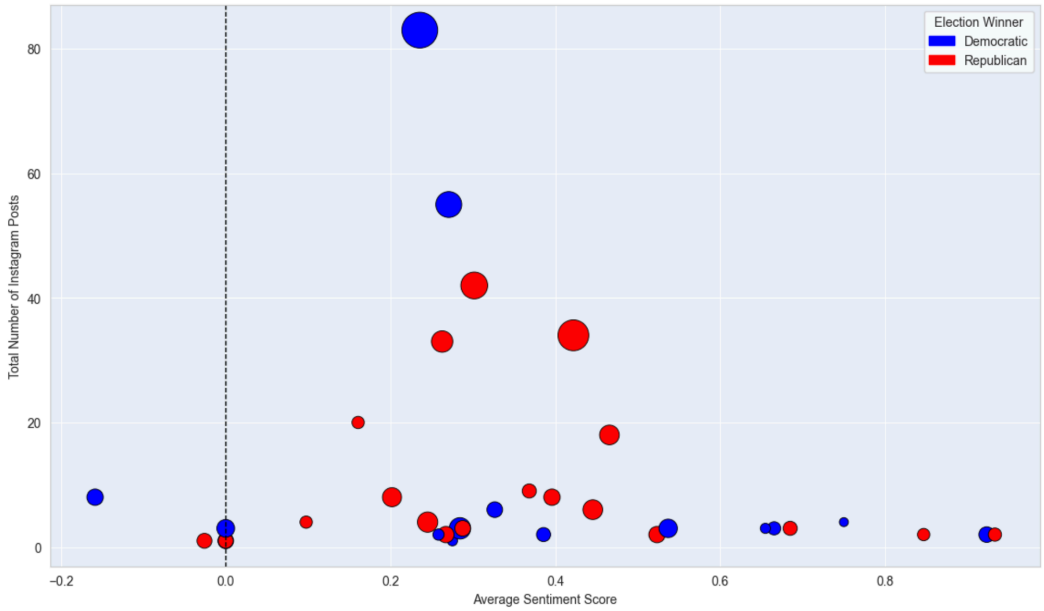

Figure 3.3: State Sentiment Trends and Election Results.

Figure 3.3 compares average sentiment scores on Instagram with state election outcomes in a bubble chart, where blue bubbles represent Democratic wins and red bubbles represent Republican wins. The bubble size corresponds to the number of Instagram posts related to the election in each state.

With no strong correlation between sentiment and electoral outcomes, the graph suggests that Instagram sentiment did not necessarily predict state-level voting behavior. Democratic-leaning states appear to have more dispersed sentiment scores whereas Republican-leaning states cluster around moderate-to-positive sentiment values. The higher sentiment scores in certain Republican-won states suggest that Instagram discussions may have been more positive in regions where election results aligned with users'expectations, while lower engagement in Democratic states may indicate disillusionment or a shift in focus to other platforms.

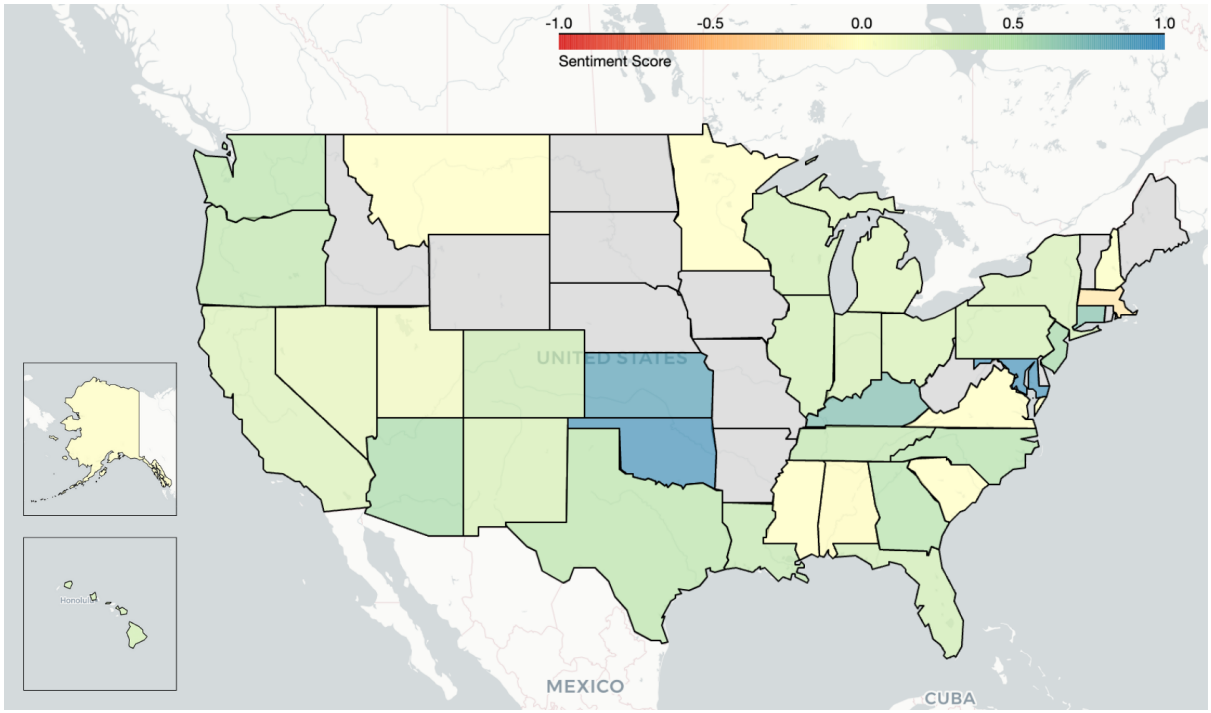

Figure 3.4: Geographic Sentiment Distribution Across the U.S.

Figure 3.4 features a choropleth map visualizes state-level Instagram sentiment post-election, with blue representing positive sentiment and red representing negative sentiment.

Positive sentiments are notable in states such as Oklahoma, Kansas, and Maryland, while negative sentiments are mainly seen in Massachusetts. The presence of gray states shows states with insufficient Instagram data to generate a sentiment score due to data limitations, but could also suggest that election-related discourse may not be as concentrated in these regions.

3.2 X Data Figures

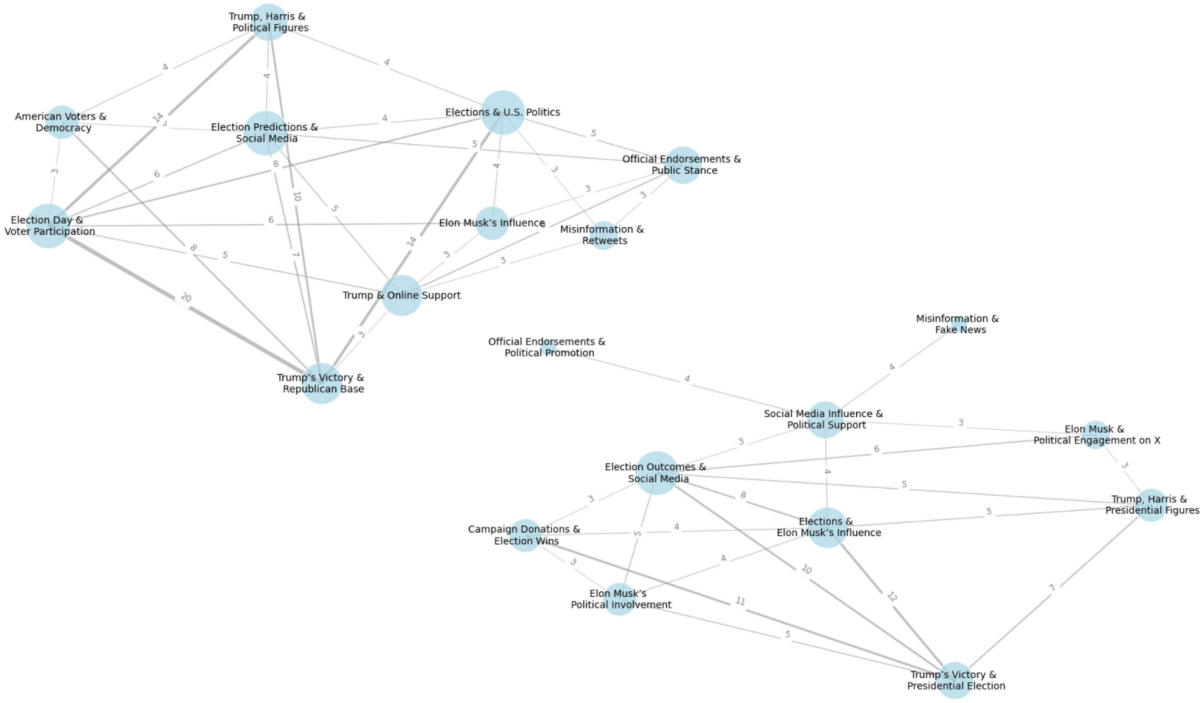

Figure 3.5: Topic Network Before (top-left) and After (bottom-right) Election Day.

Figure 3.5 showcases two co-occurrence networks regarding thematic clusters of political discussion on X before and after the election. Each node represents a dominant topic, while thicker edges indicate stronger relationships between themes, showing how different topics were interconnected in public discourse.

Before the election, a triangle of the "Election Day & Voter Participation," "Trump, Harris & Political Figures," and "Trump's Victory & Republican Base" topics were seen central to discussion. Once Election Day passed, the conversations became less interconnected, emphasizing "Campaign Donations & Election Wins" and "Election Outcomes & Social Media" alongside Trump's victory. The appearance of multiple clusters containing Elon Musk as a post-election topic suggests that discourse around the platform's moderation and current events amplified these talking points on X [40].

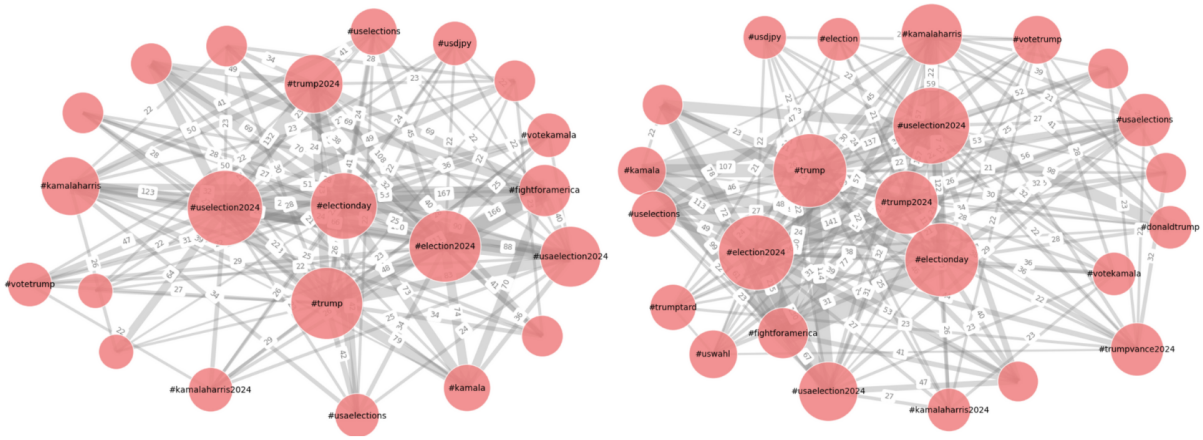

Figure 3.6: Hashtag Network Before (left) and After (right) Election Day.

Figure 3.6 depicts two network graphs that exhibits the relationships between hashtags used in election-related posts on X, where the size of each node represents the frequency of hashtag usage and the thickness of edges represents how often two hashtags appear together.

Prior to the election, hashtags such as #uselection2024, #fightforamerica, #trump2024, and #kamalaharris dominated discussions, each appearing together in posts in 100+ instances. The network exhibited a balanced presence of pro-Kamala and pro-Trump hashtags, suggesting engagement across the political spectrum. Following the election, the network became more concentrated with hashtags associated with President Trump, with #trump and #trump2024 located in the most concentrated portion of the network.

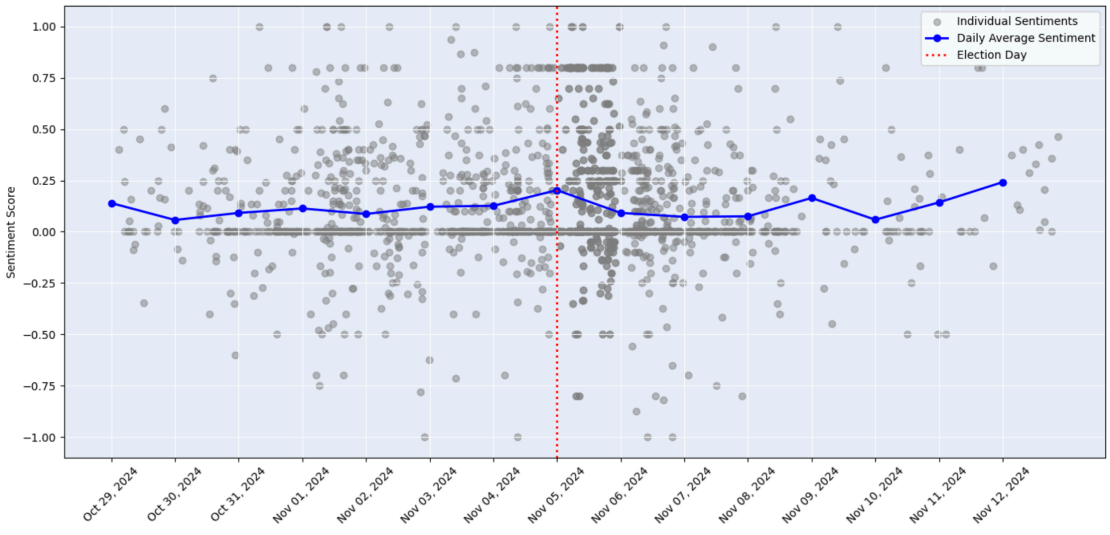

Figure 3.7: Daily Sentiments from October 29 to November 12, 2024.

Figure 3.7 tracks daily sentiment scores of individual posts as seen by the gray dots, and highlights the overall sentiment trend from October 29 to November 12, 2024 in blue. The red dotted line marks Election Day.

The graph displays an increase in the concentration of posts between November 5th and 7th, with general sentiments peaking on election day and staying relatively moderate. The observed peaks in sentiments intersect with key moments such as projected wins in Congress, the new administration being announced, and concession speeches [39].

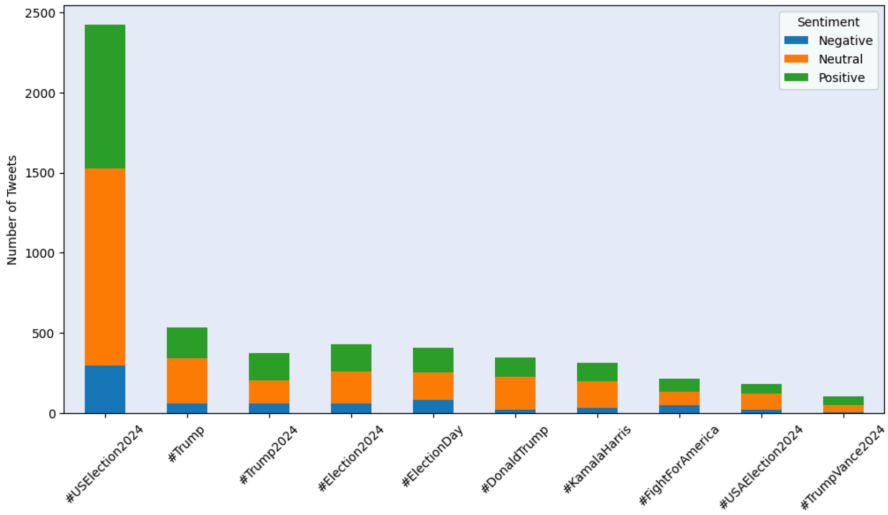

Figure 3.8: Sentiment Distribution of Top 10 Hashtags

The stacked bar chart in Figure 3.8 illustrates the sentiment distribution across the top 10 election-related hashtags on X. Given this research's main query of #USElection2024, it undoubtedly has the highest volume of mentions, with a significant portion marked as neutral and positive, while Trump-related hashtags show a more balanced mix of sentiment.

The prevalence of neutral and positive sentiment for #USElection2024 suggests that election-related discussions on X were not entirely dominated by negative discourse, possibly reflecting informational content, voter encouragement, or balanced reporting. However, Trump-related hashtags exhibit higher proportions of negative and positive sentiment, indicating that his presence in political discussions remains highly polarizing. The relatively lower volume of posts for Kamala Harris-related hashtags suggests that discussions about her were less frequent or not as contentious as Trump-centered discourse. Overall, this distribution highlights the ideological divide in political conversations on X, with highly engaged but opposing perspectives.

3.3 Combined Data Figures

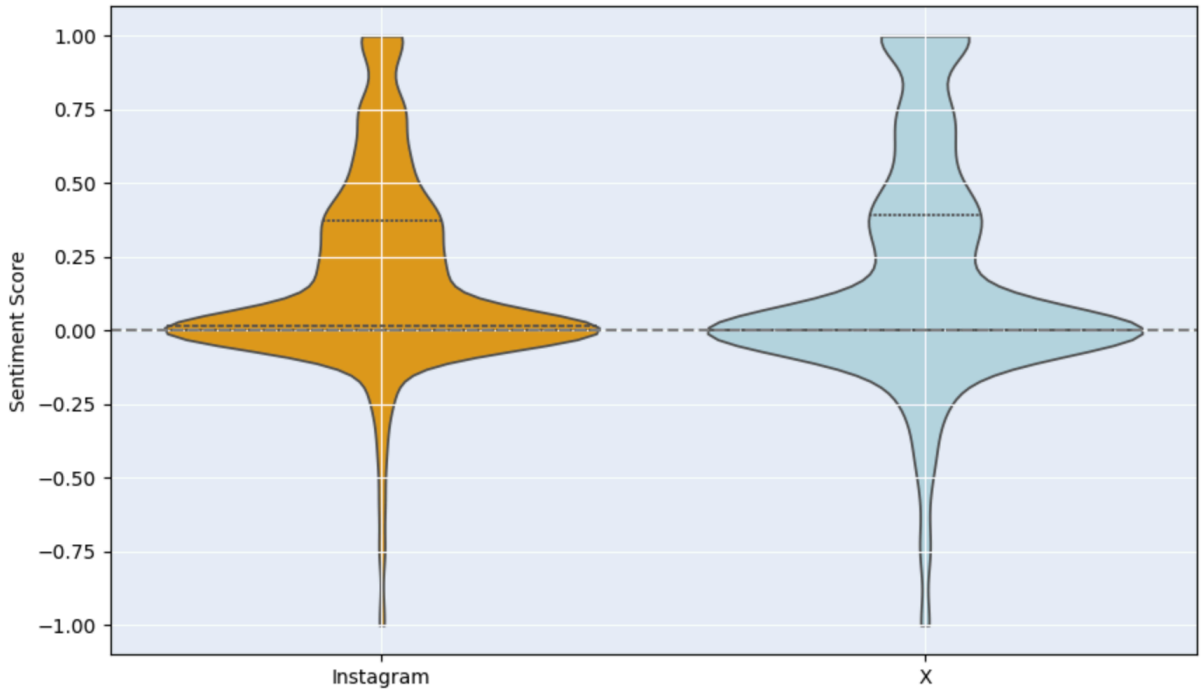

Figure 3.9: Cross-platform Sentiment Distributions.

Figure 3.9's violin plot compares sentiment distribution on Instagram and X, illustrating the spread and concentration of sentiment scores. The wider sections indicate a higher frequency of sentiment scores at that range, while the overall shape highlights the distribution of positive, negative, and neutral sentiments.

Relative to Instagram's sentiment distribution, X's distribution appears to have a broader spread of highly positive and negative sentiments. Given the dataset size discrepancies, this could suggest that X remains a platform for more reactionary discourse, with support and criticism for all candidates, policies, and outcomes. Interestingly, Instagram's distribution is highly concentrated around neutral and slightly positive sentiment, indicating that political discussions tend to be less extreme, perhaps informational or community-driven.

```
Top Topics Identified:
  Topic 1: big, american, media, usa, americans, man, social, epstein
  Topic 2: wins, today, trump, likes, election, donald, retweets, lets
  Topic 3: women, garbage, guinea, baltasar, campaign, engonga, people, world
  Topic 4: win, retweet, like, voted, votes, kings, team, save
  Topic 5: victory, good, india, trumps, want, congratulations, years, luck
  Topic 6: right, house, week, read, states, white, know, dont
  Topic 7: support, elon, icon, changed, actively, officially, youre, promoting
  Topic 8: musk, button, united, elections, state, true, fake, thought
  Topic 9: new, says, winning, news, watch, drop, modi, candidate
  Topic 10: harris, president, day, voting, real, final, supporting, hit
```

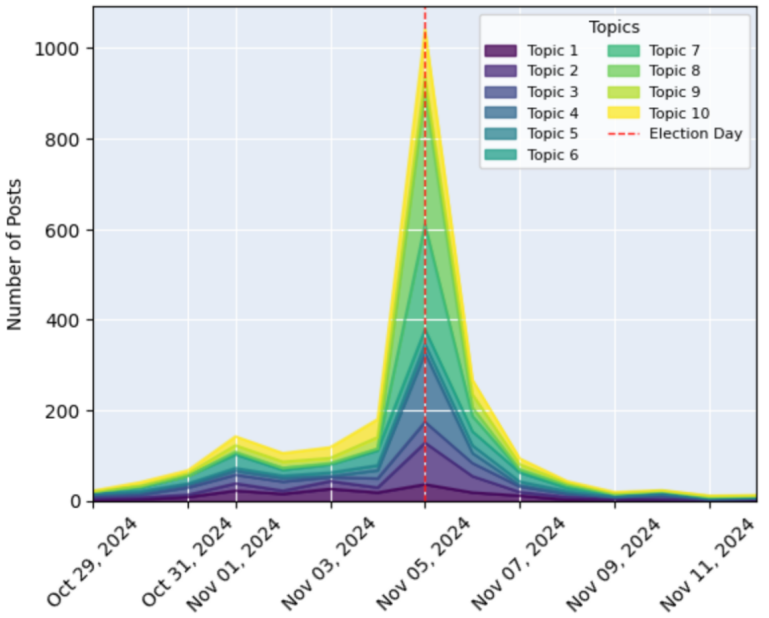

Figure 3.10: 10 Modeled Topics Across the Election Cycle.

Figure 3.10 tracks the volume of posts related to different election topics over time in an area chart, with a sharp spike on Election Day. The most discussed topics are Topics 9 and 10, which have keywords such as "resident," "voting," "news," and "drop."

Due to the non-deterministic nature of LDA [27], topic modeling produces topics that vary slightly across iterations, however, the general themes remain consistent. The post-election decline in discussions suggests that engagement tapered off once results were finalized. The prominence of keywords within the top three topics, such as "supporting," "musk," and "fake," suggests that discussions about the role of X in shaping election-related discourse remained a focal point. This aligns with concerns about platform moderation and misinformation, as well as amplifications of partisan narratives [13, 28].

Chapter 4

DISCUSSION

This paper's results indicate a shift in sentiment from pre-election optimism to moderate post-election polarization, with a moderate increase in ideological divides, differing engagement levels across platforms, and varying sentiment trends across states. Instagram, a more visually driven and community-oriented platform, exhibited less extreme sentiment than X, demonstrating plausible polarization and greater engagement with politically charged content.

Prior studies on social media and elections suggest that X is a hub for real-time political discourse [26], while Instagram fosters more engagement-based activism [11]. The findings of this paper align with this, showing that Instagram's sentiment distribution was centered around neutral and slightly positive values, while X exhibited a more volatile range of sentiments. Additionally, previous research has emphasized that social media functions as an "echo chamber" that reinforces ideological beliefs [22], a trend visible in the data, particularly in the post-election increase in highly negative sentiment in certain states and within specific political hashtags.

One of the most striking differences observed was the variation in sentiment between Instagram and X. Instagram discussions, particularly before the election, focused more on mobilization and voter encouragement. In contrast, X was more reactionary, shifting post-election discussions toward election integrity, political figures, and ideology. These findings suggest that platform affordances shape political discourse, with Instagram fostering community-driven narratives and X acting as a space for rapid, real-time debates and reactions.

The analysis of state-level sentiment trends showed that election-related discussions were most concentrated in high-population partisan states like California, New

York, Pennsylvania, Florida, and Texas. While higher sentiment scores in Republican-won states suggest more positive engagement post-election, Democratic-leaning states exhibited more significant fluctuations and declines in sentiment. This indicates that regional political climates influenced public reaction, with states where election results aligned with user expectations showing more positive sentiment.

While this paper employs a multi-method approach for sentiment analysis on social media posts, several limitations should be acknowledged. Although applicable, sentiment analysis falls short of fully understanding the subtle context of political conversations due to sarcasm, irony, and coded language that automated classifiers can miss. Data accessibility presents another limitation as all social media platforms allow users to keep their accounts private, creating gaps in key voter interactions in the dataset. The effectiveness of scraping was constrained by platform-imposed access restrictions [3], which limited the breadth of data collection as mentioned in Section 2.1. Ethical considerations were also a cause of concern, particularly regarding data privacy and algorithmic bias, as the methodology primarily relied on publicly accessible accounts. This approach inherently skewed sentiment analysis toward a subset of highly engaged users, potentially distorting broader public opinion trends [15]. Additionally, geospatial analysis is limited by the availability of location-tagged posts, which skews regional distributions toward more populated areas or away from digital footprint-conscious users who fear possible repercussions [35].

Moreover, this study is limited by its timeframe, computational power, and financial resources. The data collected was limited to the weeks neighboring the election period, which does not fully reflect the evolving political discourse over longer durations. The computational power available for this research, particularly GPU access, restricts the ability to run large-scale deep learning models that could improve sentiment classification accuracy. Alongside computation, financial constraints impacted the dataset size and model selection, as more advanced AI models remained inaccessible.

With the dataset originating from Instagram and X, the findings from this paper may not generalize to other platforms such as Facebook or Reddit, both of which cater to different user demographics and content-sharing behaviors [33]. Future research should expand platform coverage and utilize more sophisticated natural language processing techniques to increase classification accuracy.

These findings suggest several directions for future research. Longitudinal studies could track how sentiment evolves months after the election, extrapolating whether

post-election polarization is reactionary or an underlying trend for the next administration. Examining engagement over time may also reveal shifts in political discourse in response to major events and policy changes.

Expanding sentiment analysis across platforms like TikTok, Facebook, and Reddit would provide a broader perspective on political discourse. Each platform's unique environment shapes engagement differently: TikTok's rapid visual content, Facebook's community-driven discussions, and Reddit's long-form debates may foster distinct political narratives. A cross-platform approach would clarify how these differences influence sentiment and ideological reinforcement.

A comprehensive examination of misinformation and political rhetoric is imperative. Analyzing how false narratives, campaign strategies, and user-driven activism influence online sentiment could uncover the degree to which misinformation contributes to polarization. Grasping the function of political rhetoric across various platforms may further assist in developing strategies to mitigate its effects and promote more constructive discourse.

Chapter 5

CONCLUSION

This study examined the role of social media in shaping political discourse during the 2024 U.S. presidential election, analyzing sentiment trends, engagement patterns, and thematic discussions across Instagram and X. The findings indicate that platforms shape political discussions differently, with variations in sentiment distribution, engagement levels, and topical focus.

The data showed a shift in sentiment before and after the election, with pre-election discussions centered mainly on mobilization and civic engagement, followed by a decline in sentiment after Election Day. However, this decline was not uniform across platforms or states. Instagram and X both exhibited mixed sentiment trends, with neither platform showing a disproportionate increase in negative discourse. Geographic variations in sentiment further suggest that states where election outcomes aligned with expectations displayed more stable sentiment, while states with closer or more contested results saw greater fluctuation. The dominance of Trump-related discussions, both before and after the election, highlights the centrality of candidate-driven discourse in digital political engagement, whereas Biden-related discussions were notably less frequent.

One possible explanation for the dataset's absence of extreme sentiment patterns is the methodology used to collect posts. The study scraped posts using high-level political hashtags and keywords, such as #USElection2024 and #ElectionDay, rather than explicitly partisan or divisive hashtags. As a result, the data may have captured a more neutral cross-section of discussions rather than highly polarized content. Additionally, by not prioritizing posts from highly partisan sources or influencer-driven discourse, the dataset may reflect broader election conversations rather than echo chambers dominated by one ideological group.

However, this methodological approach may have skewed the dataset toward more moderate or generalized political discussions, potentially underrepresenting the intensity of partisan discourse often dominating political debates on social media [16]. Since political engagement on digital platforms is driven by high-profile influencers, partisan media accounts, and ideologically motivated communities, relying on general election-related hashtags may have excluded highly engaged, emotionally charged discussions in more insular or activist-driven digital spaces. This could explain why sentiment trends appeared relatively stable and did not exhibit extreme fluctuations, as seen in other studies focusing on more contentious online interactions [25].

Another limitation of this approach is that general election-related hashtags are frequently used by media organizations, official campaign accounts, and voters engaging in neutral discussions about the election process. As a result, the dataset may have overrepresented formal political discourse while underrepresenting grassroots, activist-driven, or highly polarized political communities. Future research could refine data collection strategies by incorporating a broader mix of hashtags, including those associated with partisan communities, viral political movements, and controversial election-related topics, to capture a more comprehensive picture of sentiment dynamics.

These findings reinforce the idea that social media is a major arena for political discourse, but engagement is not solely driven by extreme sentiment. While social media fosters broad participation in political discussions, it also reflects platform-specific behaviors and trends rather than serving as a clear indicator of ideological divides or public sentiment shifts. Future research should examine longer-term sentiment trends, cross-platform variations, and the role of misinformation in shaping political narratives, particularly as digital spaces continue to influence political decision-making and public trust in institutions. Expanding data collection methods to include a wider spectrum of political discourse, both neutral and partisan, will be crucial in developing a more nuanced understanding of how sentiment evolves in digital political environments.

REFERENCES

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. "Fake news, disinformation and misinformation in social media: a review". In: *Social Network Analysis and Mining* 13.1 (Feb. 9, 2023), p. 30. ISSN: 1869-5469. DOI: 10.1007/s13278-023-01028-5.

[2] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2 (May 1, 2017), pp. 211–236. ISSN: 0895-3309. DOI: 10.1257/jep.31.2.211.

[3] Catherine Altobelli et al. "To Scrape of Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR". In: Zenodo, Apr. 1, 2021. DOI: 10.5281/ZENODO.6411787.

[4] Quratulain Alvi et al. "On the frontiers of Twitter data and sentiment analysis in election prediction: a review". In: *PeerJ Computer Science* 9 (Aug. 21, 2023), e1517. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1517.

[5] David M Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (January Jan. 3, 2003), pp. 993–1022.

[6] Shannon Bond. "How AI deepfakes polluted elections in 2024". In: *NPR* (Dec. 21, 2024).

[7] Levi Boxell, Matthew Gentzkow, and Jesse Shapiro. *Cross-Country Trends in Affective Polarization*. w26669. Cambridge, MA: National Bureau of Economic Research, Jan. 2020, w26669. DOI: 10.3386/w26669.

[8] Samantha Bradshaw et al. "Country Case Studies Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation". In: (Jan. 13, 2021).

[9] John Bryden and Eric Silverman. "Underlying socio-political processes behind the 2016 US election". In: *PLOS ONE* 14.4 (Apr. 9, 2019). Ed. by Haroldo V. Ribeiro, e0214854. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0214854.

[10] Hassan Nazeer Chaudhry et al. “Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020”. In: *Electronics* 10.17 (Aug. 27, 2021), p. 2082. ISSN: 2079-9292. DOI: 10.3390/electronics10172082.

[11] Natalie Davidson. “InstaDAMN –The Power of Instagram’ s Platform As An Instigator and Indicator For Offline Political Participation Among Young Adults”. In: *Honors Theses* Paper 1395 (2023).

[12] Elon Musk [@elonmusk]. *Rate limits increasing soon to 8000 for verified, 800 for unverified & 400 for new unverified*. Twitter. July 1, 2023. URL: https://x.com/elonmusk/status/1675214274627530754 (visited on 03/05/2025).

[13] Neil Fasching et al. “Persistent polarization: The unexpected durability of political animosity around US elections”. In: *Science Advances* 10.36 (Sept. 6, 2024), eadm9198. ISSN: 2375-2548. DOI: 10.1126/sciadv.adm9198.

[14] James FitzGerald. “Seven swing states set to decide the 2024 US election”. In: *BBC News* (Oct. 31, 2024).

[15] Jens Foerderer. *Should we trust web-scraped data?* Aug. 4, 2023. DOI: 10.48550/arXiv.2308.02231. arXiv: 2308.02231[econ].

[16] Kiran Garimella et al. “Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship”. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18*. the 2018 World Wide Web Conference. Lyon, France: ACM Press, 2018, pp. 913–922. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186139.

[17] Sandra González-Bailón et al. “Asymmetric ideological segregation in exposure to political news on Facebook”. In: *Science* 381.6656 (July 28, 2023), pp. 392–398. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade7138.

[18] Ismail Harrando, Pasquale Lisena, and Raphael Troncy. “Apples to Apples: A Systematic Evaluation of Topic Models”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. RANLP 2021. Ed. by Ruslan Mitkov and Galia Angelova. Held Online: INCOMA Ltd., Sept. 2021, pp. 483–493.

[19] Alexander Hoyle et al. *Are Neural Topic Models Broken?* Oct. 28, 2022. DOI: 10.48550/arXiv.2210.16162. arXiv: 2210.16162[cs].

[20] Clayton Hutto and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 16, 2014), pp. 216–225. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v8i1.14550.

[21] Brandon Joyce and Jing Deng. “Sentiment analysis of tweets for the 2016 US presidential election”. In: *2017 IEEE MIT Undergraduate Research Technology Confer-*

ence (URTC). 2017 IEEE MIT Undergraduate Research Technology Conference (URTC). Cambridge, MA: IEEE, Nov. 2017, pp. 1–4. ISBN: 978-1-5386-2534-7. DOI: 10.1109/URTC.2017.8284176.

[22] Florian Justwan et al. "Social media echo chambers and satisfaction with democracy among Democrats and Republicans in the aftermath of the 2016 US elections". In: *Journal of Elections, Public Opinion and Parties* 28.4 (Oct. 2, 2018), pp. 424–442. ISSN: 1745-7289, 1745-7297. DOI: 10.1080/17457289.2018.1434784.

[23] Moaiad Khder. "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application". In: *International Journal of Advances in Soft Computing and its Applications* 13.3 (Dec. 30, 2021), pp. 145–168. ISSN: 27101274, 20748523. DOI: 10.15849/IJASCA.211128.11.

[24] Lenka Labudová. "Current Leading Social Media Platforms Used by Marketers and its Benefits". In: *European Conference on Social Media* 11.1 (May 28, 2024), pp. 394–401. ISSN: 2055-7221, 2055-7213. DOI: 10.34190/ecsm.11.1.2383.

[25] Michalis Mamakos and Eli J Finkel. "The social media discourse of engaged partisans is toxic even when politics are irrelevant". In: *PNAS Nexus* 2.10 (Sept. 29, 2023). Ed. by David Rand, pgad325. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgad325.

[26] Colleen McClain, Monica Anderson, and Risa Gelles-Watnick. *2. How X users view, experience the platform*. Pew Research Center. June 12, 2024. URL: https://www.pewresearch.org/internet/2024/06/12/how-x-users-view-experience-the-platform/ (visited on 02/27/2025).

[27] Jacob Murel and Eva Kavlakoglu. *What is Latent Dirichlet allocation | IBM*. Apr. 22, 2024. URL: https://www.ibm.com/think/topics/latent-dirichlet-allocation (visited on 03/07/2025).

[28] Maria Papageorgiou. *Social Media, Disinformation, and AI: Transforming the Landscape of the 2024 U.S. Presidential Political Campaigns - The SAIS Review of International Affairs*. Jan. 14, 2025. URL: https://saisreview.sais.jhu.edu/social-media-disinformation-and-ai-transforming-the-landscape-of-the-2024-u-s-presidential-political-campaigns/ (visited on 02/22/2025).

[29] Ashley Payne. "The New Campaign: Social Networking Sites in the 2008 Presidential Election". In: *Mahurin Honors College Capstone Experience/Thesis Projects* (Jan. 1, 2009).

[30] Miftahul Qorib, Rahel S Gizaw, and Junwhan Kim. "Impact of Sentiment Analysis for the 2020 U.S. Presidential Election on Social Media Data". In: *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*. ICMLT 2023: 2023 8th International Conference on Machine Learning Technologies. Stock-

holm Sweden: ACM, Mar. 10, 2023, pp. 28–34. ISBN: 978-1-4503-9832-9. DOI: 10.1145/3589883.3589888.

[31] Shaina Raza, Mizanur Rahman, and Shardul Ghuge. *Analyzing the Impact of Fake News on the Anticipated Outcome of the 2024 Election Ahead of Time*. Jan. 6, 2024. DOI: 10.48550/arXiv.2312.03750. arXiv: 2312.03750[cs].

[32] Ermelinda Rodilosso. "Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization". In: *Philosophy & Technology* 37.2 (June 2024), p. 71. ISSN: 2210-5433, 2210-5441. DOI: 10.1007/s13347-024-00758-4.

[33] Hamidreza Shahbaznezhad, Rebecca Dolan, and Mona Rashidirad. "The Role of Social Media Content Format and Platform in Users' Engagement Behavior". In: *Journal of Interactive Marketing* 53.1 (Feb. 2021), pp. 47–65. ISSN: 1094-9968, 1520-6653. DOI: 10.1016/j.intmar.2020.05.001.

[34] Xiaoling Shu and Yiwan Ye. "Knowledge Discovery: Methods from data mining and machine learning". In: *Social Science Research* 110 (Feb. 2023), p. 102817. ISSN: 0049089X. DOI: 10.1016/j.ssresearch.2022.102817.

[35] Reka Solymosi et al. "Privacy challenges in geodata and open data". In: *Area* 55.4 (Dec. 2023), pp. 456–464. ISSN: 0004-0894, 1475-4762. DOI: 10.1111/area.12888.

[36] Julie Uldam and Anne Vestergaard. "Introduction: Social Media and Civic Engagement". In: *Civic Engagement and Social Media*. London: Palgrave Macmillan UK, 2015, pp. 1–20. ISBN: 978-1-349-49288-6 978-1-137-43416-6. DOI: 10.1057/9781137434166_1.

[37] Bhimavarapu Usharani. "Artificial Intelligence and the 2024 US Presidential Election: A Double-Edged Sword of Voter Engagement and Disinformation". In: *Democracy and Democratization in the Age of AI*. Ed. by Kittisak Wongmahesak et al. IGI Global, Feb. 21, 2025, pp. 73–102. ISBN: 979-8-3693-8749-8 979-8-3693-8751-1. DOI: 10.4018/979-8-3693-8749-8.ch005.

[38] Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. "The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative". In: *Information Systems Frontiers* 26.2 (Apr. 2024), pp. 663–688. ISSN: 1387-3326, 1572-9419. DOI: 10.1007/s10796-023-10390-w.

[39] Adrienne Vogt, Matt Meyer, and Tori B. Powell. "Live updates: Latest on the 2024 election and Trump' s presidential transition | CNN Politics". In: *CNN* (Nov. 9, 2024).

[40] David Wright and Alex Leeds-Matthews. "Elon Musk spent more than $290 million on the 2024 election, year-end FEC filings show | CNN Politics". In: *CNN* (Feb. 1, 2025).

[41] Jinyi Ye, Luca Luceri, and Emilio Ferrara. *Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X Approaching the 2024 U.S. Presidential Election*. Nov. 12, 2024. DOI: 10.48550/arXiv.2411.01852. arXiv: 2411.01852[cs].

[42] Douglas C Youvan. "Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application". In: (2024). Publisher: Unpublished. DOI: 10.13140/RG.2.2.33567.98726.

SIGNATURE WORK NARRATIVE

My signature work project is an interdisciplinary project focused at the intersection of data science and politics by exploring political sentiments using computational analysis. Throughout this project, I integrated foundational skills from many of my major requirement courses, with *STATS 102*, *INFOSCI 103*, and *COMPDSGN 490* having the most impact. Each course provided their respective skills, design insights, and interdisciplinary perspectives that have been foundational in developing my project. Besides my previous skills, learning new skills such as sentiment analysis, topic modeling, and unique data visualizations has showcased the importance of understanding computational tools and how they can quantify unseen patterns in data, such as within social media posts to promote civic dialogue.

STATS 102: Introduction to Data Science is an introductory course taught in Python that showed me how to load, clean, manipulate, visualize, analyze, and interpret data with Python packages such as pandas, numpy, and seaborn. Although I had previous programming experience in Java prior to taking the course in Spring 2023, the hands-on experience of *STATS 102* problem sets and the final project helped me develop the skills to preprocess datasets and visualize data to increase the interpretability of data. My final project in the course involved analyzing factors impacting student exam performance using various machine learning models, working with categorical variables, label encoding, and feature importance analysis to interpret the results. Given that this was my first experience working with packages in Python, the class wasn't just an introduction to data analysis, as I was also able to realize the potential of computational tools to qualitatively visualize complex real-world data, which has translated to my Signature Work. Without the statistical reasoning and data-driven mindset I developed in *STATS 102*, this project wouldn't have gone beyond basic observations.

Within *INFOSCI 103: Computation, Society & Culture*, I explored how to interpret novel technologies and their impact on society and culture to negotiate modern challenges. Taking this course during Fall 2024 alongside other INFOSCI major require-

ments in Computation and Design, I was able to paint a holistic image for the goals for my major. While this major provides opportunities to blend technical expertise with social responsibility, we also needed to ensure that we could develop and apply computational methods in an ethical, innovative, and impactful way across fields. By tackling relevant technological problems like filter bubbles, echo chambers, and communication theories, this class provided a framework to critically examine the biases inherent within decision-making algorithms. During class discussions, we examined how social media platforms shape public discourse, influence political polarization, and reinforce biases through recommendation algorithms. These insights shaped my approach to sentiment analysis, as I sought to critically evaluate the role of algorithmic amplification in political discussions. This project not only focused on extracting sentiment from posts but also considered how sentiment varied across different political communities, revealing potential biases in digital discussions.

My senior seminar course, *COMPDSGN 490*, focused on the theme of *Protean Techno-Society: Interdisciplinary Perspectives*, designed to synthesize all three tracks within the Computation and Design major to critically examine the historical and cultural impacts of technological developments. This course broadened my perspective on interdisciplinary teamwork by using group projects as a medium to tackle concepts and practical techniques while confronting social, environmental, economic, and political issues. By working with peers in the social policy and digital media tracks, I refined my ability to communicate findings to non-specific audiences, which I translated into my signature work paper. The course emphasized the historical and cultural contexts of technological advancements, prompting me to consider the long-term societal impacts of political sentiment analysis and AI-driven public discourse monitoring. Additionally, learning to iteratively design methodologies to approach abstract prompts led me to adopt a more structured approach to my project.

Reflecting on my Signature Work, the process was equally challenging and rewarding. I initially had goals to develop an interactive dashboard that used user inputs to scrape social media networks to create entity maps based on the correlation of topics, however API access posed a key limitation that created a financial barrier to the project. Thus, given the concepts I had previously picked up in my thematic courses, I switched my project focus to sentiment analysis and data visualizations surrounding the 2024 United States presidential election. Outside the classroom, my professional experience in edge intelligence and decentralized learning research within robotics contributed in ways I didn't expect to. While previous experiences primarily focused on optimizing distributed data processing and federated learning, the technical skills

I developed, such as efficient model training and classification, proved handy in fine-tuning sentiment models.

This project also reinforced my commitment to ethical AI and responsible data analysis. While web scraping provided an alternative, such methods raise essential ethical considerations regarding data privacy, consent, and platform policies. While fleshing out my methodology, I examined the trade-offs between accessibility and ethical responsibility to ensure that my approach remained within legal and academic research guidelines.

More broadly, as sentiment analysis becomes increasingly influential in shaping public narratives, it is crucial to scrutinize how these tools are used in media, policy-making, and governance. My findings highlighted the challenges of algorithmic bias, the risks of misinterpretation when applying computational methods to social issues, and the necessity of transparency in data-driven political discourse analysis. These insights have deepened my interest in the broader implications of AI-driven decision-making and its role in shaping democratic processes while reinforcing the importance of ethical considerations in computational research.

I see my Signature Work project as more than a technical exercise. My project represents my interdisciplinary journey at Duke Kunshan University, combining computational techniques and sociopolitical inquiry to engage with data not just as numbers but as reflections of societal discourse on social media. By bridging technical expertise with critical analysis, I have strengthened my ability to conduct rigorous computational research while considering the societal implications of technology like algorithmic biases and data misinterpretation.. Looking forward, I am eager to apply the insights I've gained through this exercise to my future studies to bridge the intersection of technology and society, specifically AI and data analysis. This project has reinforced my passion for studying AI and data analysis in a way that prioritizes ethical considerations and societal well-being.

In many ways, my Signature Work project reflected my broader intellectual pursuits—applying computational methods to real-world challenges while maintaining a critical awareness of their implications. The interdisciplinary methodology I refined through this project will continue to guide my future research endeavors, allowing me to contribute meaningfully to the evolving dialogue on technology and society.

Appendix A

PYTHON PACKAGES USED

- pandas: Essential for loading, cleaning, and manipulating datasets.
- NumPy: Provides support for numerical operations.
- JSON: Handles JSON-formatted data from APIs.
- ast (Abstract Syntax Trees): Converts string representations of Python literals into objects.
- re (Regular Expressions): Facilitates text cleaning and pattern-based filtering.
- Random: Helps with random sampling in exploratory data analysis.
- dateutil parser: Parses dates from text to create datetime objects.
- collections Counter: Counts word frequencies and hashtags.
- collections defaultdict: Organizes word frequencies and hashtags.
- itertools combinations: Assists in network analysis by generating entity combinations.
- Requests: Makes HTTP requests for web-based data.
- NLTK SentimentIntensityAnalyzer: Performs sentiment scoring with VADER model.
- NLTK stopwords: Removes common words that don’t contribute to meaning-making.
- NLTK tokenize: Splits text into individual words or phrases for analysis.
- TextBlob: Secondary sentiment analysis tool used for polarity and subjectivity scoring.

- scikit-learn TfidfVectorizer: Converts text into numerical features based on word importance.
- scikit-learn CountVectorizer: Converts text into frequency-based numerical representations.
- scikit-learn NMF: A matrix factorization machine learning technique that decomposes a document-term matrix into two lower-dimensional matrices.
- scikit-learn LatentDirichletAllocation: A probabilistic machine learning model that assumes each document is a mixture of topics and each topic is a mixture of words.
- langdetect: Detects the language of a given text to filter non-English posts.
- Matplotlib Pyplot: Main library for generating static visualizations.
- Matplotlib dates: Formats date-based plots to visualize trends over time.
- Matplotlib patches: Creates custom visual elements within figures.
- Seaborn: Enhances visualizations with advanced statistical plots.
- Plotly express: Enables interactive visualizations for dynamic sentiment trend analysis.
- Folium: For interactively mapping sentiment across the U.S.
- branca: Helper library for Folium to add custom pop-ups and elements to maps.
- NetworkX: Visualizes word co-occurrences in social media discourse.
- WordCloud wordcloud: Generates word clouds from text data.

Appendix B

DATA ACQUISITION AND METADATA OVERVIEW

The main Instagram dataset was obtained by targeting a combination of election-related hashtags: #election, #presidentialelection, #republican, #democrats, #vote, and #president. The resulting dataset, *result.csv*, comprises a total of 51,022 posts, capturing various metadata points including but not limited to post URLs, profile username and full names, engagement metrics, publish dates, captions, location data, and relevant queried hashtags.

To gain a deeper understanding of Instagram engagement patterns, an additional dataset was generated, labeled *posts.csv*. This dataset contains 4,356 posts, drawn as a sample from *result.csv*, with extended metadata fields of tagged users, post IDs, type of post, video URLs and duration, and play counts.

X data was collected by scraping posts containing the hashtag #uselection2024 to attempt to keep relevance to presidential election discourse. A total of 2,429 posts were retrieved with metadata such as captions, publish dates, profile information, and URLs.