Offensive Language on Twitter During the 2024 U.S. Election

Sahibnoor Singh

2025-03-17

Supervised by Paul Bauer

Table of contents

1 Abstract

The rapid expansion of social media platforms, particularly Twitter, has profoundly reshaped political communication by providing a dynamic environment for real-time public discourse and engagement. However, alongside this increased connectivity, offensive content—including hate speech, insults, and derogatory remarks—has proliferated, posing significant threats to social cohesion, individual psychological well-being, and the integrity of democratic processes. Understanding the patterns, distribution, and implications of offensive language, especially in politically charged contexts such as elections, is therefore essential.

This study systematically investigates offensive language dynamics on Twitter within the context of the 2024 U.S. Presidential Election. Leveraging the publicly available Offensive Language Identification Dataset (OLID), we developed robust machine learning models—Logistic Regression, Random Forest, Multinomial Naive Bayes, and XGBoost—to detect offensive content based on textual features extracted through TF-IDF vectorization. Among the evaluated models, the XGBoost classifier demonstrated the highest predictive accuracy (77.19%) along with balanced precision and recall metrics, effectively capturing nuanced linguistic indicators of offensive sentiment.

To extend our analysis beyond model validation, the trained XGBoost model was applied to a novel and substantial corpus of approximately 250,000 tweets specifically related to the 2024 U.S. elections. Results revealed significant patterns in offensive sentiment directed toward prominent political accounts, including GOP, KamalaHQ, JoeBiden, elonmusk, and GuntherEagleman, highlighting targeted and polarized interactions influenced by political ideology and public controversies. Temporal analyses consistently demonstrated a decreasing trend in offensive sentiment frequency as the election approached, suggesting evolving societal norms, increased moderation, or changing user engagement behaviors.

Moreover, detailed party-based analyses exposed substantial differences in offensive language use across political affiliations. Tweets referencing the Republican Party consistently exhibited the highest levels of offensive sentiment throughout the study period, whereas smaller or less prominent parties maintained comparatively stable, lower levels. Further examination of temporal sentiment dynamics indicated fluctuations associated with political events and controversies, particularly evident in interactions involving major political figures and accounts.

These findings carry significant implications for multiple stakeholders. For policymakers and platform regulators, our results provide insights into targeted moderation strategies, essential for preserving constructive democratic discourse. For computational linguistics and natural language processing research communities, this study illustrates effective methodologies for detecting and analyzing offensive content within politically sensitive contexts, contributing to the broader field by testing model generalizability and effectiveness across diverse textual data streams.

Ultimately, by integrating computational modeling techniques with sociopolitical theories—including Agenda-Setting Theory, Social Identity Theory, and Uses and Gratifications Theory—this research offers actionable insights aimed at fostering healthier, more inclusive public discourse online. Our methodological approach and analytical findings serve as a foundation for future investigations into offensive language detection, emphasizing the importance of computational tools in mitigating online harm and promoting respectful communication during critical political events.

2 Introduction

The rapid expansion of social media platforms, particularly Twitter, has significantly transformed political communication and public discourse. Twitter has become a crucial platform for real-time information exchange, where politicians, media outlets, and citizens interact. However, this increased connectivity also facilitates the spread of offensive content such as hate speech, insults, and derogatory remarks (Zampieri et al., 2019). The consequences of offensive language extend beyond incivility, posing risks to social cohesion, individual psychological well-being, and the integrity of democratic processes.

Offensive language encompasses insults, profanity, or direct attacks targeting individuals or groups based on identity, beliefs, or affiliations (Zampieri et al., 2019). Its prominence on Twitter intensifies during major political events, such as elections, where rhetoric and emotions escalate significantly.

2.1 Motivation and Theoretical Framework

Studying offensive language on Twitter during elections is crucial due to its potential impact on:

- **Public Opinion:** Offensive political content can distort democratic discourse by amplifying divisive perspectives.
- **Voter Engagement:** Hostile communication may deter civic participation, particularly among marginalized groups.
- **Online Radicalization:** Persistent offensive content can normalize extremist views and potentially lead to offline violence.
- **Platform Governance:** Insights into offensive language can help platforms improve moderation practices during politically sensitive periods.

The theoretical foundations guiding this investigation include:

- **Agenda-Setting Theory:** Influential social media users shape public debate, potentially legitimizing offensive rhetoric.
- **Social Identity Theory:** Offensive language often reinforces in-group solidarity and hostility towards out-groups.
- **Uses and Gratifications Theory:** Users may employ offensive communication to fulfill emotional or psychological needs.

2.2 Empirical Context and Significance

Previous elections demonstrate how spikes in offensive language coincide with major political events, controversies, or candidate behavior. During the 2020 U.S. Presidential Election, offensive tweets surged around debates and policy announcements, underscoring how offensive language reflects broader political dynamics. Such phenomena reveal that offensive language online has tangible offline implications, including reduced well-being among targeted groups and potential radicalization.

2.3 Research Questions

This study systematically addresses the following research questions within the context of the 2024 U.S. Presidential Election:

1. **Can we build an accurate predictive model for offensive language detection?**
    - Leveraging the Offensive Language Identification Dataset (OLID) to train and validate machine learning models.
2. **How does offensive language evolve temporally during the election?**
    - Tracking fluctuations in offensive tweets throughout the election period to identify significant events and trends.
3. **How is offensive language distributed across the U.S. Party?**
    - Analyzing offensive tweets amoung different political parties such as Republican Party, Democratic Party,Independent, Green Party,and Libertarian Party.
4. **How is offensive language distributed among Twitter users?**
    - Assessing whether offensive tweets predominantly originate from a small subset of users or are widespread, informing targeted moderation strategies.

2.4 Broader Implications

This research offers significant implications for stakeholders:

- **Policy and Regulation:** Inform targeted interventions and regulations to enhance constructive public discourse.
- **Platform Governance:** Refine content moderation algorithms and policies to mitigate harmful online behavior.
- **Public Awareness:** Enhance critical understanding among journalists, educators, and the general public about offensive communication online.
- **Methodological Advancement:** Contribute to computational linguistics and NLP literature by testing model generalizability in event-specific contexts.

By integrating computational analysis with sociopolitical theory, this study provides actionable insights to foster respectful and inclusive online discourse, especially during politically sensitive events.

3 Data and Methods

3.1 Data Loading

In this section, we begin by importing the necessary libraries for data handling, visualization, and model development. We utilize pandas and numpy for data manipulation, matplotlib and seaborn for creating visual representations, and wordcloud for generating word cloud images of the text data. In addition, several scikit-learn modules (including those for model selection, feature

extraction, and various classifiers such as Logistic Regression, Random Forest, and Support Vector Machines) are imported to facilitate the development and evaluation of our classification models. We also import spaCy and its built-in English stop words to support text processing tasks, as well as regular expressions (re) for string manipulation. To maintain a clean output, warnings are suppressed during execution.

Next, two datasets are loaded using pandas: the first dataset, OLID_Data, contains tweets with binary labels in the subtask_a column, where offensive tweets are represented as 1 and non-offensive tweets as 0. This dataset will be used for training our binary classification model. The second dataset, Act_Elec_Data, comprises tweet data related to U.S. elections without associated labels; it will serve as the test dataset for predicting offensive language using the model trained on the OLID data. Finally, the first few rows of each dataset are printed to verify that the data have been loaded correctly and to inspect the structure and content before proceeding with further analysis and model development.

3.2 Data Overview

- **OLID Dataset**: This study is based on a dataset OLID Huggingface containing 14,100 tweets. Each tweet has been assigned a binary label for subtask a, indicating whether it is offensive (1) or not offensive (0). Of the 14,100 tweets, 9,460 (approximately 67.09%) are labeled as not offensive, and 4,640 (about 32.91%) are labeled as offensive. Two additional columns, subtask_b and subtask_c, provide more specific annotations but only for a subset of the data (specifically, the offensive tweets). On average, each tweet in this dataset has a length of about 127 characters, with the shortest tweet being 10 characters long and the longest tweet extending to 580 characters. This variation in text length ensures a broad representation of tweet lengths for tasks involving text classification, lexicon analysis, and other natural language processing objectives.
- **2024-us-election-dataset**: To facilitate **offensive vs. non-offensive** sentiment analysis, a large corpus of tweets from **June through October** was collected and consolidated into a single CSV file. This resulting file contains approximately **250,000** tweets, each with a consistent column structure modeled after the USC-X-24-US-Election GitHub repository. Among these columns, the **tweet text** is of primary importance, as it serves as the input for a sentiment classification model. The chosen model for this task—initially trained and evaluated on the **OLID dataset**—is now being applied to predict whether each tweet in this new corpus is offensive or non-offensive, effectively extending the trained model's capabilities to a broader, election-focused data stream.

3.3 Methodology

Exploratory data analysis of Offensive Language Identification Dataset (OLID):

- **Data Filtering**
  Rows containing missing values in the `cleaned_tweet` column were removed. This ensured that all subsequent analyses focused solely on entries with valid, usable text data.

- **Sentiment Distribution Visualization**
  A count plot was produced to depict the number of tweets classified under each sentiment in `subtask_a`. The seaborn library was employed to generate a bar chart showing how many tweets were labeled as offensive (`1`) or not offensive (`0`).
- **Word Cloud Generation**
  Offensive and non-offensive tweets were separated based on `subtask_a`. Each subset's text was concatenated to form two large strings. The WordCloud library was then used to create word clouds for each string, illustrating the most prominent words appearing in offensive versus non-offensive tweets. Matplotlib's subplot functionality was utilized to display the two word clouds side by side.
- **Tweet Length Calculation**
  The number of words in each tweet was determined by splitting the `cleaned_tweet` column into individual tokens (separated by whitespace) and counting the resulting tokens. These counts were stored in a new feature (`tweet_length`) for each entry, serving as a quantitative measure of tweet size.
- **Distribution of Words per Tweet**
  A histogram was generated with seaborn's `histplot` to visualize the distribution of `tweet_length` across all tweets. The chart included binning (for categorizing tweet lengths) and a kernel density estimation (`kde=True`) for a smoother visualization of how word counts are distributed.
- **Comparison of Tweet Length by Sentiment**
  A box plot was created to compare `tweet_length` between the offensive and non-offensive classes. Seaborn's `boxplot` function was used, with `subtask_a` on the x-axis (to distinguish sentiment classes) and `tweet_length` on the y-axis (to display the distribution of word counts).

These procedures collectively provided a systematic means of summarizing and visualizing the dataset as shown in Section 4.1 , focusing on sentiment labels, lexical content, and tweet length.

Data Preprocessing of Offensive Language Identification Dataset (OLID):

- **Overview**: Data cleaning is a critical preprocessing step that ensures a dataset is structured, consistent, and free of unnecessary noise before further analysis. The dataset used in this study—an updated version of the Offensive Language Identification Dataset (OLID)—originally contained 14,100 entries, each with raw tweet text, a preprocessed version of the text (cleaned_tweet), and classification labels (subtask_a). Initial inspection revealed 48 duplicate entries, 70 missing values in the cleaned_tweet column, irrelevant columns (subtask_b and subtask_c), and extraneous textual elements such as URLs, user mentions, hashtags, and special characters.
- **Duplicate Removal**: All 48 duplicate tweets were detected and removed, ensuring each record in the dataset was unique. This step reduced potential bias that could arise from repeated observations.
- **Handling Missing Values**: The 70 missing entries identified in the cleaned_tweet column were reprocessed from the raw tweet text. Any tweets without meaningful textual content after reprocessing were discarded, completely resolving the missing-data issue.

- **Column Pruning** The two columns subtask_b and subtask_c were removed, as they were irrelevant to the binary classification task (subtask_a). Eliminating these non-essential features streamlined the dataset and focused the analysis on the primary outcome.
- **Text Normalization** All tweet text was converted to lowercase. Unwanted textual components—such as URLs, user mentions (@user), hashtags, punctuation marks, special characters, and digits—were removed using the Python Regular Expressions (re) library. This process helped standardize the data, and any extra whitespace introduced by these removals was normalized for clarity.
- **Tools and Libraries Pandas**: Used for data manipulation, removing duplicates, and resolving missing values. Regular Expressions (re library): Employed to detect and eliminate unwanted textual patterns (e.g., URLs, mentions, hashtags). Natural Language Toolkit (NLTK): Considered for tasks like stop word removal and lemmatization, but alternative approaches were used due to processing constraints.

Feature Engineering on Offensive Language Identification Dataset (OLID): Feature engineering is a crucial step in text-based classification tasks, as it transforms raw textual data into numeric representations suitable for machine learning algorithms. In the present study, two approaches—**TF-IDF Vectorization** and **Word Embeddings**—were applied to the `cleaned_tweet` column in order to capture both term-level importance and semantic relationships.

- **TF-IDF Vectorization**: Term Frequency–Inverse Document Frequency (**TF-IDF**) quantifies how important a word is in a given document relative to its frequency in the entire corpus. This method assigns higher weights to terms that appear frequently in one tweet but are relatively rare across all tweets, thereby highlighting discriminative features for classification.
  - **Library**: Scikit-learn's `TfidfVectorizer`
  - **Key Parameters**:
    - `max_df`: Restricts overly frequent terms (e.g., `max_df=0.95`)
    - `min_df`: Filters out rarely used terms (e.g., `min_df=2`)
    - `ngram_range`: Captures context beyond single words (e.g., `(1, 2)` for unigrams and bigrams)

    Example : Consider the tweet, *"someone vetaken piece shit volcano"*, labeled as offensive (`subtask_a = 1`). Within the TF-IDF matrix, tokens such as "vetaken" or "volcano" would likely receive elevated weights if they appear infrequently throughout the corpus, thus distinguishing this tweet from others.
- **Word Embeddings**: Unlike TF-IDF, word embeddings aim to represent words in a lower-dimensional, continuous vector space that captures semantic and syntactic relationships. Words with similar contexts or meanings end up close together in this vector space, enriching text classification tasks with nuanced linguistic information.
  - **Libraries**: Gensim or spaCy (for applying or training embeddings such as Word2Vec or GloVe)
  - **Key Parameters**:

    * **Embedding Dimension**: Typical choices range from 100 to 300 dimensions

    * **Window Size (Word2Vec)**: Determines the context range (e.g., 5 or 10)

    * **Aggregation Method**: Converts multiple word vectors from a tweet into a single tweet-level representation (e.g., averaging)

  Example : For the same tweet (*"someone vetaken piece shit volcano"*), embeddings identify semantic closeness to related words. If "volcano" is recognized from a pre-trained GloVe model, it may cluster near "lava," "eruption," or "crater." Words outside the model's vocabulary (e.g., "vetaken," possibly misspelled) can be assigned random or infrequent vectors, highlighting potential data-specific challenges.

- **Integration and Considerations**

  – **Dimensionality**: TF-IDF vectors are often high-dimensional and sparse, whereas embeddings yield dense but lower-dimensional representations. Depending on the model, further dimensionality reduction (e.g., PCA, Truncated SVD) may be beneficial.

  – **Computational Resources**: Training embeddings from scratch can be resource-intensive if the corpus is large. Pre-trained embeddings (e.g., GloVe, Word2Vec from Google News) can accelerate experimentation and reduce hardware requirements.

  – **Model Compatibility**: Traditional classifiers (e.g., Support Vector Machines, Logistic Regression) typically handle TF-IDF vectors effectively. Deep neural networks and more advanced architectures often gain greater performance benefits from word embeddings.

Description and Feature Engineering of **2024-us-election-dataset** : The dataset analyzed for this study consists of tweets related to the 2024 U.S. Presidential Election. Initially, the dataset contained 29 columns; however, only three columns were deemed relevant for this analysis: the `text` column, representing the tweet content; the `date` column, indicating when the tweet was posted; and the `in_reply_to_screen_name` column, providing context about the reply structure on Twitter. To gain preliminary insights, an exploratory analysis was conducted using Python's Pandas library to preprocess data and Matplotlib for visualization, specifically plotting the distribution of tweet lengths (in terms of the number of characters), which helped identify general patterns and variability within the textual data.

Subsequently, to enrich the original dataset, an additional feature named `party` was introduced, classifying tweets into five political party categories: **Republican Party, Democratic Party, Independent, Green Party,** and **Libertarian Party**.

Built Party Classification Model: To create the `party` labels, an auxiliary dataset titled *"2024 U.S. Election Sentiment on X"* sourced from Kaggle was employed. This dataset already included a labeled column called `party` alongside a column named `tweet_text`. Utilizing this Kaggle dataset, a multiclass Random Forest classifier was developed using Scikit-learn:

- **Input Feature:** `tweet_text` (tweets from the Kaggle dataset).
- **Output Label:** Encoded `party` labels using Scikit-learn's LabelEncoder (Republican, Democratic, Independent, Green, Libertarian).

The model was implemented and evaluated within a Jupyter Notebook environment. After training and evaluating the model, it was applied to the primary dataset (`us-election-2024`) to predict party affiliation based solely on the content present in the `text` column. Consequently, a new column named `party` was added to the primary dataset, populated with the predicted party labels.

This process enabled further downstream analyses, leveraging party-specific patterns and trends within the textual data to uncover insights pertinent to the 2024 U.S. Presidential Election discourse.

Training : Different machine learning models are trained on Offensive Language Identification Dataset (OLID). The study employed four classification algorithms—**Logistic Regression**, **Random Forest**, **Naïve Bayes**, and **XGBoost**—to detect offensive language using TF-IDF features extracted from the `cleaned_tweet` column. An 80–20 train–test split was applied, stratified on the `subtask_a` label to preserve class balance. The models were configured with the following hyperparameters:

- **Logistic Regression**: `max_iter=1000, random_state=42`
- **Random Forest**: `n_estimators=100, random_state=42`
- **Naïve Bayes**: `MultinomialNB` (default settings)
- **XGBoost**: `use_label_encoder=False, eval_metric="logloss", random_state=42`

All models were evaluated against a test set of 2,806 samples, using **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrices** for comprehensive performance assessment.

4 Results

4.1 Exploratory data analysis

The Offensive Language Identification Dataset (OLID) was analyzed to examine the distribution of sentiment labels in subtask_A as shown in Section 6.1 , which classifies tweets as offensive (1) or not offensive (0).

- **Distribution of Sentiment Labels**: The analysis revealed an imbalance in class distribution, with a higher proportion of non-offensive tweets compared to offensive tweets. This class imbalance suggests potential challenges in classification tasks, as machine learning models trained on this dataset may exhibit bias toward the majority class.
- **Word Cloud Analysis**: To gain further insights into the textual characteristics of tweets in the dataset, word clouds were generated separately for offensive and non-offensive tweets. The word cloud for offensive tweets showed a concentration of words commonly associated with negative or aggressive expressions, indicating the presence of strong sentiment or abusive language. Conversely, the word cloud for non-offensive tweets displayed a broader and more diverse vocabulary, with words that are neutral or contextually positive. These observations highlight key lexical differences between the two classes, which could serve as valuable features for automated offensive language detection.

- **Tweet Length Analysis**: A statistical analysis of tweet length was performed by computing the number of words per tweet. The distribution of tweet lengths, visualized using a histogram, exhibited a right-skewed distribution, indicating that most tweets in the dataset contain relatively few words. Additionally, a box plot comparison of tweet length by sentiment label revealed that offensive tweets tend to be slightly longer, on average, than non-offensive tweets. This suggests that users expressing offensive content may use more words per tweet, potentially to emphasize their statements. These findings provide insights into the linguistic characteristics and distributional patterns of offensive and non-offensive tweets within the OLID dataset. Such observations can inform feature engineering strategies for classification models and contribute to the broader understanding of offensive language detection in social media texts.

The 2024-us-election-Dataset was analyzed as shown Section 6.2 in to examine the following:

- **Distribution of Tweet Lengths:** The distribution of tweet lengths was analyzed to examine the variation in the number of words per tweet in the dataset. The histogram of tweet lengths revealed a **right-skewed distribution**, indicating that most tweets contain relatively few words, with a sharp decline in frequency as the tweet length increases. The majority of tweets had fewer than **15 words**, while a small number of tweets exceeded **30 words**. The presence of a long tail in the distribution suggests that while most tweets are concise, there are occasional longer messages that may influence linguistic modeling.
- **Political Party Classification Analysis**: The dataset also included a classification of tweets based on predicted political party affiliations. A bar chart depicting the distribution of predicted party labels revealed a **significant class imbalance**. The **Republican Party** had the highest number of predicted tweets, followed by the **Democratic Party**, while other parties such as the **Independent, Green, and Libertarian parties** were significantly underrepresented. This imbalance may impact classification performance and suggests that models trained on such data may be biased toward the majority class.
- **Temporal Trends in Political Mentions**: A time series analysis was conducted to track the frequency of predicted party mentions over time. The results indicated **fluctuations in political discourse**, with the **Republican Party consistently having the highest number of mentions** across different time periods. The **Democratic Party experienced some variation in mentions**, with noticeable peaks and declines over time. The other parties remained relatively stable with minimal variations in mention frequency. These trends suggest that political engagement on social media is dynamic, with certain periods experiencing higher discussion intensity, likely influenced by external political events. These findings provide insights into both the linguistic properties of tweets and the distributional characteristics of predicted political affiliations. The observed trends can inform future modeling efforts in offensive language detection and political discourse analysis.

4.2 Model Evaluation with TF-IDF Vectorization

Different machine learning models are trained on **Offensive Language Identification Dataset (OLID)**. Four machine learning models—**Logistic Regression, Random Forest, Multinomial Naive Bayes, and XGBoost**—were trained and evaluated on the **Offensive Language Identification Dataset (OLID)** using **TF-IDF vectorization** to represent textual data. The TF-IDF

method incorporated unigram and bigram tokens, excluded English stop words, and removed extremely frequent (top 5%) and rare (occurring fewer than twice) terms.

The **Logistic Regression** classifier achieved an accuracy of **76.30%**, as shown in Section 6.3 , demonstrating strong performance in identifying non-offensive tweets (precision: **0.76**, recall: **0.95**, F1-score: **0.84**) but exhibited lower recall (**0.38**) for offensive tweets despite high precision (**0.80**), resulting in a moderate F1-score (**0.51**) for offensive content.

The **Random Forest** model achieved slightly higher accuracy (**76.87%**), as shown in Section 6.3 , with improved recall (**0.44**) for offensive tweets compared to Logistic Regression. Precision for offensive tweets remained satisfactory (**0.76**), resulting in a higher F1-score (**0.56**) than Logistic Regression. Performance metrics indicate balanced precision and recall, making Random Forest a reliable choice among the evaluated models.

The **Multinomial Naive Bayes** classifier obtained an accuracy of **73.13%**, as shown in Section 6.3 . However, despite excellent recall (**0.97**) for non-offensive tweets, the model had notably low recall (**0.24**) for offensive tweets, although precision (**0.82**) was high. This discrepancy highlights a challenge in effectively identifying offensive content, reflected in a modest F1-score (**0.38**) for the offensive class.

The **XGBoost** classifier demonstrated competitive performance, achieving the highest accuracy (**77.19%**) among all models evaluated, as shown in Section 6.3 . XGBoost delivered balanced precision (**0.79**) and improved recall (**0.42**) for offensive tweets compared to Logistic Regression and Naive Bayes, resulting in a higher F1-score (**0.55**) for offensive content. The strong performance indicates XGBoost's capability to capture subtle textual cues associated with offensive language more effectively.

Confusion Matrix Analysis: Confusion matrices further illuminated model performance characteristics, as shown in Section 6.3 . Logistic Regression and Naive Bayes primarily misclassified offensive tweets as non-offensive, contributing to low recall for offensive labels. Random Forest and XGBoost significantly reduced false negatives for offensive tweets, demonstrating improved sensitivity to offensive language detection, although further enhancement in recall remains desirable.

Overall, these results suggest that the combination of **TF-IDF vectorization** with ensemble-based models, particularly **XGBoost and Random Forest**, yields robust classification results for offensive tweet detection, yet highlights ongoing challenges in balancing recall and precision due to class imbalance. Future directions could explore more sophisticated text representation methods or class-balancing techniques to further enhance detection performance.

4.3 Extending the XGBoost Model to the Election2024 Dataset

The previously evaluated XGBoost model, trained on the **Offensive Language Identification Dataset (OLID)** using TF-IDF vectorization, was employed to predict tweet sentiments in a novel dataset related to the **2024 U.S. Elections** (**Election2024 dataset**). The tweets were vectorized using the original TF-IDF model, allowing for consistent feature representation, after which sentiment predictions (0: non-offensive, 1: offensive) were generated.

Offensive Sentiment Across Top Five Frequent Reply-To Accounts : Analysis of offensive sentiment frequency, as shown in Section 6.4 ,among the top five most frequently replied-to Twitter accounts (**GOP, KamalaHQ, JoeBiden, elonmusk, GuntherEagleman**) showed that the

highest counts of offensive tweets were directed towards the **GOP** and **KamalaHQ** accounts, followed by **JoeBiden**, **elonmusk**, and **GuntherEagleman**. This distribution highlights the accounts which elicited more aggressive or polarized interactions, suggesting a targeted presence of offensive sentiment possibly driven by ideological differences or public controversies surrounding these accounts.

Offensive Sentiment Trend Over Time : A temporal analysis, as shown in Section 6.4 ,demonstrated a noticeable decline in the count of offensive tweets over the observed period from June to September 2024. Initially higher offensive activity gradually decreased as the election approached, potentially reflecting changing user engagement dynamics, moderation of discourse intensity, or evolving societal norms during the pre-election months.

Offensive Sentiment Trends Over Time for Frequent Reply-To Accounts : Temporal sentiment trends, as shown in Section 6.4 , were further analyzed for the top five frequent reply-to accounts. Offensive sentiment directed towards the **GOP** and **KamalaHQ** accounts displayed significant fluctuations, with initial peaks followed by sharp decreases. Conversely, interactions towards **JoeBiden**, **GuntherEagleman**, and **elonmusk** exhibited more moderate and relatively stable levels of offensive sentiment. Such differences indicate distinct user engagement patterns, possibly correlated with each account's political stance, online activities, or societal controversies occurring during this period.

Party-Based Offensive Sentiment Trends Over Time : Party-specific analysis, as shown in Section 6.4 , of offensive sentiment trends indicated that tweets mentioning the **Republican Party** consistently had the highest number of offensive tweets throughout the study period, although this number steadily declined over time. Mentions of the **Democratic Party** initially exhibited a modest peak, then similarly decreased. Other parties such as the **Green, Libertarian, and Independent parties** experienced minimal offensive sentiment, remaining relatively stable throughout the analyzed timeframe. This highlights the polarized discourse surrounding major political parties compared to smaller or less prominent parties.

Temporal Analysis of Sentiment Trends : Temporal trends in the average sentiment across tweets, as shown in Section 6.4 , indicated a gradual decline in predicted offensive sentiment over the analyzed time period. This decreasing trend may reflect shifts in public discourse or moderation in contentious interactions as the election approached.

Sentiment Trends by Political Party : The average predicted sentiment was further examined over time, as shown in Section 6.4 , based on the predicted political affiliation. The temporal analysis illustrated diverse trends among political affiliations. Specifically, tweets related to the Republican Party demonstrated a rise in offensive sentiment approaching election dates, potentially indicative of heightened political polarization or contentious discourse. Conversely, tweets referencing the Democratic Party showed relatively stable offensive sentiment levels, while parties with lower tweet frequency (e.g., Green, Independent, Libertarian) exhibited fluctuating yet generally declining offensive sentiment patterns.

Temporal Sentiment Dynamics for Key Accounts : Examining temporal sentiment dynamics specifically across the top five most frequently replied-to accounts revealed significant variability. For instance, interactions directed toward elonmusk and the GOP showed distinct peaks in predicted offensive sentiment, suggesting heightened periods of controversy or public scrutiny. Conversely, other figures like JoeBiden and KamalaHQ maintained relatively stable and lower levels of offensive sentiment throughout the observed timeframe.

The findings from this extended analysis , as shown in Section 6.4 , underline the robust applicability and utility of the OLID-trained XGBoost model for detecting offensive sentiment beyond its initial training domain.

5 Conclusion

This study has systematically explored the dynamics of offensive language on Twitter within the context of the 2024 U.S. Presidential Election, employing advanced statistical modeling techniques and computational methodologies relevant to contemporary data science. By utilizing the Offensive Language Identification Dataset (OLID), the research developed robust predictive models leveraging multiple machine learning algorithms—Logistic Regression, Random Forest, Multinomial Naive Bayes, and XGBoost—to classify tweets as offensive or non-offensive. Among these classifiers, the XGBoost model achieved superior accuracy (77.19%), demonstrating balanced performance across precision and recall metrics. This finding underscores the effectiveness of ensemble-based algorithms, particularly gradient boosting methods, for addressing classification tasks characterized by class imbalance.

Further extending the validated XGBoost model to a novel and considerably larger dataset—the "2024-us-election-dataset"—provided additional empirical validation and substantive insights into the practical applicability of the model beyond its initial training context. By maintaining the consistency of feature representations through TF-IDF vectorization, the model effectively generalized to predict offensive content within politically charged tweets. This capability reinforces the versatility and reliability of machine learning approaches when addressing real-world social media discourse challenges, particularly those related to political polarization and online interactions.

In analyzing offensive sentiment across different dimensions, several notable findings emerged. The distribution analysis highlighted that offensive tweets were disproportionately directed towards high-profile political accounts, notably GOP and KamalaHQ. This observed pattern indicates pronounced ideological polarization and intensified adversarial exchanges directed at politically influential accounts, suggesting that prominent political figures are focal points for contentious and potentially harmful discourse. These results have implications for understanding online hostility dynamics, platform moderation practices, and the targeted impact of offensive rhetoric.

Temporal analyses contributed further insights by demonstrating a clear decreasing trend in offensive tweet counts over the election timeline, spanning from June through September 2024. This reduction in offensive activity over time may reflect shifts in user engagement behaviors, heightened awareness of content moderation policies, or evolving societal norms regarding acceptable political discourse as election day approaches. These temporal dynamics offer valuable insights for policymakers and platform managers, suggesting the effectiveness of proactive moderation strategies implemented during election periods or increased public awareness about the implications of offensive online interactions.

Party-specific analyses uncovered distinct patterns of offensive language distribution. Tweets referencing major political parties, especially the Republican Party, consistently exhibited higher levels of offensive content compared to other parties. In contrast, smaller or less prominent political parties, such as the Green, Libertarian, and Independent parties, showed substantially lower and more stable offensive sentiment levels. This distinction underscores the significant influence of

party affiliation on the intensity and nature of offensive interactions online, pointing to the broader sociopolitical context and partisan identities as drivers of online offensive discourse.

Detailed temporal analyses for individual political parties and specific high-profile Twitter accounts further revealed nuanced variations in offensive sentiment patterns. For instance, interactions directed towards GOP and KamalaHQ accounts showed significant initial peaks in offensive sentiment, followed by sharp declines, suggesting that specific political events, controversies, or campaigns markedly influenced these fluctuations. Conversely, interactions targeting JoeBiden, GuntherEagleman, and elonmusk exhibited more consistent and moderate levels of offensive sentiment, indicating differing user engagement dynamics based on account-specific characteristics and broader political contexts.

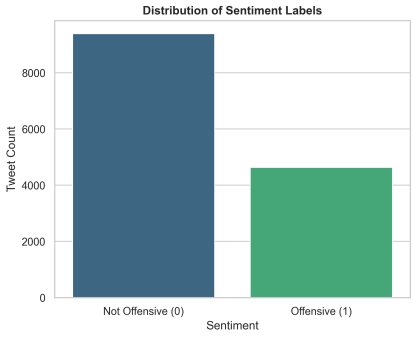
The comprehensive computational and statistical approach adopted in this research provides substantial methodological advancements relevant to natural language processing (NLP), text classification, and computational social science. The effective application of TF-IDF vectorization combined with advanced machine learning techniques demonstrates the power of statistical and computational methodologies in analyzing large-scale textual datasets. Additionally, this study highlights the importance of addressing class imbalance and model interpretability in text classification tasks, essential for reliable and equitable deployment of AI-driven moderation tools in real-world applications.

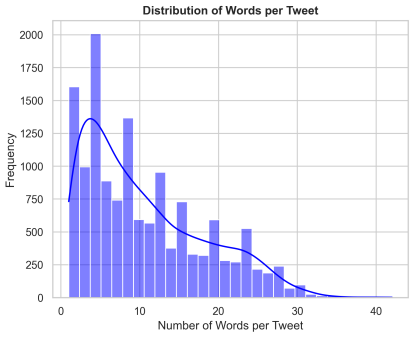
From a broader sociopolitical perspective, this research contributes significantly to understanding offensive language dynamics within contemporary digital political discourse. The findings emphasize the critical need for nuanced, context-aware moderation strategies tailored to politically sensitive periods, such as elections, to mitigate potential harms arising from online offensive content. Policymakers, platform administrators, and stakeholders in political communication can leverage these insights to foster healthier, more inclusive digital environments, enhancing democratic discourse and civic engagement.

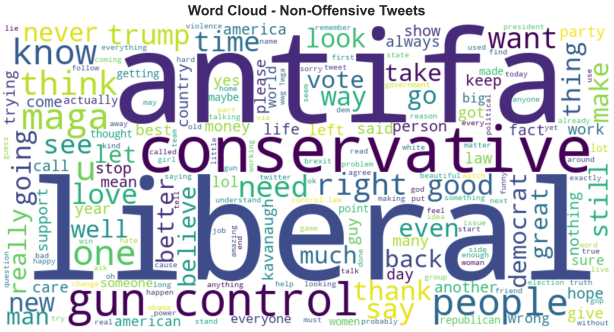
In conclusion, the integrated use of advanced statistical models, computational analysis, and theoretical frameworks in this research underscores its multidisciplinary contributions. The demonstrated robustness and versatility of predictive modeling techniques reinforce the critical role of data science and statistical methodologies in addressing complex social phenomena like offensive online communication. Future research could further expand on these findings by exploring additional feature extraction methods, deep learning architectures, and cross-platform comparative analyses to continually enhance the understanding and mitigation of offensive language in digital spaces.

6 Appendix

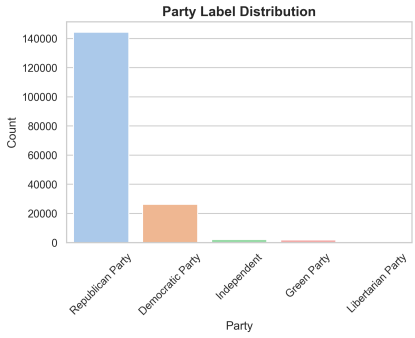
6.1 Plots of Exploratory data analysis using OLID Dataset

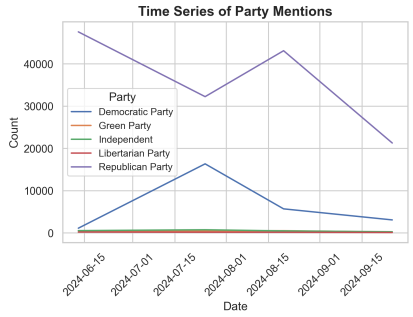


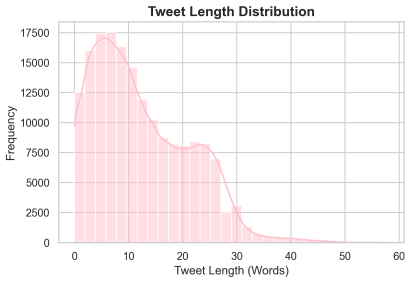




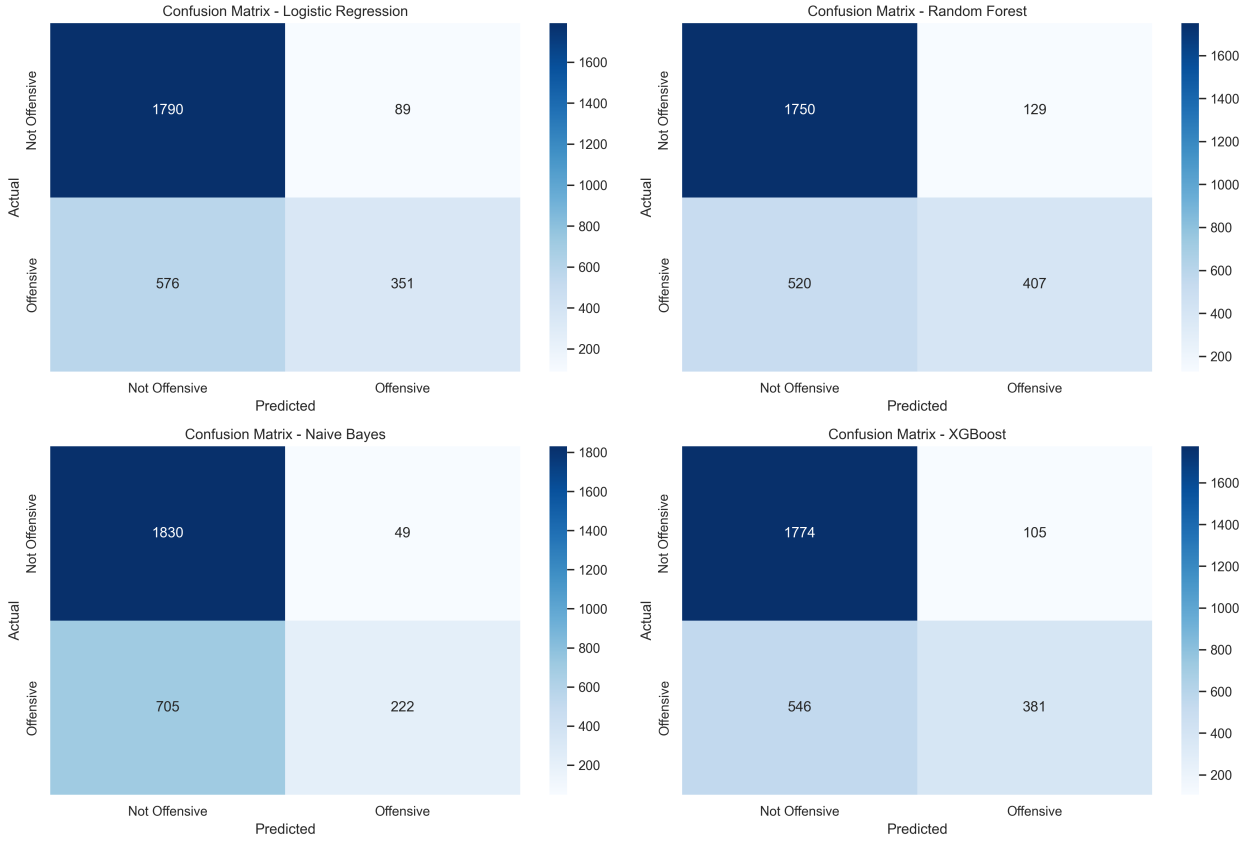
6.2 Plots of Exploratory data analysis using 2024-us-election-Dataset

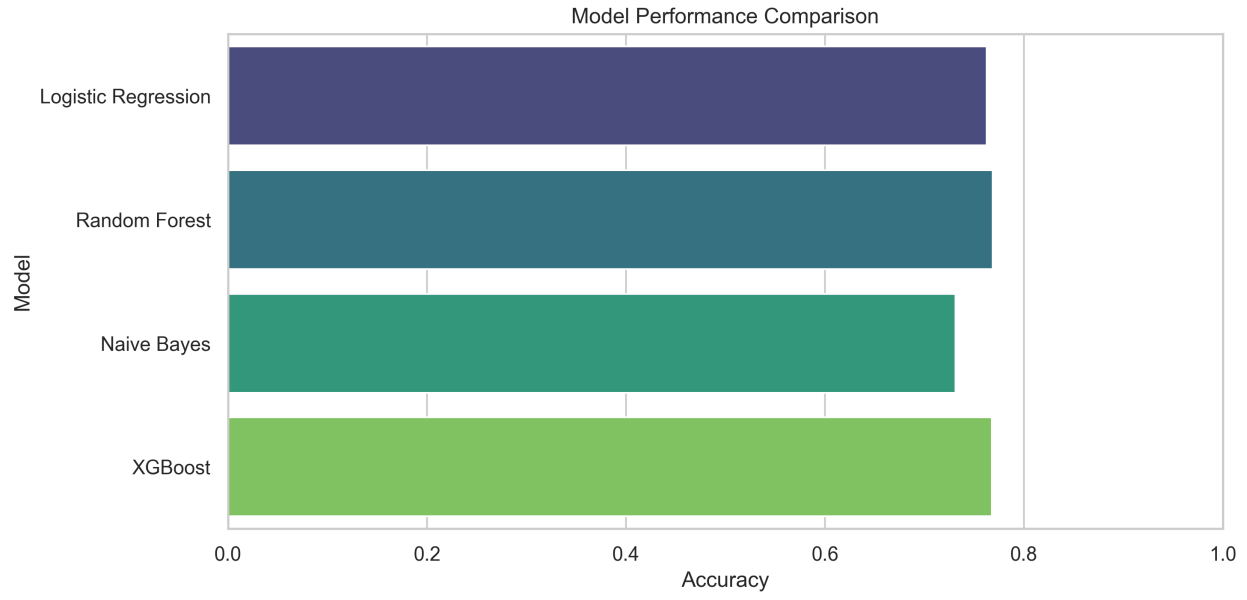




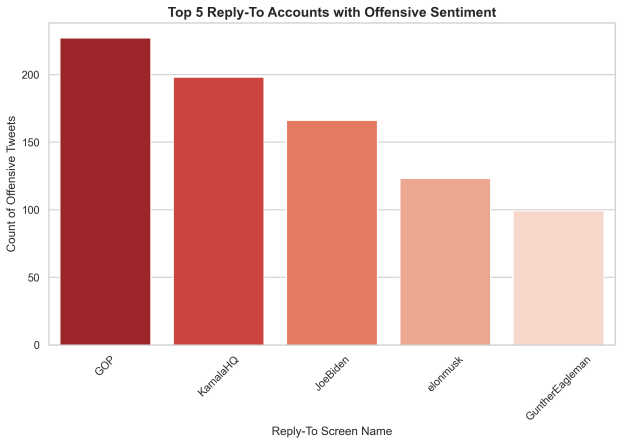


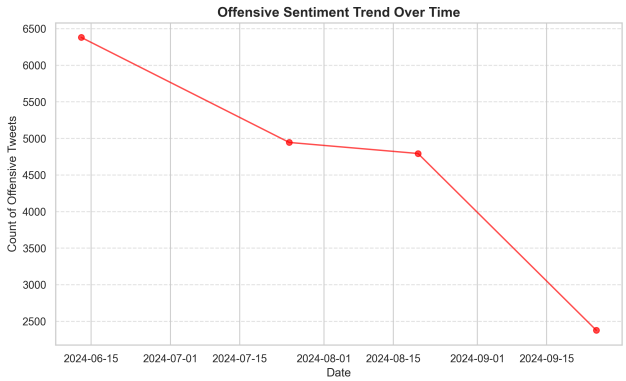
6.3 Plots of Model Evaluation

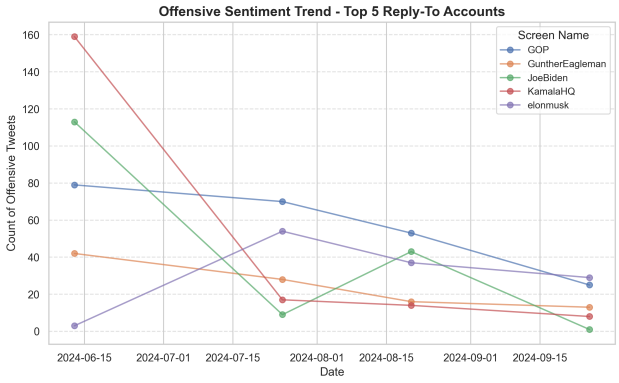


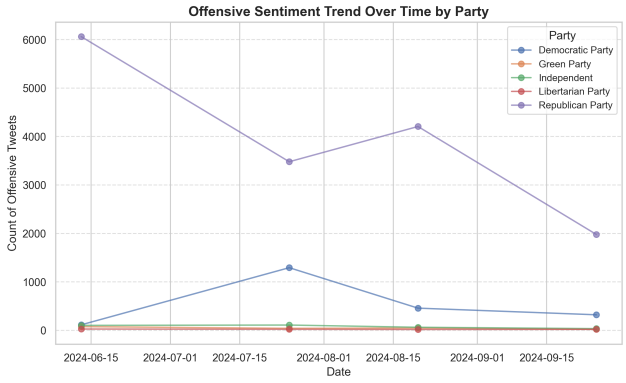


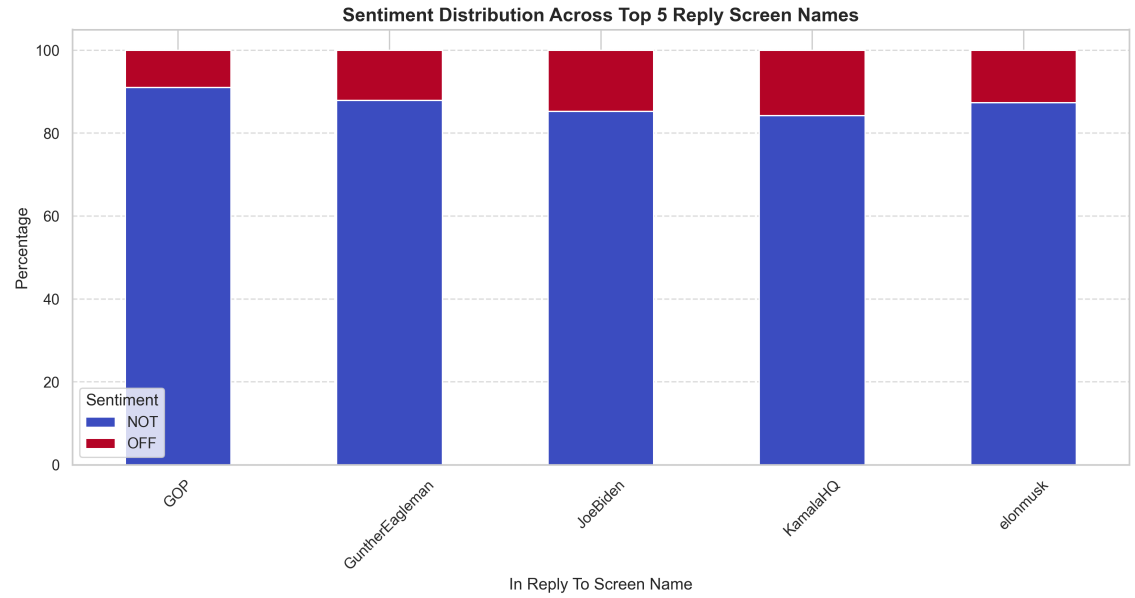
6.4 Plots on XGBoost Model to the Election2024 Dataset

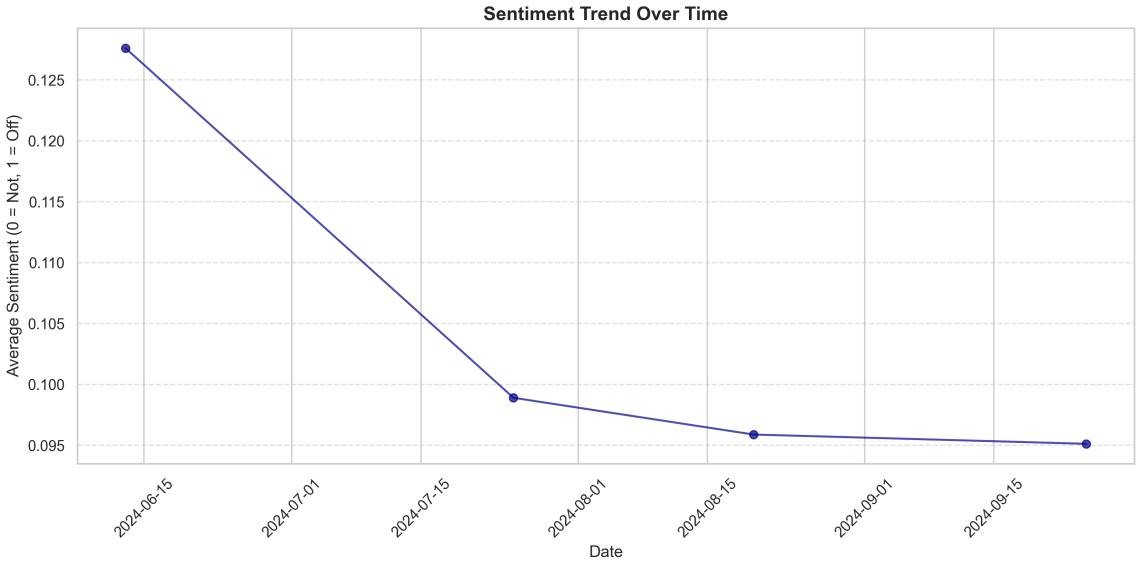


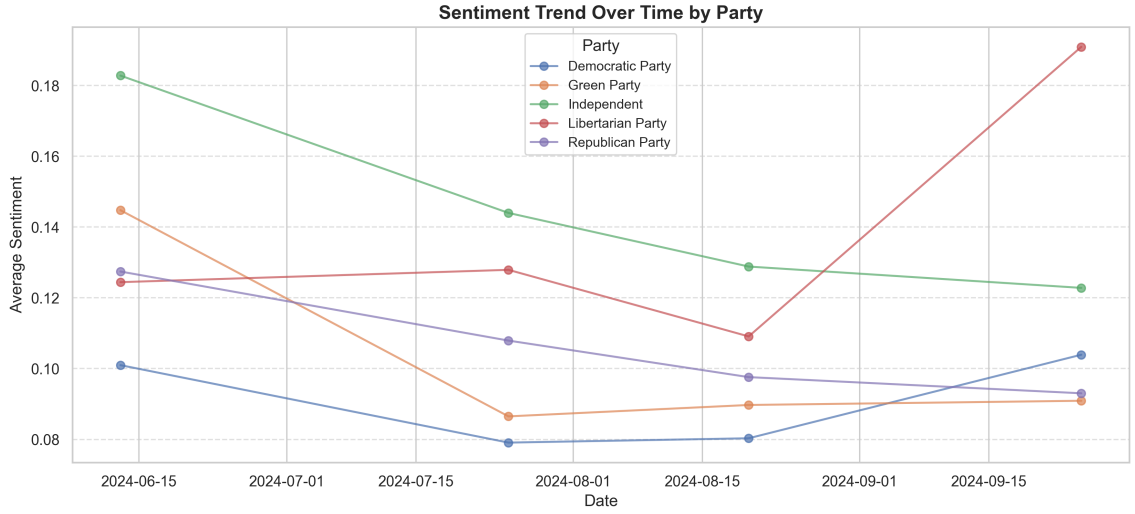







Sentiment Distribution Across Top 5 Reply Screen Names
Percentage
100
80
60
40
20
0
Sentiment
NOT
OFF
GOP
GuntherEagleman
JoeBiden
KamalaHQ
elonmusk
In Reply To Screen Name


Sentiment Trend Over Time
Average Sentiment (0 = Not, 1 = Off)
0.125
0.120
0.115
0.110
0.105
0.100
0.095
2024-06-15
2024-07-01
2024-07-15
2024-08-01
2024-08-15
2024-09-01
2024-09-15
Date


Sentiment Trend Over Time by Party
Party
Democratic Party
Green Party
Independent
Libertarian Party
Republican Party
Average Sentiment
0.18
0.16
0.14
0.12
0.10
0.08
2024-06-15
2024-07-01
2024-07-15
2024-08-01
2024-08-15
2024-09-01
2024-09-15
Date

7 References

1. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 75–86. https://doi.org/10.18653/v1/S19-2010
2. McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2), 176–187. https://doi.org/10.1086/267990
3. Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Brooks/Cole.
4. Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *Public Opinion Quarterly*, 37(4), 509–523. https://doi.org/10.1086/268109
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. https://doi.org/10.1145/2939672.2939785
7. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.
8. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
9. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
10. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162
11. USC-X-24-US-Election GitHub Repository. (2024). Retrieved from https://github.com/sinking8/usc-x-24-us-election
12. OLID Dataset on Huggingface. (2024). Retrieved from https://huggingface.co/datasets/christophsonntag/O
13. Kaggle: 2024 U.S. Election Sentiment on X. (2024). Retrieved from https://www.kaggle.com/datasets/2024-us-election-sentiment-on-x
14. Seaborn: Statistical Data Visualization. (2024). Retrieved from https://seaborn.pydata.org/
15. Matplotlib: Visualization with Python. (2024). Retrieved from https://matplotlib.org/
16. WordCloud for Python. (2024). Retrieved from https://amueller.github.io/word_cloud/
17. Pandas: Python Data Analysis Library. (2024). Retrieved from https://pandas.pydata.org/
18. NumPy: Scientific Computing with Python. (2024). Retrieved from https://numpy.org/

19. spaCy: Industrial-Strength Natural Language Processing. (2024). Retrieved from https://spacy.io/
20. Regular Expressions in Python. (2024). Retrieved from https://docs.python.org/3/library/re.html
21. NLTK: Natural Language Toolkit. (2024). Retrieved from https://www.nltk.org/
22. Gensim: Topic Modeling for Humans. (2024). Retrieved from https://radimrehurek.com/gensim/
23. Scikit-learn: Machine Learning in Python. (2024). Retrieved from https://scikit-learn.org/stable/
24. XGBoost Documentation. (2024). Retrieved from https://xgboost.readthedocs.io/en/latest/
25. Python Software Foundation. (2024). Python Language Reference. Retrieved from https://www.python.org/doc/