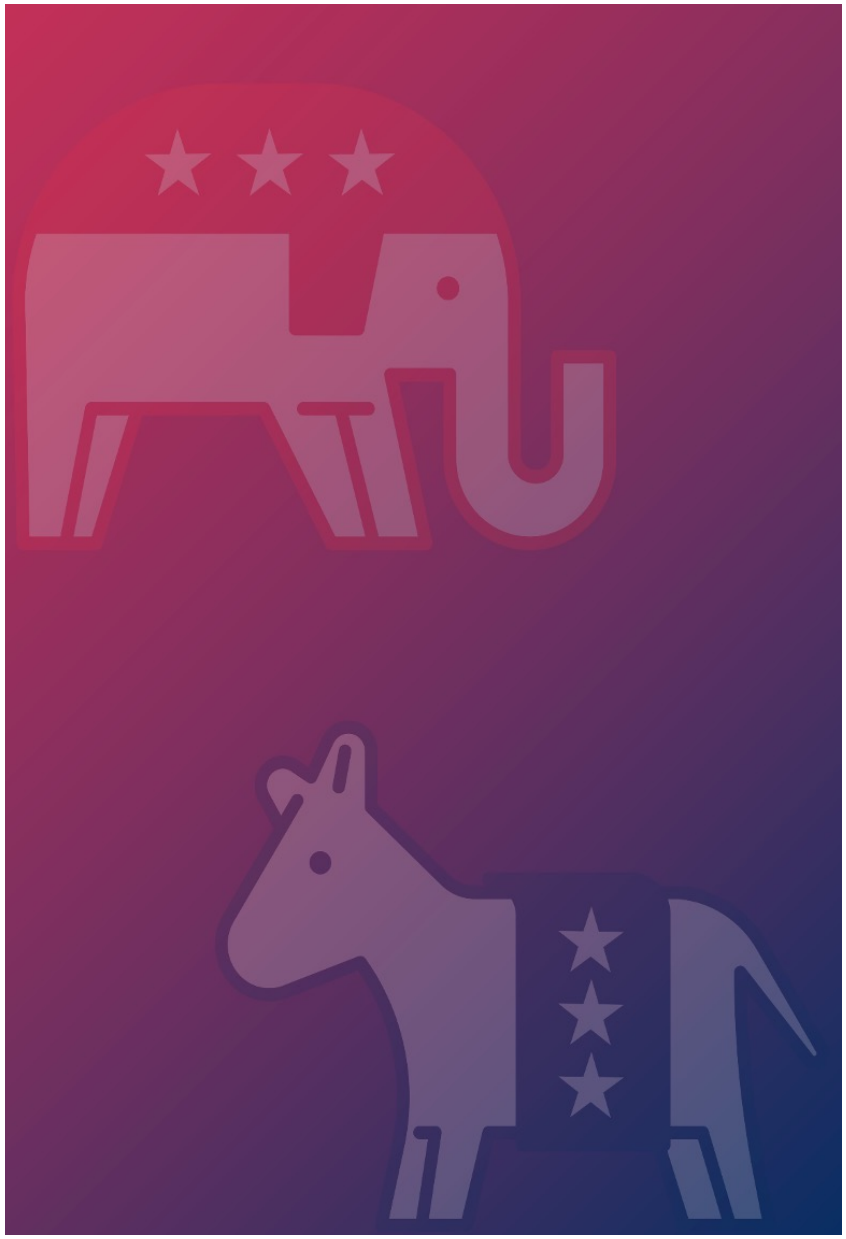


Exploring Political Polarization through Reddit

LSE DATA SCIENCE SOCIETY

TEAM 4

Table of Contents

Ideology Drives Polarization in Media

The Harvard Political Review, one of the leftist media, forecasted Harris securing the national popular vote by a margin of approximately 52% to 48%, translating to 319 electoral votes over former President Donald Trump's 219.

The "world's most accurate economist" from the Fox News is predicting an Election Day victory for former President Trump, with Republicans also predicted to take back control of Congress in a "clean sweep."

REPORT | JANUARY 24, 2020

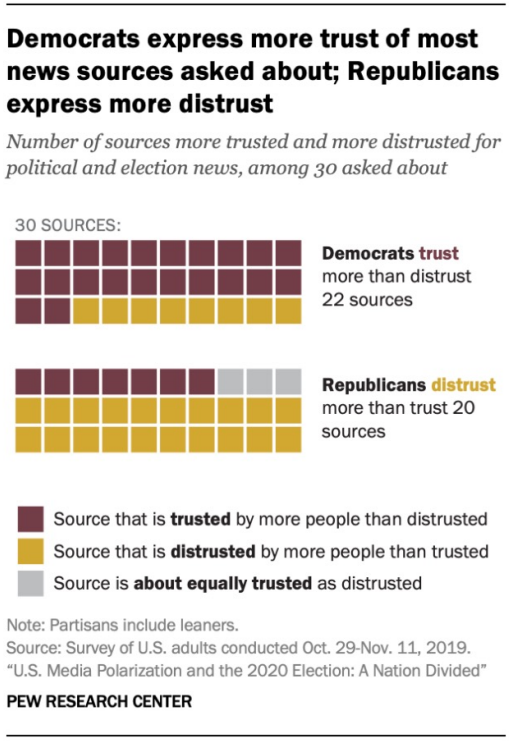
SHARE

U.S. Media Polarization and the 2020 Election: A Nation Divided

Deep partisan divisions exist in the news sources Americans trust, distrust and rely on

BY MARK JURKOWITZ, AMY MITCHELL, ELISA SHEARER AND MASON WALKER

Ideology Drives Media Credibility Polarization



Pew Research Center

Polarization is Growing:

The gap in media trust between political groups has widened significantly, reflecting increasing ideological divides.

Impact:

Polarized media credibility drives Democrats and Republicans to favour ideologically aligned news sources, which reinforces their tendency to develop entirely different viewpoints on the same issues.

Ideology adds another layer to party-line divides of most trusted and distrusted news sources

% who trust each source for political and election news (first five shown)

Democrat/Lean Dem				Republican/Lean Rep			
LIBERAL		MODERATE/ CONSERVATIVE		MODERATE/ LIBERAL		CONSERVATIVE	
CNN	70%	CNN	65%	Fox News	51%	Fox News	75%
New York Times	66	ABC News	63	ABC News	47	Hannity (radio)	43
PBS	66	NBC News	61	CBS News	42	Limbaugh (radio)	38
NPR	63	CBS News	60	NBC News	41	ABC News	24
NBC News	61	PBS	48	CNN	36	CBS News	23

% who distrust each source for political and election news (first five shown)

Democrat/Lean Dem				Republican/Lean Rep			
LIBERAL		MODERATE/ CONSERVATIVE		MODERATE/ LIBERAL		CONSERVATIVE	
Fox News	77%	Fox News	48%	CNN	43%	CNN	67%
Limbaugh (radio)	55	Limbaugh (radio)	34	MSNBC	32	MSNBC	57
Breitbart	53	Hannity (radio)	28	HuffPost	30	New York Times	50
Hannity (radio)	50	Breitbart	22	BuzzFeed	29	NBC News	50
NY Post	27	BuzzFeed	20	Fox News	29	CBS News	48

Note: Order of outlets does not necessarily indicate statistically significant differences.
Source: Survey of U.S. adults conducted Oct. 29-Nov. 11, 2019.
"U.S. Media Polarization and the 2020 Election: A Nation Divided"

PEW RESEARCH CENTER

About two-thirds of liberal Democrats (66%) trust The New York Times, for example. In comparison, just 10% of conservative Republicans trust the Times, while 50% outright distrust it.

Rush Limbaugh, meanwhile, is the third-most trusted source among conservative Republicans (38%) but tied for the second-most distrusted source among liberal Democrats (55%).

Echo chamber (media)

Article Talk

From Wikipedia, the free encyclopedia

In news media and social media, an **echo chamber** is an environment or ecosystem in which participants encounter beliefs that amplify or reinforce their preexisting beliefs by communication and repetition inside a closed system and insulated from rebuttal.[2][3][4] An echo chamber circulates existing views without encountering opposing views, potentially resulting in confirmation bias. Echo chambers may increase social and political polarization and extremism.[5] On social media, it is thought that echo chambers limit exposure to diverse perspectives, and favor and reinforce presupposed narratives and ideologies.[4][6]

Motivation

Explore how polarization manifests in language, news preferences, and sentiment within two ideologically distinct Reddit communities: 'Democrats' and 'Republicans'

Objectives

1. News Source Preferences

Identify differences in news sources used within each subreddit to reflect ideological divides.

2. Sentiment Analysis

Analyze the sentiment behind the post/comments.

3. Linguistic Difference

Analyze word frequency in each subreddit.

4. Classifier to Quantify Polarization

Test the hypothesis that if distinct communication styles exist, subreddit classification will be easy.

Data Acquisition

Data Collection Method:

Utilized the Python package **PRAW (Python Reddit API Wrapper)** to fetch data programmatically.

Retrieved **1000 posts from each subreddit**, focusing on the "hot" posts, ranked by the number of upvotes and comments.

Time Window:

The time window 2023/12 ~ 2024/12 was chosen as the focus due to the significant event of the U.S. presidential election, and also PRAW only allows recent data.

Data Source:

Data was collected from the *'democrats'* and *'republicans'* subreddits, two of the most active political communities on Reddit, known for high engagement in terms of upvotes and posts.

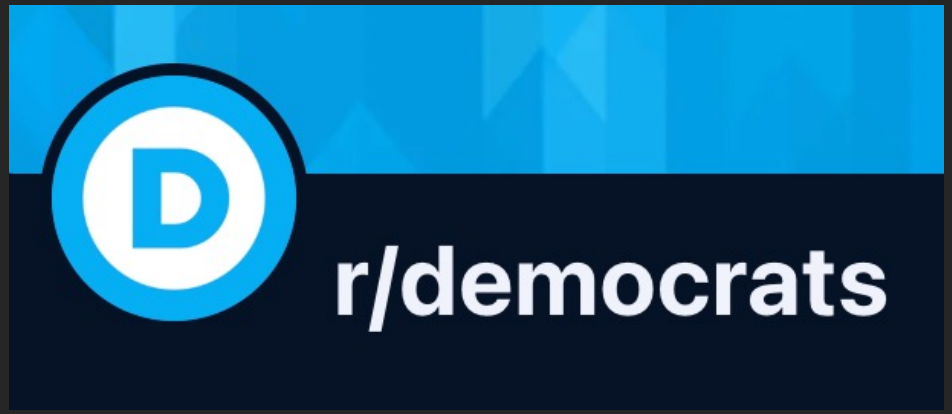
Created Oct 4, 2008

Public

483K
Democratic voters

298
Online

Top 1%
Rank by size



Republican

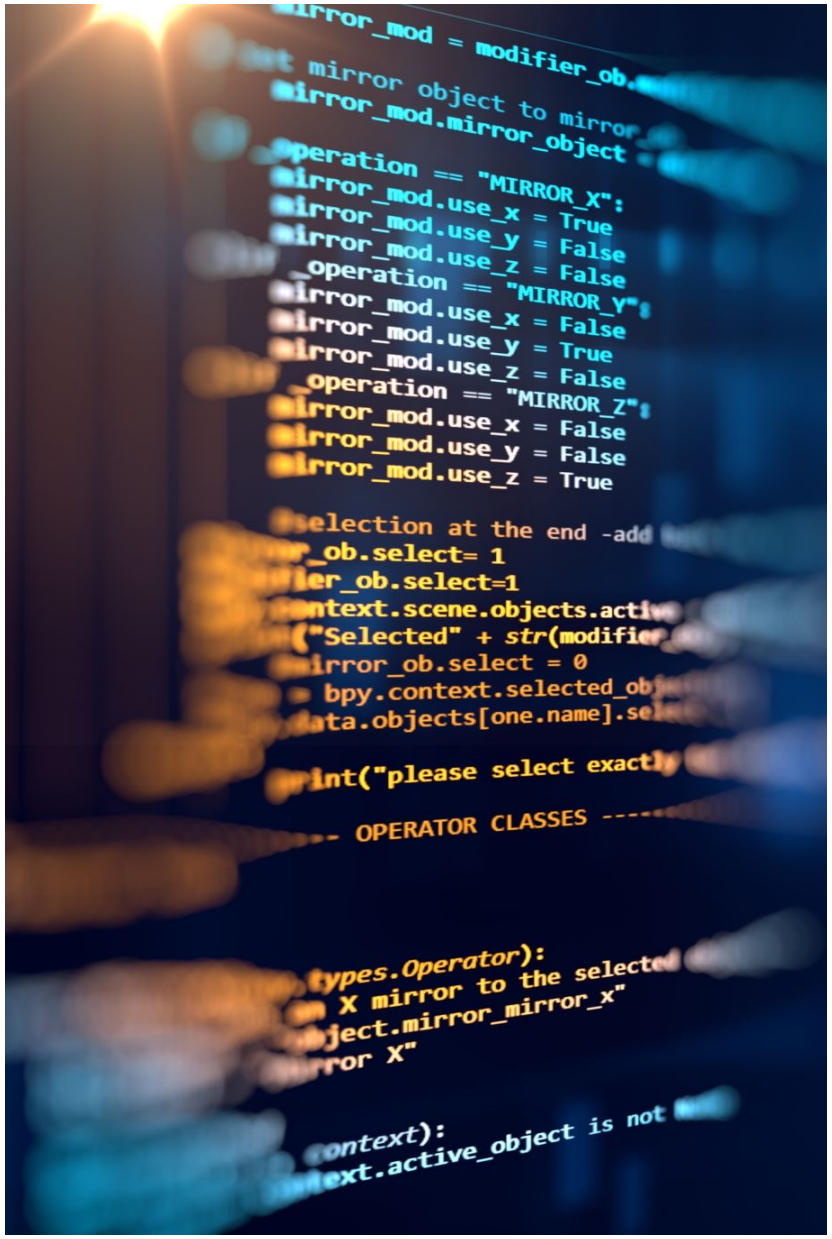
🇺🇸 RIP r/TheDonald 🇺🇸 Let's make Reddit GREAT again!

Created Oct 10, 2008

Public

207K
Members

194
Online



Features

Each post includes the following features:

- **Title**: The title of the post.
- **Score**: The number of upvotes for a comment or submission.
- **ID**: The unique identifier for the post.
- **URL**: A URL for link or image in the post.
- **Number of Comments**: Total comments on the post.
- **Created**: Timestamp indicating when the post was created.
- **Body**: The main content of the post (if any).

	title	score	id	url	comms_num	created	body
194	trump is now starting a tarrif war with Canada...	3376	1h0077o	https://www.reddit.com/gallery/1h0077o	1229	1.732585e+09	so along with China tariffs hurting American b...
175	We are so screwed	2763	1h0p075	https://i.redd.it/0p8vr0yupb3e1.jpeg	1113	1.732661e+09	Posted 5 minutes ago. (https://bsky.app/profi...
102	Biden pardons son	3421	1h4iyre	https://www.nbcnews.com/news/amp/rcna182369	1003	1.733102e+09	I'm curious what my fellow Dems think about th...
176	As of 11/26	6268	1h0hx8d	https://i.redd.it/vqmblnt2aa3e1.png	912	1.732643e+09	
165	"We will suffer" is the shitty part cause "we"...	4701	1h14tnw	https://i.redd.it/e9oinmia9g3e1.jpeg	651	1.732716e+09	

Another Data: Fetching Comments Data

	id	comments	label
0	1gxpeav	['Not just her. Mike Rounds and Kevin Cramer I...	Democrats
1	1gxgra8	['Still hard to believe', 'I feel your pain. I...	Democrats
2	1gxd8ay	['Translation - rumors were true and suppresse...	Democrats
3	1gwtt2i	['We need to repurpose "DEI". One suggestion.....	Democrats
4	1gwtdqn	['Hahahahaha all that tap dancing from him and...	Democrats
...	...	...	...
1995	187ix87	["/r/Republican is a partisan subreddit. This...	Republicans
1996	187472x	["/r/Republican is a partisan subreddit. This...	Republicans
1997	1863tdd	["/r/Republican is a partisan subreddit. This...	Republicans
1998	184gtav	["/r/Republican is a partisan subreddit. This...	Republicans
1999	182ymxi	["/r/Republican is a partisan subreddit. This...	Republicans

[2000 rows x 3 columns]

First, every comments from each post will be extracted (again, with PRAW) and labeled as their origin subreddit.

```python
# Convert the 'comments' column from string to list
df['comments'] = df['comments'].apply(ast.literal_eval)

# Explode the 'comments' column to separate each text into its own row
df_exploded = df.explode('comments').reset_index(drop=True)

print(df_exploded)
```

```
            id                                               comments  \
0       1gxpeav  Not just her. Mike Rounds and Kevin Cramer I b...
1       1gxpeav  Can’t Biden just order the FBI to conduct a ba...
2       1gxpeav  The Republicans don't need her vote, though. T...
3       1gxpeav       Eh. The bar gets lower by the day I suppose.
4       1gxpeav  Don’t worry everybody, Susan Collins will expr...
...         ...                                                ...
226478  182ymxi  The perpetrators werre one race, and the victi...
226479  182ymxi  Is he a super progressive prosecutor or someth...
226480  182ymxi  >Is he a super progressive prosecutor or somet...
226481  182ymxi  Ah, I see. So in CA, the lying in wait is a sp...
226482  182ymxi  Yeah, I worked in Mississippi an Tennessee, an...

             label
0         Democrats
1         Democrats
2         Democrats
3         Democrats
4         Democrats
...             ...
226478  Republicans
226479  Republicans
226480  Republicans
226481  Republicans
226482  Republicans

[226483 rows x 3 columns]
```

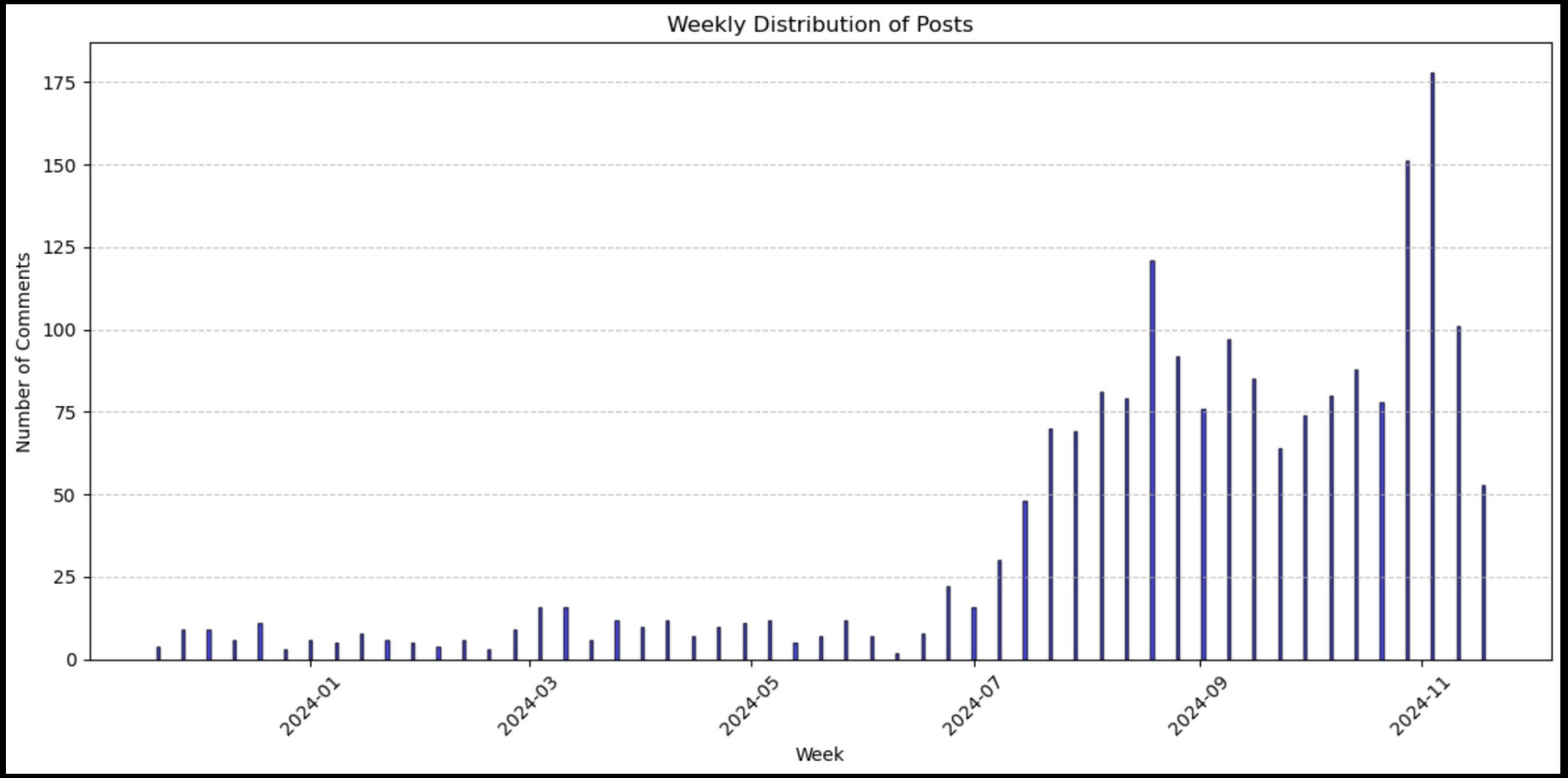
Using the explode function, we can split a list of strings in a single cell into individual strings, creating separate rows for each string.

```python
# List of phrases to check (bot-suspicious)
phrases_to_remove = [
    "**Join:**\n\n* /r/KamalaHarris\n\n*",
    "**Take action:",
    "/r/Republican is a partisan subreddit."
]

# Filter out rows where 'comments' start with any of the specified phrases
df_filtered = df_exploded[~df_exploded['comments'].str.startswith(tuple(phrases_to_remove))].reset_index(drop=True)

print(df_filtered)
```

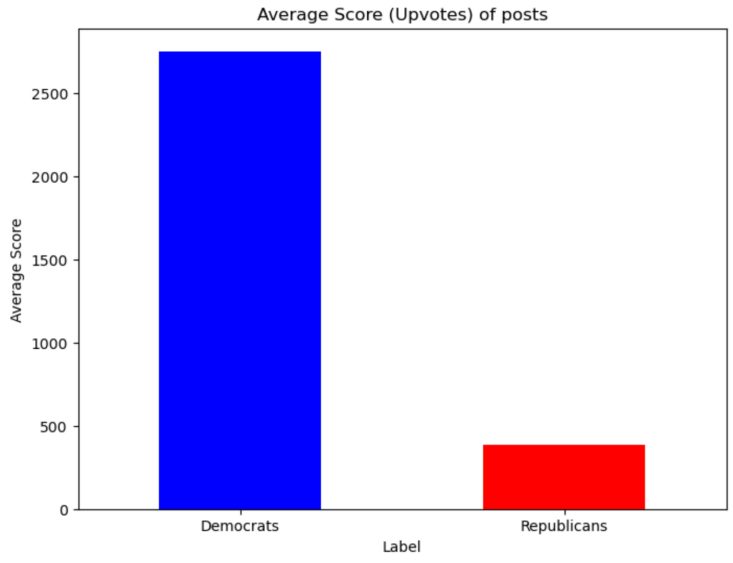
Manually checked bot-suspicious comments and excluded them.

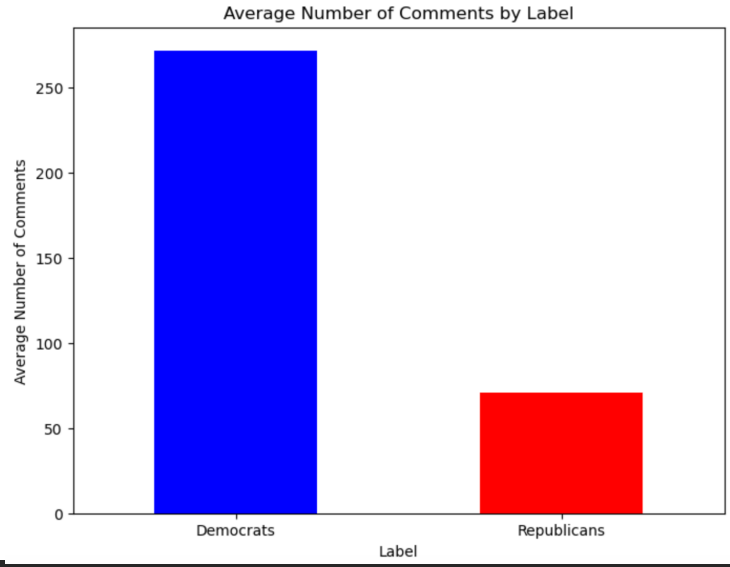
```python
# Step 1: Remove redundant and too-short comments
df_filtered = df_filtered[df_filtered['comments'].str.len() >= 20]  # Keep only comments with 20+ characters

# Step 2: Count the number of Republican comments
num_republicans = df_filtered[df_filtered['label'] == 'Republicans'].shape[0]

print(f"Number of Republican comments: {num_republicans}")

# Step 3: Randomly sample the same number of Democrat comments
df_democrats_sampled = df_filtered[df_filtered['label'] == 'Democrats'].sample(n=num_republicans, random_state=42)

# Step 4: Combine the sampled Democrats and Republicans
df_balanced = pd.concat([df_democrats_sampled, df_filtered[df_filtered['label'] == 'Republicans']])

# Step 5: Shuffle the dataset
df_balanced = df_balanced.sample(frac=1, random_state=42).reset_index(drop=True)

# Check the balance
print(df_balanced['label'].value_counts())
```

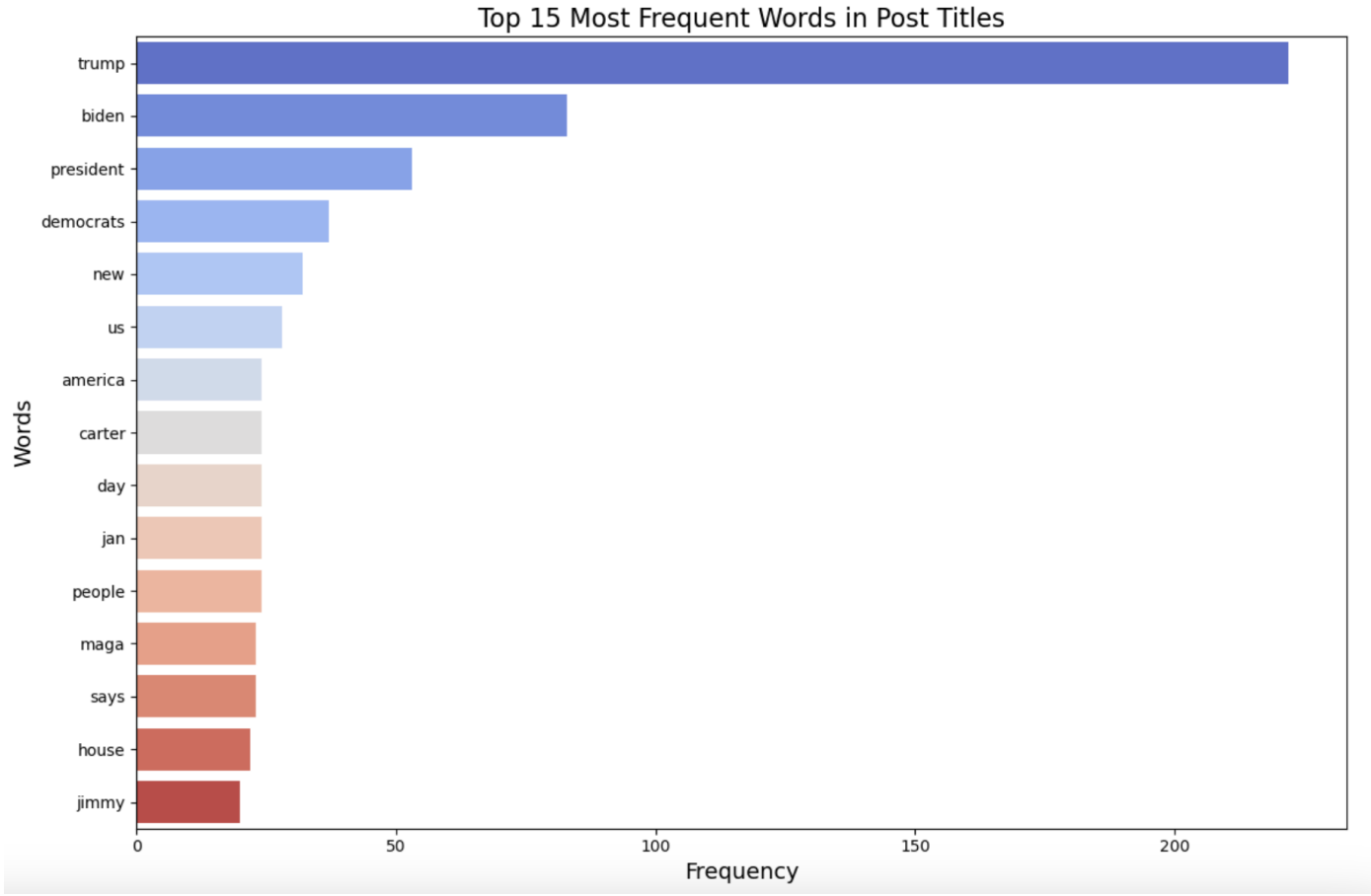
We filtered out comments with fewer than 20 characters to remove redundant or overly short entries. To balance the dataset, we randomly sampled an equal number of Democrat comments to match the count of Republican comments, combined them, and shuffled the resulting dataset.


Weekly Distribution of Posts
Number of Comments
0
25
50
75
100
125
150
175
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
Week

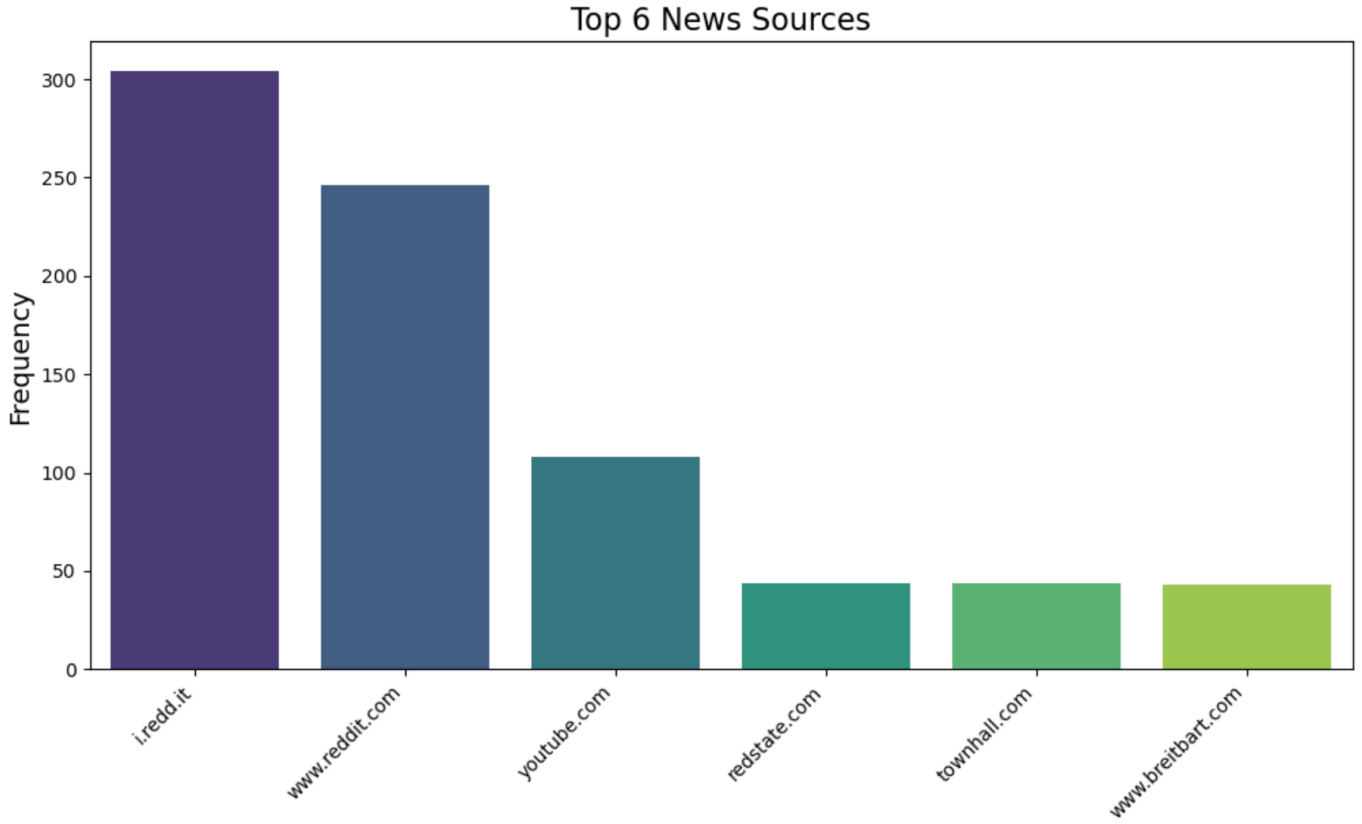
The fact that the Democrats subreddit is much larger has led to a higher average of upvotes and number of comments compared to the Republicans subreddit.

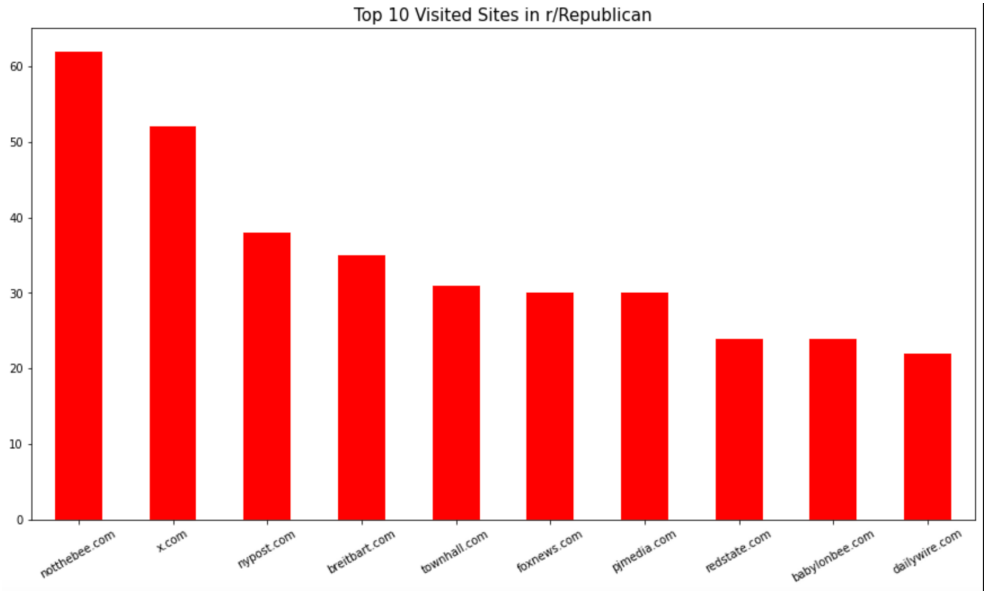


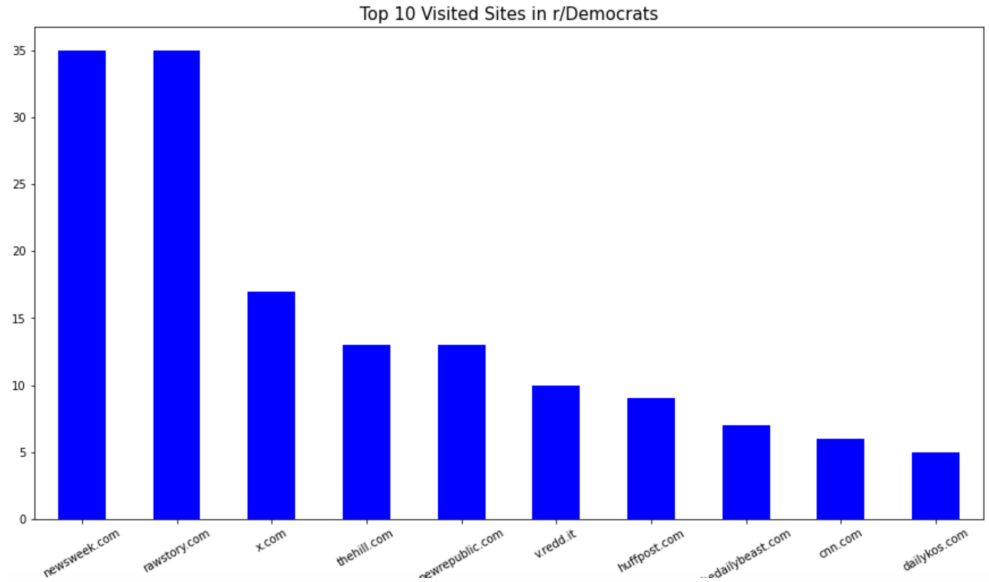



Top 15 Most Frequent Words in Post Titles
trump
biden
president
democrats
new
us
america
carter
day
jan
people
maga
says
house
jimmy
Words
0
50
100
150
200
Frequency

Key Insights 1 : Where do they get the news from?







(Excluded reddit.com and youtube)

Republicans favor conservative-leaning outlets such as notthebee.com, foxnews.com, and breitbart.com.

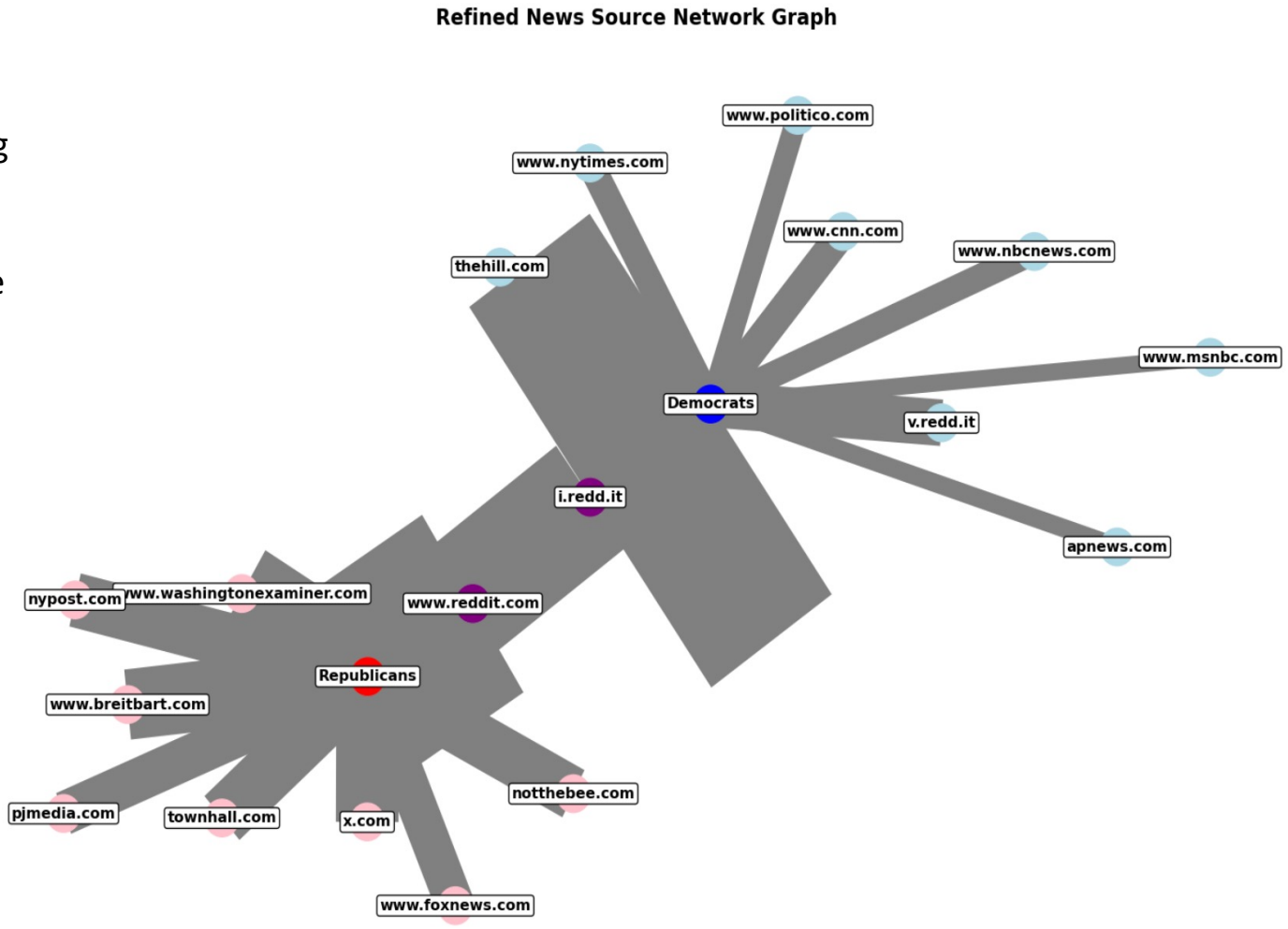
Democrats prefer progressive sources like newsweek.com, rawstory.com, and huffpost.com.

Only x.com (Twitter) was the common news source, and rest of them are different!

These preferences highlight ideological polarization in reddit.

Network Source Relationships:

- Extract top news domains for Democrats and Republicans, identifying shared sources to analyze media influence.
- Construct a networkx graph to visualize relationships between domains and subreddits.
- Used distinct colors and layouts to highlight insights into political discourse.



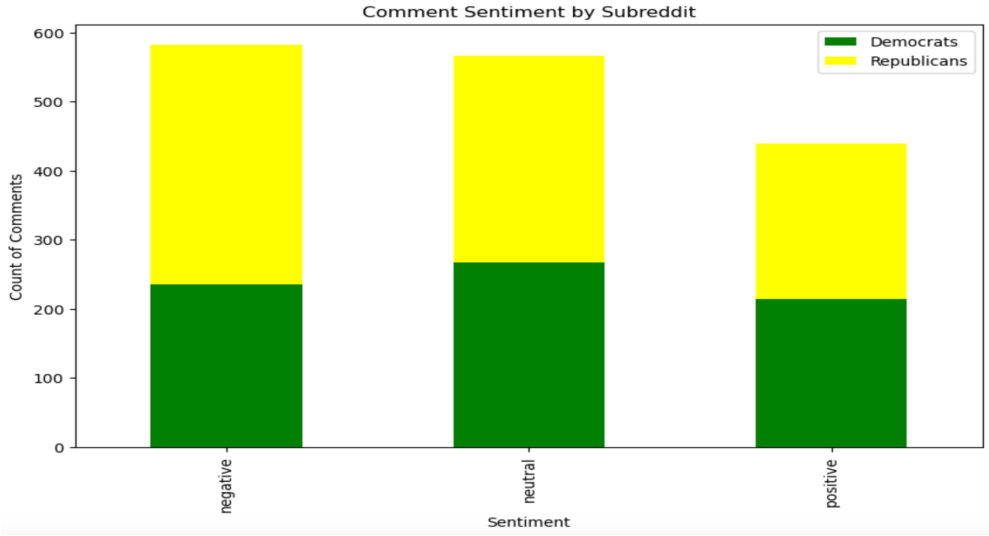
Key Insight 2 : Sentiment Analysis

What Drives the Overall Sentiment in Political Discourse?

VADER-Based Sentiment Scoring:

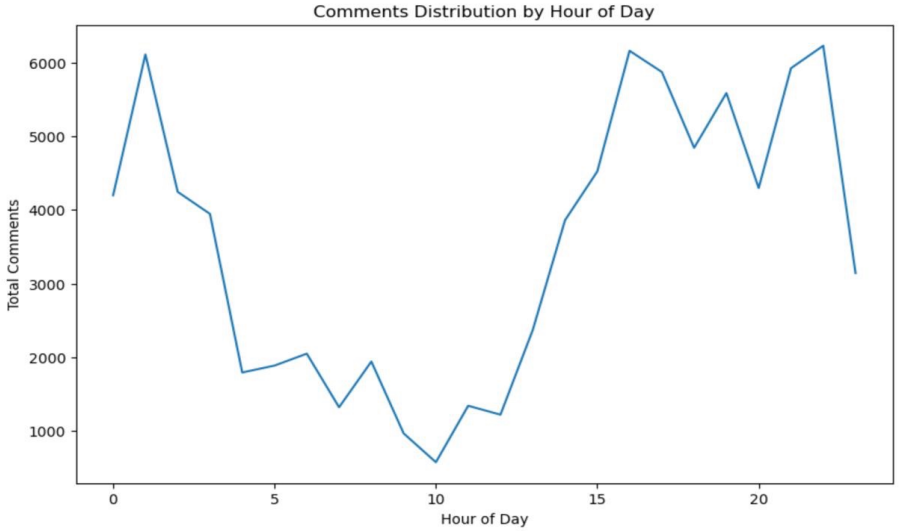
Used VADER to calculate sentiment scores for your text data.

then classify these scores into sentiment categories based on defined thresholds.

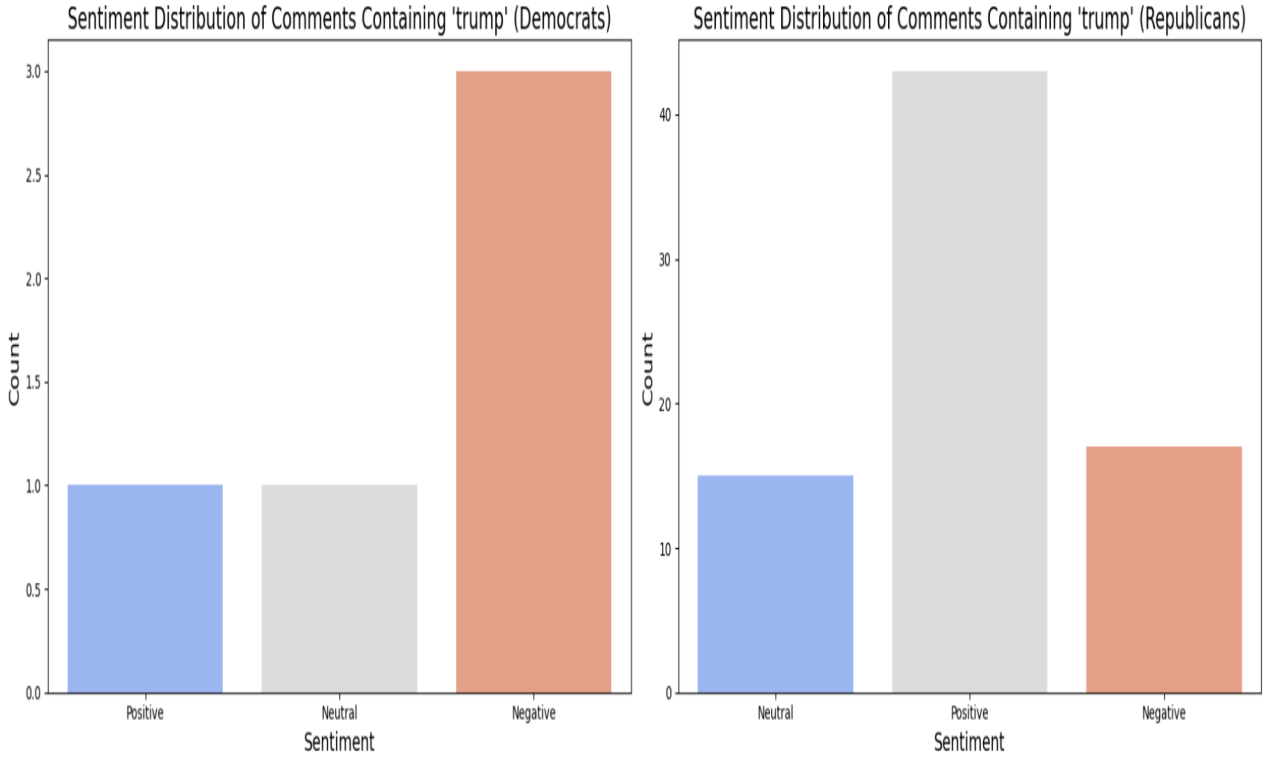


Time-Based Sentiment Trends:

- Aggregate sentiment scores by hours of the day
- plot these averages to visualize trends over time.

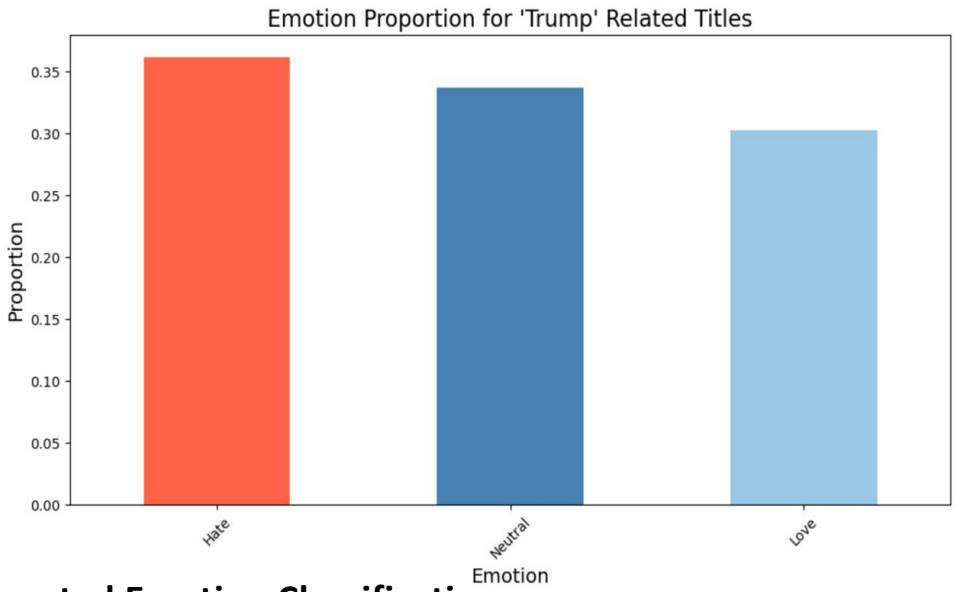


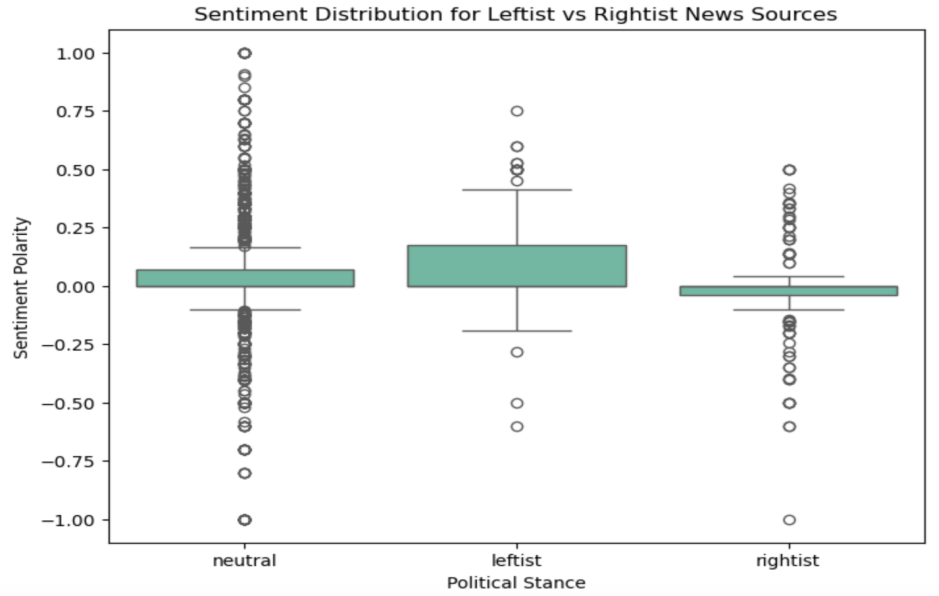
```
Most Frequent Word from Title (Democrats): trump
                                                  title  \
27   The "What Trump Has Done" sub reddit is lookin...
644  You can't reasonably hold an opinion about som...
660                                 Defend Democracy
680              We should just go silent for a while
709  I say to the Democrats "do everything you can ...

                                                   body sentiment
27   I have spent the last 16 years documenting wha...  Positive
644  DISCLAIMER: I edited the bottom image by chang...   Neutral
660  The Media failed US, they sane wash Trump, to ...  Negative
680  Hear me out. With Trump going hog wild as a Pr...  Negative
709  If he's trying extend an olive 🫒 branch, don't...  Negative

Most Frequent Word from Title (Republicans): trump
                                                 title  \
4    Donald Trump’s PRESIDENTIAL ACTIONS, Executive...
19   "President Trump and first lady Melania Trump ...
29   You have to give Trump credit. He is doing exa...
33   Trump pardons 1,500 Capitol rioters as he sign...
43   If on inauguration day, you spent your time tr...
44   So did the liberals once again get petty and s...
47               Billionaires at Trump’s inauguration
52   Trump to rename Gulf of Mexico, Mount Denali o...
53               Why are there so many libs on Reddit?
56   Trump's day so far: Tea with Biden, a Melania ...
```



Temporal Mood Analysis

What Emotional Forces Shape the Democrat vs. Republican Narrative?





Targeted Emotion Classification:

- Loaded our dataset and preprocessed the text.
- Then use CountVectorizer to transform the text into numerical features.
- Train a classifier (we chose Naive Bayes) on these features to predict emotions.

NLP for Public Opinion:

- Initialize VADER sentiment analysis.
- apply it to our opinion text to obtain sentiment scores.
- Classify these scores into positive, negative, or neutral categories.

How Do Leftist and Rightist News Sources React to the Same Political Scandal?

Key Insight 3 : Word Cloud

To visualize the most frequently used words in both subreddits, our first idea was to create a word cloud using the **titles of posts**. The process involved:

1. Combining all post titles into a single text.

2. Tokenizing the text and filtering out non-alphabetic words and custom stopwords.

3. Computing the word frequencies to identify the top 30 most common words.

The resulting word cloud highlights the dominant themes and topics discussed in the subreddit, excluding generic stopwords to focus on meaningful terms.

```
{"like", "just", "one", "get", "harris","trump",
 "kamala","donald","biden","joe","says","said","let"
,"go","tim","would","could","time","think","look",'see'
, "line",'today',"people","first","much",'day','us','par
'make','new','ever','voted','president',"election",'vote
'take','need','voter','votes','voted','voting','voters',
'say','real','gets','get','got','really',
 'house', 'put','days','still',
 "years","another",'other','breaking',
 "back",'going','states','state',
'democrats','democrat','republicans','republican'})
```

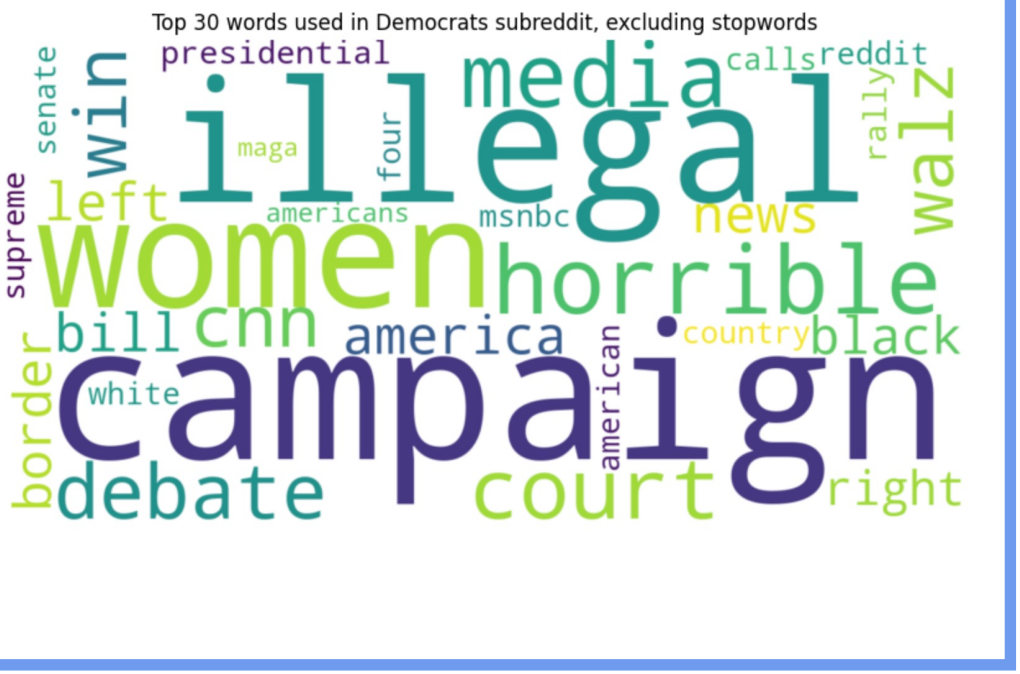
Stopwords we used

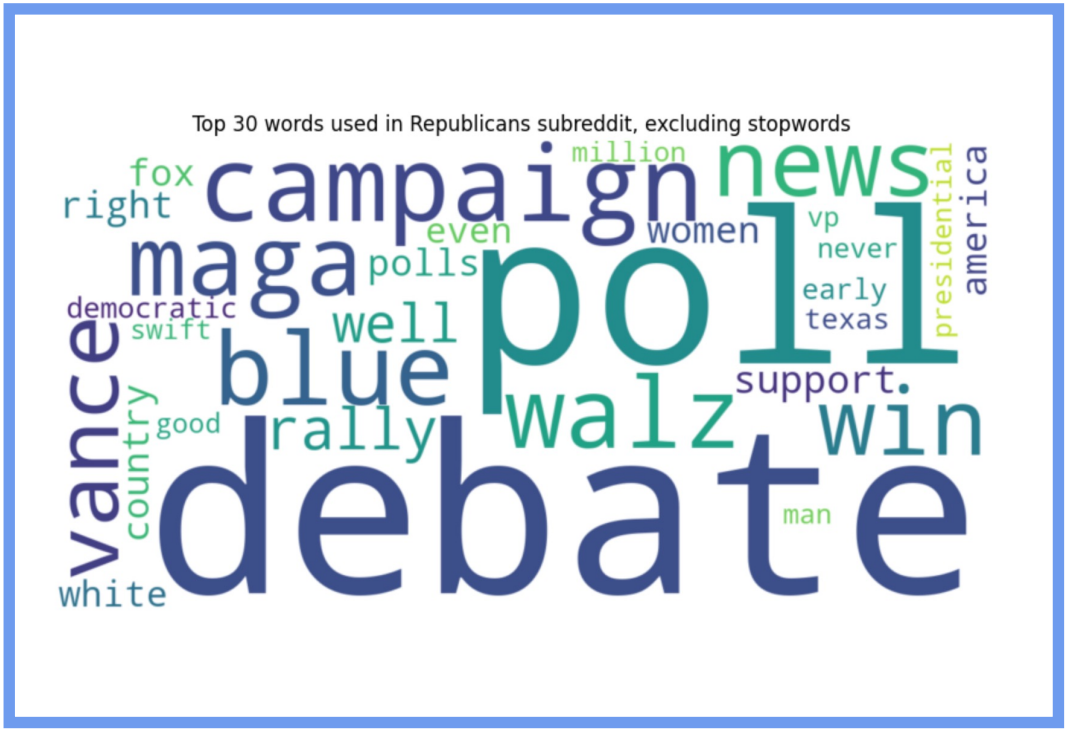
To generate meaningful word clouds, we excluded irrelevant words using a custom stopwords list. This list included:

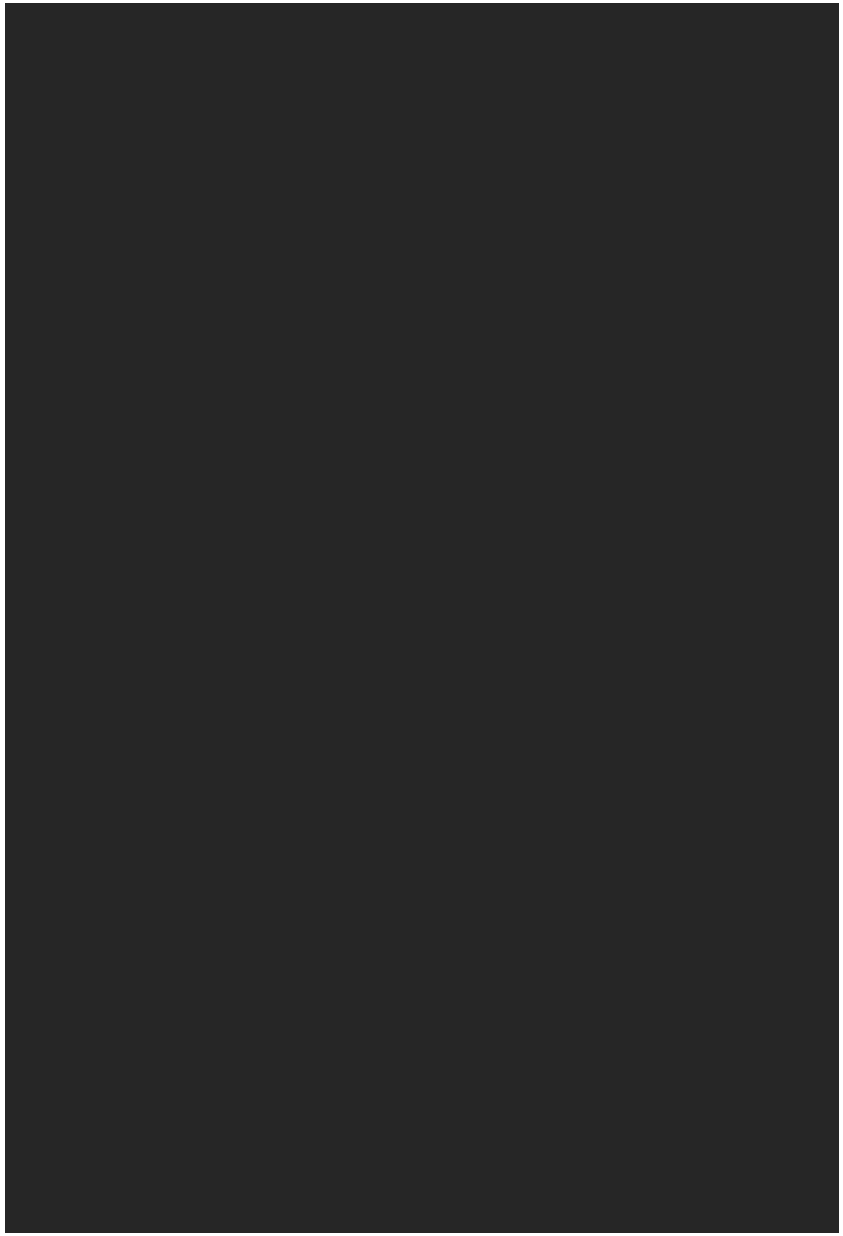
- Common English stopwords provided by the NLTK library.

- Additional domain-specific words frequently used in both subreddits (e.g., "Biden", "Trump", "election").

- Punctuation and generic terms that do not add value to the analysis (e.g., "like", "just").


Top 30 words used in Democrats subreddit, excluding stopwords
presidential
illegal
media
win
calls
reddit
senate
maga
four
rally
walz
left
americans
msnbc
news
supreme
women
horrible
bill
cnn
america
country
black
border
white
campaign
american
debate
court
right


Top 30 words used in Republicans subreddit, excluding stopwords
million
fox
campaign
news
right
even
women
vp
maga
polls
never
poll
democratic
swift
well
early
texas
presidential
america
blue
support.
vance
rally
walz
win
country
good
debate
man
white



Focus on Key Issues:

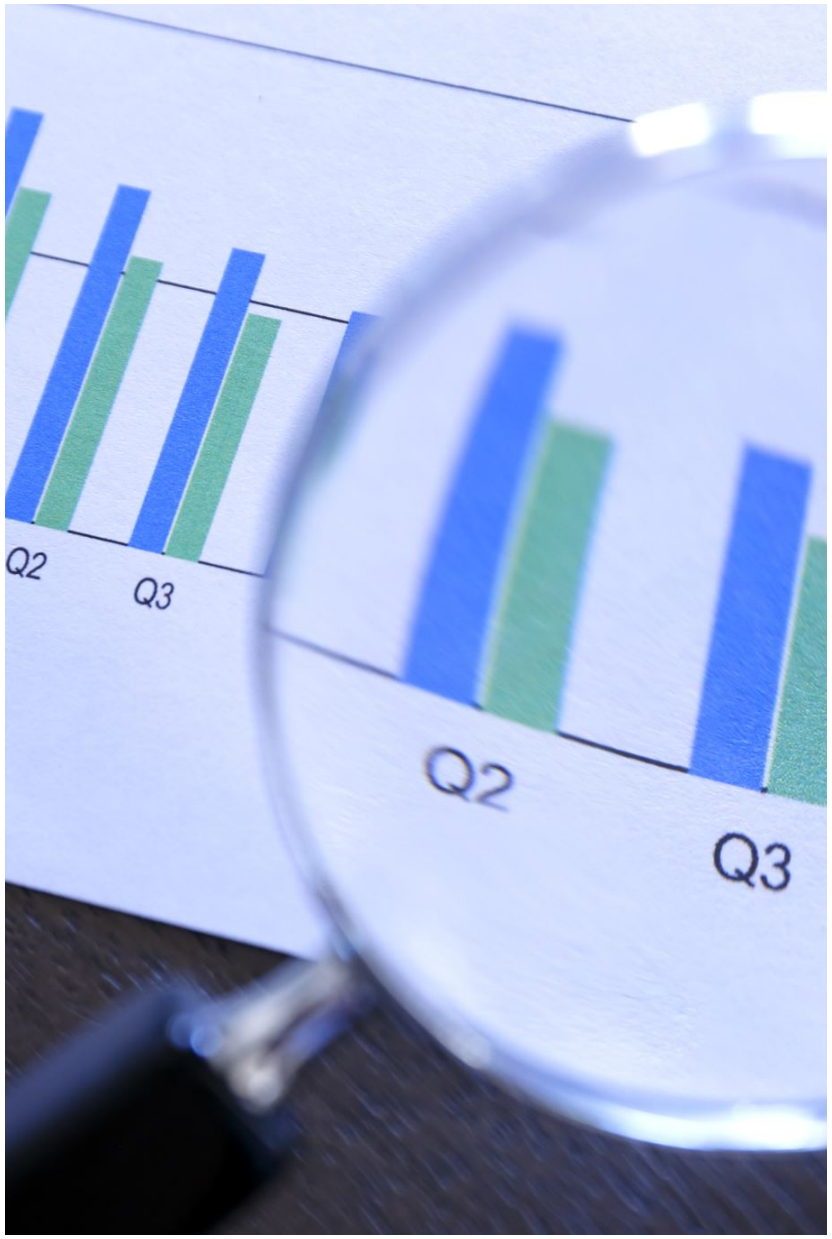
The *Republicans* word cloud emphasizes terms like "poll," "maga," and "debate," highlighting political strategy and campaign events, while the *Democrats* word cloud includes terms like "illegal," "horrible," and "court," reflecting a focus on criticism of policies and legal issues.

Language and Sentiment:

Republicans use more proactive and optimistic terms like "campaign," "rally," and "win," while Democrats lean towards negative terms like "horrible" and "illegal," reflecting disappointment after the election results.

Distinct Themes:

The Republican word cloud includes partisan identity markers like "maga" and "blue," whereas the Democratic word cloud features broader concerns such as "media," "women," and "court," potentially addressing election related issues (or Trump related).



Key Insight 4 : Building Classifier

- We utilized **TF-IDF (Term Frequency-Inverse Document Frequency)** to convert subreddit comments into numerical features, focusing on the top 5000 terms while excluding common stopwords.

- We split the data into training and testing sets to evaluate the model's performance.

- A **Logistic Regression** classifier was trained using the TF-IDF-transformed data to predict subreddit labels.

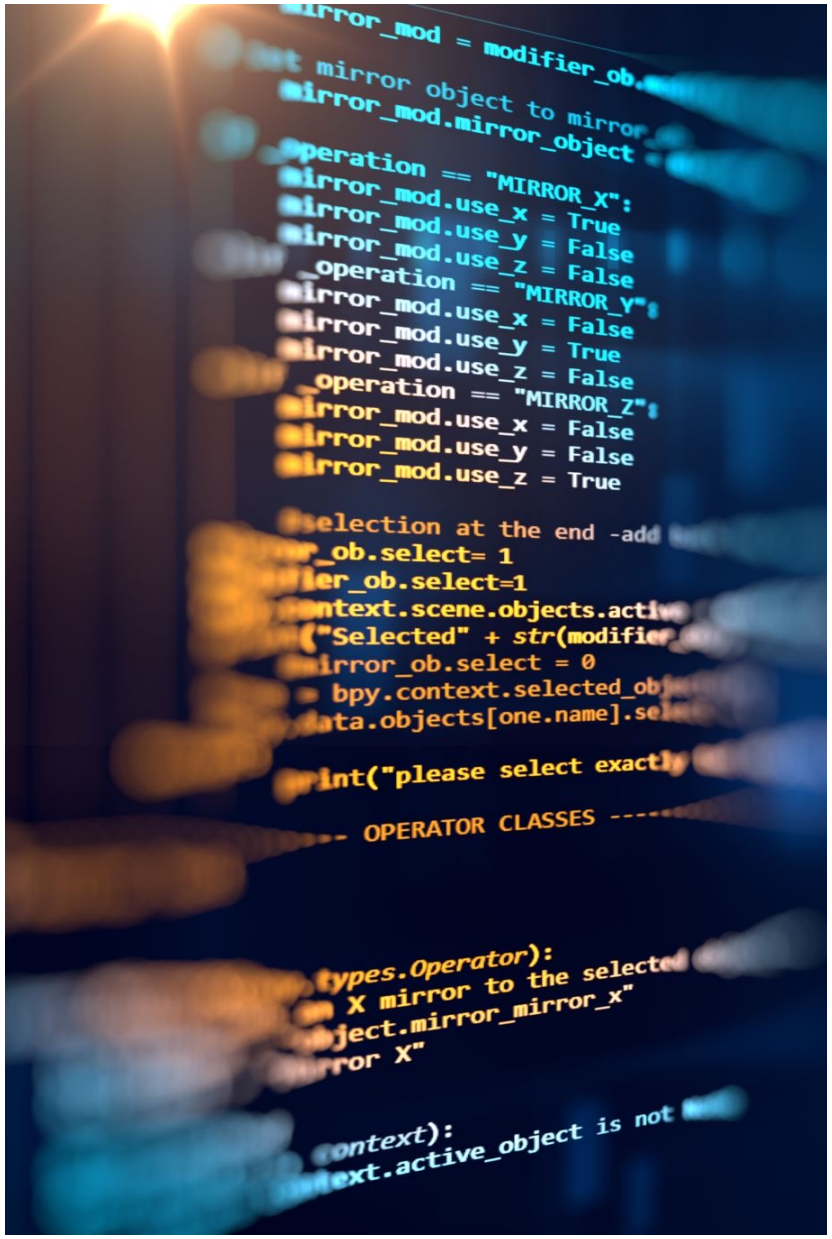
```python
# Train the Logistic Regression model with the best hyperparameters on the training set
best_model = grid_search_lr.best_estimator_

# Use the trained model to make predictions on the test set
y_pred = best_model.predict(X_test)

# Evaluate the model's performance on the test set
from sklearn.metrics import accuracy_score
test_accuracy = accuracy_score(y_test, y_pred)
print("Test Accuracy with Best Hyperparameters:", test_accuracy)
```

```
Test Accuracy with Best Hyperparameters: 0.6866478379431242
```

Implementing Deep Learning

Logistic Regression provided baseline performance but it could get better.

To address this, we adopted deep learning to leverage its ability to model sequential dependencies in textual data.



Why LSTM?

Captures Long-Term Dependencies:

LSTMs excel at processing sequential data by using gates to decide what information to remember, update, or forget at each step.

Bidirectional LSTM:

Processes text both forward and backward, capturing richer context from surrounding words.

Example:

•In the sentence *"The movie was not good"*, the LSTM's forget gate discards the initial expectation of positivity from "was" and updates its understanding based on "not," recognizing the negation and correctly identifying the sentiment as negative.

```python
# Train the model
history = model.fit(
    X_train, y_train,
    validation_split=0.2,
    epochs=5,  # Start with 5 epochs
    batch_size=32,
    verbose=1
)
```

```
Epoch 1/5
2054/2054 ━━━━━━━━━━━━━━━━━━━━ 135s 64ms/step - accuracy: 0.5094 - loss: 0.6921 - val_accuracy: 0.5460 - val_loss:
0.6892
Epoch 2/5
2054/2054 ━━━━━━━━━━━━━━━━━━━━ 135s 66ms/step - accuracy: 0.5627 - loss: 0.6825 - val_accuracy: 0.6373 - val_loss:
0.6455
Epoch 3/5
2054/2054 ━━━━━━━━━━━━━━━━━━━━ 133s 65ms/step - accuracy: 0.6697 - loss: 0.6118 - val_accuracy: 0.6823 - val_loss:
0.5861
Epoch 4/5
2054/2054 ━━━━━━━━━━━━━━━━━━━━ 133s 65ms/step - accuracy: 0.7360 - loss: 0.5256 - val_accuracy: 0.6909 - val_loss:
0.5746
Epoch 5/5
2054/2054 ━━━━━━━━━━━━━━━━━━━━ 132s 64ms/step - accuracy: 0.7617 - loss: 0.4833 - val_accuracy: 0.6949 - val_loss:
0.5855
```

```python
# Predict the labels
predictions = model.predict(X_test)
predicted_classes = (predictions > 0.5).astype(int)

# Compare predictions to ground truth
from sklearn.metrics import classification_report
print(classification_report(y_test, predicted_classes, target_names=['Republicans', 'Democrats']))
```

```
642/642 ━━━━━━━━━━━━━━━━━━━━ 10s 15ms/step
              precision    recall  f1-score   support

 Republicans       0.71      0.65      0.68     10240
   Democrats       0.68      0.74      0.71     10296

    accuracy                           0.69     20536
   macro avg       0.69      0.69      0.69     20536
weighted avg       0.69      0.69      0.69     20536
```

Result:

Higher Recall for Democrats:

Recall of 0.74 for Democrats indicates fewer false negatives, making Democrat comments easier to classify.

Performance Improvement:

Accuracy increased from Logistic Regression by 1%, small but still showed better classification performance overall.

Room for Improvement:

Further tuning (e.g., hyperparameters, feature engineering, pretrained embeddings) could enhance results.

Model Bias:

Democrats' comments may have more distinctive linguistic patterns, while Republican comments require deeper analysis.

Custom comments are tested

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 27ms/step
Comment: I strongly believe in equal rights and social justice.
Predicted Label: Democrats

Comment: Taxes are too high, and the government is inefficient.
Predicted Label: Republicans

Comment: We need to focus on climate change before it's too late.
Predicted Label: Democrats

Comment: Trump is the worst human being in the world.
Predicted Label: Democrats
```

Guess it decently works..

Conclusion: Do Results Align with the Motivation to Identify Signs of Polarization?

News Source Polarization: The news sources referenced by each subreddit are highly polarized, with very few shared sources. This reflects **polarized media credibility** and the existence of **echo chambers**.

Sentiment Analysis: Some differences in sentiment were observed between subreddits, suggesting **ideological divergence** in their reactions to similar concepts.

Word Cloud: The word clouds highlighted distinct topics and focus areas in each subreddit, **visualizing the thematic polarization** between them.

Classifier Performance: Despite initial assumptions that the polarized language would make classification easier, the task was more challenging than expected. While the classifier achieved decent accuracy, it may not be reliable for real-world use.

References

https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided

Ross Arguedas, A., Robertson, C., Fletcher, R., & Nielsen, R. (2022). Echo chambers, filter bubbles, and polarisation: A literature review.


thank
you!