Analysis done on our last submitted paper.

Experiments Conducted:

- Data Collection: Utilized Twitter's API to gather 286,000 tweets containing hashtags from actively posting user accounts over a three-month period.
- Network Analysis: Mapped out the interactions between users, specifically focusing on who replies to or retweets whom, forming a directed adjacency list of follower-leader pairs.
- Distribution Analysis: Analyzed the frequency and distribution of replies and retweets to classify Twitter users as "high influence" or "low influence" based on the volume and focus of interactions they received.

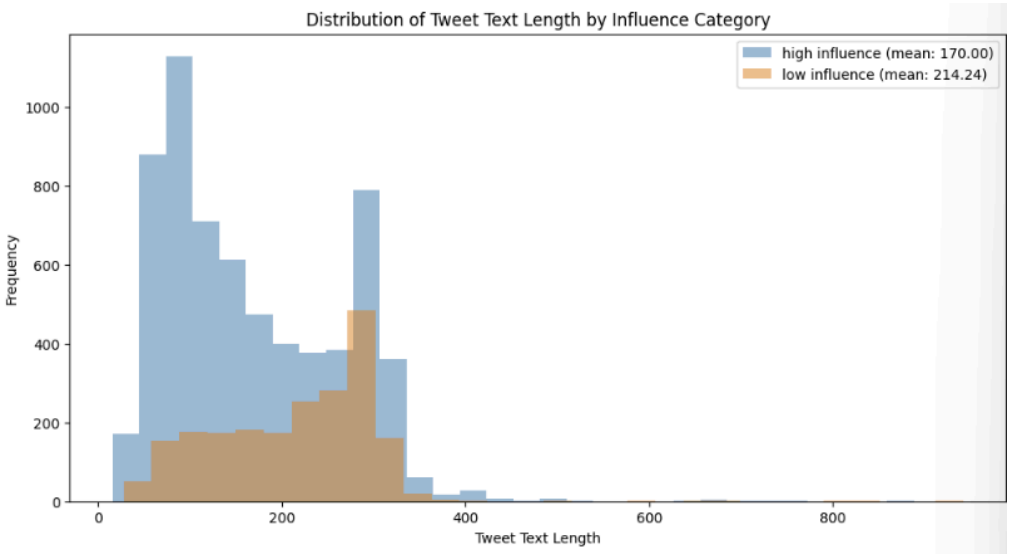
Planned Experiments for Research Paper

- Machine Learning Classification [Completed]
    - Classify Twitter users into high or low influence categories using machine learning models
- Tweet Sentiment Analysis: [Completed]
    - Determine the sentiment (positive or negative) of tweets.
    - Compare the content shared by high-influence and low-influence users.
- Graph Construction and Statistics [Completed]
    - Objective: Represent the Twitter network as a graph with users as nodes and interactions as edges.
    - Statistics: Calculate the number of nodes, edges, degree distribution, and assortativity.
- Community Detection and Impact Analysis [Completed]
    - Objective: Identify communities within the Twitter network using the Louvain algorithm.
    - Impact: Assess changes in graph structure and metrics after selectively removing communities.
- Node Classification:
    - Objective: Use graph-based machine learning models (e.g., GCN, GraphSAGE, GAT) to classify users based on influence.
    - Evaluation: Measure training loss, test accuracy, and class-wise prediction accuracy.
- Link Prediction:
    - Objective: Predict missing interactions between users using machine learning models.
    - Method: Use edge masking techniques and node features (degree, influence, community) for model training.

-

1. Most Frequent Hashtags from Dataset
    a. This dataset is between Jan 1st 2022 to April 30th 2022
    b. Some of the popular events that happened all around the world during that time Russian Invasion of Ukraine, COVID-19 Pandemic, Political Shifts in Latin America, North Korea Missile Tests
    c. Sources:
        i. https://www.cfr.org/blog/ten-most-significant-world-events-2022
        ii. https://en.wikipedia.org/wiki/Portal:Current_events/January_2022
    d. So the Hashtags in the dataset align with the most popular events happening during that time period. Both High and Low Influence users tend to discuss the most popular events happening around the world based on hashtag analysis.

2. Distribution of Tweet Text Length by Influence Category
    a. Low Influence users tend to text longer tweets compared to high influence users. Mean text length of high influence users is 170 characters, whereas that of low influence users is 214.24 characters.



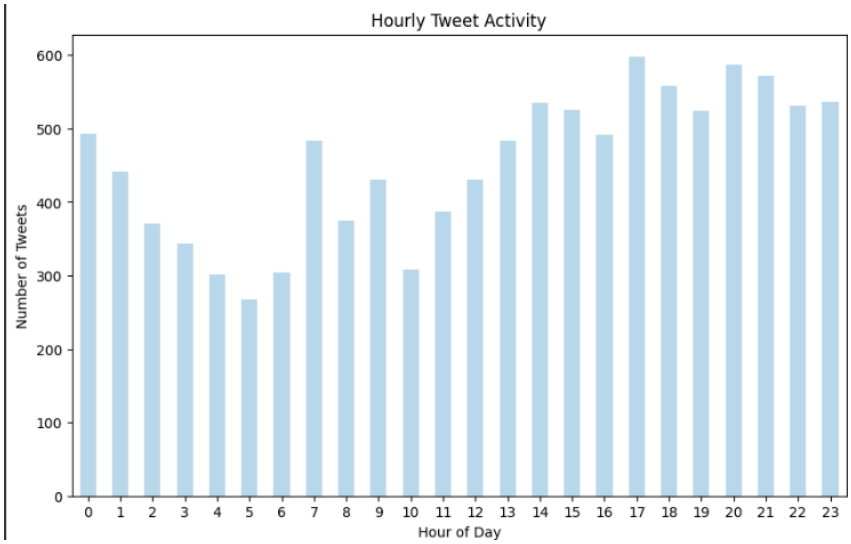

```
Text Length Statistics:

high influence Text Length:

 Mean: 170.00

 Median: 148.0
```

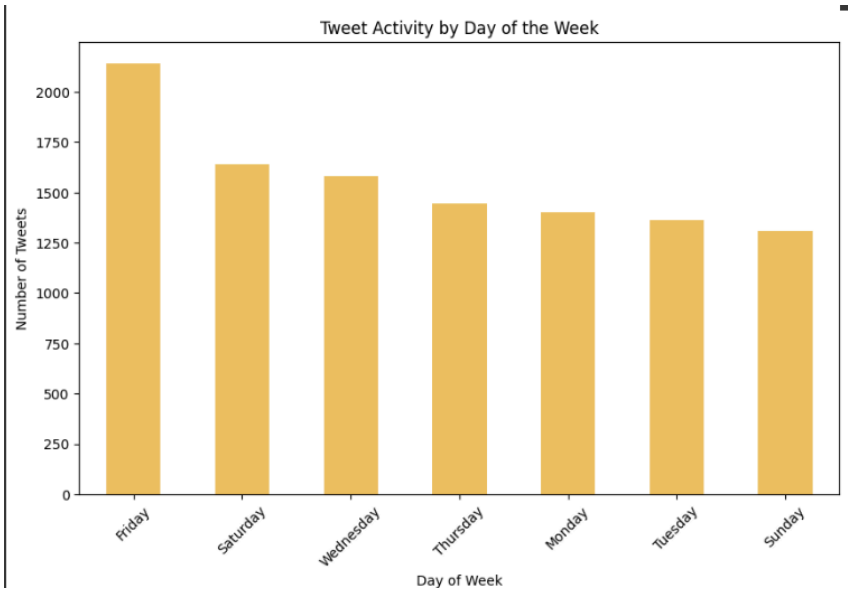
```
 Standard Deviation: 95.91


low influence Text Length:

 Mean: 214.24

 Median: 230.0

 Standard Deviation: 86.63
```

3. Distribution of Tweet Frequency by Influence Category
    a. High influence users in the network tend to have 60.93 tweets per user whereas a low influencer user just has 1.65 tweets per user. There's a huge segregation between high influence and low influence users. Thus, it proves that twitter is a platform in which high influence users are very active in posting compared to low influence users.

```
Tweet Frequency Statistics:

high influence Tweet Frequency:

 Mean: 60.93

 Median: 10.0

 Standard Deviation: 150.08


low influence Tweet Frequency:

 Mean: 1.65

 Median: 1.0

 Standard Deviation: 3.89
```

4. Tweet Activity based on Hour of Day and Day of Week
    a. Based on the Hourly Tweet Activity. The tweet activity is high on average from afternoon till mid night
    b. On Fridays the Tweet Activity is the highest





5. Using Latent Dirichlet Allocation
    a. Using Latent Dirichlet Allocation (LDA), we identified ten distinct topics within our textual dataset, revealing prevalent themes such as geopolitical issues (Topics 1 and 8 involving Ukraine and Russia), domestic politics (Topics 2 and 4 with a focus on Canadian and conservative politics), and public health (Topic 9 on COVID-19). This unsupervised topic modeling approach has provided a structured summary of the corpus, showcasing its potential to unearth underlying patterns and inform strategic decision-making based on thematic insights from large-scale textual data

Topic 1:

putinhitler standwithukraine https stopputinnow russiasanctions boycottrussia bot putinwarcriminal genocide home

Topic 2:

amp ford time abuse violence vote https doug need racism

Topic 3:

good just fbpe followbackfriday followed hope morning hi jacquesrogues lovely

Topic 4:

conservative amp like https ford nevervoteconservative cdnpoli canada people ontario

Topic 5:

https amp traitor j6 tell israel apartheid says israeli support

Topic 6:

https know government use legal week south ableg new school

Topic 7:

https amp canada trudeau ottawa canadians like omicron antisemitism ba2

Topic 8:

https ukraine russia putin russian people trump ukraineinvasion war military

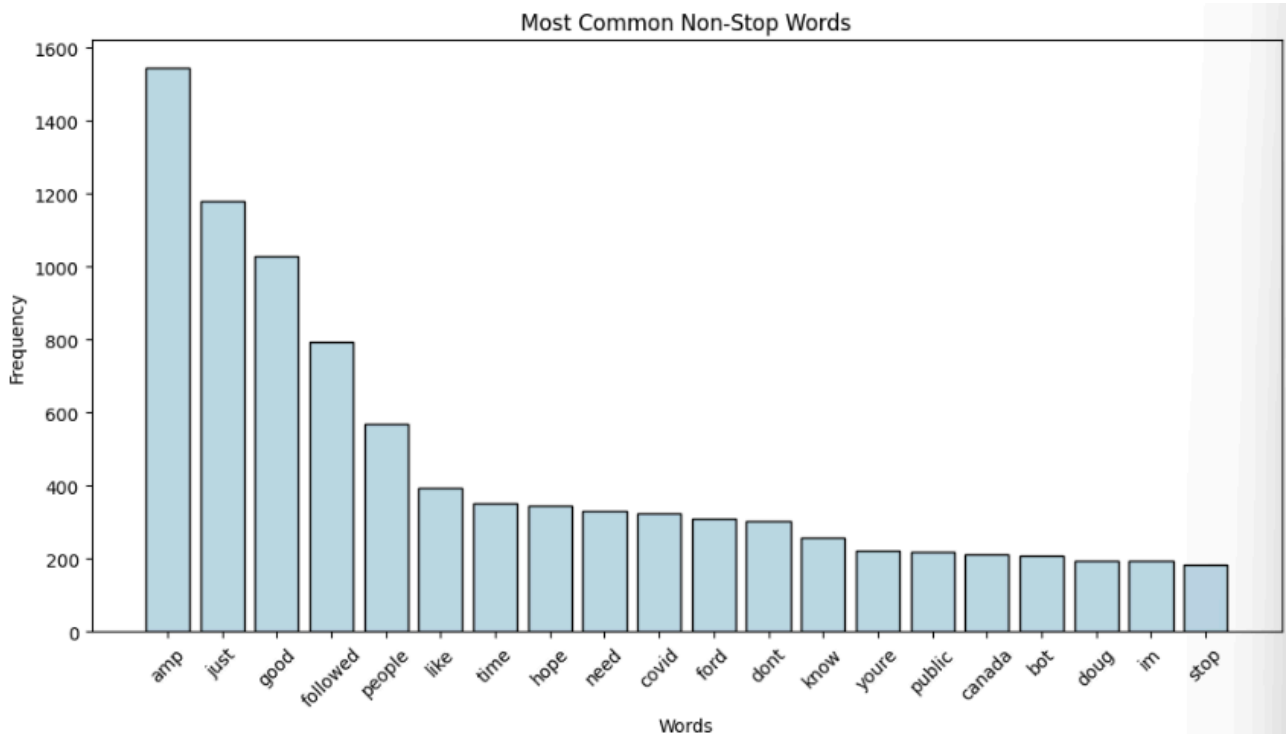
Topic 9:

amp covid19 covid https longcovid bcpoli people omicron virus health

Topic 10:

https amp ottawa police canada don like way women say

6. Most common Non-Stop Words



7. Sentiment and Subjectivity Analysis of Tweets: An Insight into Communicative Styles on Twitter

   Both High and Low Infl Users

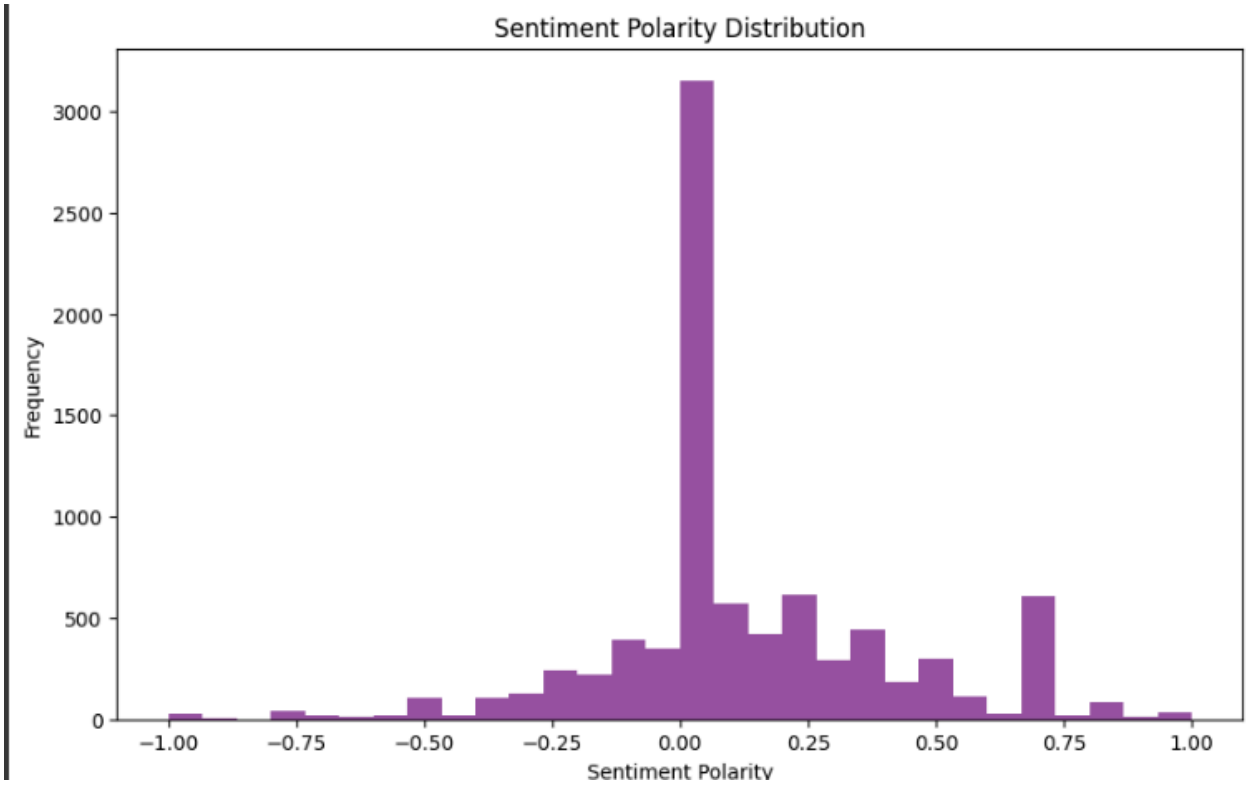
In the exploration of tweet sentiments within our dataset, we employed TextBlob to conduct a detailed sentiment analysis. The results reveal a predominant neutrality in tweet sentiments, with an average sentiment polarity score of zero. This neutrality indicates a balanced presence of both positive and negative sentiments among the tweets, suggesting a diverse range of expressions and opinions among Twitter users.

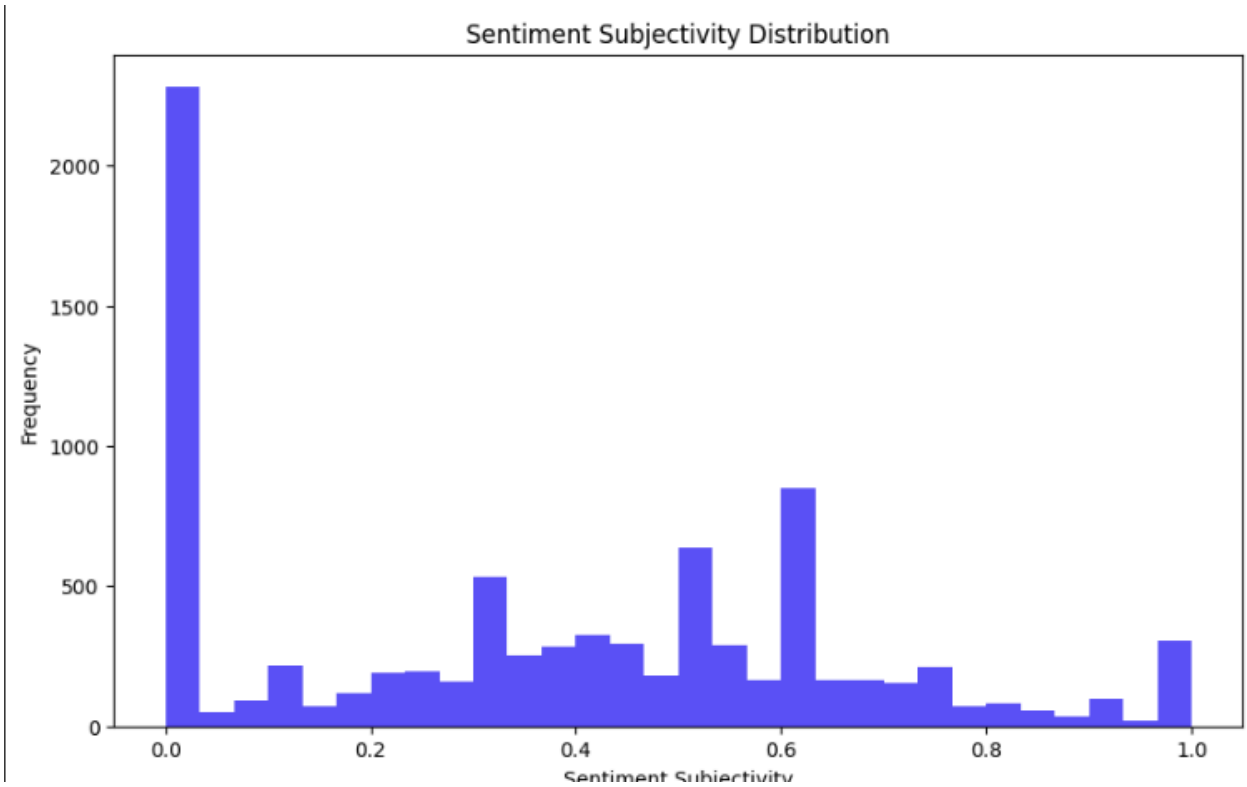
Further analysis on sentiment variability supports this observation, showing that most tweets maintain a sentiment close to neutrality. Such minimal fluctuation in sentiment polarity underscores a consistency in the tone of communications, reflecting either a general restraint in emotional expression or a balanced representation of viewpoints.

The study also extends to the subjectivity of the tweets, where our findings indicate a general trend towards objectivity, with an average subjectivity score being notably low. This suggests that the tweets predominantly share factual information or restrained personal opinions, rather than highly subjective or emotionally charged content. The range of subjectivity observed, however, does include tweets that exhibit full subjectivity, highlighting a subset of the data where personal feelings and opinions are more explicitly expressed.

A lexical analysis of the most common words within the tweets reveals a predominance of functional words such as conjunctions and prepositions. This pattern indicates that the language used across the tweets is straightforward, focusing on basic structural elements of communication rather than complex or specialized vocabulary. The frequent use of basic connectors may suggest that the discussions are general in nature, possibly centered around day-to-day topics rather than niche or targeted discussions.

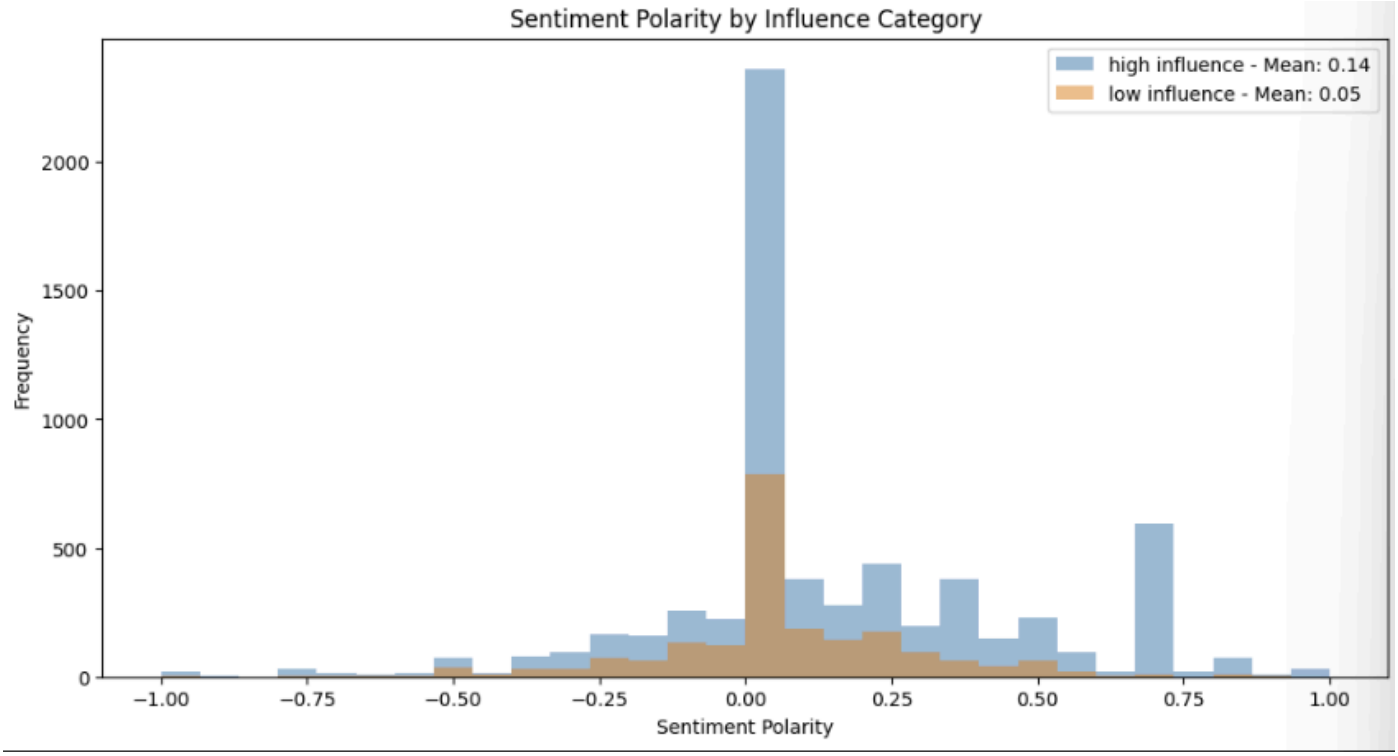
Overall, these analytical insights provide a deeper understanding of the communicative tone and style prevalent within the tweet corpus. By examining how sentiments are expressed and topics are discussed, we can gain valuable perspectives on the nature of discourse on the platform, aiding in further research into linguistic patterns and user interactions on social media. This analysis not only enhances our understanding of digital communication dynamics but also offers a foundation for developing more nuanced strategies for engaging with or studying social media content.


Sentiment Polarity Distribution
Frequency
3000
2500
2000
1500
1000
500
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Polarity


Sentiment Subjectivity Distribution
Frequency
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Sentiment Subjectivity

```
Sentiment Polarity Statistics:
count   8546
mean       0
std        0
min       -1
25%        0
50%        0
75%        0
max        1
Name: sentiment_polarity, dtype: float64

Sentiment Subjectivity Statistics:
count   8546
mean       0
std        0
min        0
25%        0
50%        0
75%        1
max        1
Name: sentiment_subjectivity, dtype: float64

Most Common Words:
the: 6696
to: 4878
a: 4179
of: 3241
and: 2883
is: 2858
in: 2721
for: 2013
you: 1821
have: 1641
```

Sentiment Analysis on High Influence and Low Influence



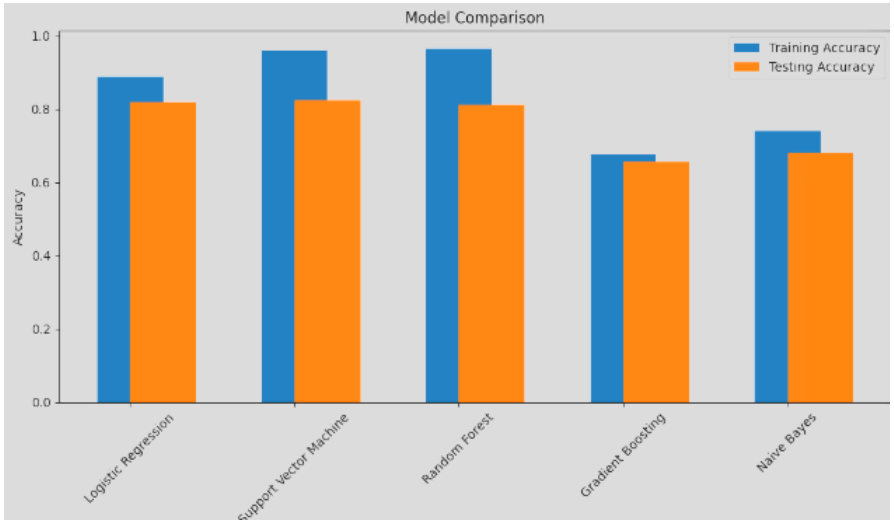

```
high influence Sentiment Analysis Metrics:
Mean Polarity: 0.1375
Median Polarity: 0.0000
Standard Deviation of Polarity: 0.3079
Proportion of Positive Sentiments: 0.4884
Proportion of Neutral Sentiments: 0.3305
Proportion of Negative Sentiments: 0.1811

low influence Sentiment Analysis Metrics:
Mean Polarity: 0.0550
Median Polarity: 0.0000
Standard Deviation of Polarity: 0.2255
Proportion of Positive Sentiments: 0.4513
Proportion of Neutral Sentiments: 0.3059
Proportion of Negative Sentiments: 0.2428
```

The sentiment analysis of tweets from users classified as high and low influence reveals distinct communication patterns and emotional expressions unique to each group. High influence users generally display a broader emotional range with a notable skew towards positive sentiments, as indicated by a mean polarity of 0.1375 and nearly half of their tweets being positive. This could suggest a strategic approach to maintaining a positive

public image or engaging a wide audience positively. On the other hand, low influence users show a narrower range of sentiment with a higher proportion of negative tweets (24.28%), suggesting a more straightforward or critical engagement style. This could reflect a lesser concern for managing public perceptions or a greater focus on expressing genuine opinions or criticisms. The median neutrality in both groups points to a substantial proportion of factual or balanced content, indicating that despite the emotional variances, a significant level of objectivity is maintained in the tweets. Understanding these dynamics can provide valuable insights for tailoring communication strategies on social media platforms, particularly in content creation, marketing, and community management.

8. Machine Learning High vs Low Infl Classification
    a. The provided experiment demonstrates a machine learning workflow for text classification, employing several algorithms to predict categories based on textual input. Initially, text data is preprocessed to handle null values and transformed into a numerical format using TF-IDF vectorization, which emphasizes words unique to specific documents. Various classification models including Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Naive Bayes are applied. Each model is trained on the transformed data, and performance is assessed through accuracy metrics on both training and testing datasets to evaluate their effectiveness and generalization capabilities. This approach helps identify the most suitable model for handling textual data in predictive tasks.
    b. Text Data
        i. The Support Vector Machine (SVM) model emerges as the best performer in this comparison, achieving the highest testing accuracy of 0.8226, indicating superior generalization to unseen data. While the Random Forest model shows a slightly higher training accuracy, suggesting strong learning from the training set, its lower testing accuracy points to potential overfitting, making SVM more reliable for practical applications.

```
Training Logistic Regression...
Logistic Regression Training Accuracy: 0.8890
Logistic Regression Testing Accuracy: 0.8185
--------------------------------------------------------
Training Support Vector Machine...
Support Vector Machine Training Accuracy: 0.9606
Support Vector Machine Testing Accuracy: 0.8226
--------------------------------------------------------
Training Random Forest...
Random Forest Training Accuracy: 0.9647
Random Forest Testing Accuracy: 0.8111
--------------------------------------------------------
Training Gradient Boosting...
Gradient Boosting Training Accuracy: 0.6751
Gradient Boosting Testing Accuracy: 0.6558
--------------------------------------------------------
Training Naive Bayes...
Naive Bayes Training Accuracy: 0.7398
Naive Bayes Testing Accuracy: 0.6806
--------------------------------------------------------
```



c. Hashtags

i. In an in-depth analysis of machine learning models trained on hashtag-based features from a dataset, we observe distinct patterns in their performance, particularly focusing on how well each model generalizes based on its training and testing accuracies.
The Support Vector Machine (SVM) and Naive Bayes models exhibit the most promising results, with testing accuracies of 0.7886 and 0.7854, respectively. These values suggest that both

models effectively handle the variations in text data encapsulated by hashtags. The SVM, in particular, shows an excellent balance between training (0.9230) and testing accuracies, indicating robustness and a strong capacity to generalize from trained data to unseen scenarios. Similarly, Naive Bayes demonstrates a significant ability to model the text data effectively, only slightly trailing behind SVM in testing performance.

Conversely, Logistic Regression offers a competitive alternative with a testing accuracy close to that of SVM and Naive Bayes, marking it as another viable option for this type of data. However, its slightly lower performance might indicate a need for further tuning or use in conjunction with other feature types to enhance its predictive power.

Random Forest, despite achieving the highest training accuracy (0.9335), falls short in testing performance with an accuracy of 0.7748. This indicates potential overfitting, where the model excels at recalling the training data but fails to maintain this performance with new data. This overfitting suggests that Random Forest might benefit from methods aimed at reducing model complexity, such as pruning.

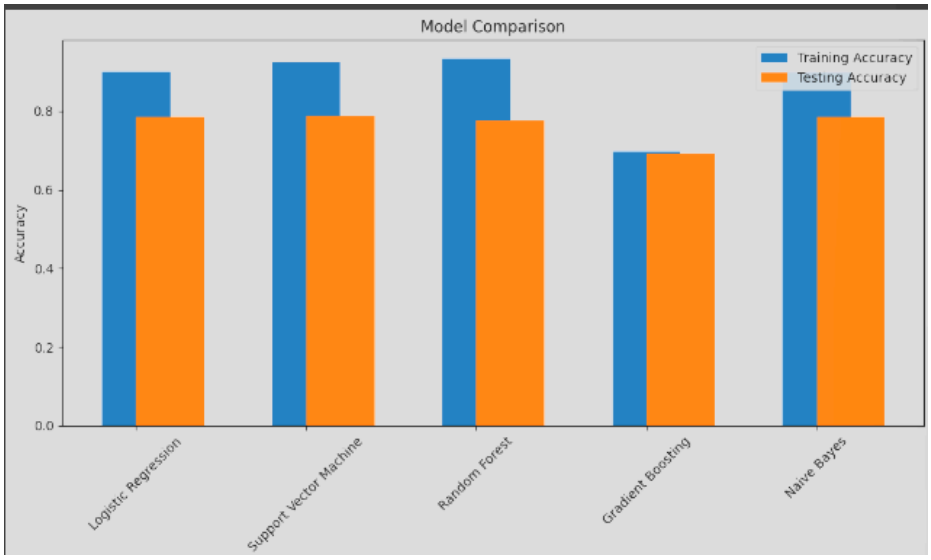
Lastly, Gradient Boosting shows the least effective performance, with a notable drop from training accuracy (0.6966) to testing accuracy (0.6926). This result could reflect the model's challenges with the sparse and potentially high-dimensional nature of text data, implying that Gradient Boosting may require substantial parameter adjustments or might be inherently less suited for this task compared to other models.

In conclusion, choosing the right model for text-based classification tasks such as these involves not only assessing accuracy but also considering how well the model avoids overfitting and adapts to new data. SVM and Naive Bayes stand out as particularly effective for this dataset, with their ability to generalize well, making them preferable choices for tasks involving natural language processing of social media data, especially when leveraging hashtags as key features.

ii. The machine learning model performance metrics reveal that the Support Vector Machine (SVM) excels, leading in accuracy, precision, recall, and F1 score, making it optimal for this dataset. Logistic Regression and Naive Bayes also perform well, particularly in accuracy and recall, indicating effectiveness in identifying relevant instances but with some false positives. Random Forest,

while robust, falls slightly behind SVM. Gradient Boosting lags considerably across all metrics, suggesting potential issues with fitting the dataset, either due to overfitting or underfitting. These insights guide selecting the right model based on precision, recall balance, and overall accuracy needs.

```
Training Logistic Regression...
Logistic Regression Training Accuracy: 0.9007
Logistic Regression Testing Accuracy: 0.7845
--------------------------------------------------
Training Support Vector Machine...
Support Vector Machine Training Accuracy: 0.9230
Support Vector Machine Testing Accuracy: 0.7886
--------------------------------------------------
Training Random Forest...
Random Forest Training Accuracy: 0.9335
Random Forest Testing Accuracy: 0.7748
--------------------------------------------------
Training Gradient Boosting...
Gradient Boosting Training Accuracy: 0.6966
Gradient Boosting Testing Accuracy: 0.6926
--------------------------------------------------
Training Naive Bayes...
Naive Bayes Training Accuracy: 0.8985
Naive Bayes Testing Accuracy: 0.7854
```



```
Logistic Regression - Accuracy: 0.7845, Precision: 0.7892, Recall: 0.7845, F1 Score: 0.7738
Support Vector Machine - Accuracy: 0.7886, Precision: 0.7902, Recall: 0.7886, F1 Score: 0.7803
Random Forest - Accuracy: 0.7748, Precision: 0.7776, Recall: 0.7748, F1 Score: 0.7639
Gradient Boosting - Accuracy: 0.6926, Precision: 0.7433, Recall: 0.6926, F1 Score: 0.6309
Naive Bayes - Accuracy: 0.7854, Precision: 0.7920, Recall: 0.7854, F1 Score: 0.7738
```