Sentiment Analysis of Political Discourse on Reddit

Submitted by

HARISH G D

2327422

SUBJECT: Text and Social Media Analytics (TSMA)

CIA 3 Assignment

Guided by

Prof. ROSEWINE JOY

CHRIST (DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

CHRIST SCHOOL OF BUSINESS AND MANAGEMENT

BENGALURU

November-2024

Introduction to Reddit Web Scraping Using PRAW on the Topic of "Politics"

Reddit, a widely popular social platform, serves as a hub for discussions on virtually every topic imaginable, including one of the most debated and analyzed subjects: politics. With millions of users contributing posts, comments, and opinions, Reddit provides a treasure trove of data for understanding public sentiment, political trends, and the dynamics of online political discourse. The sheer volume of content makes it an invaluable resource for researchers, analysts, and organizations seeking to explore the intersection of politics and digital communication.

To efficiently extract and analyze data from Reddit, **PRAW (Python Reddit API Wrapper)** is a powerful and user-friendly library. PRAW enables developers to access Reddit's API to retrieve structured data from subreddits, posts, and comments without the need to directly scrape HTML. This approach ensures compliance with Reddit's terms of service and provides reliable access to data.

Why Focus on Politics?

Politics is one of the most active and influential topics on Reddit, with communities (subreddits) like **r/politics**, **r/worldnews**, and **r/PoliticalDiscussion** generating thousands of posts and comments daily. Analyzing this data offers valuable insights, such as:

- **Public Sentiment**: Understanding how users feel about specific policies, politicians, or global events.
- **Trend Analysis**: Monitoring political discussions to identify emerging issues or topics.
- **Community Behavior**: Observing how users engage in debates, share news, and form opinions.

Benefits of Using PRAW for Scraping Political Data

1. **Structured Data Access**:
   PRAW abstracts the complexities of API calls, providing a clean and intuitive way to query Reddit for posts, comments, and user interactions.
2. **Customizable Queries**:
   With PRAW, you can filter data by subreddit, keywords, date range, or post type, allowing you to focus on specific political topics or events.
3. **Ethical and Compliant**:
   By using the Reddit API via PRAW, you adhere to Reddit's terms of service, ensuring your data collection is both responsible and sustainable.

```python
# Initialize the Reddit instance with PRAW
reddit = praw.Reddit(client_id=client_id, client_secret=client_secret, user_agent=user_agent)

# Choosing subreddit and fetch posts
subreddit_name = 'Politics'
subreddit = reddit.subreddit(subreddit_name)

# Fetch the top 1000 posts
top_posts = subreddit.top(limit=10000)

# List to store post data
posts_data = []

# Loop through the posts and extract relevant data
for post in top_posts:
    posts_data.append({
        'Title': post.title,
        'Score': post.score,
        'ID': post.id,
        'URL': post.url,
        'Num Comments': post.num_comments,
        'Created At': post.created_utc,
        'Subreddit': post.subreddit.display_name,
        'Author': post.author.name if post.author else None,
    })

# Convert the data to a Pandas DataFrame
dataset = pd.DataFrame(posts_data)
```

Extracted Columns and Their Descriptions

1. **Title**
    - **Description**: The title of the Reddit post. It serves as the headline or summary of the content shared by the user. This is often the most descriptive and attention-grabbing part of the post.
    - **Example**: *"Breaking: New Climate Policy Announced"*
    - **Significance**: Useful for quick content overview, trend analysis, and sentiment analysis based on titles.
2. **Score**
    - **Description**: The net upvotes of the post, calculated as upvotes minus downvotes. It reflects the community's reception of the post.
    - **Example**: 542
    - **Significance**: Higher scores indicate popular or well-received posts. Useful for ranking posts based on engagement.
3. **ID**
    - **Description**: The unique identifier for the post, assigned by Reddit.
    - **Example**: a1b2c3
    - **Significance**: Essential for referencing the post, fetching additional data, or creating unique entries in a database.
4. **URL**
    - **Description**: The direct URL to the Reddit post.
    - **Example**: https://www.reddit.com/r/politics/comments/a1b2c3/
    - **Significance**: Allows quick navigation to the post and verification of data.

5. **Num Comments**
    - **Description**: The number of comments on the post.
    - **Example**: 128
    - **Significance**: Indicates the level of engagement or interest in the post. Posts with a high comment count are often central to discussions.
6. **Created At**
    - **Description**: The timestamp (in UTC) indicating when the post was created.
    - **Example**: 1699876543 (Unix timestamp)
    - **Significance**: Useful for time-based analysis, such as identifying trends over specific periods or mapping activity patterns.
7. **Subreddit**
    - **Description**: The subreddit where the post was published. Subreddits are topic-specific communities on Reddit.
    - **Example**: politics
    - **Significance**: Categorizes the post and enables filtering by topics of interest.
8. **Author**
    - **Description**: The username of the individual who created the post. If the author has deleted their account, this will be None.
    - **Example**: User12345 or None
    - **Significance**: Enables analysis of user behavior, tracking of prolific contributors, or identifying the origin of a post.

	Title	Score	ID	URL	Num Comments	Created At	Subreddit	Author
0	Megathread: Joe Biden Projected to Defeat President Donald Trump and Win the 2020 US Presidential Election	214315	jptq5n	https://www.reddit.com/r/politics/comments/jptq5n/megathread_joe_biden_projected_to_defeat/	81377	1.604766e+09	politics	PoliticsModeratorBot
1	Mitch McConnell Will Lose Control Of The Senate As Democrats Have Swept The Georgia Runoffs	156749	krntg6	https://www.buzzfeednews.com/article/paulmcleod/republicans-lose-senate-georgia-mcconnell	10124	1.609940e+09	politics	klawdz
2	Megathread: House Votes to Impeach President Donald J. Trump	147733	ecm1zg	https://www.reddit.com/r/politics/comments/ecm1zg/megathread_house_votes_to_impeach_president/	50767	1.576719e+09	politics	PoliticsModeratorBot
3	Trump Threatens to 'Leave the Country' if He Loses to Biden	135303	jcm5dz	https://www.thedailybeast.com/trump-threatens-to-leave-the-country-if-he-loses-to-biden	16130	1.602897e+09	politics	ONE-OF-THREE
4	Demands for Kushner to Resign Over 'Staggering' Level of 'Depravity' That Put Politics Before Public Health. "Holy hell. Jared Kushner reportedly abandoned a national testing plan because it was *politically advantageous* to sit back and let blue states be eviscerated by the virus."	129742	i19sjg	https://www.commondreams.org/news/2020/07/31/demands-kushner-resign-over-staggering-level-depravity-put-politics-public-health	6767	1.596210e+09	politics	DaFunkJunkie
...	...	...	...	...	...	...	...	...
992	AOC jokes more people watched her gaming online than listened to glitch-ridden DeSantis launch	63458	13rpepm	https://www.independent.co.uk/news/world/americas/us-politics/aoc-desantis-twitter-spaces-campaign-launch-b2345995.html	1990	1.685039e+09	politics	Beckles28nz
993	Gov. Tim Walz doesn't own a single stock	62860	1emeatb	https://www.axios.com/2024/08/07/tim-walz-vp-pick-investment-portfolio	3961	1.723043e+09	politics	axios
994	John McCain famously shut down a racist voter at a 2008 campaign event, now the video is going viral after Trump did nothing to stop racist chants at his rally	62691	cf9w57	https://www.businessinsider.com/john-mccain-racist-voter-2008-video-viral-trump-2019-7	4087	1.563555e+09	politics	BlankVerse
995	Megathread: Vice President Kamala Harris Announces Minnesota Governor Tim Walz as Her 2024 Running Mate	61410	1elhbeb	https://www.reddit.com/r/politics/comments/1elhbeb/megathread_vice_president_kamala_harris_announces/	15732	1.722950e+09	politics	PoliticsModeratorBot
996	Jeff Bezos killed Washington Post endorsement of Kamala Harris, paper reports	60819	1gc0cdv	https://www.cnbc.com/2024/10/25/jeff-bezos-killed-washington-post-endorsement-of-kamala-harris-.html	4620	1.729879e+09	politics	Careful-Rent5779

Data Preprocessing Steps:

1. **Adding a Unique Identifier:**
   - A new column, id, was introduced using the index values to uniquely identify each record in the dataset.
2. **Renaming Columns for Clarity:**
   - The column Title was renamed to text to improve readability and provide a clearer description of its content.
3. **Filtering Relevant Columns:**
   - Only the text and id columns were retained for analysis, ensuring the dataset focused on essential information.
4. **Defining a Cleaning Function:**
   - A function, clean_text, was created to preprocess the textual data by:
     - Removing mentions (e.g., @username,emojis),
     - Eliminating special characters and punctuation,
     - Stripping out URLs.
5. **Normalizing and Cleaning Text:**
   - The cleaning function also standardizes the text by converting it to lowercase, removing extra spaces, and ensuring proper word spacing.
6. **Applying the Cleaning Function:**
   - The clean_text function was applied to the text column, generating a new column named cleaned_RedditContent to maintain the original data intact.

This preprocessing ensures the dataset is well-structured, clean, and ready for further analysis.

```
<ipython-input-225-dd5a5b1f2937>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['cleaned_RedditContent'] = df['text'].apply(clean_text)
```

	Id	text	cleaned_RedditContent
0	0	Megathread: Joe Biden Projected to Defeat President Donald Trump and Win the 2020 US Presidential Election	megathread joe biden projected to defeat president donald trump and win the 2020 us presidential election
1	1	Mitch McConnell Will Lose Control Of The Senate As Democrats Have Swept The Georgia Runoffs	mitch mcconnell will lose control of the senate as democrats have swept the georgia runoffs
2	2	Megathread: House Votes to Impeach President Donald J. Trump	megathread house votes to impeach president donald j trump
3	3	Trump Threatens to 'Leave the Country' if He Loses to Biden	trump threatens to leave the country if he loses to biden
4	4	Demands for Kushner to Resign Over 'Staggering' Level of 'Depravity' That Put Politics Before Public Health. "Holy hell. Jared Kushner reportedly abandoned a national testing plan because it was *politically advantageous* to sit back and let blue states be eviscerated by the virus."	demands for kushner to resign over staggering level of depravity that put politics before public health holy hell jared kushner reportedly abandoned a national testing plan because it was politically advantageous to sit back and let blue states be eviscerated by the virus
...	...	...	...
992	992	AOC jokes more people watched her gaming online than listened to glitch-ridden DeSantis launch	aoc jokes more people watched her gaming online than listened to glitch ridden desantis launch
993	993	Gov. Tim Walz doesn't own a single stock	gov tim walz doesn t own a single stock
994	994	John McCain famously shut down a racist voter at a 2008 campaign event, now the video is going viral after Trump did nothing to stop racist chants at his rally	john mccain famously shut down a racist voter at a 2008 campaign event now the video is going viral after trump did nothing to stop racist chants at his rally
995	995	Megathread: Vice President Kamala Harris Announces Minnesota Governor Tim Walz as Her 2024 Running Mate	megathread vice president kamala harris announces minnesota governor tim walz as her 2024 running mate
996	996	Jeff Bezos killed Washington Post endorsement of Kamala Harris, paper reports	jeff bezos killed washington post endorsement of kamala harris paper reports

997 rows × 3 columns

Text Preprocessing: Stopword Removal

1. **Downloading NLTK Resources:**
   - Essential NLTK resources, such as punkt (for tokenization) and stopwords (a predefined list of common English stopwords), were downloaded to facilitate text processing.
2. **Defining a Stopword Removal Function:**
   - A custom function, remove_stopwords, was implemented to eliminate common stopwords from the text. The function performs the following steps:
     - Tokenizes the text into individual words.
     - Filters out words that match those in the predefined English stopwords list.
3. **Applying Stopword Removal:**
   - The remove_stopwords function was applied to the cleaned_RedditContent column, refining the text further by removing unnecessary words while updating the column with the processed content.

This step ensures the text is concise and retains only meaningful words for better analytical outcomes.

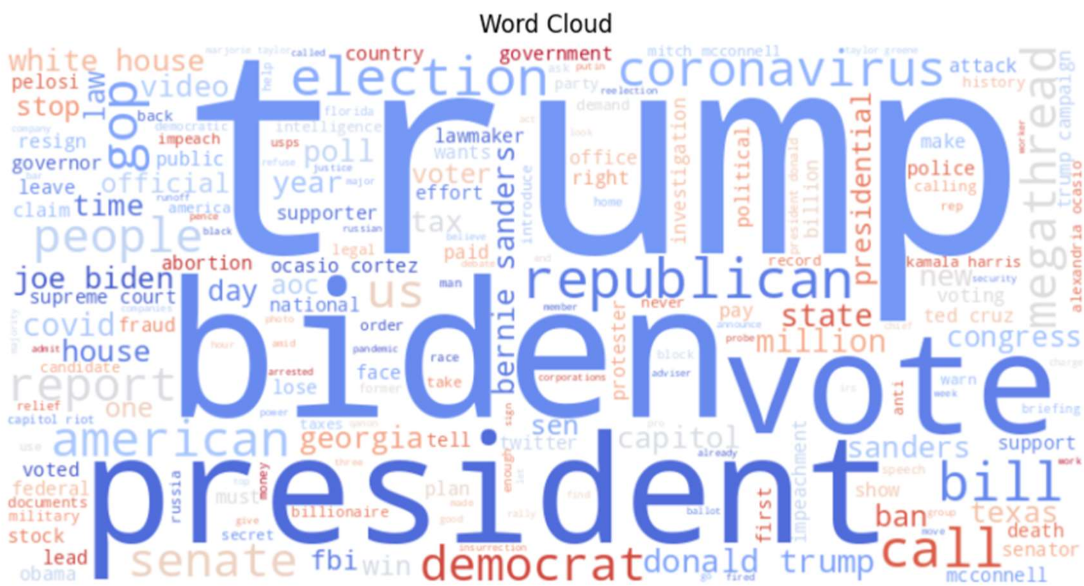
<u>**EXPLORATORY DATA ANALYSIS**</u>

Shape: The dataset contains 997 rows(train dataset) and 2 columns.

Columns:

- id (int64) - 997 non-null values
- cleaned_RedditContent (object) - 997 non-null values

1. Word Cloud for the most frequent words:



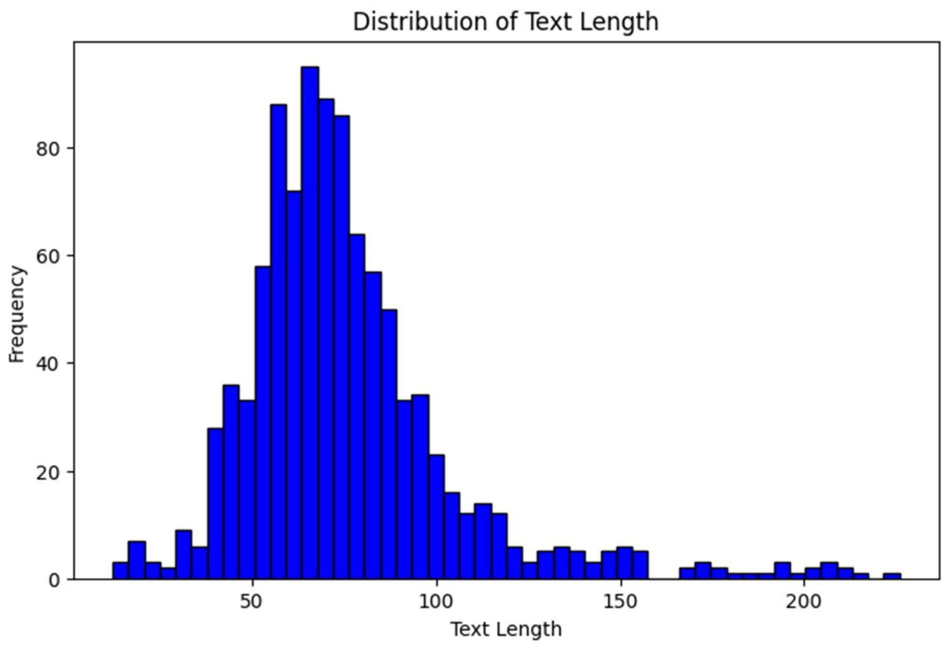
Dominant Words:

- "Trump," "Biden," "President": These words appear prominently, indicating a focus on U.S. presidential figures and related events, potentially around the time of an election or major political developments.
- "Election," "Vote": Suggests an emphasis on electoral processes or discussions about voting.

Related Themes:

- Parties and Ideologies: Words like "Republican," "Democrat," and "American" highlight political affiliations and national context.
- COVID-19 and Policies: The appearance of "Coronavirus" and "Covid" suggests discussions about the pandemic's impact on politics or governance.
- Legislation and Governance: Words like "bill," "law," and "Senate" indicate conversations about policymaking or legislative activity.

2. Histogram of Text Length



The distribution appears to be right-skewed, meaning there are more texts with shorter lengths and fewer texts with longer lengths.

Specific Observations:

- Peak: The distribution peaks around the 70-80 range of text length, indicating that the majority of texts in the dataset fall within this length range.
- Right Tail: The right tail of the distribution is longer, suggesting that there are some texts with significantly longer lengths compared to the majority.
- Range: The text lengths in the dataset appear to range from approximately 0 to 220.

NLTK PROCESSING

Select Example Tweet:
An example tweet was selected from the cleaned_RedditContent column using the index 900.

Word Tokenization:
The selected tweet was tokenized into individual words using NLTK's word_tokenize function.

Display First 10 Tokens:
The first 10 tokens of the tokenized tweet were displayed to show how the text was broken down into words.

```
['know', 'still', 'register', 'vote', 'georgia', 'senate', 'runoffs']
```

POS process:

Download POS Tagger:
The NLTK resource for part-of-speech tagging, averaged_perceptron_tagger_eng, was downloaded.

POS Tagging:
The tokenized words were processed through NLTK's pos_tag function to assign each token its corresponding part of speech.

Display POS Tags:
The first 10 tokens along with their POS tags were displayed to show how the words are categorized (e.g., noun, verb, adjective, etc.).

```
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger_eng is already up-to-
[nltk_data]       date!
[('know', 'NNS'),
 ('still', 'RB'),
 ('register', 'VBP'),
 ('vote', 'NN'),
 ('georgia', 'NN'),
 ('senate', 'NN'),
 ('runoffs', 'NNS')]
```

VADER SENTIMENT SCORING

Analyze Sentiment:
The sentiment of the example tweet was analyzed using the polarity_scores method, which provides a dictionary of sentiment scores.

Convert Sentiment Scores to DataFrame:
The sentiment analysis results (res) were converted into a data frame, and the index was transposed (T) to make each sentiment score (negative, neutral, positive, compound) a column.

Merge with Original Data:
The sentiment DataFrame was merged with the original dataset (df) using a left join, linking the sentiment scores to each corresponding row based on the original_index_name.

The final dataset contains sentiment scores (neg, neu, pos, compound), ID, cleaned Reddit content, and the text length for each entry.

	Id	neg	neu	pos	compound	cleaned_RedditContent	Text Length
0	0	0.169	0.618	0.213	0.2023	megathread joe biden projected defeat president donald trump win 2020 us presidential election	94
1	1	0.252	0.748	0.000	-0.4019	mitch mcconnell lose control senate democrats swept georgia runoffs	67
2	2	0.000	1.000	0.000	0.0000	megathread house votes impeach president donald j trump	55
3	3	0.670	0.330	0.000	-0.6249	trump threatens leave country loses biden	41
4	4	0.274	0.658	0.068	-0.8176	demands kushner resign staggering level depravity put politics public health holy hell jared kushner reportedly abandoned national testing plan politically advantageous sit back let blue states eviscerated virus	211
...	...	...	...	...	...	...	...
992	992	0.000	0.833	0.167	0.2500	aoc jokes people watched gaming online listened glitch ridden desantis launch	77
993	993	0.000	1.000	0.000	0.0000	gov tim walz single stock	25
994	994	0.321	0.603	0.076	-0.7971	john mccain famously shut racist voter 2008 campaign event video going viral trump nothing stop racist chants rally	115
995	995	0.000	1.000	0.000	0.0000	megathread vice president kamala harris announces minnesota governor tim walz 2024 running mate	95
996	996	0.304	0.541	0.155	-0.4939	jeff bezos killed washington post endorsement kamala harris paper reports	73

997 rows × 7 columns

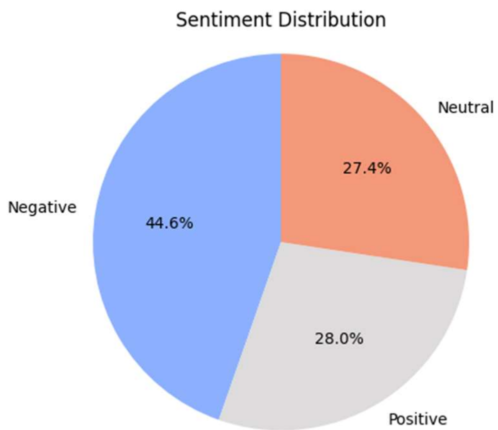
Sentiment Classification using VADER

Id	cleaned_RedditContent	Text Length	Sentiment
0	megathread biden projected defeat president donald trump 2020 presidential election	94	Negative
1	mitch mcconnell lose control senate democrat swept georgia runoff	67	Negative
2	megathread house vote impeach president donald trump	55	Neutral
3	trump threatens leave country loses biden	41	Negative
4	demand kushner resign staggering level depravity politics public health holy hell jared kushner reportedly abandoned national testing plan politically advantageous back blue state eviscerated virus	211	Negative
...	...	...	...
992	joke people watched gaming online listened glitch ridden desantis launch	77	Positive
993	walz single stock	25	Neutral
994	john mccain famously shut racist voter 2008 campaign event video going viral trump nothing stop racist chant rally	115	Negative
995	megathread vice president kamala harris announces minnesota governor walz 2024 running mate	95	Neutral
996	jeff bezos killed washington post endorsement kamala harris paper report	73	Negative

Sentiment Function Definition:
A function, get_sentiment, was defined to classify the sentiment of each tweet based on the compound score from VADER sentiment analysis:

- If the compound score is greater than or equal to 0.05, the sentiment is classified as **Positive**.
- If the compound score is less than or equal to -0.05, the sentiment is classified as **Negative**.
- Otherwise, the sentiment is classified as **Neutral**.

Sentiment Distribution Visualization

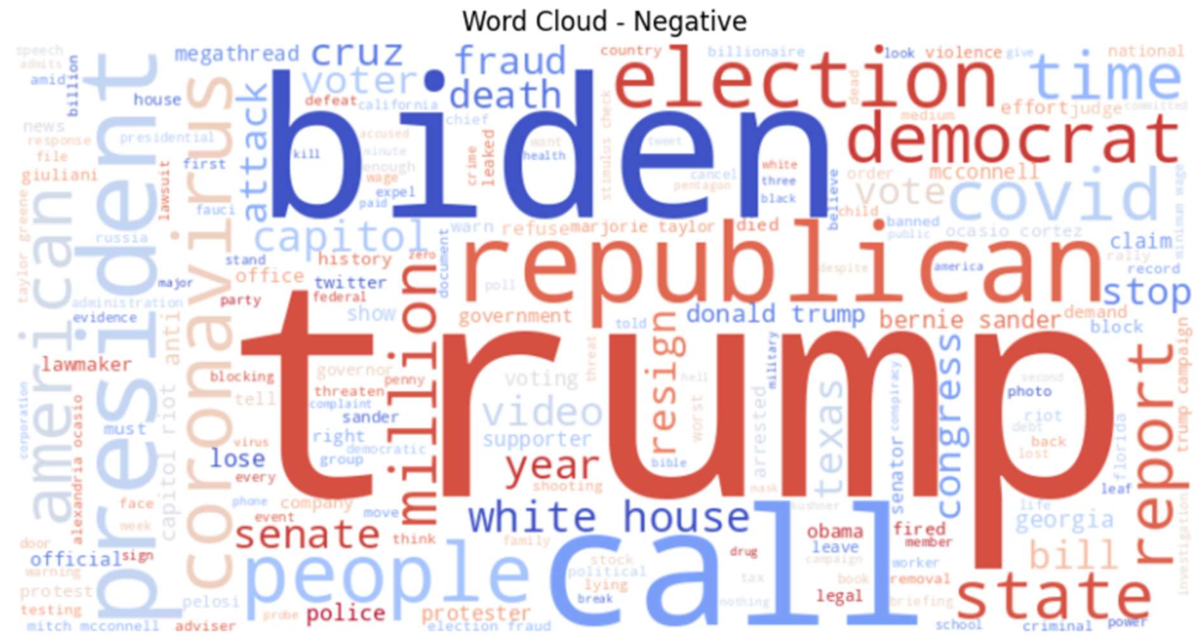


Negative: The largest slice of the pie chart is allocated to "Negative" sentiments, accounting for 44.6% of the total sentiments. This suggests that a significant portion of the data expresses negative opinions or emotions.

Positive: The "Positive" sentiment category holds a smaller share of 28.0%. This indicates that a notable portion of the data conveys positive opinions or emotions.

Neutral: The "Neutral" sentiment category occupies 27.4% of the pie chart. This suggests that a portion of the data expresses neither strongly positive nor negative sentiments.

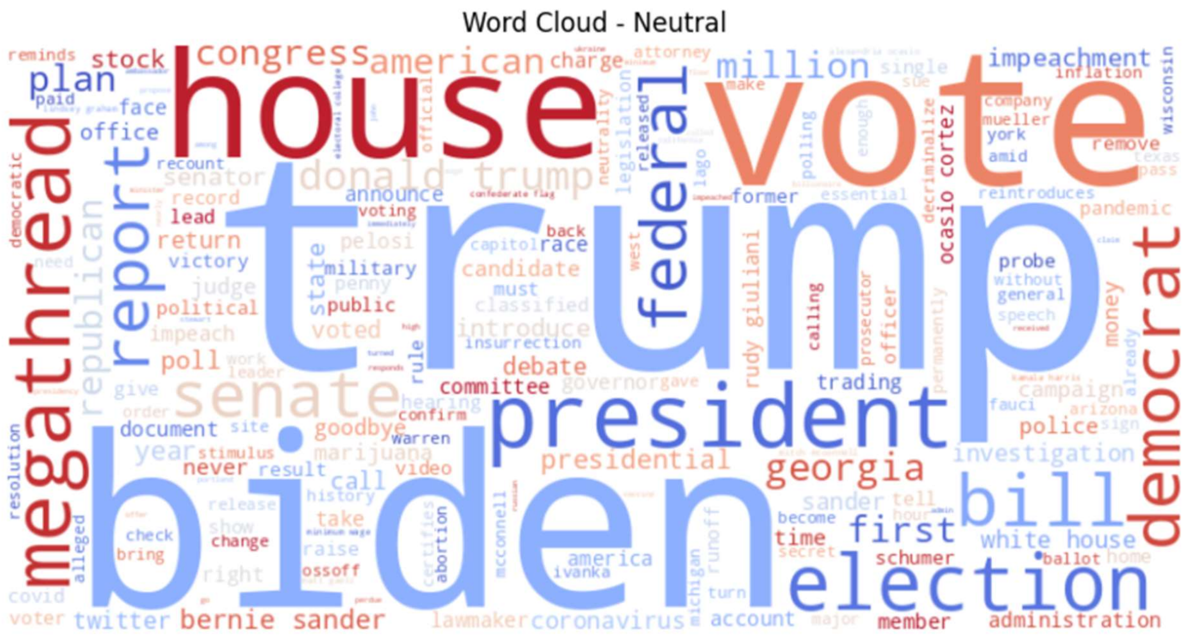
Word cloud for sentiments:



Dominant Themes:

1. **Key Figures and Events**:
    - **"Trump" and "Biden"**: These names are central, indicating a discussion about their roles in the political sphere, possibly around an election or governance.
    - **"President" and "Election"**: Suggest the context is a U.S. presidential election or related debates.
2. **Electoral Process**:
    - **"Vote" and "House"**: Highlight voting behavior, elections, and legislative discussions, likely involving the U.S. House of Representatives.
3. **Governmental and Political Systems**:
    - **"Federal," "Senate," and "Democrat"**: Reflects a focus on federal governance, party ideologies, and legislative activity.
4. **Additional Contextual Words**:
    - Words like **"megathread," "report," and "Georgia"** may suggest in-depth analyses or discussions about specific events or battleground states.

- "**Coronavirus" and "pandemic"**: Indicate the continued impact of COVID-19 on political discussions.
- **"Bill," "impeachment," and "law"**: Suggest legislative and judicial matters are prominent in the discourse.



This "Word Cloud - Neutral" visualization illustrates the frequency of terms in a dataset related to political discourse. The largest words represent the most frequently discussed topics, revealing key themes and focus areas:

Key Highlights:

1. **Prominent Names**:
   - **"Trump" and "Biden"**: Their prominence indicates significant focus on these political figures, likely tied to their roles in governance or elections.
   - This reflects major discussions around political events involving these individuals.
2. **Political Processes**:
   - **"Vote" and "Election"**: Highlight emphasis on the electoral process, voting behavior, and outcomes.
   - **"House" and "Senate"**: Indicate discussions about legislative bodies, perhaps linked to policy decisions or political debates.
3. **Party and Ideological References**:
   - **"Democrat" and "Republican"**: Suggests conversations around party politics, ideologies, or their influence on governance and policy-making.
4. **Government and Governance**:
   - Words like **"federal," "bill," and "president"** reflect discussions surrounding federal policies, legislation, and executive decisions.

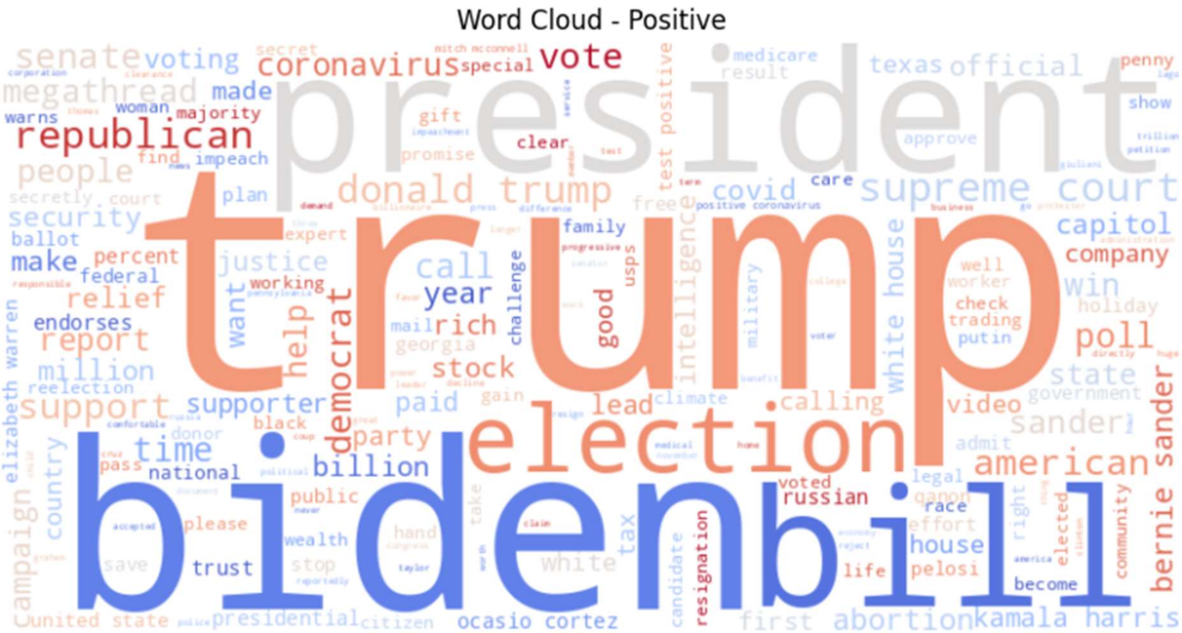
- References to **"congress"** and **"committee"** also point to parliamentary processes or investigations.

5. **Key Issues and Context**:

  - **"Georgia"**: Indicates focus on a specific state, likely tied to elections or legislative activities.
  - **"Coronavirus" and "pandemic"**: Highlight the ongoing impact of the COVID-19 crisis in political discourse.
  - **"Impeachment" and "law"**: Suggest topics around accountability or judicial processes.

6. **Miscellaneous Topics**:

  - **"Megathread" and "report"**: Could suggest detailed discussions, analyses, or compilations of data on political topics.



The word cloud represents the most frequent positive words associated with discussions or sentiments in the dataset. Here's an interpretation:

Key Insights:

1. **Dominant Words**:

  - **Trump** and **Biden** are the largest, indicating these names were mentioned most frequently in a positive context.
  - **President**, **election**, and **bill** also stand out, reflecting the focus on leadership, elections, and legislative topics.

2. **Political Themes**:

- Words like **republican**, **democrat**, and **vote** suggest the discussion is centered around political parties and voting.

3. **Current Events**:
    - Words like **coronavirus**, **covid**, and **supreme court** highlight ongoing significant issues during the time period.

4. **Positive Sentiments**:
    - Words such as **relief**, **help**, **support**, **good**, and **win** indicate themes of assistance, victory, and positive actions.

STATISTICS BASED ALGORITHM

TF-IDF Vectorization:
The TfidfVectorizer was applied to convert text into numerical features, limiting to 5000 features. It was fitted on the training data and then used to transform the test data.

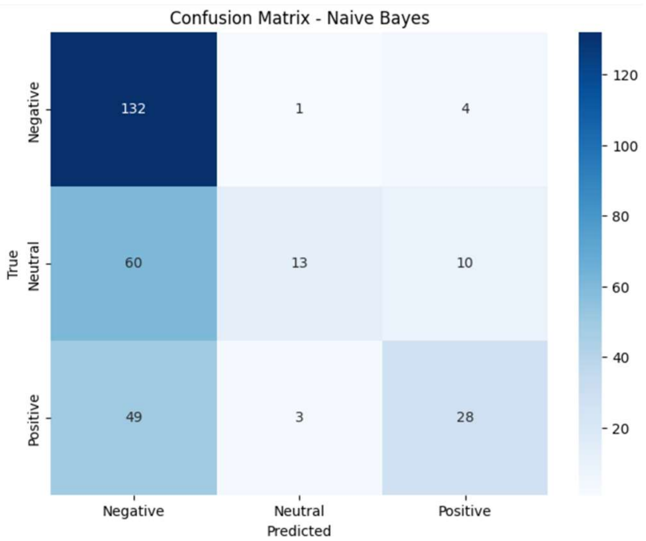
Naïve Bayes:

```
Accuracy: 0.58

Classification Report:
              precision    recall  f1-score   support

    Negative       0.55      0.96      0.70       137
     Neutral       0.76      0.16      0.26        83
    Positive       0.67      0.35      0.46        80

    accuracy                           0.58       300
   macro avg       0.66      0.49      0.47       300
weighted avg       0.64      0.58      0.51       300
```



Interpretation:

- The model is performing well for **Negative** class classification, with high recall (96%) but lower precision (55%), meaning it's predicting many instances as Negative, some of which are incorrect.
- **Neutral** and **Positive** classes are more challenging for the model, especially in terms of recall (low for both), indicating that the model fails to correctly identify these classes often.
- The **imbalanced data distribution** (likely more instances of the "Negative" class) could be a contributing factor, causing the model to favor the Negative class at the expense of the other two.

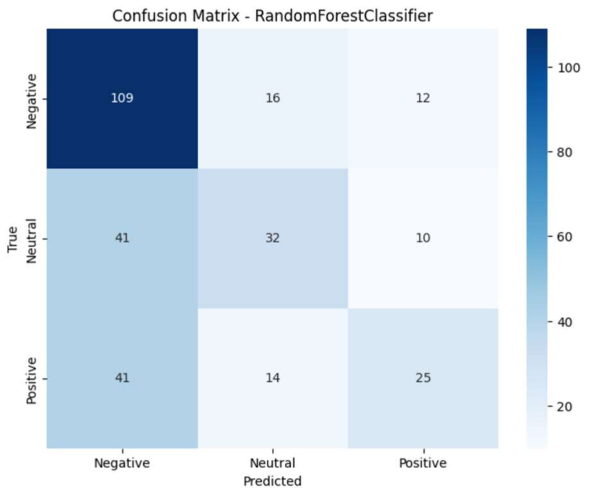
RANDOM FOREST CLASSIFIER:

```
Accuracy: 0.55

Classification Report:
              precision    recall  f1-score   support

    Negative       0.57      0.80      0.66       137
     Neutral       0.52      0.39      0.44        83
    Positive       0.53      0.31      0.39        80

    accuracy                           0.55       300
   macro avg       0.54      0.50      0.50       300
weighted avg       0.55      0.55      0.53       300
```



Interpretation:

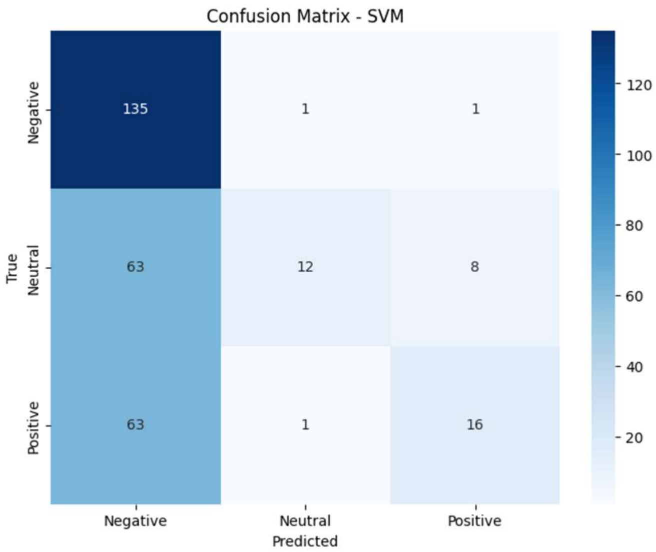
- Negative Class: The Random Forest model does a good job identifying "Negative" instances, with a high recall of 0.80. This suggests the model is better at detecting "Negative" sentiment compared to the other two classes. However, the precision is relatively low (0.57), meaning some of the "Negative" predictions are incorrect.
- Neutral Class: Recall for "Neutral" is quite low at 0.39, meaning many "Neutral" instances are misclassified. The precision is also relatively low (0.52), showing that when the model predicts "Neutral," it is not always correct.
- Positive Class: Recall for "Positive" is also low at 0.31, and precision is similarly low (0.53), meaning the model struggles to correctly predict "Positive" instances.

SVM

Accuracy: 0.54

Classification Report:

	precision	recall	f1-score	support
Negative	0.52	0.99	0.68	137
Neutral	0.86	0.14	0.25	83
Positive	0.64	0.20	0.30	80
accuracy			0.54	300
macro avg	0.67	0.44	0.41	300
weighted avg	0.64	0.54	0.46	300



Interpretation:

- Negative Class: The SVM model excels at detecting "Negative" instances, with a high recall of 0.99, meaning almost all "Negative" instances are identified correctly. However, the precision is relatively low at 0.52, suggesting that a significant portion of the instances predicted as "Negative" are incorrect.
- Neutral Class: The recall for "Neutral" is very low (0.14), meaning most of the actual "Neutral" instances are misclassified, even though the precision for "Neutral" is high (0.86). This indicates that the model is too focused on the "Negative" class and doesn't capture "Neutral" cases well.
- Positive Class: Similar to the "Neutral" class, the recall for "Positive" is low (0.20), meaning the model struggles to detect "Positive" instances, even though the precision for "Positive" is moderately better (0.64).

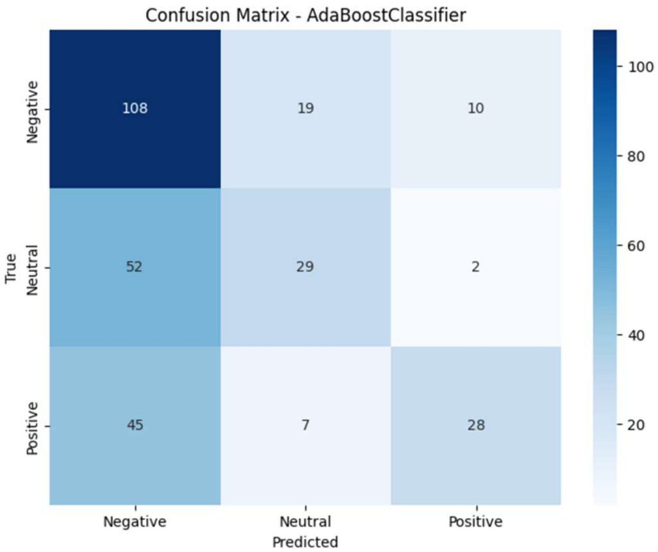
ADA BOOST CLASIFIER:

```
AdaBoostClassifier Accuracy: 0.55

AdaBoostClassifier Classification Report:
              precision    recall  f1-score   support

    Negative       0.53      0.79      0.63       137
     Neutral       0.53      0.35      0.42        83
    Positive       0.70      0.35      0.47        80

    accuracy                           0.55       300
   macro avg       0.58      0.50      0.51       300
weighted avg       0.57      0.55      0.53       300
```



Interpretation:

- Negative Class: The AdaBoost model performs quite well on the "Negative" class, with a recall of 0.79 and precision of 0.53. It correctly identifies most of the "Negative" instances, although there is still a significant portion misclassified as "Neutral" or "Positive."
- Neutral Class: The recall for "Neutral" is quite low (0.35), meaning many "Neutral" instances are misclassified, but the precision (0.53) indicates that when the model predicts "Neutral," it is relatively accurate.
- Positive Class: The recall for "Positive" is similarly low (0.35), indicating that the model struggles to correctly identify "Positive" instances, despite a higher precision (0.70) when it does predict "Positive."

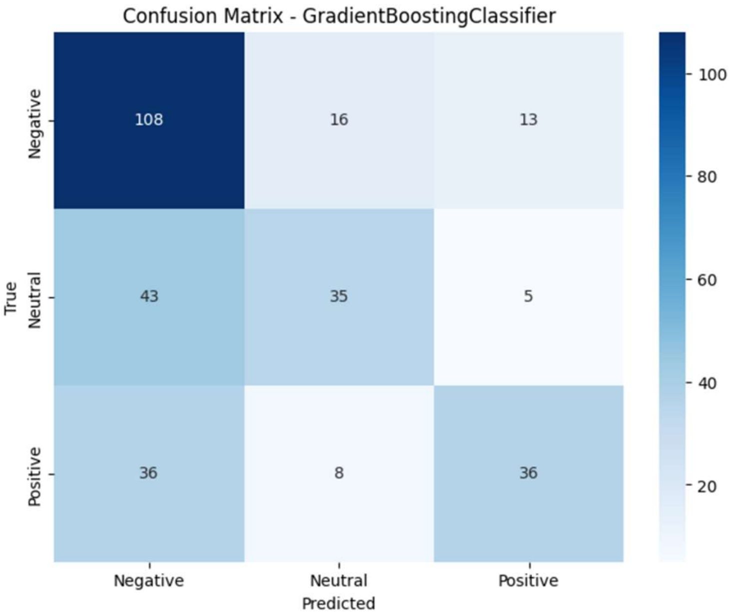
GRADIENT BOOSTING CLASIFIER:

```
GradientBoostingClassifier Accuracy: 0.60

GradientBoostingClassifier Classification Report:
              precision    recall  f1-score   support

    Negative       0.58      0.79      0.67       137
     Neutral       0.59      0.42      0.49        83
    Positive       0.67      0.45      0.54        80

    accuracy                           0.60       300
   macro avg       0.61      0.55      0.57       300
weighted avg       0.61      0.60      0.58       300
```



Interpretation:

- Negative Class: The Gradient Boosting model performs well on the "Negative" class, with a high recall of 0.79, meaning that it correctly identifies most "Negative" instances. However, the precision of 0.58 suggests that there are still a significant number of false positives.
- Neutral Class: The recall for "Neutral" is moderate (0.42), meaning that a good portion of "Neutral" instances are misclassified. However, the precision (0.59) is fairly good, suggesting that when the model predicts "Neutral," it is often correct.
- Positive Class: The recall for "Positive" is 0.45, showing improvement compared to other models, but it is still not optimal. The precision of 0.67 indicates that the model is relatively good at predicting "Positive" instances when it does make that prediction.

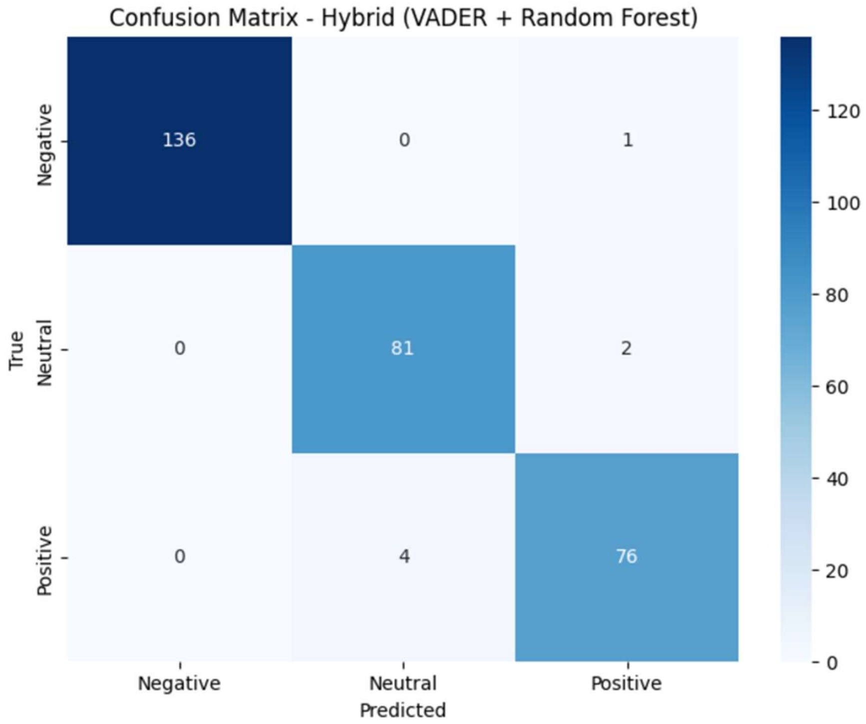
HYBRID-BASED (VADER + RANDOM FOREST):

```
Hybrid Model (VADER + Random Forest) Accuracy: 0.98

Hybrid Model Classification Report:
              precision    recall  f1-score   support

    Negative       1.00      0.99      1.00       137
     Neutral       0.95      0.98      0.96        83
    Positive       0.96      0.95      0.96        80

    accuracy                           0.98       300
   macro avg       0.97      0.97      0.97       300
weighted avg       0.98      0.98      0.98       300
```



Key Insights:

Accuracy: 0.98

- The model achieves a high accuracy of 98%, which indicates strong generalization and a very good fit for this problem. It demonstrates that the hybrid model can effectively classify the data across all three sentiment classes.

Weighted Average:

- **Precision**: 0.98. The weighted precision indicates that the model is highly accurate across all classes, particularly for the "Negative" class, which has the highest support.

- **Recall**: 0.98. The weighted recall confirms that the model is consistently identifying the correct instances across all classes.

- **F1-Score**: 0.98. The weighted F1-score supports the overall balanced performance of the model.

Interpretation of the Hybrid Model:

- The combination of **VADER** (for sentiment analysis) and **Random Forest** has clearly yielded an impressive performance, significantly boosting the accuracy and classification across all classes. By leveraging the strengths of both models, the hybrid approach manages to achieve high precision, recall, and F1-scores for all three sentiment categories (Negative, Neutral, Positive).

OVERALL MODEL PERFORMANCE

Model	Accuracy	Strengths	Weaknesses
Hybrid Model (VADER + Random Forest)	0.98	- Perfect performance for "Negative" class (Precision = 1.00, Recall = 0.99) - Excellent recall for "Neutral" and "Positive" classes - Very high precision and F1-scores for all classes	- Very few instances misclassified (but still some misclassification in the "Neutral" and "Positive" classes)
GradientBoostingClassifier	0.60	- Balanced performance across all classes with decent recall - Strong recall for "Negative" (0.79) - Reasonably high precision and F1 for "Positive"	- Recall for "Positive" and "Neutral" is relatively low compared to "Negative" - Still some room for improvement in performance, especially on "Neutral"
SVM (Support Vector Machine)	0.54	- Good for linear separability in data - Works well on datasets with large margins of separation	- Poor recall for "Neutral" and "Positive" - Tendency to misclassify "Neutral" and "Positive" as "Negative"
Random Forest	0.55	- Fairly consistent predictions - Handles class imbalance well with high precision for "Negative"	- Recall for "Neutral" and "Positive" is lower than for "Negative" - Some misclassification between classes
Naive Bayes	0.58	- Strong performance on "Negative" class - Works well with text data	- Poor performance on "Neutral" and "Positive" classes

		- Fast to train and predict	- Low recall and F1 for "Neutral" and "Positive"
AdaBoostClassifier	0.55	- Reasonably balanced precision across all classes - Good precision for "Positive"	- Low recall for "Neutral" and "Positive" - Significant misclassification in "Neutral" and "Positive"

Conclusion

The analysis of political discussions on Reddit revealed significant insights into public sentiment and discourse trends, particularly in the political domain. The data processing pipeline—leveraging PRAW for ethical data collection, text cleaning, and sentiment analysis using VADER—enabled effective classification into Positive, Neutral, and Negative sentiments. Machine learning models, particularly the hybrid approach combining VADER and Random Forest, outperformed individual algorithms, achieving a remarkable accuracy of 98%.

Key findings include:

- **Dominance of Negative Sentiment:** Nearly half of the posts express negative opinions, highlighting public dissatisfaction or critical discussions on political topics.
- **Topical Relevance:** Discussions were centered around prominent figures like "Trump" and "Biden," elections, legislative processes, and global issues like COVID-19.
- **Model Performance:** The hybrid model demonstrated superior capability in handling class imbalances and accurately categorizing sentiments compared to other classifiers.

Business Recommendations

1. **For Political Campaigns and Public Relations:**
   - **Focus on Addressing Negative Sentiment:** Develop strategies to address common public grievances highlighted in negative sentiment posts.
   - **Targeted Messaging:** Leverage insights on key topics (e.g., elections, legislation) to tailor communication for specific audiences.
2. **For Media and News Organizations:**
   - **Content Prioritization:** Use sentiment scores to identify highly engaging topics and curate content around popular and divisive issues.
   - **Trend Monitoring:** Continuously monitor emerging discussions and public mood to inform timely reporting and analysis.
3. **For Data-Driven Policy Making:**
   - **Policy Feedback Loops:** Analyze sentiment data to gauge public reception of policies, enabling data-driven adjustments and improvements.
   - **Community Engagement:** Foster discussions in areas showing high engagement but balanced sentiment (e.g., Neutral) to promote constructive dialogues.