SENTIMENT ANALYSIS
FAKE NEWS ON SOCIAL MEDIA PLATFORMS

Submitted by

GROUP – 1
MSIS BIG DATA ANALYTICS (FALL 2024)

DIVYA VEMULA
TIPPANI MANI KRISHNA
RAHUL CHAUHAN
SHAHERYAR NADEEM

Under the guidance of
Dr. LUSI YANG
Assistant Professor

ROBINSON SCHOOL OF BUSINESS

GEORGIA STATE UNIVERSITY

FALL 2024

Executive Summary

The rapid spread of fake news on social media has led to misinformation and social polarization, necessitating the development of advanced detection mechanisms. This project applies Natural Language Processing (NLP) and machine learning techniques to analyze and classify fake and real news articles, with a focus on sentiment analysis and predictive modeling.

A dataset of 20,800 news articles was sourced from Kaggle and underwent extensive preprocessing, including tokenization, stop-word removal, and lemmatization. Exploratory data analysis (EDA) revealed key linguistic patterns distinguishing fake and real news, such as the frequent use of sensationalist and politically charged terms in fake news articles. Sentiment analysis using TextBlob indicated that fake news tends to adopt a neutral or positive tone to enhance credibility.

Various machine learning models were trained and evaluated, including Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, CatBoost, and Neural Networks. Model performance was assessed using accuracy, precision, recall, and F1-score. Logistic Regression emerged as the most effective model, achieving the highest recall (94.2%) and accuracy (94.19%), making it the best for detecting fake news. CatBoost and Neural Networks also demonstrated strong performance, particularly in balanced classification.

The findings highlight the importance of integrating sentiment analysis with fake news detection models to improve accuracy and reliability. This approach can aid social media platforms and regulatory bodies in identifying and mitigating misinformation more effectively. The study also emphasizes the need for scalable, automated solutions to combat the growing challenge of fake news dissemination in digital media.

Table of Contents

1. Introduction

1.1 Background of Study

With the widespread use of social media and digital platforms, misinformation has become a significant societal concern. Fake news, designed to mislead and trigger strong emotional responses, is often used to influence public opinion. While many fake news detection models focus on distinguishing real from fake news, understanding the sentiment in fake news articles can offer deeper insights into their psychological impact. This study explores sentiment patterns in fake news and how they differ from real news.

1.2 Purpose of the Study

The study aims to explore:

- The differences in sentiment between fake and real news articles.
- The most frequently used words and themes in fake news articles.
- The role of topic modeling and sentiment analysis in understanding fake news narratives.
- The effectiveness of various machine learning models in classifying news articles.

1.3 Need for Recommendation for Sentiment Analysis

The increasing prevalence of fake news on social media has significant implications for public trust, political discourse, and social cohesion. While traditional fact-checking mechanisms can help identify misinformation, they are often time-consuming and difficult to scale. This study emphasizes the need for automated solutions that go beyond simple fake news classification to provide actionable recommendations.

Why is Recommendation Important?

- **Enhancing Fake News Detection**: Understanding the **sentiment behind fake news** can provide deeper insights into how misinformation manipulates public opinion.
- **Improving Content Moderation**: By identifying fake news articles that carry extreme emotional sentiments to improvise and clean content moderation for further review.
- **Protecting Financial & Political Integrity**: Fake news is commonly used for financial scams and political manipulation (e.g., misinformation campaigns during elections).
- **Preventing Misinformation Spread**: By identifying common linguistic patterns in fake news, social media platforms can design **automated interventions**, such as reducing the reach of misleading content or tagging articles with credibility indicators.

1.4 Aim of the Project

The primary goal of this project is to leverage natural language processing (NLP) and machine learning techniques to:

- Analyze the sentiment of fake news.
- Identify structural differences between fake and real news.
- Develop and evaluate machine learning models for fake news classification.

1.5 Objectives of the Project

- Conduct data cleaning and preprocessing.
- Perform exploratory data analysis (EDA) with visualizations.
- Implement topic modeling to identify dominant themes.
- Apply sentiment analysis using TextBlob.
- Train and evaluate various machine learning classifiers.
- Compare model performance and identify the best classifier.

2. Literature Study

2.1 Machine Learning Models and Process

Machine learning techniques are widely used for text classification tasks, including fake news detection. The process typically involves:

1. **Data Preprocessing:** Tokenization, stop-word removal, lemmatization, and vectorization using TF-IDF.
2. **Feature Extraction:** Extracting text-based features such as word count, character count, and URL frequency.
3. **Model Selection:** Training classifiers such as Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, CatBoost, and Neural Networks.
4. **Evaluation Metrics:** Accuracy, precision, recall, F1-score, and confusion matrices are used to assess model performance.

Machine Learning Models Explored

- **Logistic Regression:** A linear model for binary classification.
- **Random Forest:** An ensemble method using multiple decision trees.
- **Gradient Boosting:** A boosting technique that improves weak learners.
- **Decision Tree:** A tree-based model that makes predictions using feature splits.
- **CatBoost:** A gradient boosting model optimized for categorical features.
- **Neural Network (MLP):** A deep learning approach with multiple hidden layers.

2.2 Tools and Python Packages Used in the Analysis

The project utilized the following Python libraries:

- **Pandas & NumPy:** Data manipulation and numerical operations.
- **Matplotlib & Seaborn:** Data visualization.
- **Plotly & WordCloud:** Interactive graphs and word clouds.
- **NLTK & Gensim:** Text preprocessing and topic modeling.
- **TextBlob:** Sentiment analysis.
- **Scikit-learn:** Machine learning model implementation.
- **CatBoost:** Gradient boosting classifier.

3. Dataset Overview

3.1 Data Source

The dataset was obtained from Kaggle: [Detecting Fake News step by step]

- **Total Articles:** 20,800 news articles.
- **Attributes:** id, title, author, text, label.
- **Label Distribution:**
    - 50.06% Fake News (label = 0).
    - 49.93% Real News (label = 1).

label	proportion
1	50.0625
0	49.9375

3.2 Data Fields

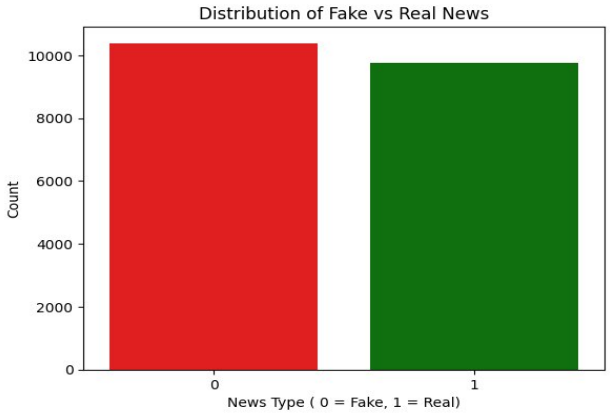
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      20800 non-null  int64 
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
```

- **id:** Unique identifier for each article.
- **title:** The headline, which may contain clues about sensationalism or bias.
- **author:** Information on the article's creator; missing values were imputed with "Unknown".
- **text:** The full content of the news article, critical for NLP analysis.
- **label:** A binary indicator where 0 represents fake news and 1 represents real news. The dataset is balanced with roughly 51.6% fake news and 48.4% real news.

	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

3.3 Data Cleaning and Preprocessing

1. Removed rows with missing values in 'id', 'title', and 'text' - To Ensure complete articles remain and maintain data integrity.
2. Replaced missing 'author' values with 'Unknown' - To Preserves all articles.
3. Replaced missing 'label' values with the most frequent label- Prevents data loss while maintaining label distribution.
4. Removed duplicate articles based on 'title' and 'text' - to ensure unique news articles remain.



4. Exploratory Data Analysis and Insights

4.1 Text Length Analysis with Boxplot and Histogram

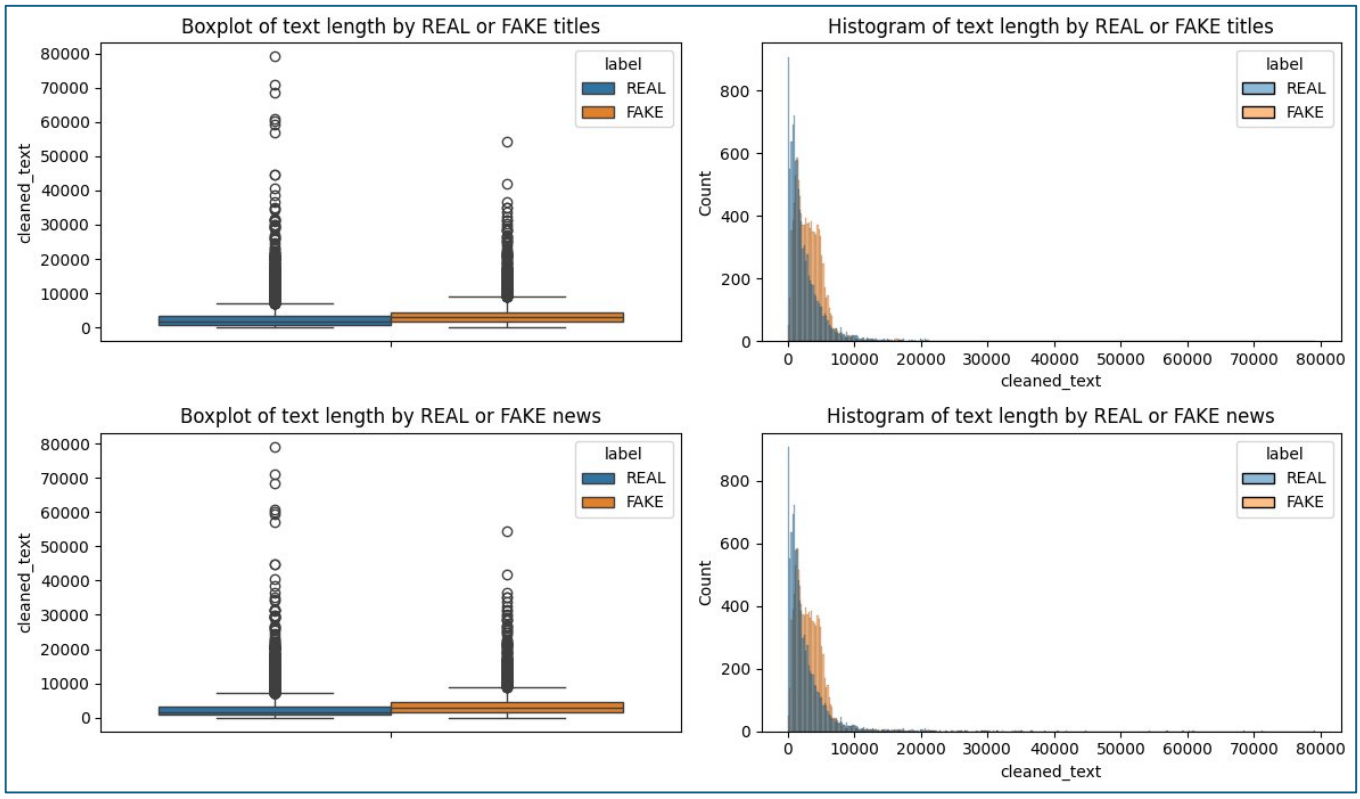
The boxplots indicate that the length of fake and real news titles, as well as full news articles, is quite similar. However, fake news articles tend to be slightly shorter on average. The histograms further support this observation by showing a right-skewed distribution, meaning that most news articles (both real and fake) have shorter text lengths, with a few significantly longer ones acting as outliers. This suggests that text length alone is not a strong indicator for differentiating real and fake news.

Boxplots (Left Side):

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

Histograms (Right Side):

- Most articles, whether real or fake, are relatively short.
- A few articles are significantly longer, but they are rare.
- 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.



4.2 Density Plots

The above chart consists of six density plots representing the distribution of various textual features for **real** and **fake news** articles. The **green** and **red** lines correspond to **fake** and **real** news, respectively. Below is an interpretation of each plot:

1. **Word Count Label Distribution (Top Left)**
    - Most of both **fake** and **real** news articles have relatively low word counts, with a right-skewed distribution.
    - There is significant overlap between fake and real news, indicating that word count alone may not be a strong differentiator.
    - Some articles (outliers) have a much higher word count, though they are rare.

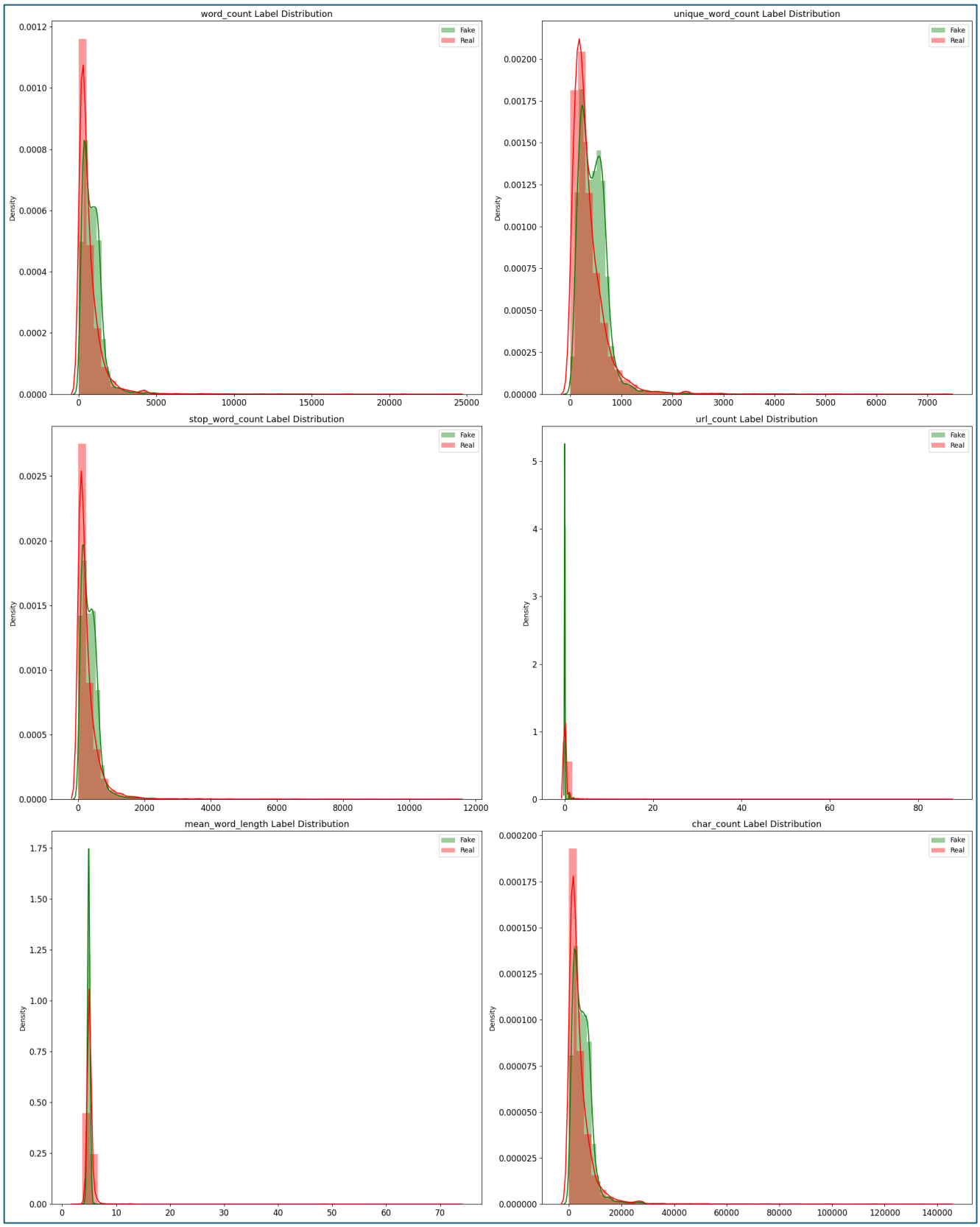
2. **Unique Word Count Label Distribution (Top Right)**
    - Like word count, unique word count follows a right-skewed distribution.
    - Fake news has slightly fewer unique words on average compared to real news, but the difference is minor.
    - A large portion of articles have relatively few unique words, meaning much of the text consists of repeated or common words.

3. **Stop Word Count Label Distribution (Middle Left)**
    - The distribution of stop words is also right-skewed, with most articles containing relatively few stop words.
    - The overall distribution of stop word count is similar for fake and real news, suggesting that stop word usage is not a significant distinguishing factor.

4. **URL Count Label Distribution (Middle Right)**
    - The plot indicates that most articles contain **zero or very few URLs**.
    - Some fake news articles appear to include slightly more URLs than real news, but the difference is minimal.
    - This suggests that while URL presence might play a role in distinguishing fake news, it is not a dominant factor.

5. **Mean Word Length Label Distribution (Bottom Left)**
    - Both real and fake news articles peak around **4-6 characters per word**, showing little distinction.
    - There is some variation, but the distributions are nearly identical, implying that mean word length is not a strong feature for classification.

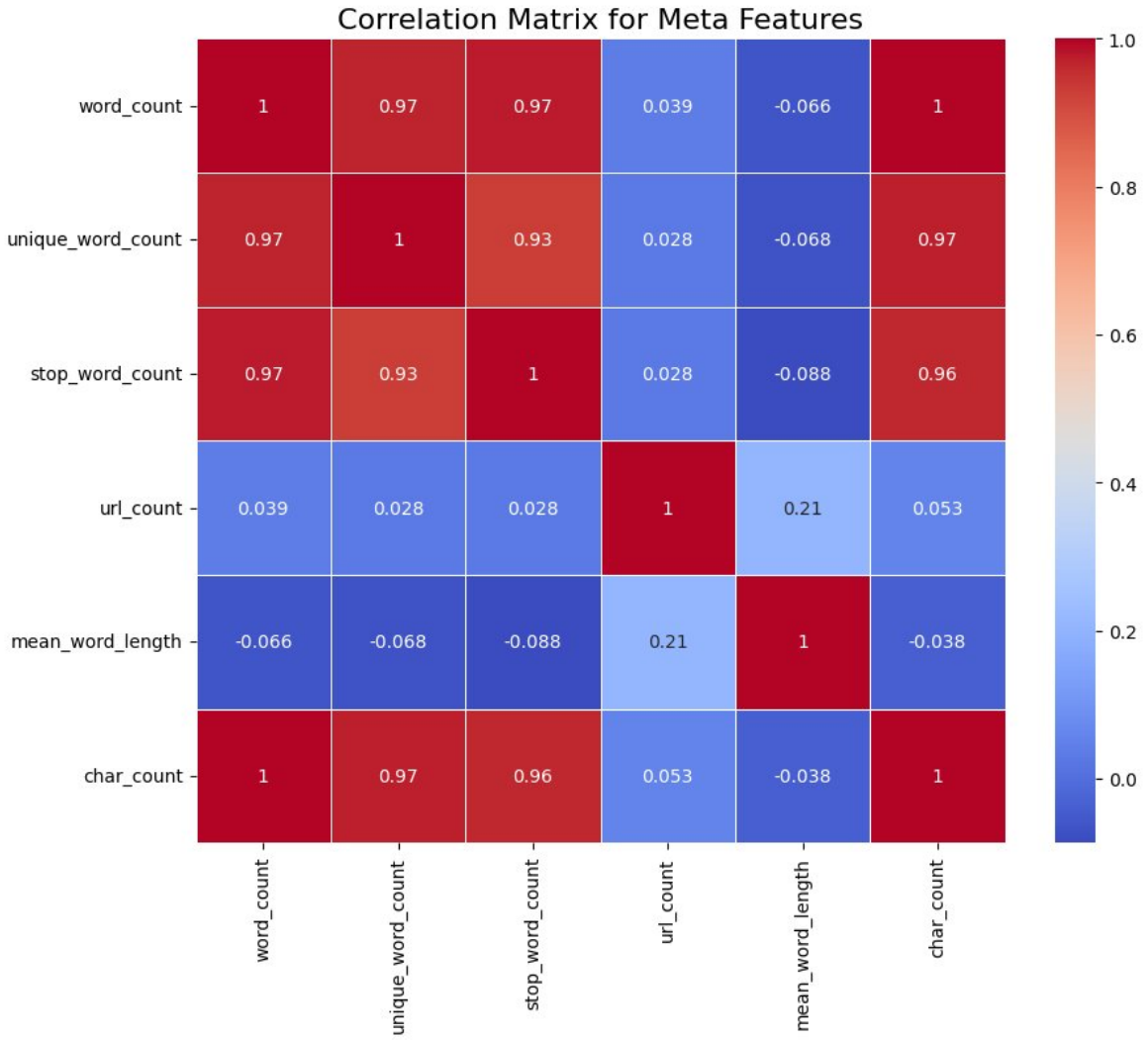
6. **Character Count Label Distribution (Bottom Right)**
    - Character count follows a similar right-skewed pattern to word count.
    - Real and fake news have overlapping distributions, with fake news tending to have slightly fewer characters on average.
    - Longer articles are less frequent, but the trend appears consistent across both categories.


word_count Label Distribution
unique_word_count Label Distribution
stop_word_count Label Distribution
url_count Label Distribution
mean_word_length Label Distribution
char_count Label Distribution
Fake
Real
Density

4.3 Correlation Plot on Textual Features

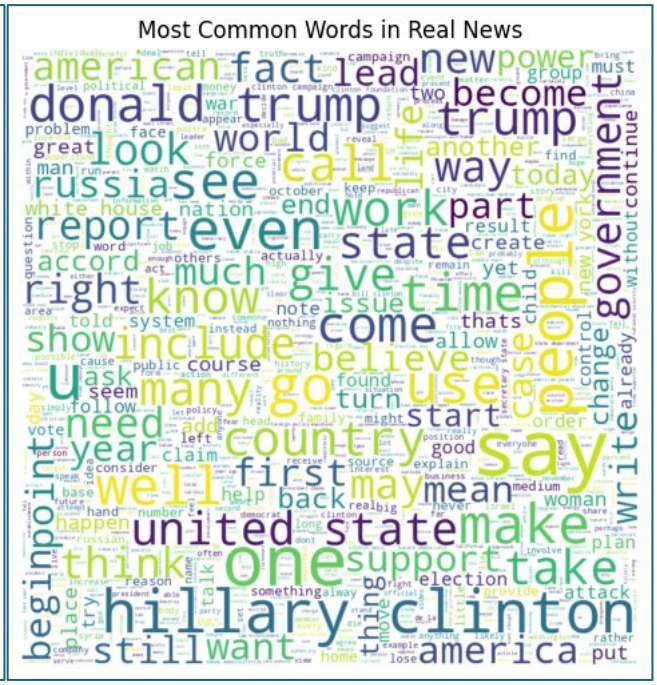
The heatmap shows the correlation between different meta-features extracted from the dataset.

- Strong correlations are observed among word count, unique word count, stop word count, and character count, indicating that longer texts naturally contain more words, unique words, and stopwords. Highly correlated features may be redundant for machine learning models.
- URL count has a weak correlation with other features, suggesting that the presence of URLs is independent of text length. **url_count** could be a key independent feature in distinguishing real vs. fake news.
- Mean word length has a slight negative correlation with word count and stop word count, meaning longer texts tend to have shorter average word lengths.
- Understanding these relationships improves feature selection and model efficiency.



4.4 Word Clouds

- Fake news tends to be more sensational, politically biased, and indirect in its language, often using emotionally charged or ambiguous wording.
- Real news focuses on factual reporting, governance, and neutrality, using words that emphasize structured information rather than persuasion.
- Although both categories cover political topics, real news maintains a more balanced and fact-driven approach, whereas fake news appears more opinionated and dramatic.



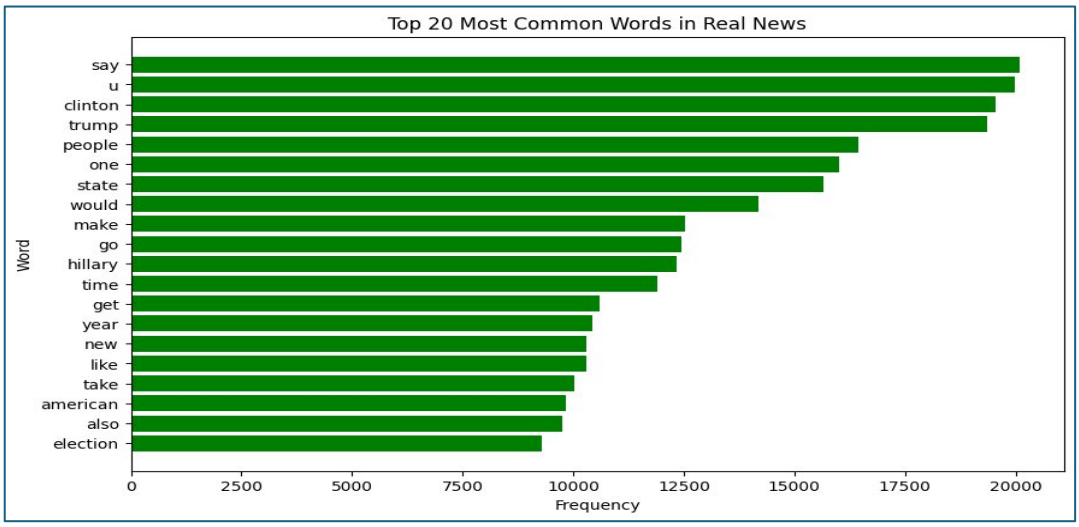
Differences Between Fake and Real News

Aspect	Fake News	Real News
Political Bias	Overemphasizes political figures and controversial topics. Frequently mentions *Trump, Clinton, government, president, democrat, republican.*	Covers politics with a more structured approach, focusing on *policy, governance, and elections* rather than individuals.
Sensationalism	Uses emotionally charged and ambiguous words like *attack, issue, claim, support, lead, call, report.* Often focuses on **uncertain or controversial statements.**	Prefers neutral and fact-based words like *fact, report, policy, explain, write, include,* indicating an emphasis on structured and verified reporting.
Use of Entities and Locations	Highlights specific places such as *United States, New York, White House,* often associating them with political controversies.	Mentions locations in a broader news context, covering both national and international topics (e.g., *Russia, America, world*).
Linguistic Characteristics	Relies on words like *say, claim, call,* which suggest indirect reporting, potentially increasing misinformation.	Uses *report, fact, explain, accord* to indicate credible and structured reporting.
Tendency for Misinformation	Frequent use of speculative and persuasive words suggests an attempt to **influence opinions** rather than provide balanced information.	More factual and **objective**, with an effort to **present data-backed news** instead of speculation.

4.5 Plots for Most Common Words

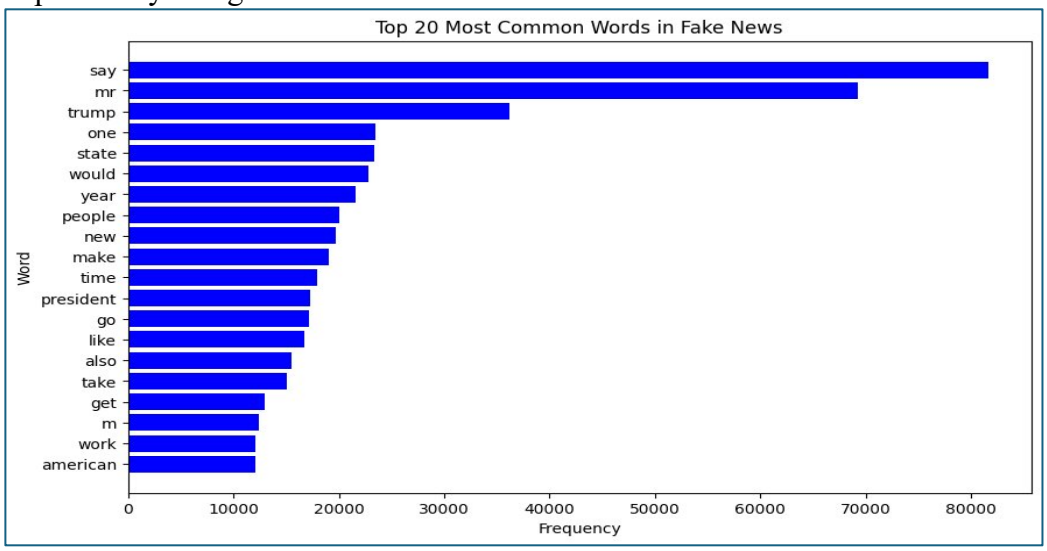
4.5.1 Top 20 Most Common Words in Real News:

- Words like *say, Clinton, Trump, people, state,* and *government* dominate the list, reflecting a structured approach to political and social reporting.
- The presence of words such as *fact, election,* and *policy* further reinforce the idea that real news aims to provide verified information and coverage of political events.



4.5.2 Top 20 Most Common Words in Fake News:

- Fake news articles show *say* as the most frequently used word, followed by *Mr., Trump, one, state,* and *president.*
- The prevalence of *Mr.* in fake news is noteworthy, as it suggests a tendency to use formal-sounding language to appear more credible.
- The significant focus on political figures and institutions indicates that fake news articles often revolve around politically charged narratives.



5. Data Analysis before Machine Learning Modelling

5.1 Data Pre-processing (Text Pre-processing)

The text preprocessing pipeline consists of multiple steps aimed at cleaning and standardizing textual data for NLP tasks such as sentiment analysis or fake news detection. The process involves:

1. **Tokenization:** Converts text to lowercase and removes special characters and retains only alphanumeric words. It uses a tokenizer to split text into meaningful tokens.
2. **Stopword Removal:** Splits the text into individual words by removing common stopwords (e.g., "the," "is," "and") using NLTK's English stopwords list.
3. **Part-of-Speech Tagging:** Assigns a part-of-speech (POS) tag to each word and maps tags to WordNet categories (adjectives, nouns, verbs, adverbs).
4. **Lemmatization:** Converts words to their root form (e.g., "running" → "run") and then uses POS tagging to ensure correct lemmatization with.
5. **Full Text Preprocessing:** Removes HTML tags and then converts text to lowercase and keeps only alphabetic characters (removes numbers and punctuation).

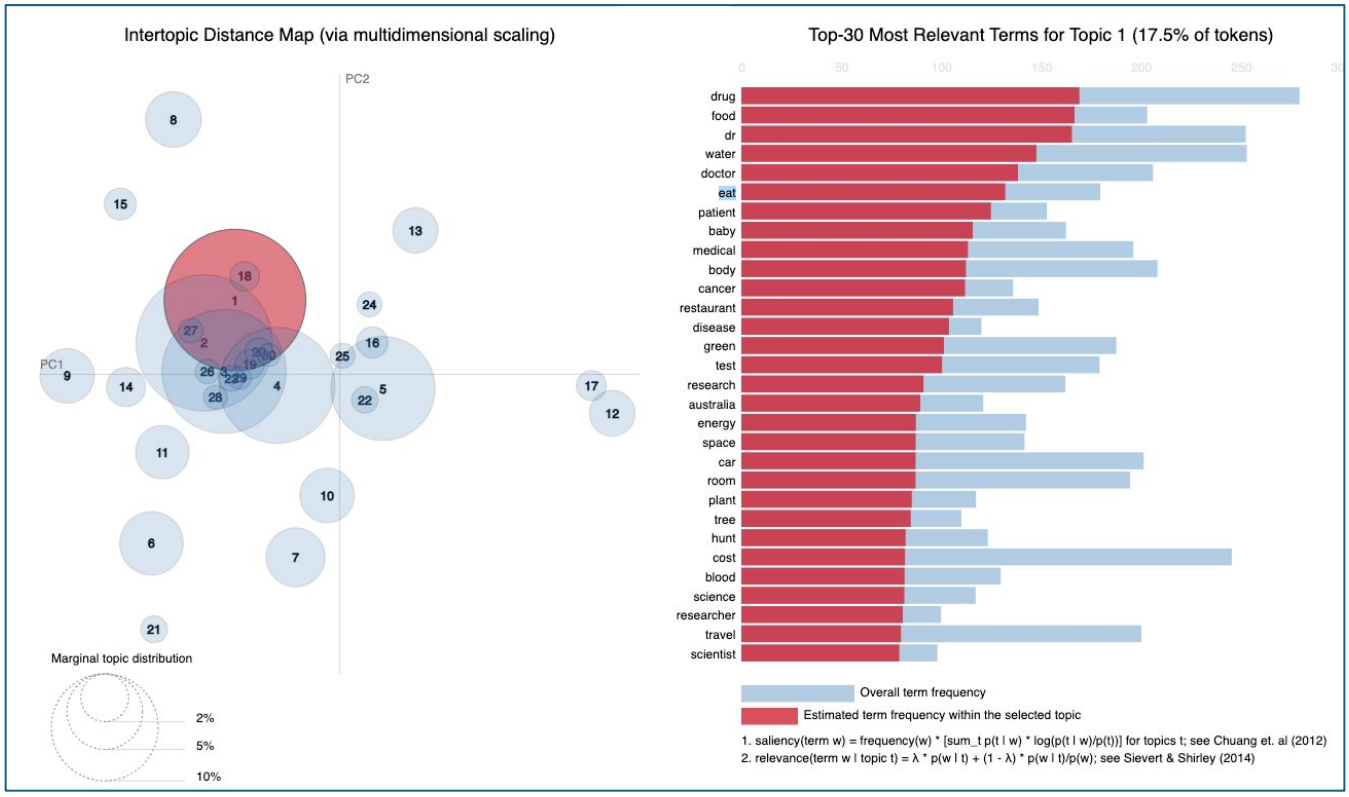
	id	title	author	text	label	word_count	unique_word_count	stop_word_count	url_count	mean_word_length	char_count	cleaned_text
0		House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1	820	432	356	0	5.001220	4930	house dem aide didnt even see comeys letter ja...
1		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		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...

5.2 Topic Modelling with TF-IDF

Topic modeling helps in identifying hidden themes in a collection of documents. In this analysis, we use Latent Dirichlet Allocation (LDA) to classify fake news articles into distinct topics. Additionally, TF-IDF transformation is applied to emphasize important words, and outlier detection is conducted to spot articles with unusual word distributions.

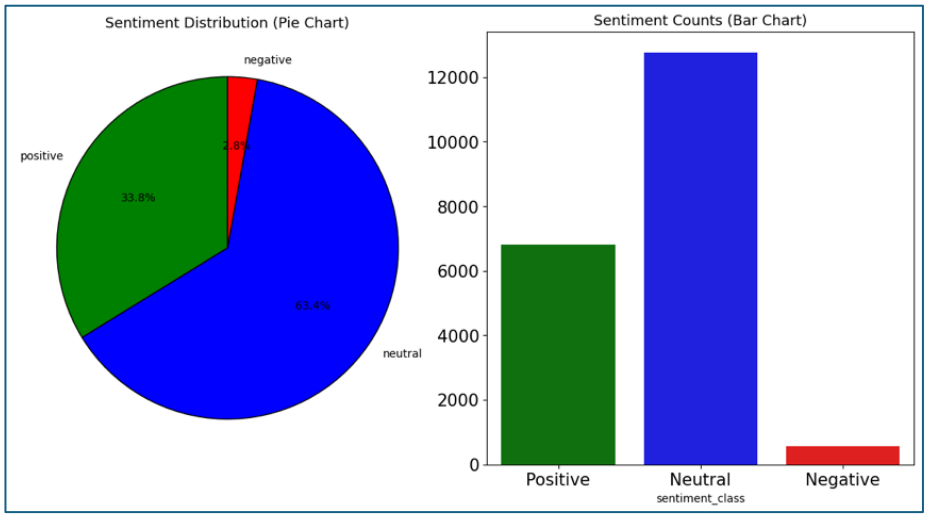
- **Topic Distribution (Left):** Overlapping circles indicate related topics, with larger circles representing more dominant topics.

- **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.



5.3 Sentiment Analysis with TextBlob

The sentiment analysis approach helps in understanding the overall **tone of news content**. The visualization provides an intuitive representation of sentiment trends, allowing for further analysis on potential biases or content trends within the dataset.



- Neutral sentiments dominate (63.4%)**, with** more than 12,000 occurrences indicating that most of the analyzed content lacks strong emotional polarity.
- Positive sentiments make up 33.8%, with more than 7,000 occurrences which suggests a significant portion of the content conveys optimism or favorable opinions.
- Negative sentiments are the least (2.8%), implying that very few texts express strong negative emotions, reinforcing that negative expressions are minimal in the dataset.

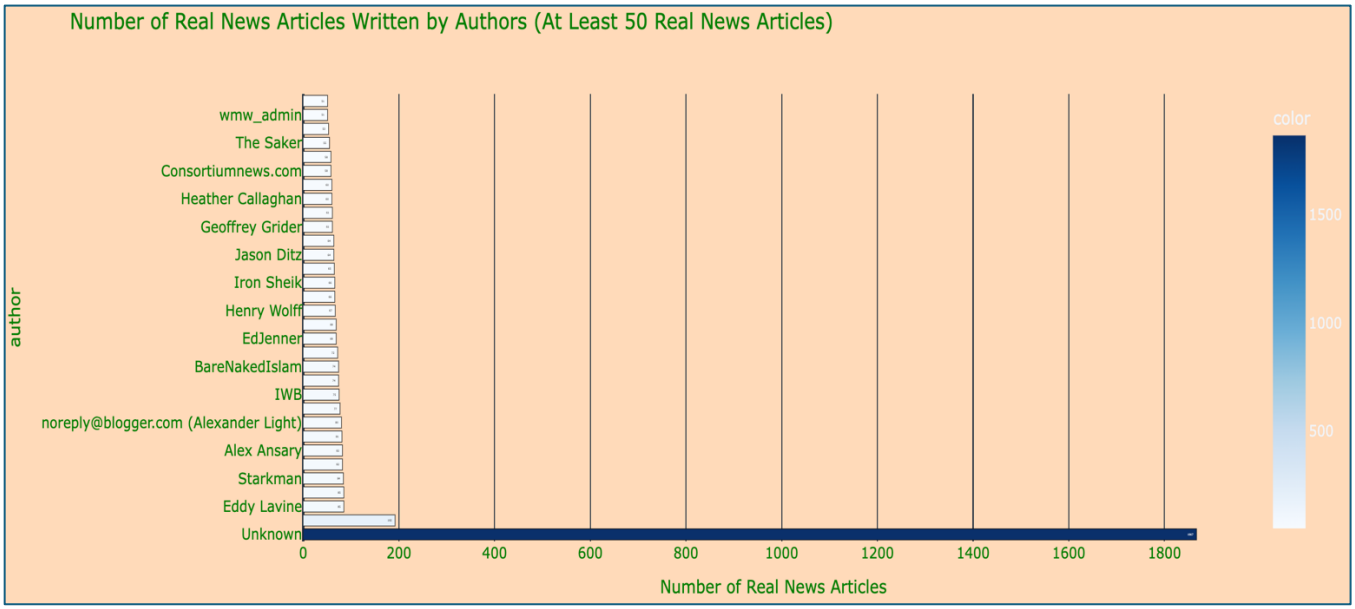
The dataset is overwhelmingly neutral, meaning the texts analyzed are either factual, non-emotional, or balanced in tone. Positivity is significantly higher than negativity, suggesting a general tendency toward favorable or optimistic expressions. Low negativity indicates that negative or harsh opinions are rare, which might suggest a bias in the dataset, or the general nature of the content being analyzed.

5.4 Distribution of Article Lengths

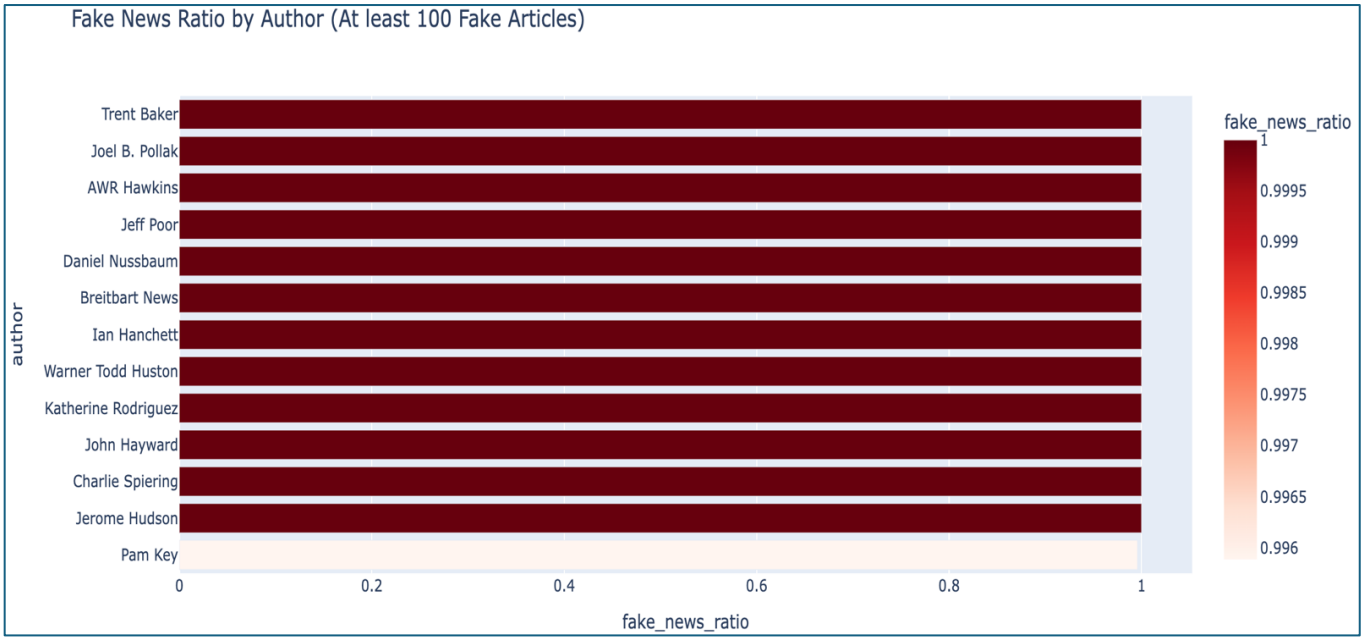
The analysis provides insights into the distribution of articles among authors, highlighting those who predominantly contribute to fake news or real news.

A Treemap Visualization It calculates the number of articles per author and filters those with more than 100 articles**.** This suggests that a large portion of news lacks clear authorship, and article length varies significantly among contributors.

Below Horizontal Chart calculates their **fake news ratio** (fake articles/total articles) showing that some authors (e.g., Trent Baker, Joel B. Pollak, Breitbart News) publish nearly **100% fake news**.



The fake news ratio for all authors is extremely close to 1 (almost 100% fake). Trent Baker, Joel B. Pollak, AWR Hawkins, and Jeff Poor have a fake news ratio of almost 1, meaning nearly all their articles are classified as fake. Pam Key has the lowest fake news ratio but is still very close to 1. The presence of Breitbart News as an "author" suggests that this dataset might classify entire news outlets similarly to individual authors.



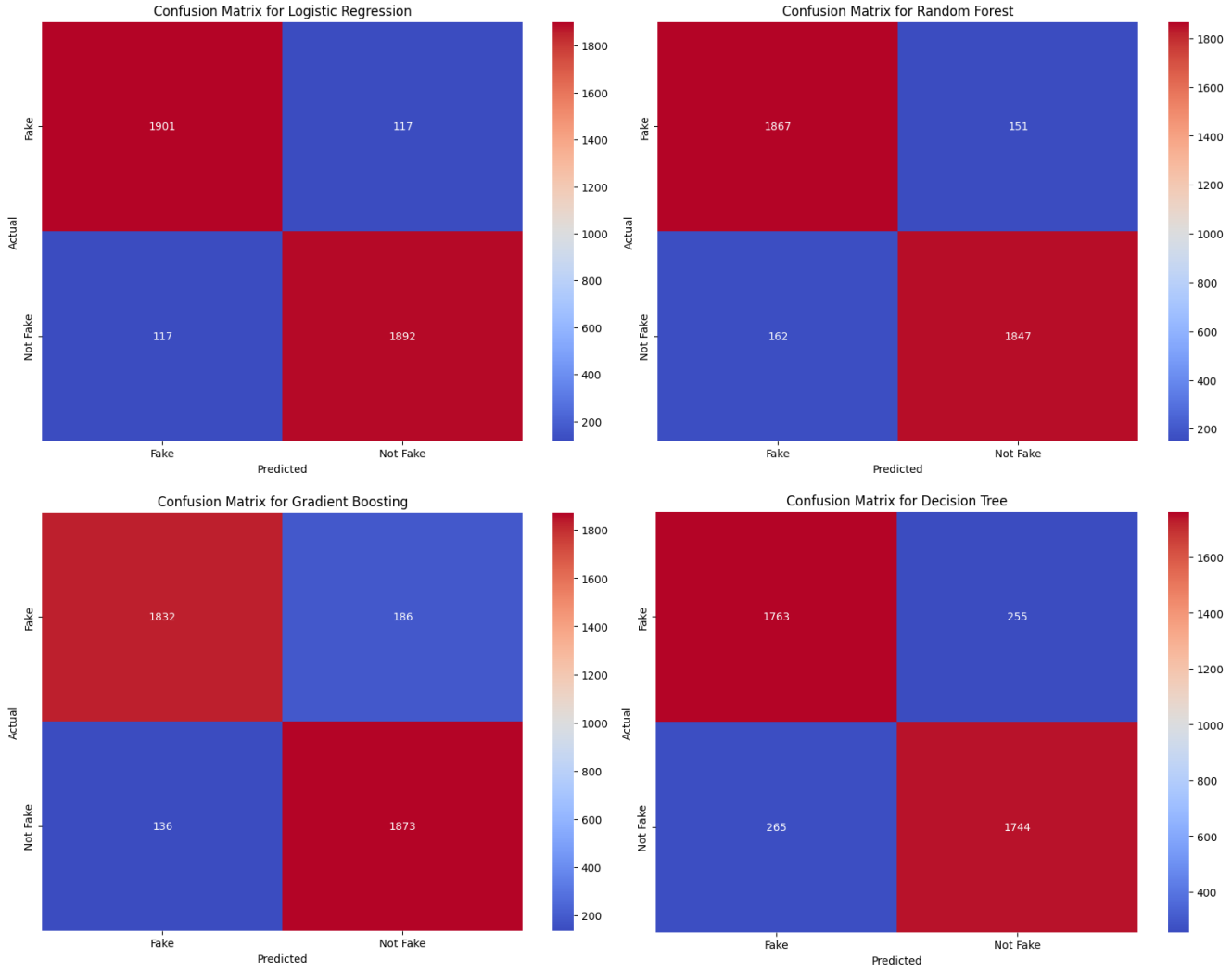
6 Machine Learning Models for Classification

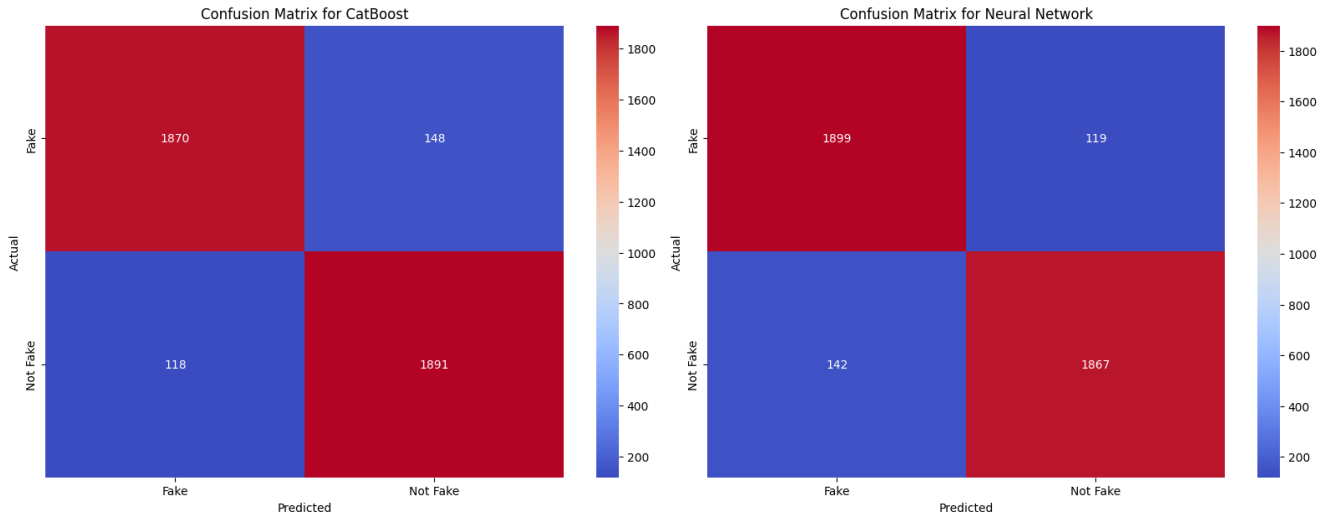
6.1 TF-IDF Vectorization and Feature Extraction

The dataset consists of 20,133 news articles processed using TF-IDF (Term Frequency-Inverse Document Frequency)**.** The text data was transformed into a 5000-dimensional feature matrix, capturing unigrams, bigrams, and trigrams to enhance context understanding. This feature representation was used for training machine learning models.

6.2 Machine Learning Models

- The dataset was split into 80% training data and 20% testing data.
- The following are the confusion matrices of the six models that were trained to classify news as fake or real with confusion matrices.





6.2.1 Model Evaluation Metrics

Model	Accuracy (%)	Precision	Recall	F1-score
Logistic Regression	94.19	0.942	0.942	0.942
Random Forest	92.23	0.925	0.92	0.922
Gradient Boosting	92	0.907	0.931	0.919
Decision Tree	87.09	0.874	0.869	0.871
CatBoost	93.39	0.926	0.94	0.933
Neural Network (MLP)	93.52	0.941	0.93	0.936

6.3 Models Performance Comparison and Conclusion

6.3.1 Best Model for Fake News Detection

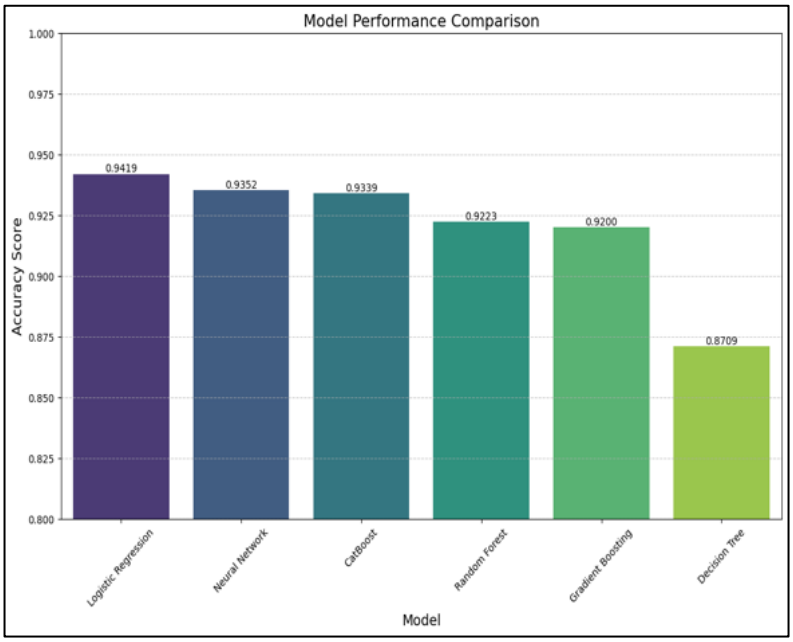
- Logistic Regression and CatBoost performed best in Fake News Detection.
- Logistic Regression achieved the highest recall (94.2%), meaning it correctly identified more fake news articles.
- Neural Network (MLP) had the highest F1-score (0.936), making it a well-balanced model.

6.3.2 Best Model for Real News Detection

- Logistic Regression had the highest accuracy (94.19%), making it the most reliable in classifying both real and fake news correctly.
- Precision for real news (TN / (TN + FN)) was highest for Logistic Regression.
- Gradient Boosting and CatBoost also performed well in detecting real news, with Gradient Boosting showing strong recall (93.1%).

6.3.3 Model Selection Summary

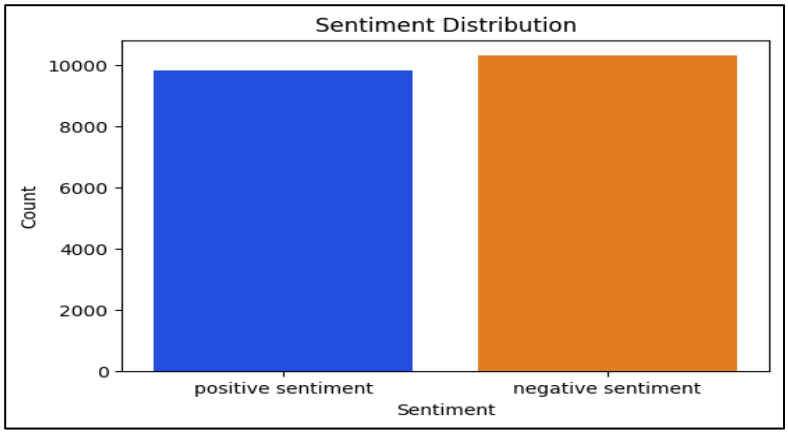
- If you prioritize minimizing fake news misclassification (i.e., recall for fake news) → Logistic Regression is the best.
- If you want a balance between detecting real and fake news effectively (F1-score, accuracy) → Neural Network (MLP) or Logistic Regression are the best choices.



Decision Trees performed the worst, with lower recall and precision, meaning they are not suitable for this task.

7 Sentiment Analysis using Machine Learning

The sentiment analysis function processes news articles by transforming text into a **TF-IDF matrix**, normalizing it, and using a **pre-trained classifier** to predict sentiment as either **positive or negative**. The predictions are stored in a new column, and a **count plot** visualizes the distribution of sentiments across the dataset. This approach helps in identifying sentiment trends and potential biases in news articles efficiently.

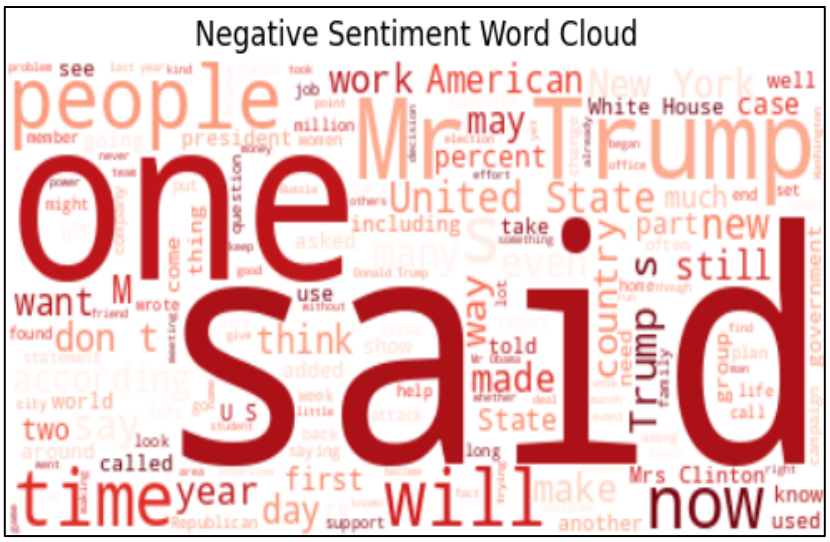


7.1 Word Clouds for Sentiment Analysis

7.1.1. Positive Sentiment Word Cloud

- The words “people,” “time,” and “will” suggest discussions related to future actions, collective experiences, or general political optimism.
- Words like “election,” “country,” “support,” and “American” indicate that positively classified articles might focus on national progress, unity, or political campaigns.
- The presence of “Donald Trump” and “Hillary Clinton” suggests that political figures are often mentioned in a positive sentiment context, possibly in support-based discussions.

7.1.2. Negative Sentiment Word Cloud



- The largest word "said" suggests that negatively classified texts rely heavily on reported speech or quotes, which might reflect criticism, scandals, or negative news coverage.
- “Mr. Trump” being a dominant phrase indicates that articles mentioning him frequently might carry negative sentiment.
- Words like “work,” “now,” and “many” may suggest topics around economic struggles, dissatisfaction, or urgency.
- “United States” being prominent hints at discussions on national issues or policies with negative framing.

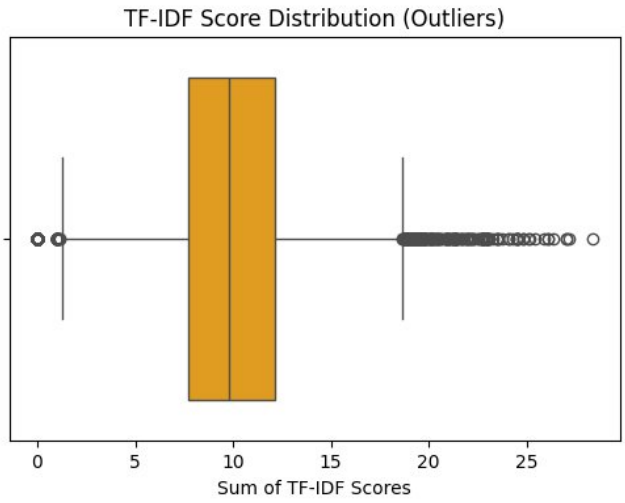
7.1.3. Potential Implications:

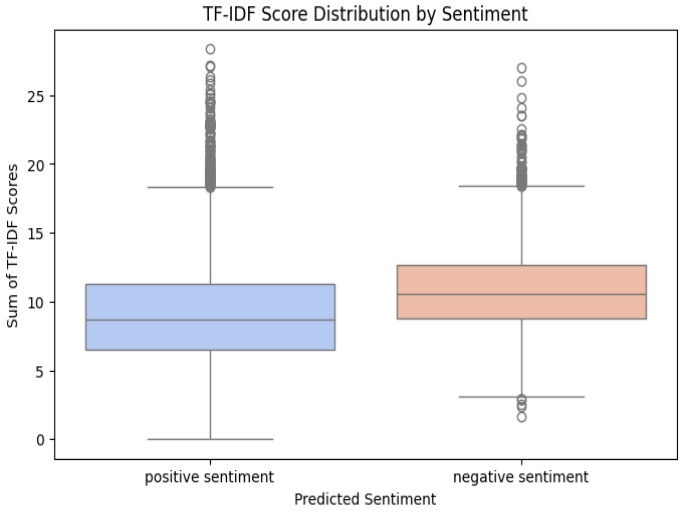
The negative sentiment texts seem to revolve around reported speech, controversies, and urgency, while positive sentiment texts emphasize future actions, national identity, and political support. The presence of political figures in both suggests that the dataset is heavily political, and sentiment polarity may be influenced by reporting styles or framing biases.

7.2 TF-IDF Distribution with outliers and Sentiment

TF-IDF Score Distribution with Outliers**,** shows the presence of numerous outliers suggests that certain documents have significantly higher or lower TF-IDF scores, indicating unique term distributions. A high number of outliers on the right side (higher TF-IDF sums) indicate that some documents contain many unique or highly weighted words.

TF-IDF Score Distribution by Sentiment**,** compares the sum of TF-IDF scores between positive and negative sentiment classifications exhibiting a range of TF-IDF values with some extreme outliers, but negative sentiment appears to have slightly higher median TF-IDF scores. This suggests that negatively classified documents might contain a distinct set of frequently weighted terms compared to positive sentiment document.





8 Fake vs. Real News Detection using Cosine Similarity

The evaluates the textual similarity between fake and real news articles, as well as between positive and negative sentiment classifications using TF-IDF vectorization **and** cosine similarity.

Fake vs. Real News Similarity

- The cosine similarity score between fake and real news is 0.8139, indicating a high degree of similarity.
- This suggests that fake news often mimics real news in vocabulary, structure, and writing style, making detection more challenging.

Positive vs. Negative Sentiment Similarity

- The cosine similarity score between positive and negative sentiment is 0.8012**,** showing moderate similarity.
- While these sentiments differ in meaning, overlapping vocabulary and common words contribute to their similarity.

9 Conclusion

This project focused on developing a machine-learning model to detect fake news on social platforms, using techniques such as TF-IDF vectorization, 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.