Sentiment Analysis of Fake News Social Media Platforms

Team

Divya Vemula
Mani Krishna Tippani
Rahul Chauhan
Shaheryar Nadeem

Objective:

The spread of fake news undermines public trust and influences critical decisions in politics, finance, and society.

- This project aims to develop a machine learning model to detect fake news using the ***fake_or_real_news.csv*** dataset from Kaggle.
- By analyzing linguistic patterns and structural differences, we apply natural language processing (NLP) techniques and ML algorithms to automate fake news detection.
- The findings can enhance fact-checking systems, aiding media organizations and social platforms in mitigating misinformation.

Approach:

- Data collection, cleaning and feature extraction
- EDA with visualizations (word clouds, density plots)
- Text preprocessing (tokenization, stop word removal, lemmatization)
- Topic Modeling
- Sentiment Analysis (Text Blob)
- Classification Modeling & Evaluation

Dataset Overview:

- 20,800 news articles (20,800 rows and 5 columns)
- **Source:** Kaggle – Fake vs. Real News dataset
- **Attributes:** id, title, author, text, label

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1

label	proportion
1	50.0625
0	49.9375

Label Details:

- 50.06% Fake News (label = 0)
- 49.93% Real News (label = 1)

Data Cleaning & Preprocessing

Handling Missing Values:

- Dropped rows → 'id', 'title' or 'text' missing → To Ensure complete articles remainand maintain data integrity.
- 'author' → replaced with 'Unknown' → To Preserves all articles.
- 'label' → replaced with most frequent label → Prevents data loss while maintaining label distribution.

Removing Duplicates:

- Removed duplicates based on 'title' and 'text' to ensure unique news articles remain.

NLP Techniques

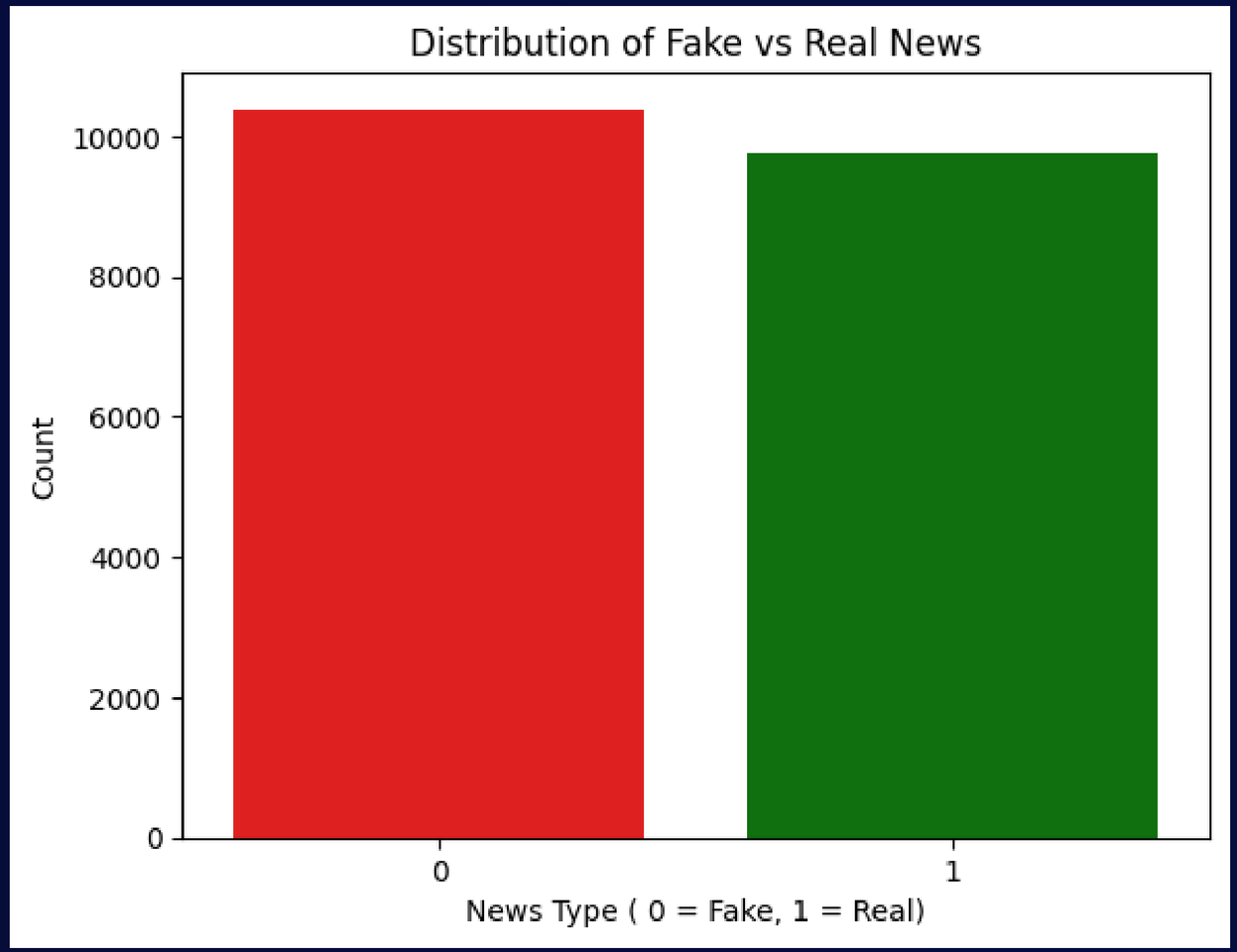
- Tokenization, Stop Word Removal, Lemmatization, Parts-Of-Speech tagging, Text Lowering

Data After Cleaning & Preprocessing

```
[ ] news_df.head()
```

	id	title	author	text	label	word_count	unique_word_count	stop_word_count	url_count	mean_word_length	char_count	cleaned_text
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1	820	432	356	0	5.001220	4930	house dem aide didnt even see comeys letter ja...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	710	417	310	0	4.836620	4160	ever get feel life circle roundabout rather he...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	1266	659	536	0	5.059242	7692	truth might get fire october tension intellige...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	557	277	236	0	4.788151	3237	video civilian kill single u airstrike identif...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	154	102	59	0	5.071429	938	print iranian woman sentence six year prison i...

20,800 news articles → 20133 after pre-processing



- A relatively balanced distribution of real and fake news articles.
- Fake news slightly outnumbers real news over 10,000.
- Real news around 9800 of the total dataset.

•**Strong Correlations:**

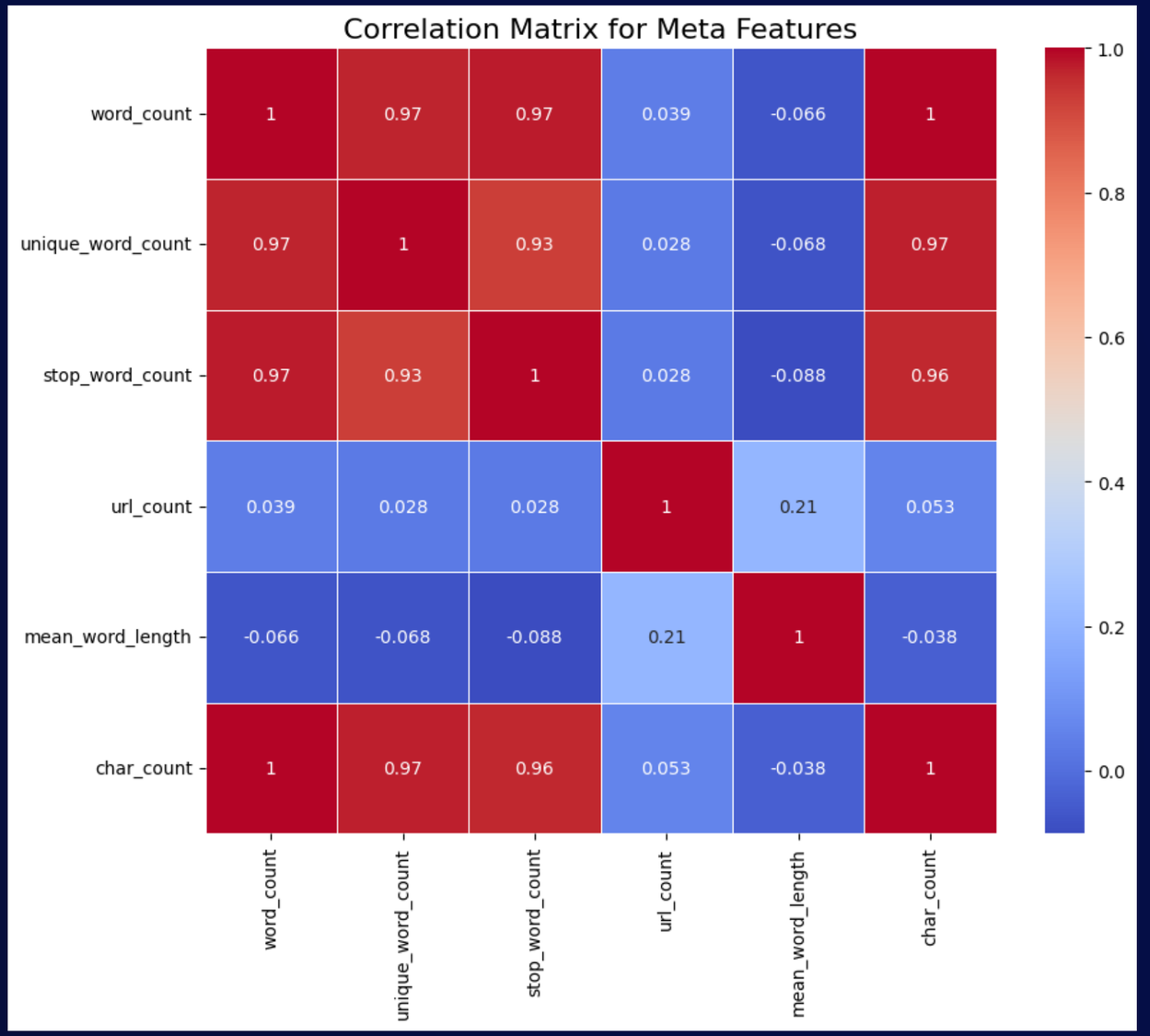
- **word_count, unique_word_count, stop_word_count, and char_count** are highly correlated, meaning longer texts naturally have more unique words, stopwords, and characters.

•**Weak Correlations:**

- **url_count** has little correlation with other features, suggesting URLs are independent of text length.
- **mean_word_length** has a slight negative correlation with word_count and stop_word_count, indicating longer texts tend to have shorter average word lengths.

•**Why It Matters:**

- Highly correlated features may be redundant for machine learning models.
- **url_count** could be a key independent feature in distinguishing real vs. fake news.
- Understanding these relationships improves feature selection and model efficiency.



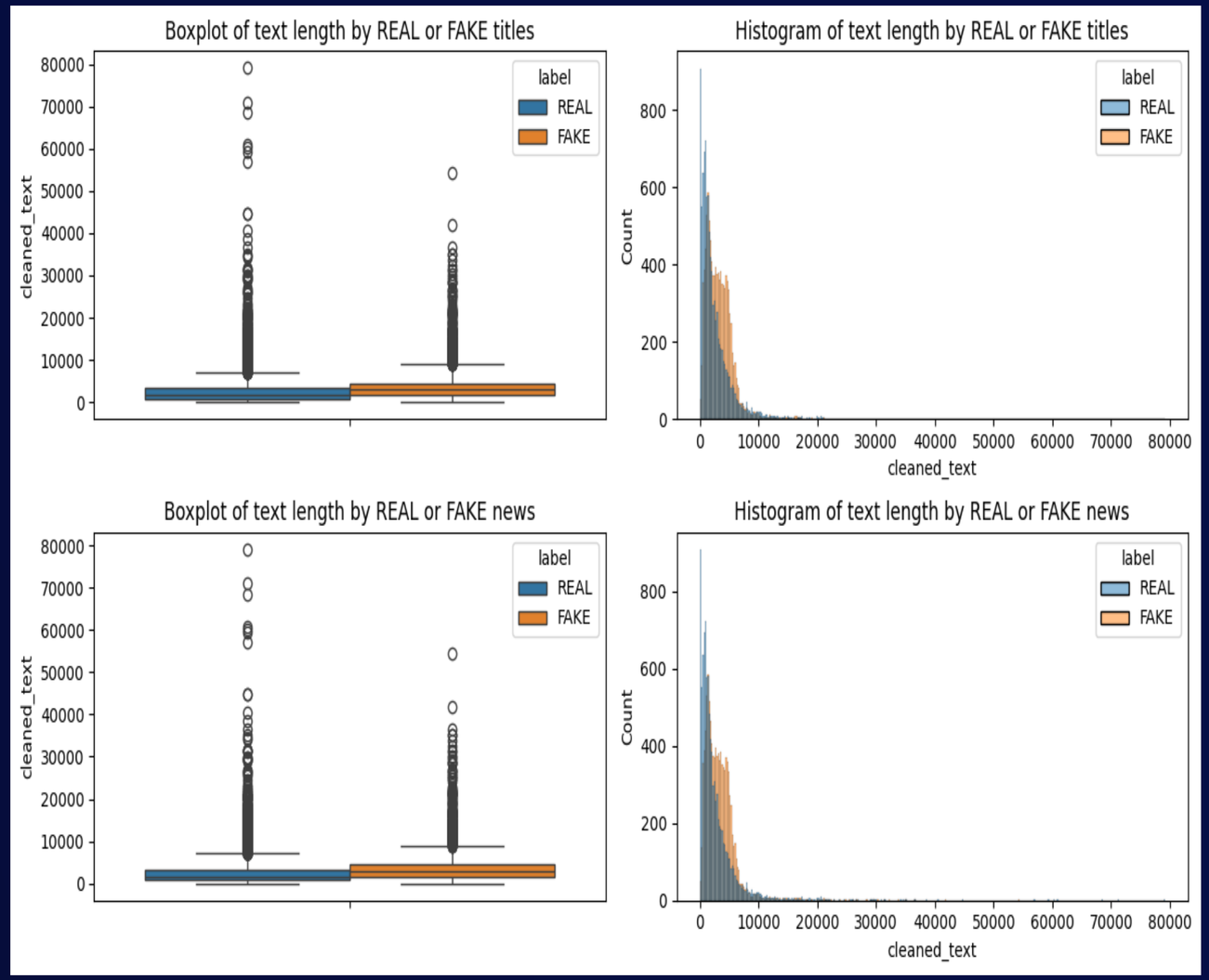
Analysis of Outliers in Real and Fake News

Boxplots (Left Side):

1. The average length of real and fake news articles is quite similar.
2. Fake news articles tend to be slightly shorter.
3. Some articles (both real and fake) are much longer than usual, which appear as outliers.

Histograms (Right Side):

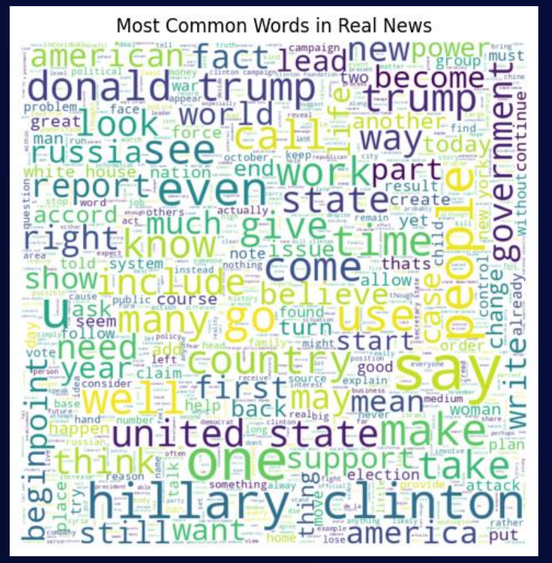
1. Most articles, whether real or fake, are relatively short.
2. A few articles are significantly longer, but they are rare.
3. The overall pattern of text length is similar for both real and fake news, meaning text length alone is not a reliable way to tell them apart.



Analysis of Common Words in Real and Fake News

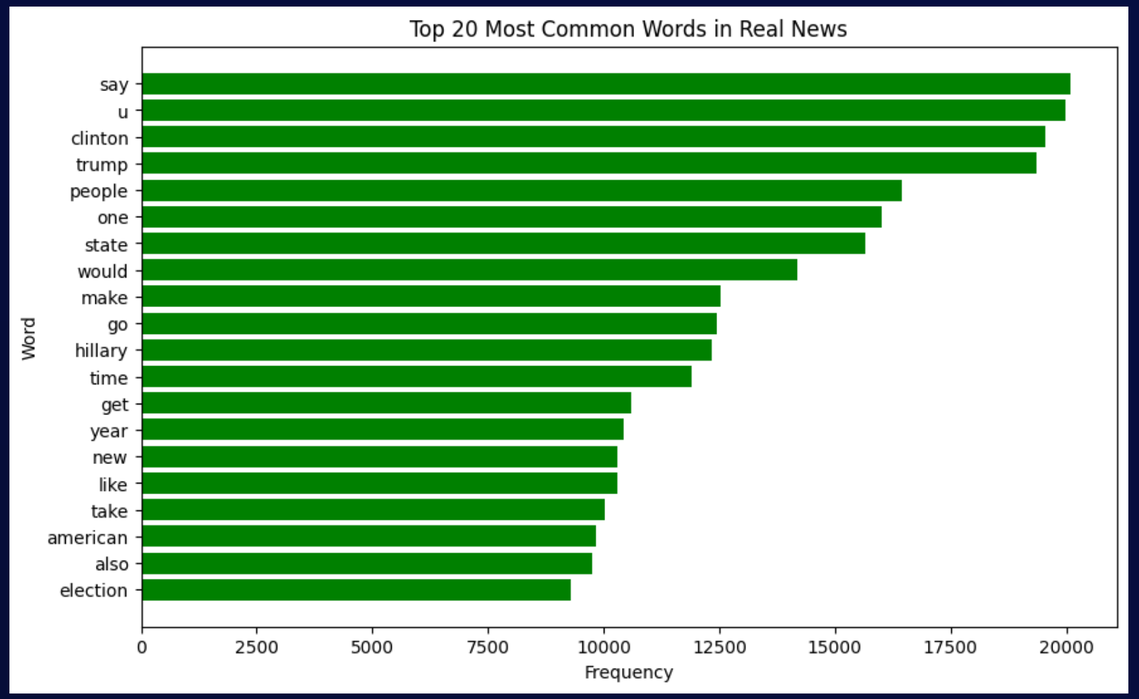
Fake News Observations

1. Politics: The presence of words like "Trump," "Clinton," "government," "president," "country," "state," "democrat," "republican" suggests a significant portion of fake news revolves around political topics.
2. Sensational Language: Words like "attack," "issue," "group," "support," "report," "call," "lead," "work" indicate a tendency to focus on controversial or attention-grabbing topics.
3. Entities and Locations: Terms like "united states," "new york," "white house" suggest a focus on U.S. political events.



Real News observation

1. **Politics: The presence of words like "Trump," "Clinton," "government," "state," "united," "country," "Russia," "support" suggests that real news articles extensively cover political topics, elections, and governance.**
2. **Neutral and Informational Tone: Words such as "report," "include," "accord," "fact," "explain," "write," "power" suggest that real news aims to present information rather than sensationalize.**
3. **General News Coverage: Terms like "people," "world," "many," "work," "way," "right" indicate discussions on societal and global issues beyond just politics.**

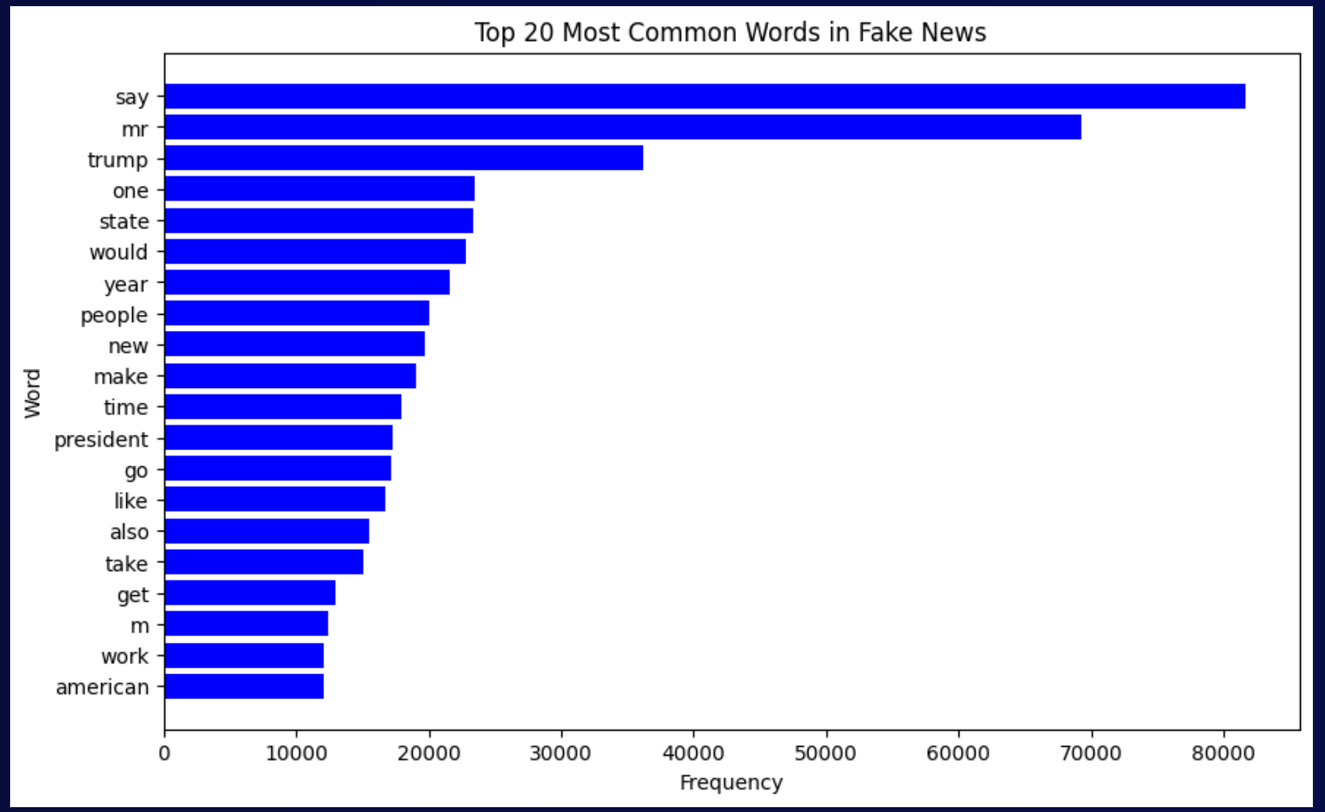


- **Both fake and real news share common words**, such as *say, trump, state, people, year, make, time*—indicating that basic vocabulary alone may not distinguish between them.

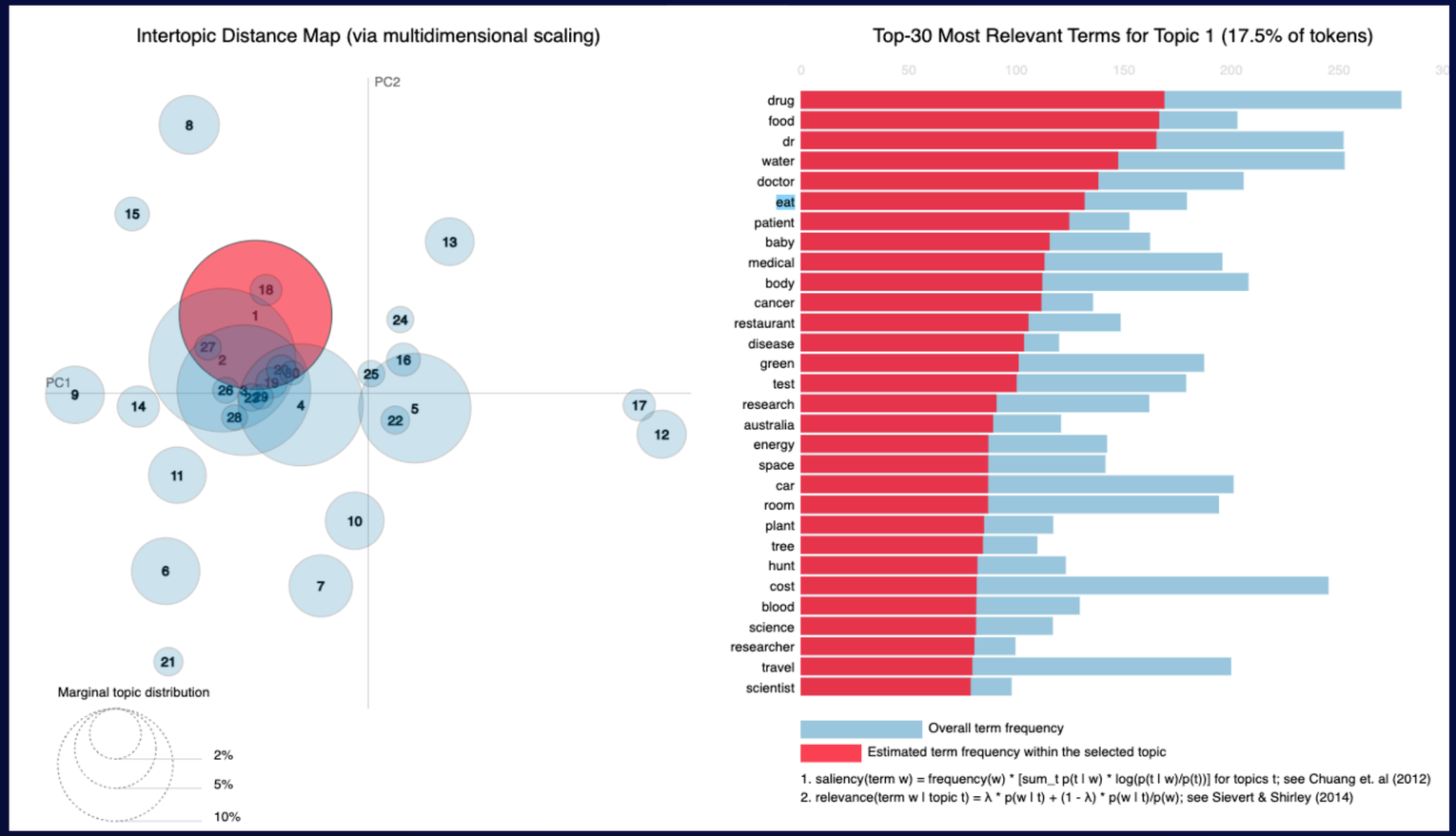
- **Fake news emphasizes “mr” and “president” more frequently**, suggesting a tendency to focus on political figures or authoritative personas.

- **Real news uses "Clinton" and "election" more often**, potentially reflecting more balanced coverage of political events.

- **Overall, the frequency of words in fake news is more extreme**, with *"say"* appearing significantly more than in real news—hinting at a possible reliance on quotes or hearsay to appear credible.



Topic Modeling With TF-IDF

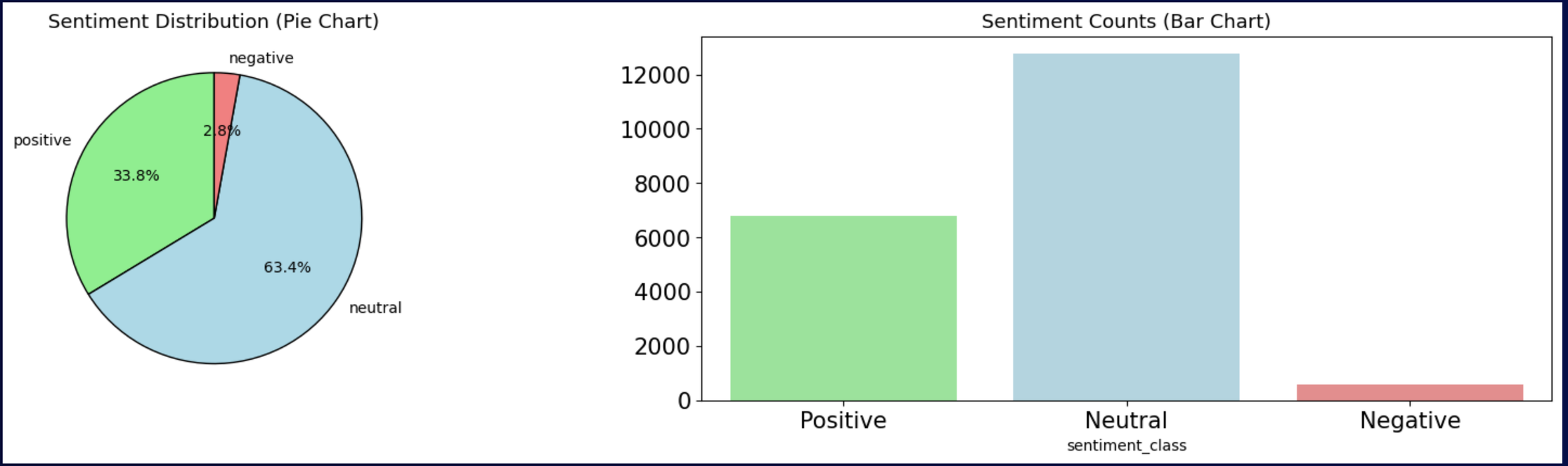


Topic Distribution (Left): Overlapping circles indicate related topics, with larger circles representing more dominant topics.

Key Terms (Right): Lists the most relevant words for *Topic 1*, where "drug," "food," and "doctor" appear most frequently.

Term Frequency: Red bars show term frequency within the topic, while blue bars represent overall dataset frequency.

Sentiment Analysis



Pie Chart (Left):

•The majority of articles (63.4%) have a *neutral* sentiment.
•33.8% of articles exhibit a *positive* sentiment.
•Only 2.8% of articles are classified as *negative*.

Bar Chart (Right)

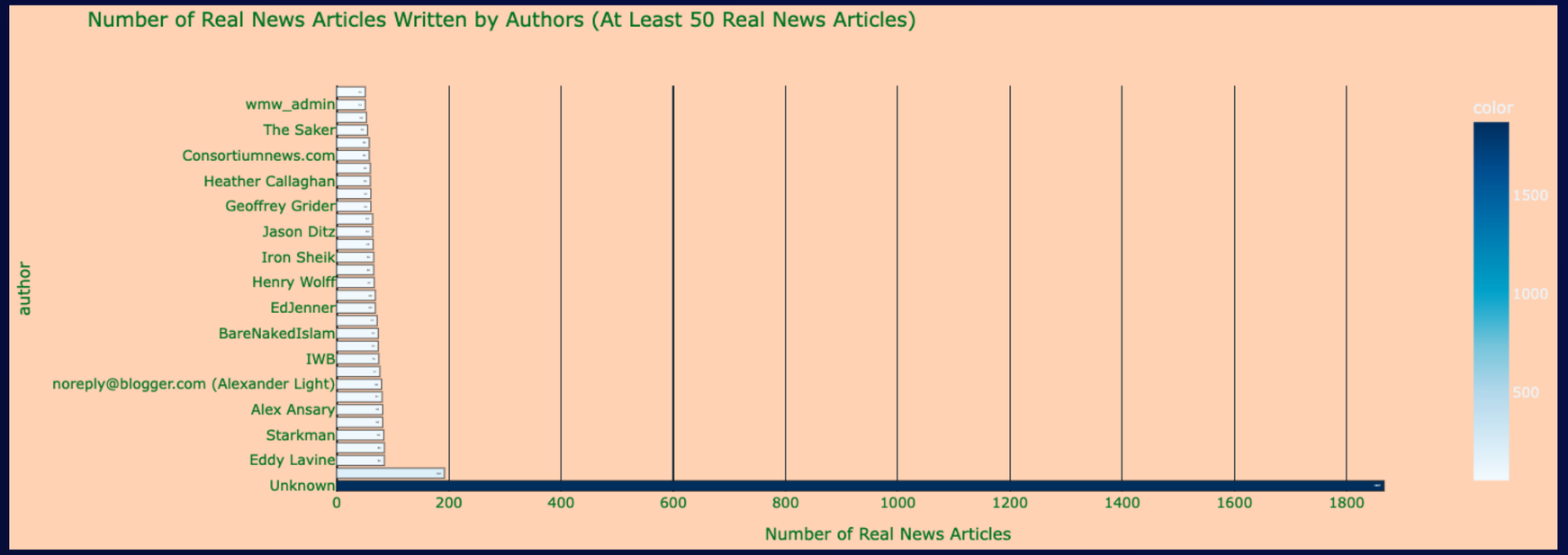
- Neutral sentiment articles is the highest, exceeding 12,000.
- Positive sentiment articles make up a significant portion, around 7,000.
- Negative sentiment articles are minimal in comparison.

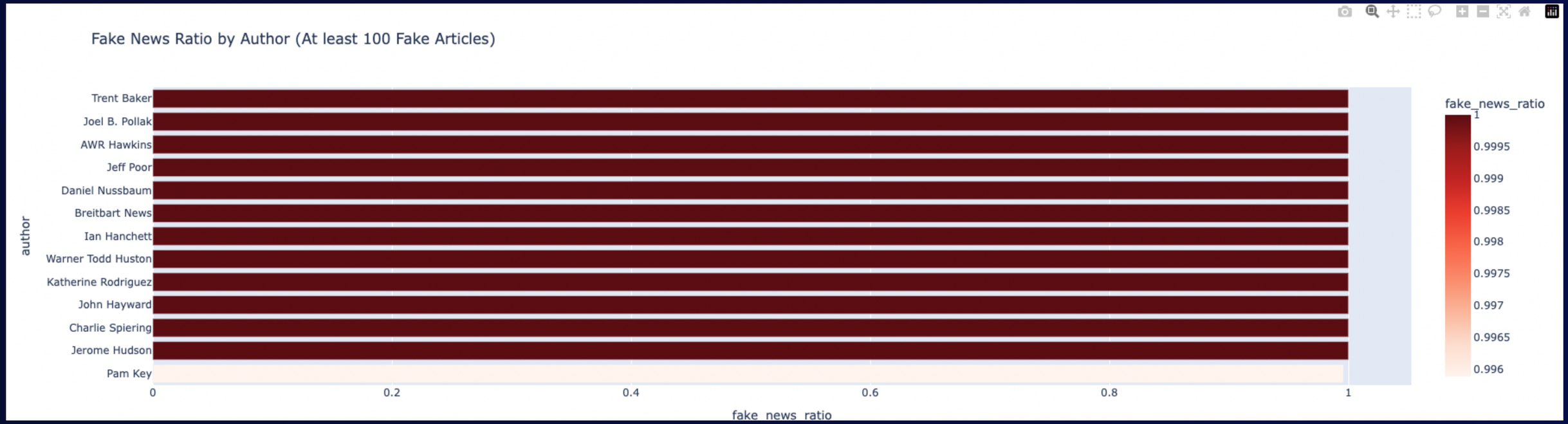
This suggests that most news articles in the dataset maintain a neutral tone, with a considerable portion leaning positive, while negative sentiment is relatively rare.

Distribution of Article by Authors

- This treemap visualizes the distribution of articles by authors who have written more than 100 articles.
- This suggests that a large portion of news lacks clear authorship, and article length varies significantly among contributors.

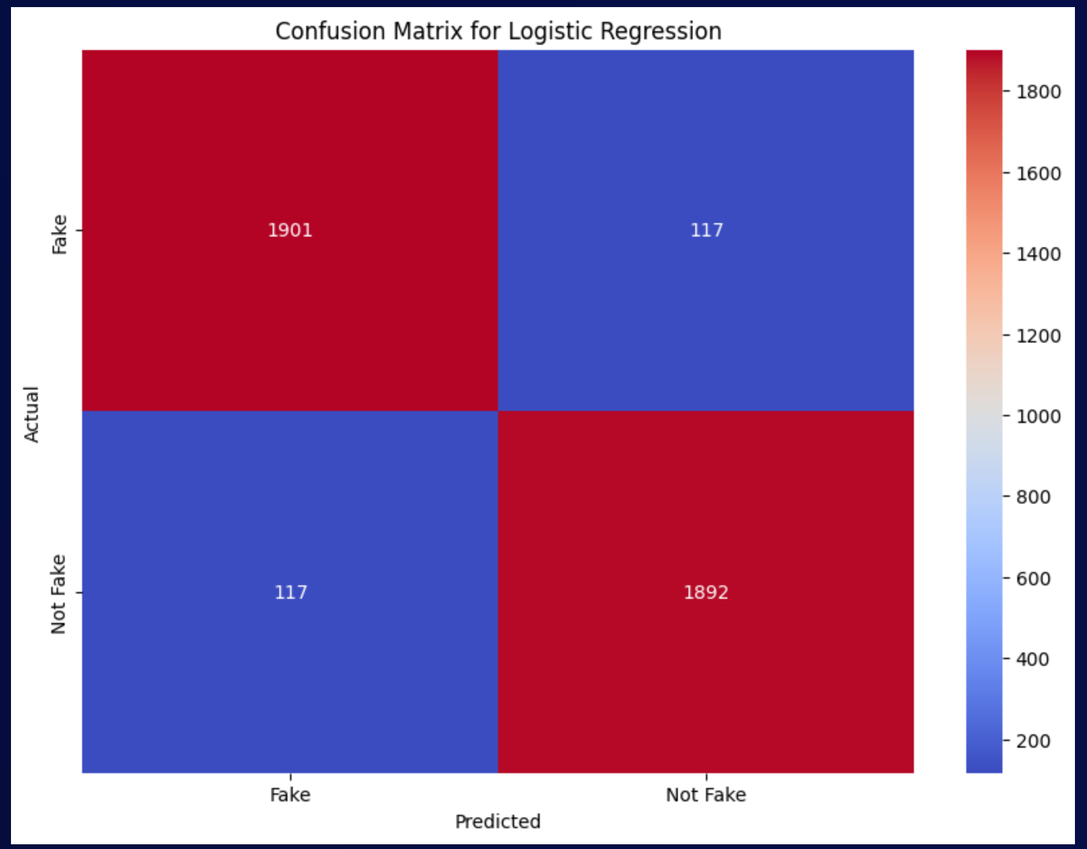
Comparision of Fake and Real News Articles





Model Building

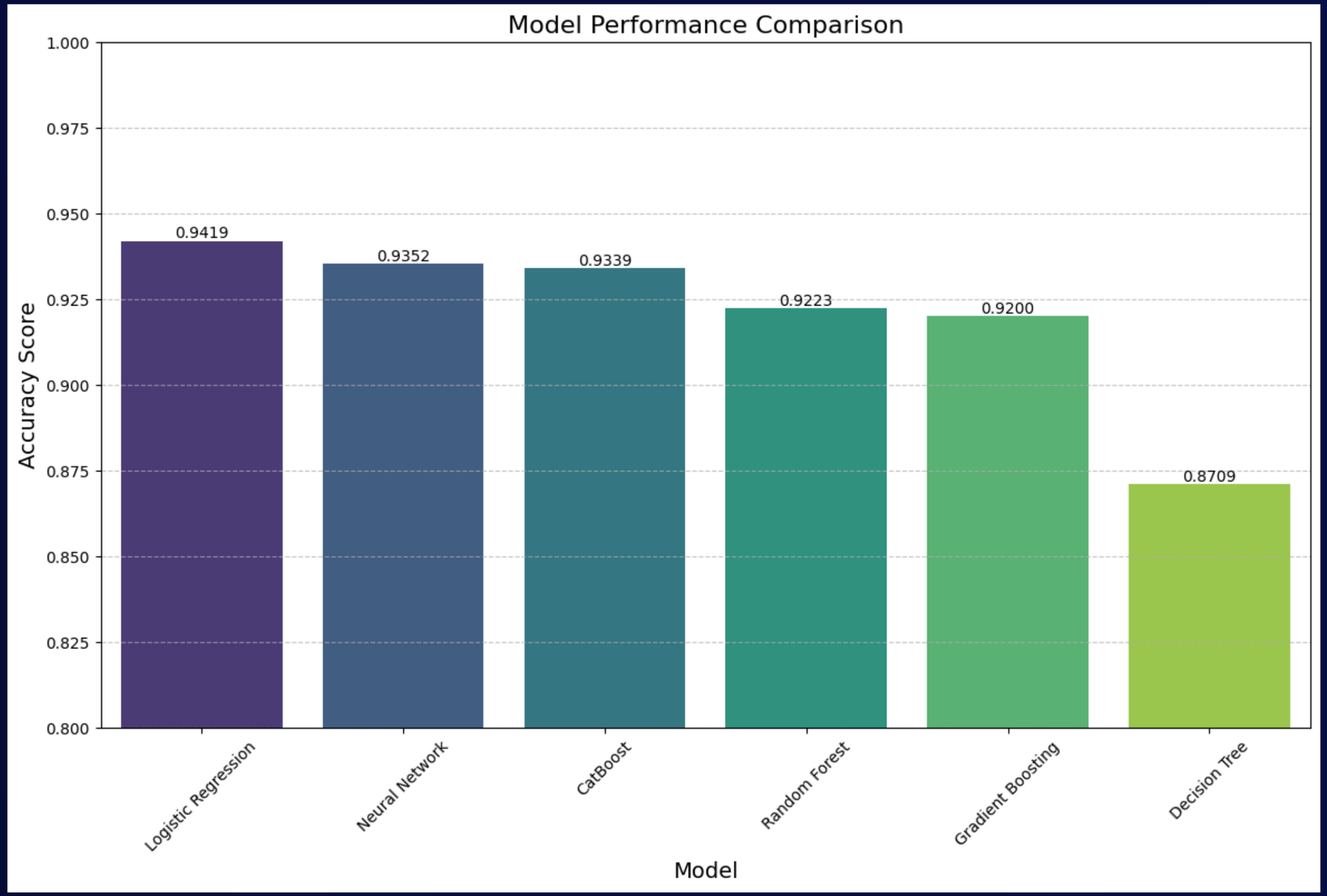
- **Train-Test Split:** 80% training, 20% testing
- **Models Explored:**
  - Logistic Regression (Best Model with 94% accuracy)
  - Random Forest
  - Gradient Boosting
  - Decision Tree
  - CatBoost
  - Neural Network (MLP)
- **Evaluation Metric:** Accuracy score, confusion matrix, classification report



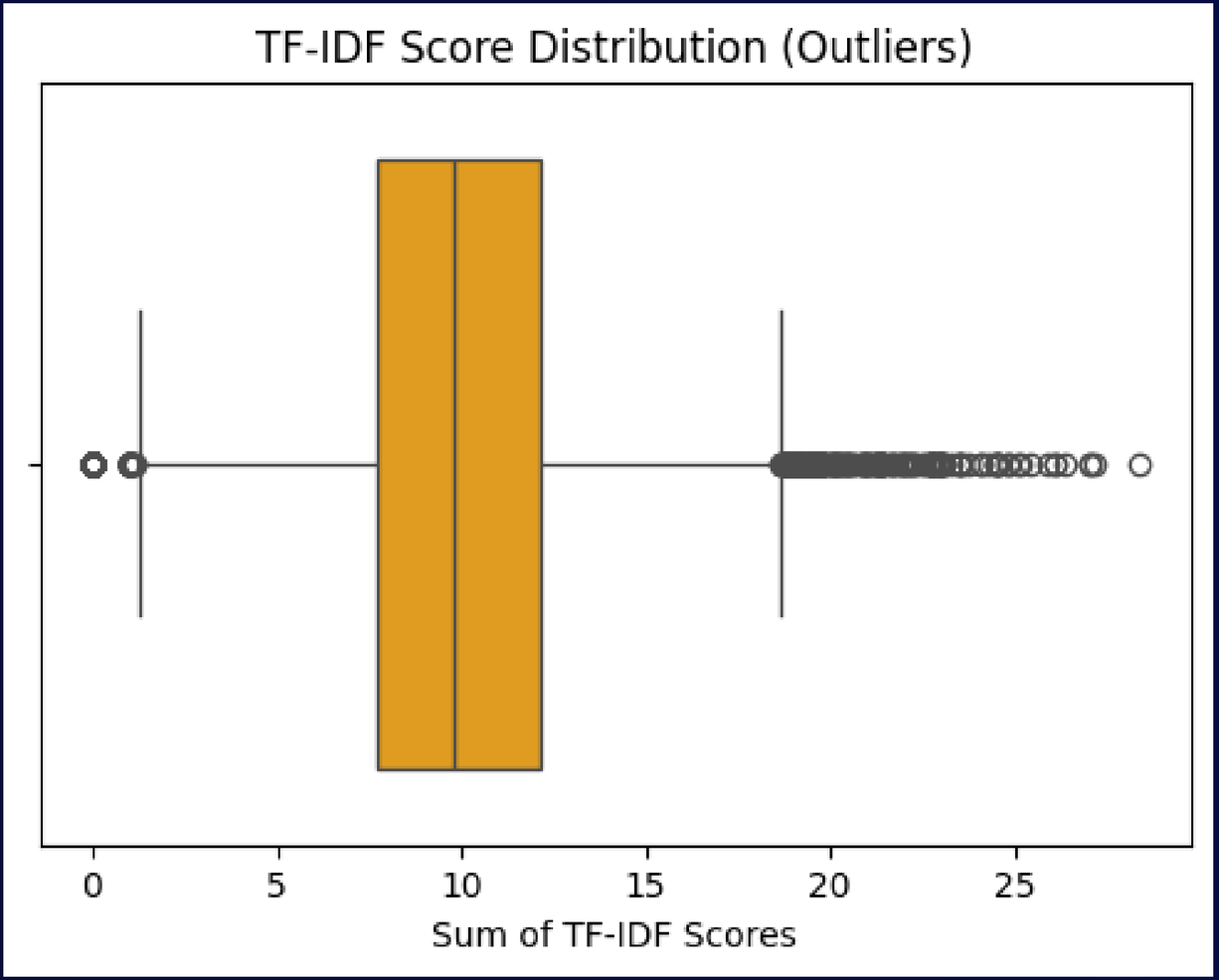
Classification Report for Logistic Regression:

	precision	recall	f1-score	support
Fake	0.94	0.94	0.94	2018
Not Fake	0.94	0.94	0.94	2009
accuracy			0.94	4027
macro avg	0.94	0.94	0.94	4027
weighted avg	0.94	0.94	0.94	4027

Analysis of Models Performance



Analyzing TF-IDF Score Distribution



A high number of outliers on the right side (higher TF-IDF sums) indicate that some documents contain a large number of unique or highly weighted words.

Interpreting Cosine Similarities

```
Cosine similarity between Fake and Real news: 0.8139112532952245
Cosine similarity between Positive and Negative sentiment: 0.8012404270384239
```

Key Takeaways:

- This relatively high similarity (0.81) suggests that fake and real news share a significant portion of their vocabulary and structure.

- The high similarity (0.80) implies that positive and negative sentiment texts also have substantial overlap in their word usage.

Overall, Fake and real news, as well as positive and negative sentiment texts, share a strong structural and linguistic foundation. This suggests that language models can capture meaningful patterns across both domains.

Conclusion

This project focused on developing a machine-learning model to detect fake news on social platforms, using techniques such as TF-IDF vectorization, Text Blob, topic modeling, and cosine similarity to enhance the detection accuracy. The top-performing model was Logistic Regression (94.19%). Sentiment analysis helped identify subtle differences between positive and negative sentiment, revealing shared vocabulary that could be further refined.

The computational efficiency of the machine-learning models makes them suitable for resource-constrained environments, providing a scalable solution for fake news detection. The model’s potential applications include use by organizations like the Anti-Fake News Center for faster detection.