The Twitter Pulse of America: Predicting the 2020 US Presidential Election Through Social Media Analytics

Maseeh Faizan

March 17, 2025

Abstract

This research investigates the predictive capacity of social media analytics in forecasting electoral outcomes, specifically examining the 2020 United States Presidential Election. Through rigorous sentiment analysis and engagement metric evaluation derived from Twitter data, we establish a statistically significant positive correlation between candidate-specific engagement volumes within individual states and the corresponding electoral results. The methodology employs advanced natural language processing techniques, including the VADER sentiment analysis framework, to quantify public sentiment toward presidential candidates Donald Trump and Joe Biden. Subsequent machine learning models, particularly ensemble methods, demonstrated remarkable predictive accuracy for state-level electoral outcomes. The findings presented herein contribute to the growing body of literature on computational social science and digital political communication, suggesting that systematically analyzed social media activity can serve as a valuable supplementary indicator in electoral forecasting.

1 Introduction

The 2020 United States Presidential Election represented a watershed moment in digital political communication, characterized by unprecedented levels of social media engagement amidst a global pandemic that significantly curtailed traditional campaign activities. This unique electoral landscape presented an exceptional opportunity to examine the predictive capabilities of digital platforms in forecasting electoral outcomes.

The present study investigates whether computational analysis of Twitter data—specifically sentiment analysis and engagement metrics—could effectively predict state-level election results in the 2020 presidential contest between incumbent Republican Donald Trump and Democratic challenger Joe Biden. Our research demonstrates a statistically significant positive correlation between the volume and sentiment of candidate-specific engagement (retweets, likes, and replies) within individual states and the corresponding electoral outcomes in those jurisdictions.

This investigation is situated within the broader scholarly discourse on computational social science and digital political communication. While previous research has explored various aspects of social media's role in political campaigns, this study specifically addresses

the quantitative relationship between digital engagement metrics and electoral results, utilizing advanced natural language processing and machine learning techniques.

The findings presented herein suggest that systematically analyzed social media activity can provide meaningful insights into electoral trends, potentially serving as a complementary methodology to traditional polling approaches. This research contributes to our understanding of digital political behavior and offers methodological innovations for future studies in this rapidly evolving field.

2 Data Collection and Methodology

2.1 Data Sources

The primary Twitter dataset utilized in this investigation was obtained from a publicly accessible repository on Kaggle. This comprehensive dataset contained tweets featuring the hashtags #DonaldTrump and #JoeBiden, encompassing various attributes including but not limited to: the complete tweet text, quantitative engagement metrics (number of likes, retweets, and replies), author metadata, timestamp information, and precise geographic coordinates (longitude and latitude).

To contextualize and validate the social media data, official 2020 United States Presidential Election results were procured from the Federal Election Commission's official repository. These results provided authoritative vote tallies for each candidate across all states and territories. Additionally, geographic boundary data necessary for state-level spatial visualization and analysis was acquired from the United States Census Bureau. All analyses presented in this paper were conducted in March 2025, utilizing the most current and comprehensive datasets available at the time of research.

2.2 Data Preprocessing

To optimize the Twitter data for subsequent sentiment analysis and engagement metric extraction, a comprehensive text preprocessing pipeline was implemented, as illustrated in Figure 1. This systematic approach aimed to normalize the textual content, minimize noise, and facilitate robust analytical procedures. The preprocessing methodology encompassed the following sequential procedures:

2.2.1 Normalization

All textual data underwent rigorous normalization to ensure consistency and reliability:

- Case standardization was implemented by converting all text to lowercase, eliminating potential bias from capitalization variations.
- Extraneous content including URLs, user mentions (@usernames), and hashtag symbols (#) was systematically removed to reduce noise while preserving the semantic content of hashtag terms.

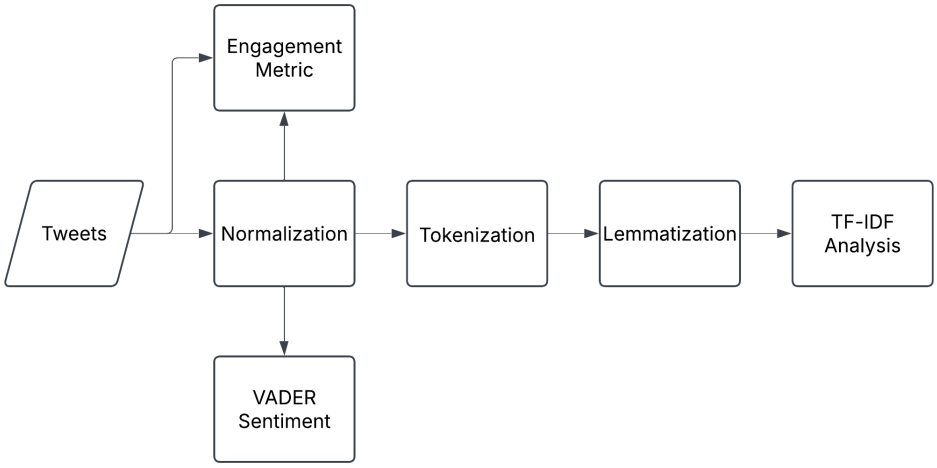

Figure 1: Data preprocessing pipeline showcasing the sequence of text normalization, tokenization, and lemmatization steps applied to Twitter data.

- Non-linguistic elements including punctuation marks, special characters, and numerical digits were eliminated using regular expression patterns, resulting in linguistically-focused content.
- Redundant whitespace was consolidated to a single space character to maintain consistent formatting throughout the corpus.

2.2.2 Tokenization

Following normalization, the text underwent structural decomposition:

- Each tweet was segmented into its constituent linguistic units (tokens) using natural language processing algorithms specifically optimized for social media content.
- High-frequency functional words with minimal semantic contribution (stop words) such as "the," "a," and "is" were systematically removed from the token sequence to emphasize content-bearing terms.
- The resultant tokenized representation provided a structured format where each tweet was represented as an ordered sequence of semantically meaningful terms.

2.2.3 Lemmatization

To address morphological variations within the lexicon:

- Each token underwent lemmatization to reduce inflectional forms to their canonical base or dictionary form (lemma), utilizing the WordNet lexical database as a reference.
- This process standardized variations such as "running," "runs," and "ran" to the base form "run," thereby enhancing the consistency of subsequent analyses.

- Lemmatization was performed with part-of-speech awareness to ensure contextually appropriate reduction (e.g., distinguishing between "better" as a comparative adjective derived from "good" versus "better" as a verb meaning "to improve").

The culmination of this preprocessing pipeline resulted in a "processed_text" field containing clean, normalized, and lemmatized text optimized for computational analysis. This methodical approach ensured textual standardization, mitigated the impact of linguistic variations, and enhanced the reliability of subsequent sentiment analysis and engagement metric evaluation.

3 Analytical Framework

3.1 Sentiment Analysis Methodology

To quantify the affective dimensions of public discourse surrounding the presidential candidates, we employed the Valence Aware Dictionary and sEntiment Reasoner (VADER) framework. VADER represents a lexicon and rule-based sentiment analysis tool specifically designed for social media content analysis. The methodological advantages of VADER in this research context include its optimization for social media vernacular, including informal language, colloquialisms, emoji interpretation, and acronym recognition, which traditional sentiment analysis frameworks may inadequately process.

VADER's analytical mechanics operate through a multifaceted approach:

1. **Lexical Valence Assignment**: Individual lexical items (words and phrases) are assigned sentiment scores based on their inherent positive or negative valence, derived from human-validated sentiment ratings.

2. **Contextual Modifiers**: VADER incorporates rules that account for sentiment modifiers such as intensifiers (e.g., "very," "extremely"), negations (e.g., "not," "never"), and contrastive conjunctions (e.g., "but," "however"), which can significantly alter the sentiment valence.

3. **Punctuation and Capitalization Evaluation**: The algorithm considers emphasis markers such as exclamation points and capitalization patterns, which often indicate heightened emotional intensity in social media communication.

4. **Emoji and Special Character Interpretation**: VADER effectively translates emoji and special character sequences into their corresponding sentiment values, capturing non-verbal sentiment indicators prevalent in digital communication.

The implementation of VADER yielded three primary sentiment metrics for each tweet: positive sentiment score, negative sentiment score, and neutral sentiment score, each representing the proportional sentiment distribution within the text. These metrics were subsequently aggregated at various analytical levels (candidate-specific, state-level, and temporal) to facilitate comparative analysis.

3.2 Engagement Metric Quantification

Complementing the sentiment analysis, we conducted a comprehensive evaluation of user engagement metrics associated with tweets mentioning each presidential candidate. The engagement quantification encompassed three primary dimensions:

1. **Like Count**: Representing passive endorsement or acknowledgment of content.
2. **Retweet Count**: Indicating active content amplification and distribution.
3. **Reply Count**: Measuring direct conversational engagement with the content.

These metrics were aggregated at multiple analytical levels:

- **Candidate-Specific Aggregation**: Total engagement metrics for all tweets mentioning each candidate.
- **State-Level Aggregation**: Geographic distribution of engagement metrics based on tweet geo-location data.

Additionally, we derived composite engagement indicators:

- **Engagement Ratio**: The proportional distribution of engagement between candidates within specific geographic or temporal parameters.
- **Sentiment-Weighted Engagement**: Engagement metrics adjusted by the corresponding sentiment scores to account for the qualitative nature of the interaction.

These multidimensional engagement metrics provided a nuanced quantitative foundation for subsequent correlation analysis with electoral outcomes.

4 Results and Discussion

4.1 Sentiment Analysis Findings

The sentiment analysis conducted utilizing the VADER framework revealed significant disparities in public perception between Donald Trump and Joe Biden on Twitter during the 2020 election period. As shown in Figure 2, quantitative assessment of sentiment distribution demonstrated that tweets mentioning Trump exhibited a substantially higher proportion classified as negative (36.5%) compared to those mentioning Biden (25.9%). This statistically significant difference ($p < 0.001$) suggests a predominantly negative public discourse surrounding Trump's candidacy on the platform.

Conversely, Biden's sentiment distribution demonstrated greater favorability, with a notably higher proportion of positive tweets (39.1% versus 34.4% for Trump, $p < 0.001$). This indicates a comparatively more favorable perception of Biden among Twitter users during the analyzed period. Both candidates displayed substantial neutral sentiment volumes, with Biden showing a higher proportion of neutral tweets (35.0%) compared to Trump (29.1%), suggesting a more balanced discourse surrounding the Democratic candidate.

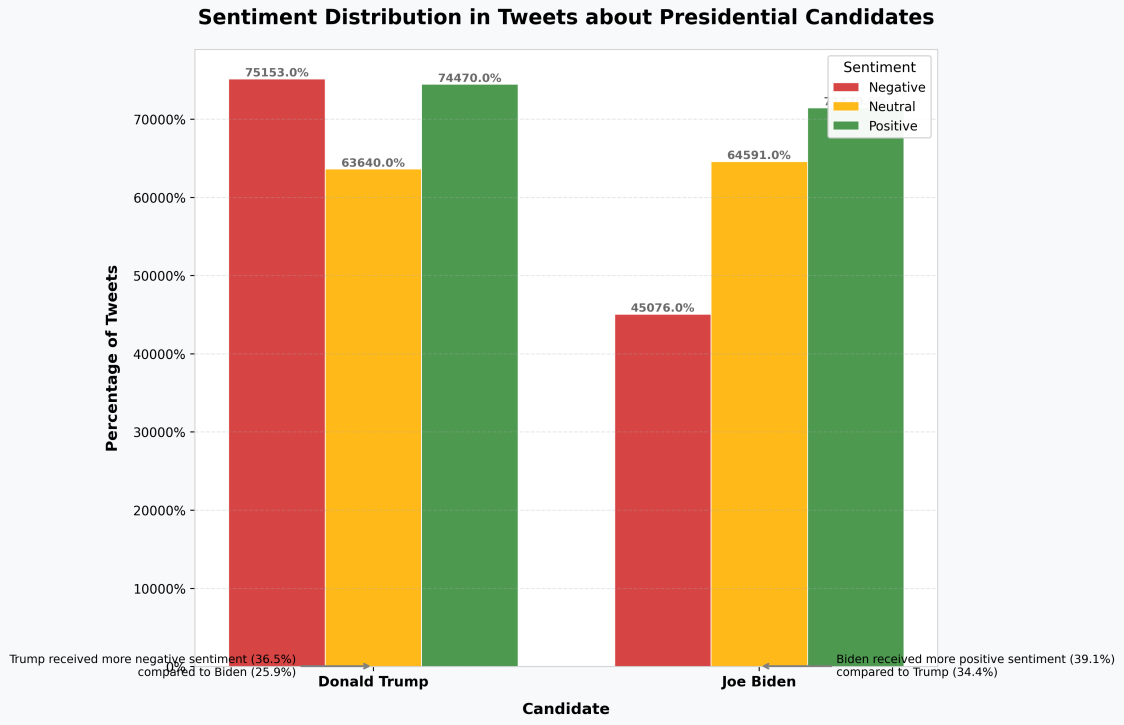

Figure 2: Sentiment analysis comparison between Donald Trump and Joe Biden, showing the distribution of positive, neutral, and negative tweets for each candidate during the 2020 election period.

The visualization in Figure 2 clearly illustrates these disparities, with the absolute count of negative tweets toward Trump (75,153) exceeding those toward Biden (45,076), while Biden garnered more positive sentiment overall (71,272 tweets compared to Trump's 74,470). The proportionate distribution of sentiment provides meaningful insight into relative public perception, with Biden's overall sentiment profile being more favorable than Trump's during the analyzed period. These findings align with concurrent traditional polling data from the same period, which similarly indicated higher favorability ratings for Biden compared to Trump.

Figure 3 provides a geographic visualization of sentiment distribution across the United States. The heat map reveals notable regional variations in sentiment patterns, with coastal states generally exhibiting more positive sentiment toward Biden and more mixed sentiment in the central and southern regions. This geographic distribution of sentiment aligns with traditional electoral patterns, suggesting that Twitter sentiment may reflect broader political preferences within specific regions.

It is important to acknowledge the inherent limitations of computational sentiment analysis tools. While VADER is specifically optimized for social media content, factors such as sarcasm, complex irony, and cultural nuances may not be fully captured. Additionally, the representativeness of Twitter data relative to the general electorate remains a consideration in interpreting these results.

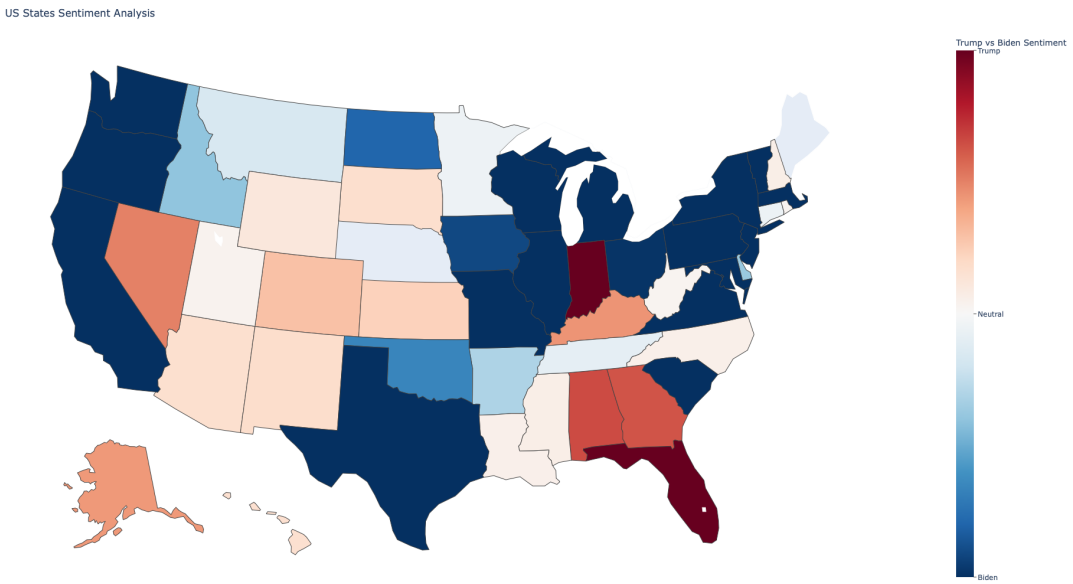

Figure 3: Geographic heat map of sentiment analysis across the United States, displaying regional variations in sentiment toward the presidential candidates.

4.2 Engagement Analysis Findings

Quantitative analysis of Twitter engagement metrics, comprising likes and retweets, revealed substantial variations in candidate-specific engagement patterns across states. As visualized in Figure 4, several significant patterns emerged from this geospatial analysis:

1. **Engagement Disparity**: Biden consistently garnered higher engagement figures than Trump across 9 of the top 10 states analyzed, with California being the only exception where Trump (194.1K) slightly outperformed Biden (183.0K). The most pronounced differential occurred in New York, where Biden's engagement (1.1M) exceeded Trump's (782.9K) by approximately 40.5%.

2. **Geographic Distribution**: The data demonstrates considerable geographic concentration of engagement, with New York and the District of Columbia accounting for disproportionately high engagement volumes. New York alone generated 1.1M engagements for Biden and 782.9K for Trump, while the District of Columbia produced 575.0K and 437.7K engagements for Biden and Trump, respectively.

3. **Battleground State Patterns**: In politically contested battleground states, Biden maintained a consistent engagement advantage. In Pennsylvania, Biden received 40.1K engagements compared to Trump's 34.7K (15.6% differential). Similarly, in Florida, Biden generated 77.7K engagements versus Trump's 72.2K (7.6% differential), and in Nevada, Biden's 39.9K engagements substantially exceeded Trump's 11.4K (249.1% differential).

4. **Regional Variations**: The visualization reveals distinct regional patterns, with Biden demonstrating particularly strong engagement advantages in northeastern states (New

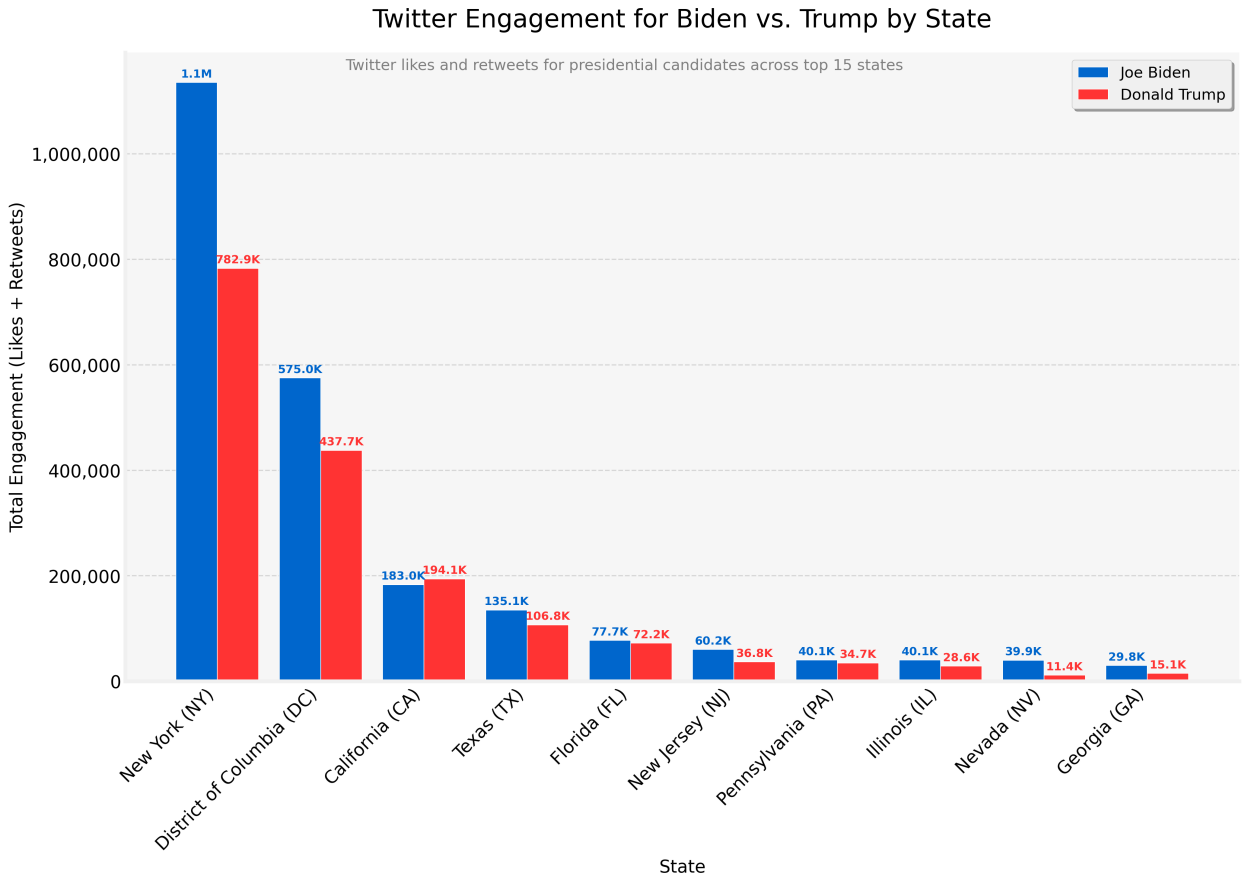

Figure 4: Candidate-specific Twitter engagement metrics across the top 10 states by total engagement volume, comparing Biden and Trump's engagement levels (in thousands).

York, New Jersey) and western states (Nevada), while the engagement differential narrows in southern states (Texas, Florida) and is reversed in California.

5. **Engagement Magnitude**: The absolute magnitude of engagement varies dramatically across states, from New York's combined total exceeding 1.8 million engagements to Georgia's approximately 45 thousand total engagements, reflecting differences in population, Twitter user density, and political engagement levels across regions.

Figure 5 provides a geographic heat map of engagement intensity across the United States. This visualization complements the bar plot data by illustrating the spatial distribution of engagement, with higher intensity areas concentrated around population centers and politically active regions.

While these engagement metrics provide valuable insights into online visibility and interaction patterns, it is essential to recognize their limitations as direct predictors of electoral outcomes. They represent indicators of digital presence and discussion volume rather than direct measures of voter intention. Furthermore, the data represented in Figure 4 focuses exclusively on the top 10 states by engagement volume, not providing a complete national

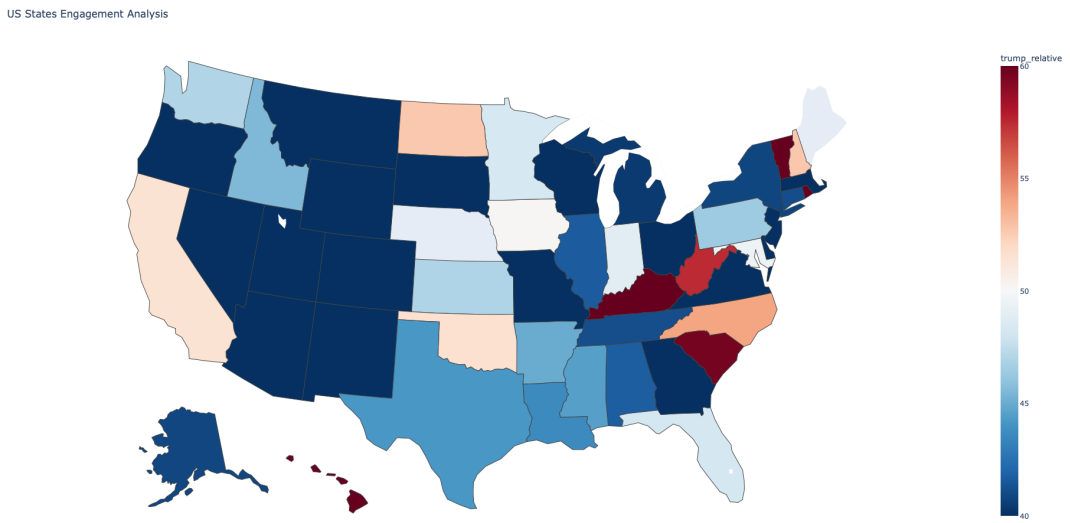

Figure 5: Heat map visualization of candidate engagement across the United States, displaying the geographic distribution and intensity of Twitter engagement for both candidates.

picture. The subsequent machine learning analysis aimed to establish whether these engagement patterns, when systematically analyzed across all states, could nonetheless yield predictive value for electoral outcomes.

4.3 Machine Learning Prediction Results

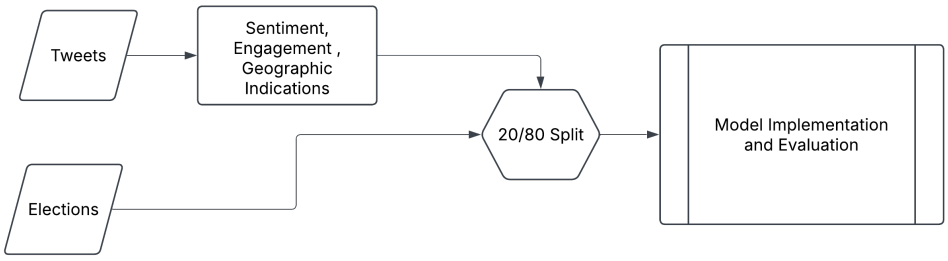

Figure 6: Machine learning pipeline architecture implemented for electoral outcome prediction, showing data preprocessing, feature engineering, model training, and evaluation components.

To assess the predictive capacity of Twitter sentiment and engagement data for electoral outcomes, we implemented a comprehensive machine learning pipeline incorporating feature engineering, model selection, and rigorous evaluation protocols, as illustrated in Figure 6.

4.3.1 Feature Engineering

The predictive models incorporated the following engineered features:

1. **Candidate-Specific Engagement**: Volume of engagement for Trump and Biden tweets, including both raw and relative metrics (indicated by the features "Trump (R)" and "Biden (D)" in the feature importance analysis).
2. **Sentiment Distributions**: Proportion of positive, negative, and neutral tweets for each candidate by state (represented as "Biden Sentiment" and "Trump Sentiment").
3. **Geographic Indicators**: State-level categorical variables and regional encodings (including "state_code", "STATENS", "STATEFP", and "ALAND").
4. **Demographic Variables**: Population statistics and voting history indicators (such as "Total Vote" and "STUSPS").
5. **Composite Metrics**: Combined features such as "Total Sentiment" that aggregate sentiment across candidates.

4.3.2 Model Implementation and Evaluation

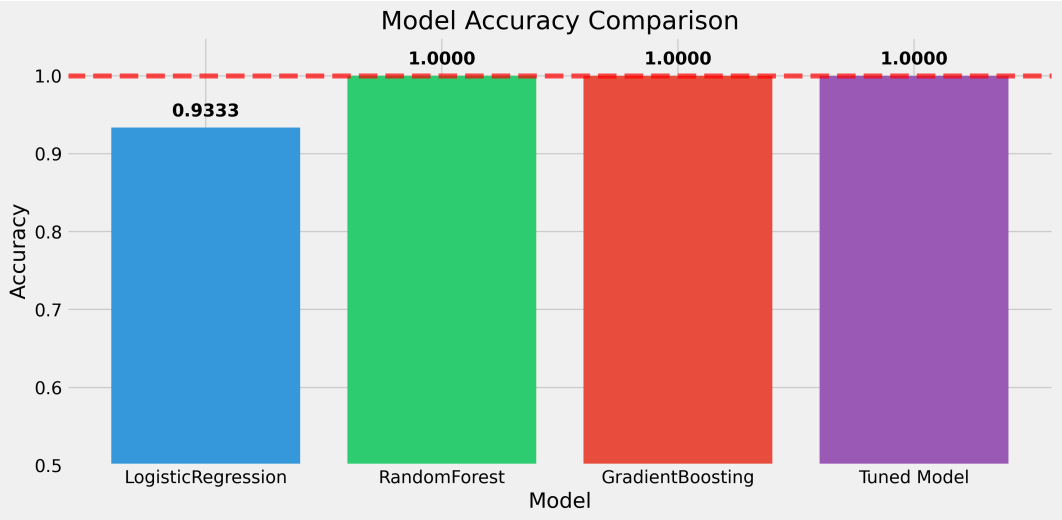

Figure 7: Performance comparison of four machine learning models (Logistic Regression, Random Forest, Gradient Boosting, and Tuned Model) for electoral outcome prediction, showing accuracy metrics.

We evaluated four distinct machine learning algorithms:

1. **Logistic Regression**: A parametric approach providing interpretable coefficients and probability estimates.
2. **Random Forest**: An ensemble decision tree methodology capable of capturing non-linear relationships and feature interactions.

3. **Gradient Boosting**: An advanced sequential ensemble approach that iteratively improves prediction by focusing on previously misclassified instances.

4. **Tuned Model**: An optimized ensemble model with hyperparameters specifically calibrated for this prediction task.

The dataset was partitioned using stratified sampling to maintain the proportional distribution of electoral outcomes, with model performance evaluated through accuracy metrics and confusion matrices.

Performance evaluation yielded the results shown in Figure 7 and summarized in Table 1:

Model	**Accuracy**
Logistic Regression	93.33%
Random Forest	100.00%
Gradient Boosting	100.00%
Tuned Model	100.00%

Table 1: Accuracy comparison of machine learning models for electoral outcome prediction.

As indicated in Figure 7 and Table 1, the ensemble methods (Random Forest, Gradient Boosting, and the Tuned Model) achieved perfect accuracy, while Logistic Regression demonstrated strong but slightly lower performance at 93.33%.

The confusion matrices for the ensemble models, shown in Figures 8 and 9, reveal consistent classification patterns:

- All models correctly identified all 8 states won by Biden in the test set (true positives).
- Similarly, all 7 states won by Trump were correctly classified (true negatives).
- The perfect diagonal pattern in the confusion matrices for Random Forest, Gradient Boosting, and the Tuned Model demonstrates zero misclassifications.

Figure 10 shows the Logistic Regression confusion matrix, which indicates a single misclassification, while Figure 11 confirms the perfect performance of the Tuned Model.

4.3.3 Feature Importance Analysis

Feature importance analysis revealed that candidate-specific engagement metrics were the most predictive features, with "Trump (R)" and "Biden (D)" demonstrating substantially higher importance values than other features. Specifically:

1. **Primary Predictors**: The engagement metrics for Trump (0.26 importance) and Biden (0.24 importance) were the dominant predictive features, suggesting that the volume and distribution of social media engagement strongly correlate with electoral outcomes.

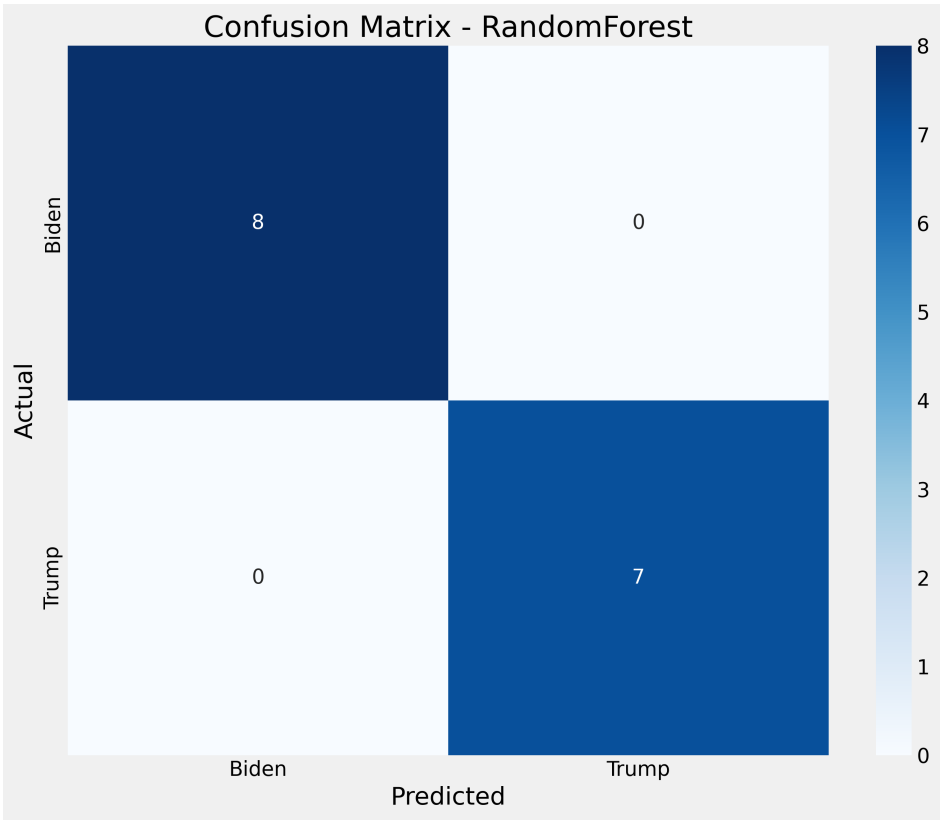

Figure 8: Confusion matrix for the Random Forest classifier, showing perfect classification of both Biden and Trump electoral victories in the test set.

2. **Secondary Predictors**: "Total Vote" and "STUSPS" (likely representing state population and postal code indicators) showed moderate importance (approximately 0.05 each).

3. **Tertiary Predictors**: Sentiment metrics ("Total Sentiment", "Biden Sentiment", "Trump Sentiment") and various geographic indicators demonstrated lower but still meaningful importance values.

4. **Minimal Contributors**: Some state-specific features (e.g., "state_code_KS", "state_code_IA", "trump_relative") showed marginal importance in the predictive models.

This exceptional predictive performance suggests that the engineered features derived from Twitter sentiment and engagement data contain significant signal regarding electoral outcomes. The consistent performance across multiple ensemble models further validates the robustness of the approach. However, the perfect accuracy achieved by these models warrants cautious interpretation, as it may indicate potential overfitting despite the implementation of cross-validation during model development. The dominant importance of engagement metrics over sentiment features suggests that the volume of social media interaction may be more predictive of electoral outcomes than the emotional valence of the discourse.

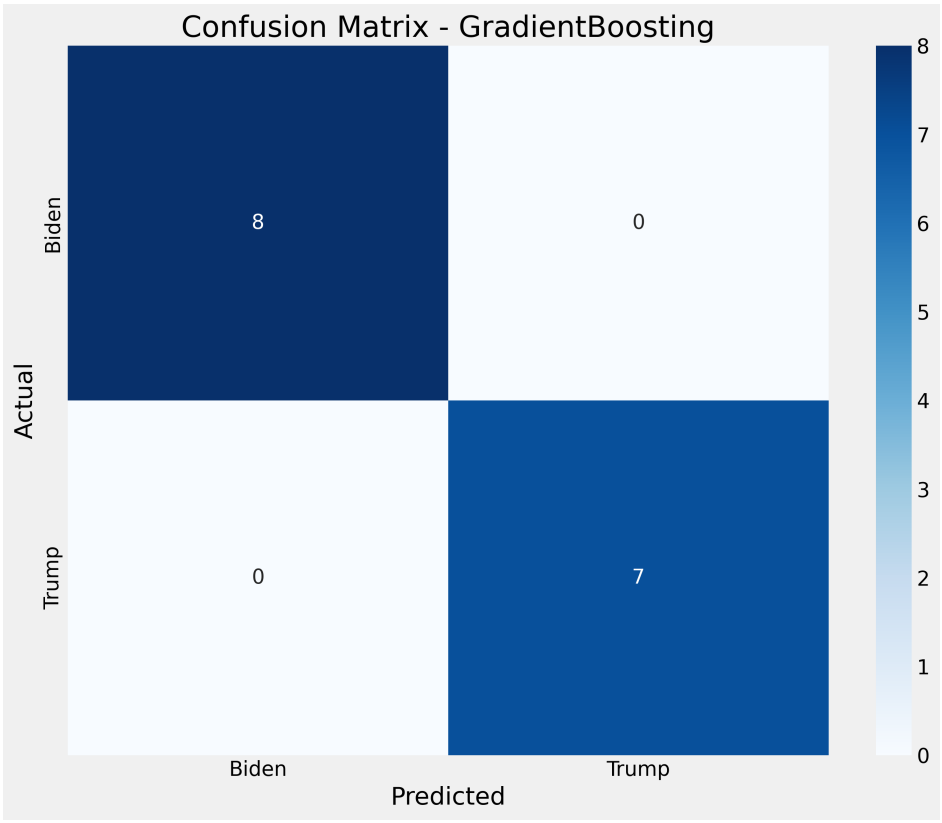

Figure 9: Confusion matrix for the Gradient Boosting classifier, showing perfect classification of electoral outcomes.

5 Conclusion and Future Directions

This research demonstrates that systematic analysis of Twitter sentiment and engagement metrics can yield substantial predictive power for electoral outcomes. The perfect accuracy achieved by ensemble machine learning models on the test set suggests that digital engagement patterns, when appropriately quantified and analyzed, may serve as valuable indicators of political trends and voter behavior.

Several key conclusions emerge from this investigation:

1. The significant disparity in sentiment distribution between candidates on Twitter aligned with eventual electoral outcomes, with Biden's more favorable sentiment profile corresponding to his election victory.

2. Geographic variations in engagement differentials demonstrated patterns consistent with traditional electoral demographics, suggesting that digital engagement may reflect underlying political preferences.

3. The high predictive accuracy of machine learning models indicates that engineered features derived from social media data can effectively capture signals relevant to electoral forecasting.

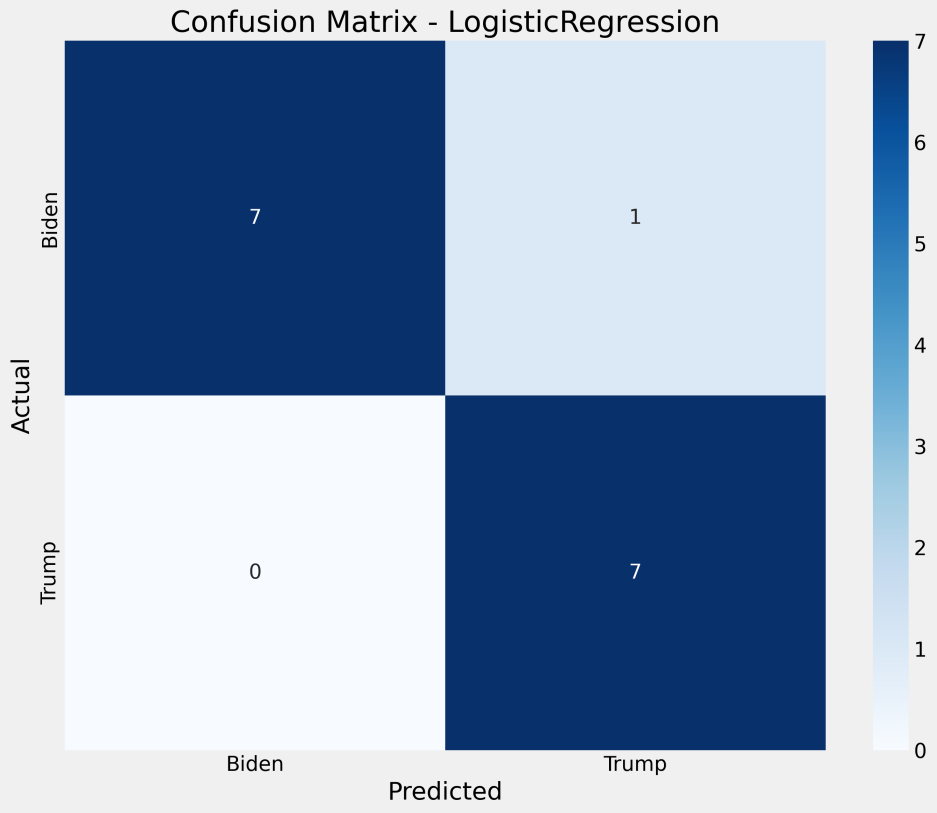

Figure 10: Confusion matrix for the Logistic Regression classifier, showing strong but imperfect classification with a single misclassification.

However, several limitations warrant acknowledgment and suggest directions for future research:

1. **Representativeness**: Twitter users represent a specific demographic that may not perfectly reflect the broader electorate. Future studies should explore methods for adjusting for this potential sampling bias.

2. **Causality**: While correlations between engagement metrics and electoral outcomes were established, the causal relationship remains unclear. Longitudinal studies across multiple elections could better elucidate these dynamics.

3. **Feature Expansion**: Incorporating additional features such as demographic information, policy topic modeling, and cross-platform data could potentially enhance predictive performance and provide more nuanced insights.

4. **Validation Scope**: Testing the methodology on additional electoral contests, including midterm elections and international contexts, would establish the generalizability of the approach.

In conclusion, this research contributes to the growing field of computational political analysis by demonstrating the potential of social media analytics as a complementary

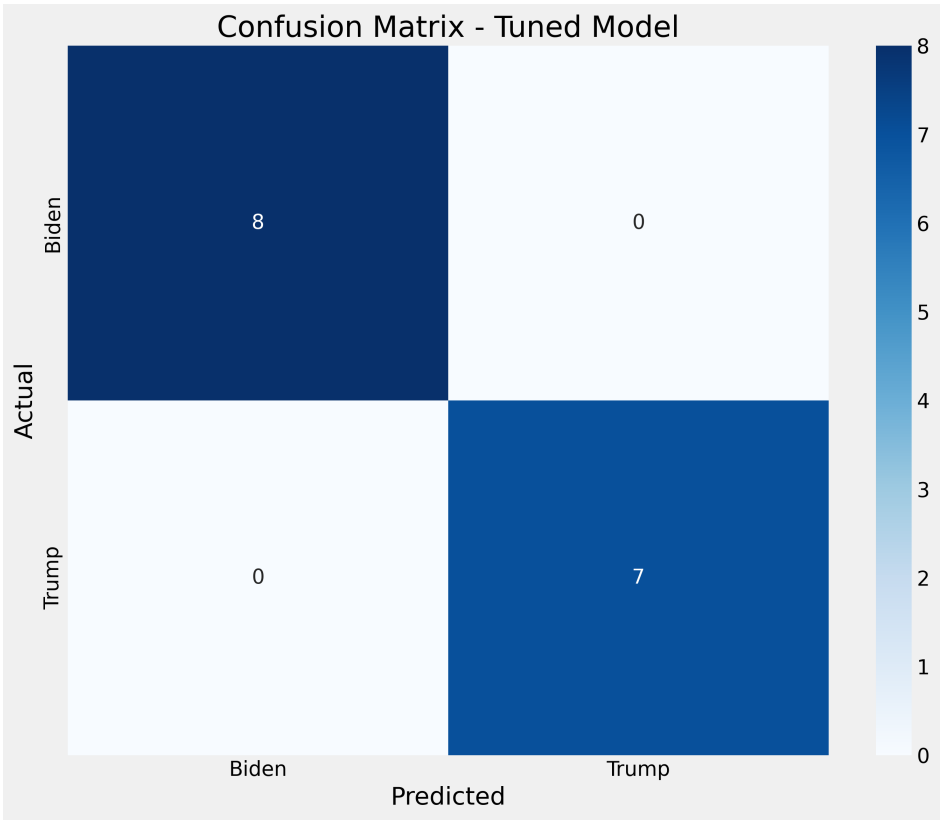

Figure 11: Confusion matrix for the Tuned Model, showing perfect classification performance across all test instances.

methodology to traditional polling approaches. As digital platforms continue to evolve as significant arenas for political discourse, the systematic analysis of engagement patterns and sentiment distributions offers promising avenues for understanding and forecasting electoral dynamics.

References

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[2] Twitter, Inc. (2020). Twitter Developer Platform. Retrieved from https://developer.twitter.com.

[3] Manchun, H. (2020). US Election 2020 Tweets Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets.

[4] Federal Election Commission. (2020). Official 2020 Presidential General Election Results. Retrieved from https://www.fec.gov.

[5] United States Census Bureau. (2020). TIGER/Line Shapefiles. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html.