Project Report for IST-736: Text Mining

Fake News Detection with Machine Learning

By Arti Ravi Garg, Maybel Herrera and Rahul Gurujapu

Introduction

In today's digital world, fake information spreads quickly through social media and online sources, and it's very difficult to distinguish between truth and deception. Fake news can manipulate public opinion, fuel political polarization, and create unnecessary panic or confusion. It affects individuals by changing their beliefs based on false news and can even impact major events such as elections, public health, and economic stability. Detecting fake news is necessary to ensuring an informed society and preventing the wrong consequences of wrong information. By using data-driven approaches and AI models like text mining, we can distinguish between fake and real news.

Motivation

Fake news has become a norm, rendering it hard to separate fact from fiction. Whether it's an uplifting animal rescue, an assault, or a political opinion, false information can falsify the difference between truth and falsehood. As technology advances, the need to identify fake news is more critical than ever. Lacking effective detection tools, misinformation can sway public opinion, affect key decisions, and undermine media credibility. A model must be developed to detect false news.

Goal

The purpose of this project is to design a successful model that will accurately distinguish between fake news and real news using text mining. Through analysis of linguistic patterns, sentiment, and word frequency, the model will identify deceitful content and distinguish it from true sources. This process is aimed at enhancing the ability to detect disinformation, providing a good model to combat the spread of fabricated news during a period where AI-generated content enables lying more than ever in the past.

About the Data

The primary dataset, **"news_articles.csv,"** contains 10 attribute with 2047 instances:

Attribute names are:

- Author
- title
- text
- language
- Type
- label
- hasImage

- date
- time
- Timezone

To maintain focus on key aspects, the following columns were excluded:

- title_without_stopwords
- text_without_stopwords
- site_url
- main_img_url

The publication date field was converted into a standardized datetime format.This transformation allowed for more precise analyses based on:

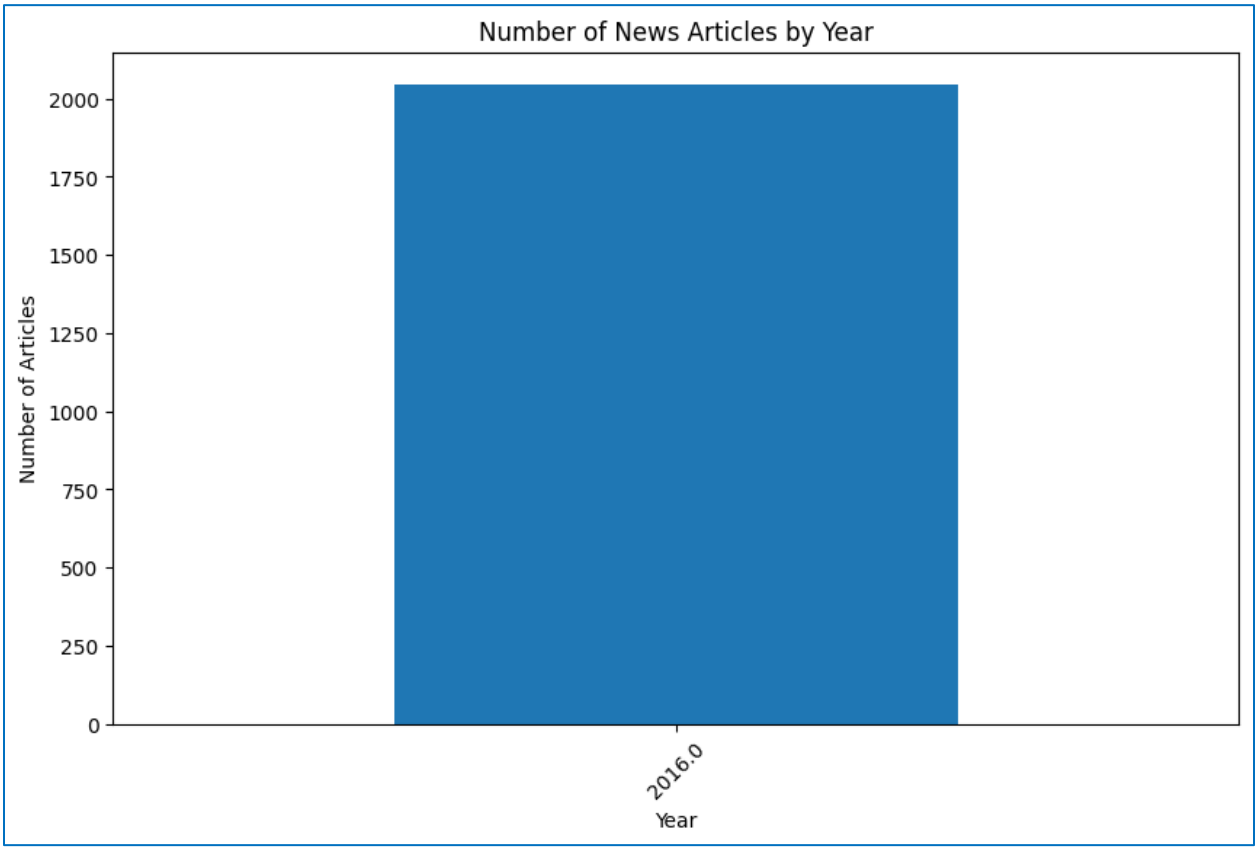
- Year
- Time of day

Exploratory Data Analysis

To understand the type of news articles that could be part of this dataset, we began by exploring time frame the articles were published.

Information regarding the frequency of article publications over the years was extracted and plotted in a bar chart, illustrating how article counts varied by year. The bar chart in the visualization represents the number of articles published per year. However, from the chart, it appears that the dataset is limited to a single time period. It contains only one

distinct year (2016).



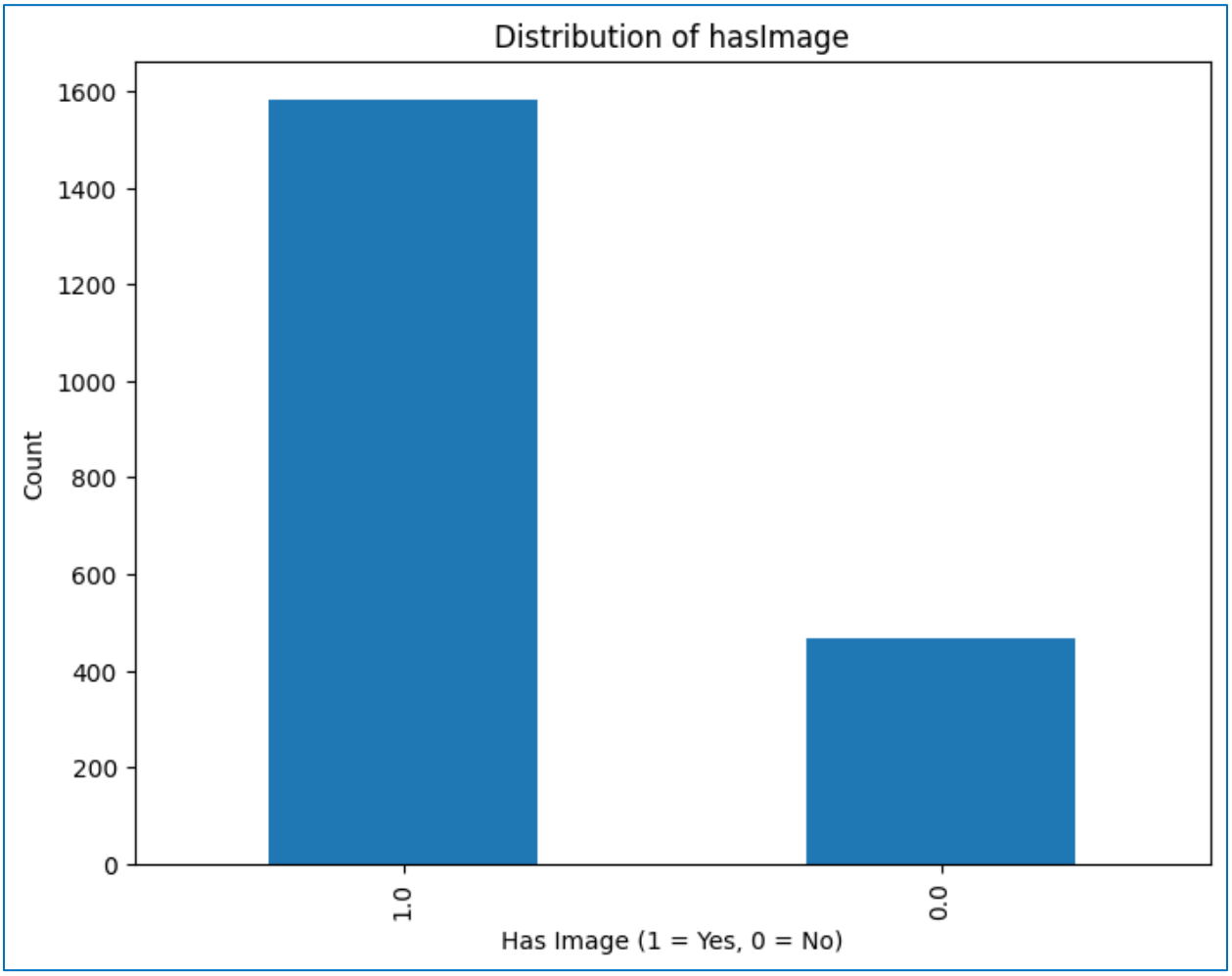
Additionally, another bar plot indicated how often images appeared in articles, offering insight into the relative importance of visual elements. Understanding the presence of images in news articles is crucial in analyzing content engagement and presentation quality. This analysis explores the distribution of articles with and without images.

The bar chart presents the distribution of articles based on whether they contain images. The key observations are:

A significant majority of articles (represented by 1.0 on the x-axis) contains images.

A smaller proportion of articles lack images (0.0 on the x-axis).

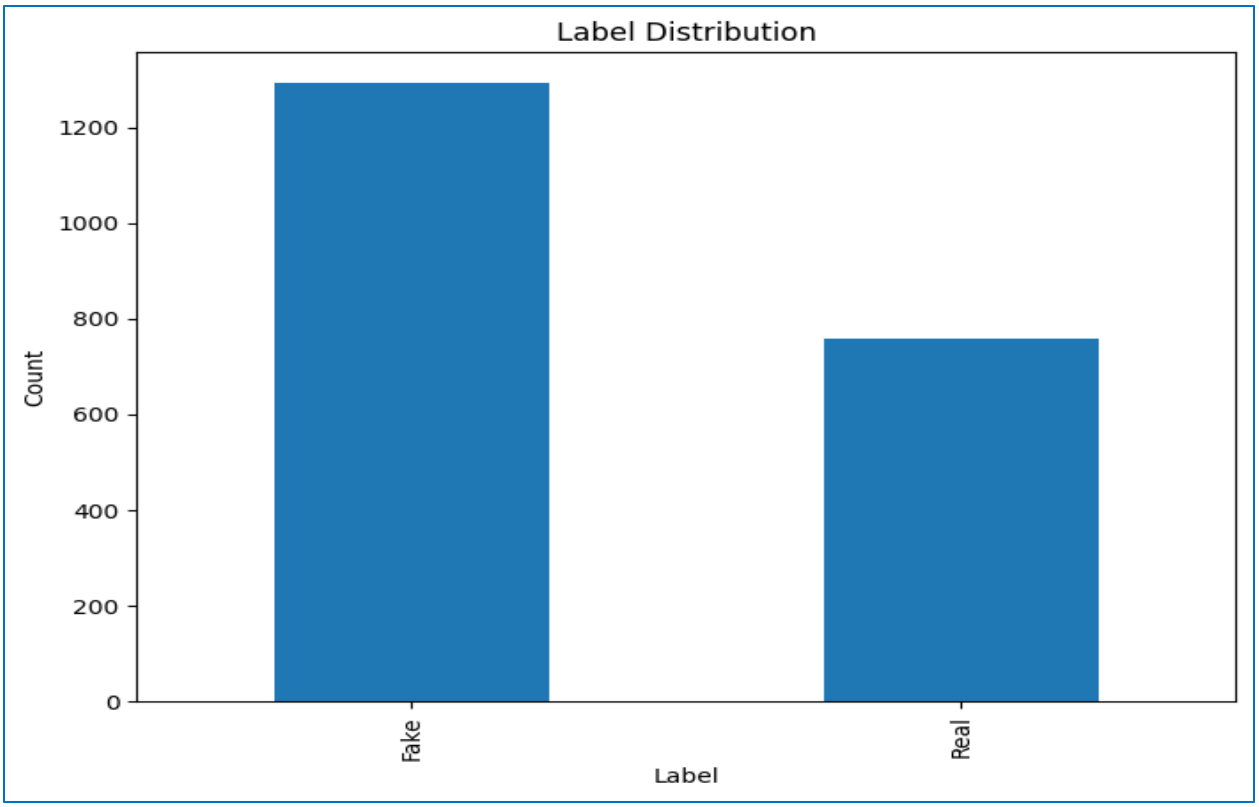
The ratio suggests that images are commonly used in the dataset, potentially indicating a preference for visual elements in news publication.



To evaluate the dataset's balance, the number of articles labeled as real versus fake was compared visually in a bar chart. The bar chart indicates that fake news articles outnumber real news articles significantly.

The fake news category has more than 1,200 articles, whereas the real news category has fewer than 800 articles.

This imbalance suggests that the dataset contains nearly twice as many fake news articles as real ones.

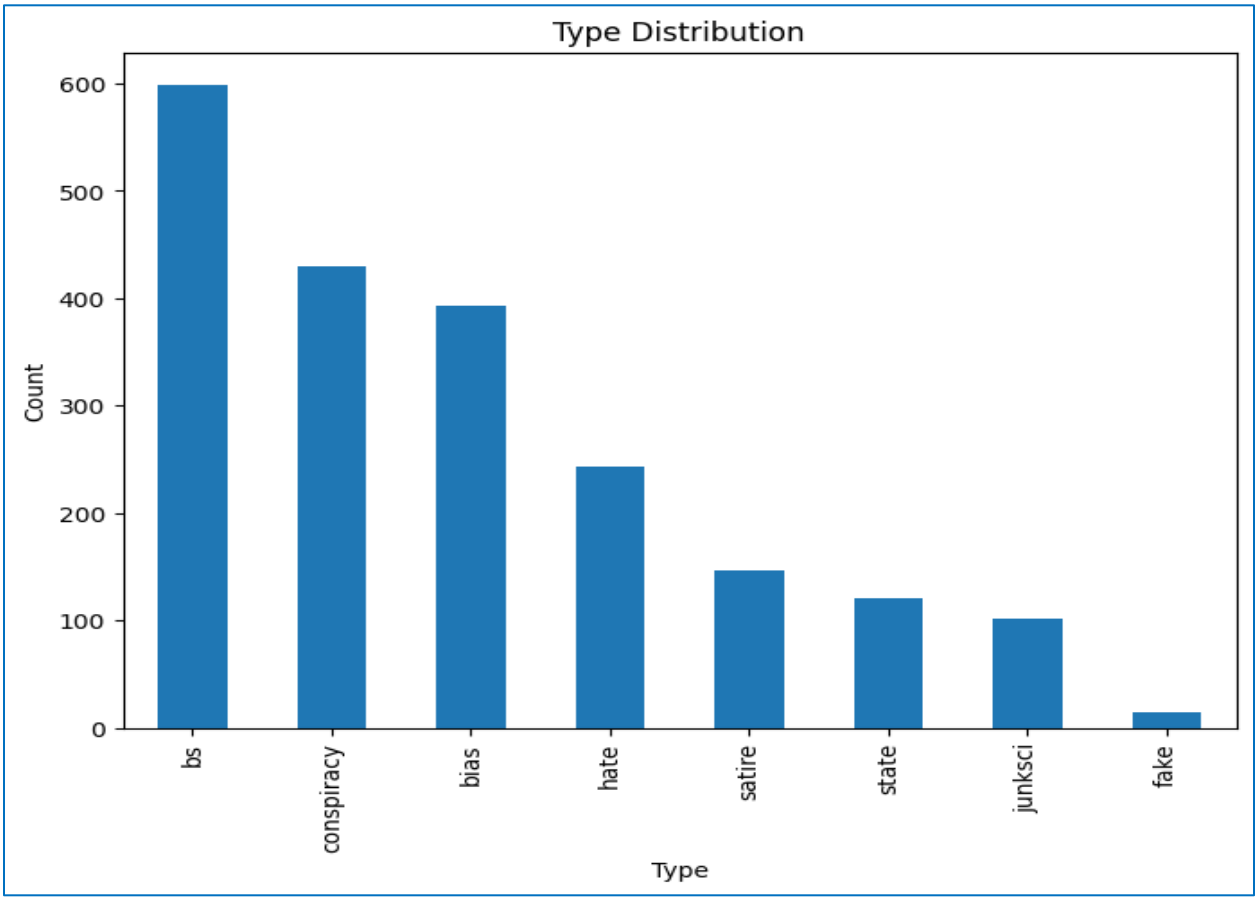


The range of news categories was also reviewed, using a bar plot to highlight the prevalence of different types of news items. This analysis examines the distribution of different types of misinformation in the dataset. The categories include baseless statements, conspiracy, bias, hate, satire, state-sponsored, junk science (junksci), and fake news. Understanding this distribution helps in identifying the most common misinformation trends.

Findings:

- "baseless" is the most frequent category, with over 600 occurrences, suggesting a prevalence of misleading or deceptive content.
- Conspiracy-based misinformation follows closely, with around 450 instances, highlighting a significant presence of unfounded theories.
- Bias-based articles are also frequent, showing that subjective and one-sided reporting is a notable issue.
- Hate speech appears frequently in the dataset, with over 200 instances, emphasizing the need for monitoring and mitigating harmful narratives.
- Satire and state-sponsored misinformation are present in moderate amounts, suggesting some influence of political or humorous misinformation.
- Junk science (pseudoscience) is relatively low but still appears in the dataset.

- Fake news is the least frequent, indicating that outright fabrications may be less common compared to misleading or biased content.

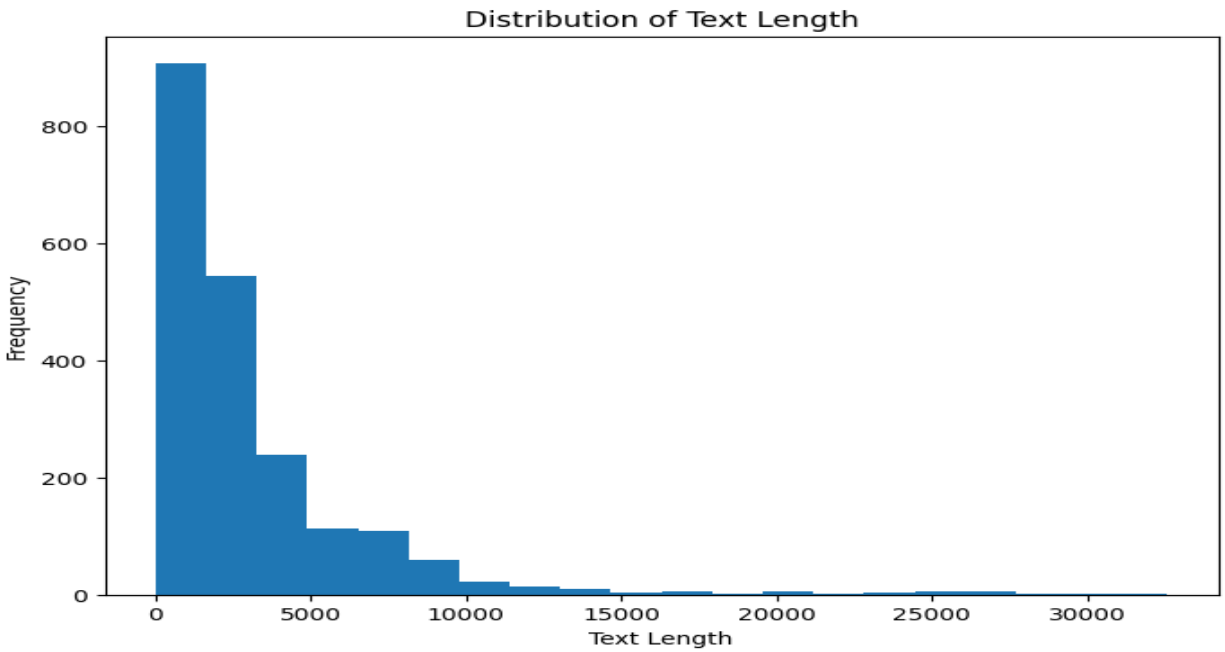


Finally, the text lengths of articles were examined through a histogram, revealing how some articles were notably short and others significantly longer. Understanding the length of articles helps in determining content characteristics, such as whether certain types of articles tend to be longer or shorter, and if any preprocessing adjustments are needed for text analysis tasks.

Findings:

- The histogram shows a right-skewed distribution, indicating that most articles are relatively short, while a few are significantly long.
- The majority of articles have a text length below 5000 characters, with the highest concentration near the lower end of the scale.
- Some articles extend beyond 10,000, 20,000, and even 30,000 characters, but these are rare.
- The steep drop-off in frequency suggests that extremely long articles are outliers.

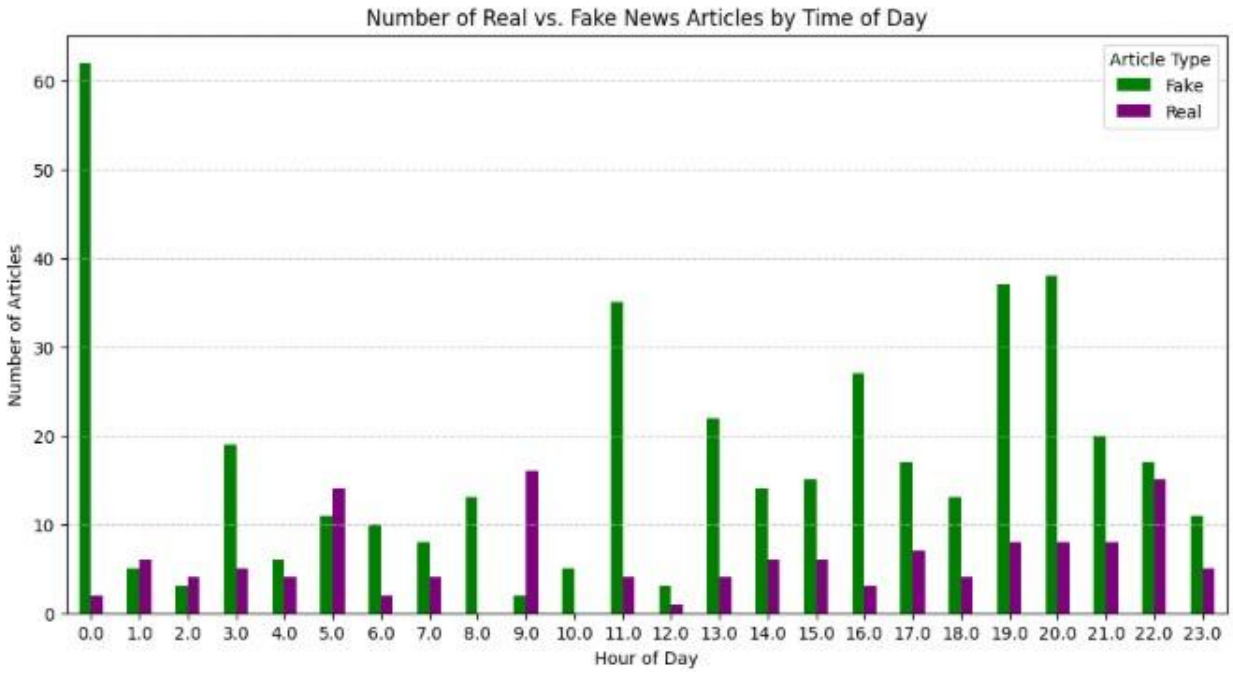
The dataset contains a mix of short and long articles, with most clustering at the lower end of the spectrum. This right-skewed distribution suggests that while short articles dominate, a few lengthy ones significantly extend the scale



The bar charts examine the distribution of news articles by time of day, focusing on:

- The distinction between real and fake news based on the hour of publication.
- The overall distribution of articles across different periods (morning, afternoon, evening, and night).

By analyzing publication times, we can identify trends in misinformation dissemination and assess whether certain time periods are more prone to publishing specific types of content.



Key Findings:

- Fake news articles dominate in nearly all hours, particularly during midnight (0:00), 11:00, 18:00, and 19:00. Real news articles are more evenly distributed throughout the day, but their count remains lower than fake news in most time slots.
- A notable spike in fake news occurs at midnight (0:00), where over 60 fake news articles are published. This is significantly higher than any other hour.
- Real news articles show some activity between 5:00–7:00 and 9:00–11:00, but they never surpass fake news in volume.
- Fake news publication seems to peak late at night and in the early evening (18:00–20:00), suggesting a pattern in when misinformation is most actively disseminated.

The dominance of fake news at midnight and in the evening might indicate strategic posting times to maximize engagement or bypass fact-checking. Real news is more balanced but still falls short in terms of volume, possibly due to differences in editorial processes.

The findings suggest that social media algorithms and reader engagement patterns may favor fake news at night, making it a key area for monitoring misinformation.

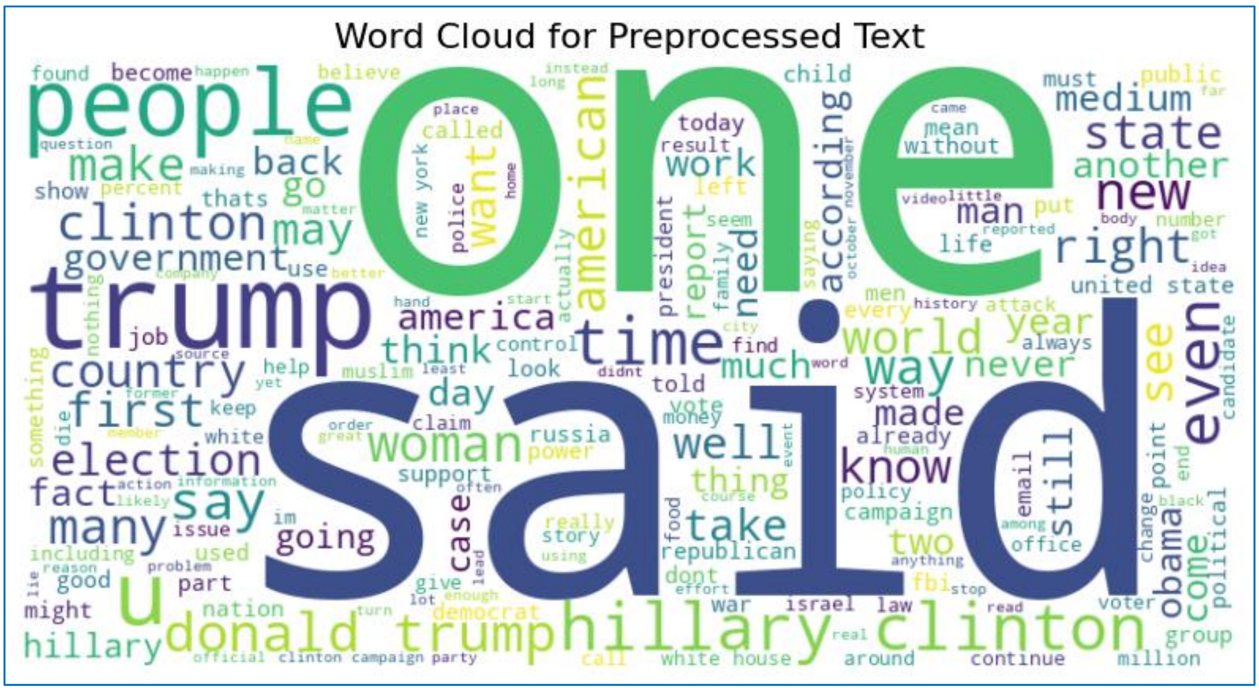
Data Preprocessing

Before modeling, the data underwent several preprocessing steps to ensure its quality and consistency.

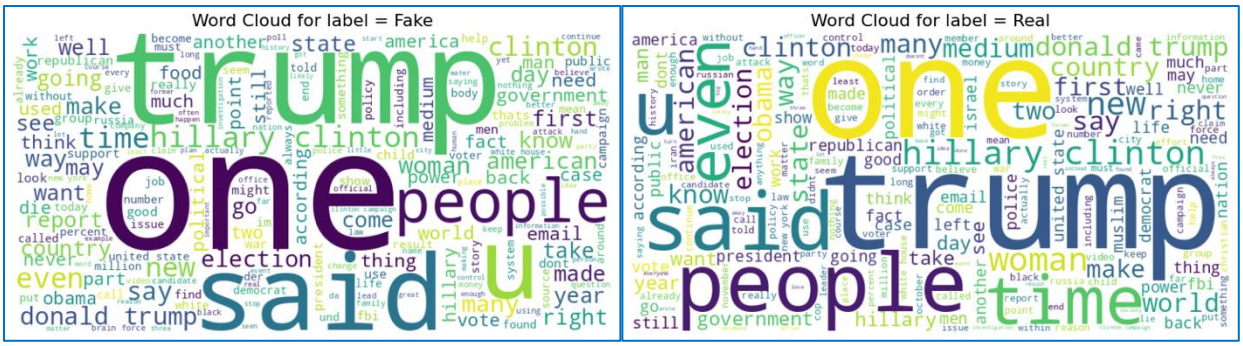
1. **Lemmatization** was applied to convert words into their base forms, reducing the effect of morphological variations.
2. Common words that do not contribute significantly to semantic meaning, known as **Stop words**, were removed to minimize noise. We also tried to update the stopwords list by adding "news", "said", "people", "trump","clinton", "hillary","donald" words in stop words but it does not increase the accuracy for our models so we go a head with original "english" stop words for this project.
3. **Special characters** were eliminated, ensuring that extraneous punctuation and symbols did not distort modeling results.

Visualization of Textual Data

In addition to the EDA, further insights were gained by visualizing the text itself. Word clouds captured the most frequently used terms in all articles,



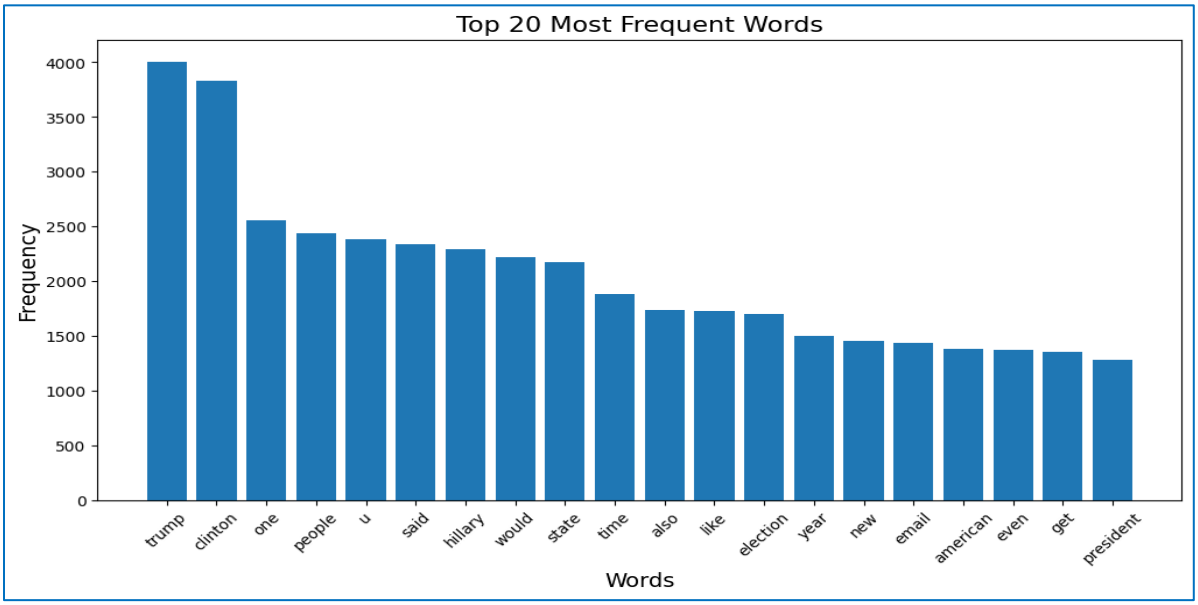
whereas separate word clouds for real and fake news showed how vocabulary usage might differ across these two categories.



We can observe that words like "Trump," "one," "said," and "people" appear in both real and fake news word clouds, but with varying frequencies. This suggests that the occurrence of these words differs between real and fake news, which could help in detecting the authenticity of the news. Additionally, other distinguishing words, such as "even" and "time," are more prominent in the real news word cloud, which could further aid in identifying real news.

Top 20 words in the data set-

The bar chart displays the top 20 most frequently occurring words in the dataset, offering insights into key vocabulary patterns.

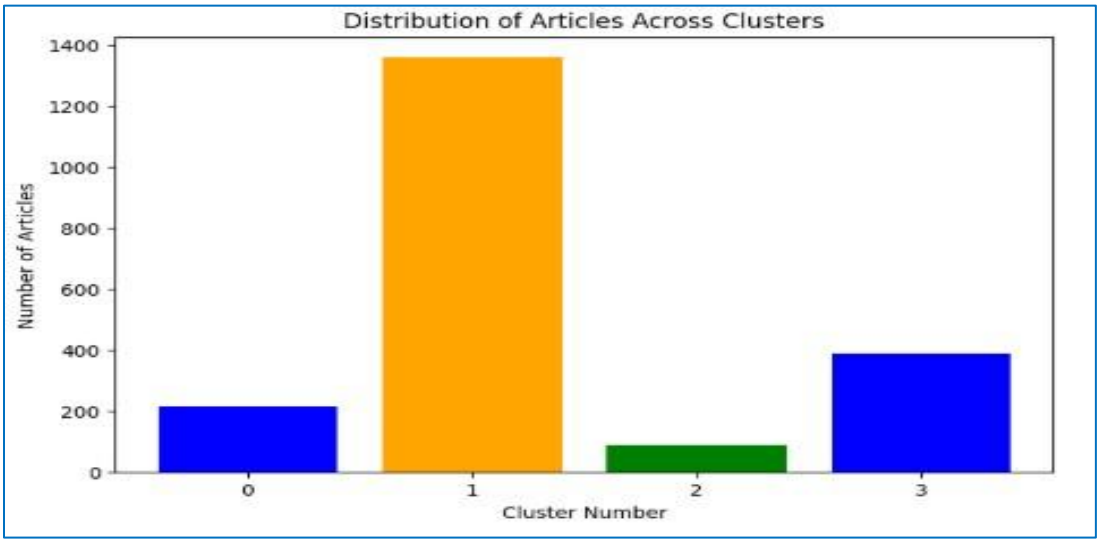


Key Observations:

1. **Dominant Words**: The most frequently used words are "trump" and "clinton," each appearing over 4000 times. This suggests that the dataset is heavily focused on topics related to these figures, likely political discussions or news articles.
2. **Other High-Frequency Words**: Words such as "people," "election," "state," and "email" indicate themes of governance, public discourse, and possibly controversies (e.g., Clinton's email scandal).
3. **Neutral and Contextual Words**: Some frequently occurring words like "one," "said," "would," and "like" serve as connectors or common phrases in general text.
4. **Potential Themes**: The presence of words like "election," "president," and "american" reinforces the idea that the dataset may be related to U.S. politics, elections, and governance.
5. **Balanced Representation**: While "Trump" and "Clinton" are most dominant, the dataset also includes broader contextual terms, suggesting coverage of related political events rather than a single-topic focus.

Clustering and Topic Modeling

K-means clustering was performed to determine how news articles might group together based on lexical similarities. A bar chart illustrated the number of articles in each cluster



Cluster 0 : This cluster seems focused on the **Hillary Clinton email investigation** and related controversies involving the **FBI** and **James Comey** during the 2016 U.S. election.

Cluster1: This appears to be a more general and broader topic, potentially centered on **human interest stories**, societal issues, or general news.

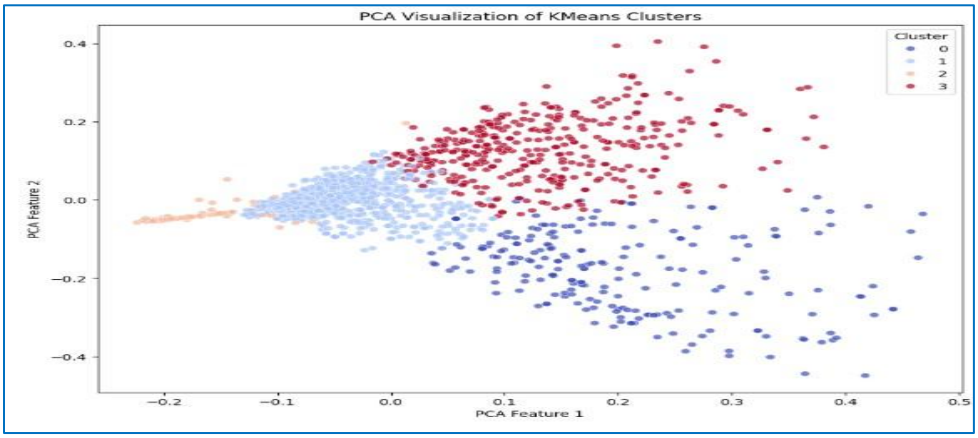
Cluster 2: This cluster has a mix of **German words** ("der," "die," "und," "zu," "da") and terms related to **brain health** or pseudoscience (like "supercharge," "activation," "neural").

Cluster 3: This cluster is clearly focused on the **2016 U.S. Presidential Election**, particularly the rivalry between **Donald Trump** and **Hillary Clinton**.

Out of the four clusters, two primarily focus on the **2016 U.S. election**. This concentration on a specific event could potentially limit the model's ability to generalize for broader **fake news detection** beyond this context.

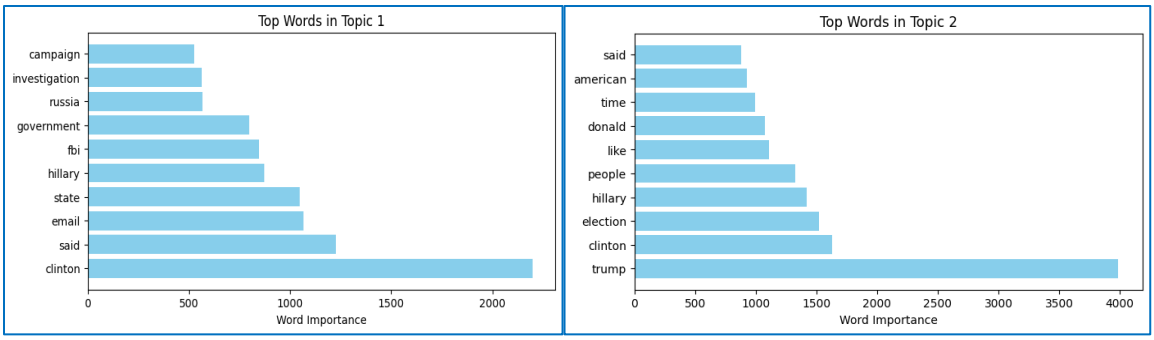
```
Top Words per Cluster:

Cluster 0:
clinton, fbi, email, hillary, investigation, comey, campaign, foundation, podesta, election

Cluster 1:
people, said, time, state, year, woman, new, like, article, world

Cluster 2:
der, die, und, brain, da, infowars, zu, supercharge, activation, neural

Cluster 3:
trump, election, clinton, donald, hillary, republican, vote, voter, president, said
```

A two-dimensional representation using Principal Component Analysis visually demonstrated how these clusters were distributed in the feature space.

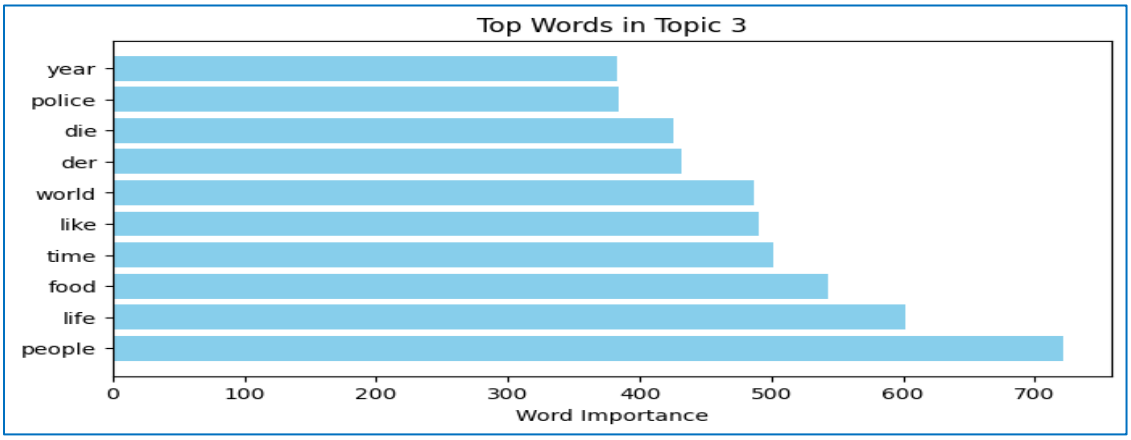


Clusters 0 and 3 with their dense locations on the 2016 U.S. election appear to make up disparate groupings. Cluster 2's tight clustering suggests more specialized or niche material, likely caused by linguistic differences or specific jargon about health news. Cluster 1 is more dispersed, however, suggesting more general news articles.

Latent Dirichlet Allocation (LDA) was also employed as a topic modeling technique, identifying key topics present in the corpus. Topic-word distributions indicated which terms were most representative of each discovered topic, shedding light on the major themes contained in both real and fake news articles.



The dense intersection of words pertaining to the election, particularly Topics 1 and 2, maintains the intuition that the majority of the dataset revolves around the 2016 U.S. election. This dominance can potentially impact the model's capacity to generalize, as it may be over-trained on this specific event.



Conversely, Topic 3 appears to have broader, more general subjects.

Topic Themes:

Looking at the top words in each topic, we can interpret its main theme.

What is Topic 1 about?

The top 10 Key Words are:

clinton, email, said, state, hillary, fbi, government, investigation, russia, campaign

This topic likely clusters articles discussing the FBI's investigation into Clinton's emails, political debates, and election narratives. The possible topic theme in these cluster of news articles could be Hillary Clinton Email Investigation & 2016 Election Controversy

What is Topic 2 about?

The top 10 key words are:

trump, clinton, election, hillary, people, donald, american, time, said, like

This topic likely clusters articles discussing the rivalry between Trump and Clinton, the election process, and political narratives in the U.S. The possible topic theme in these clusters of news articles could be 2016 U.S. Presidential Election & Political Discourse

What is Topic 3 about?

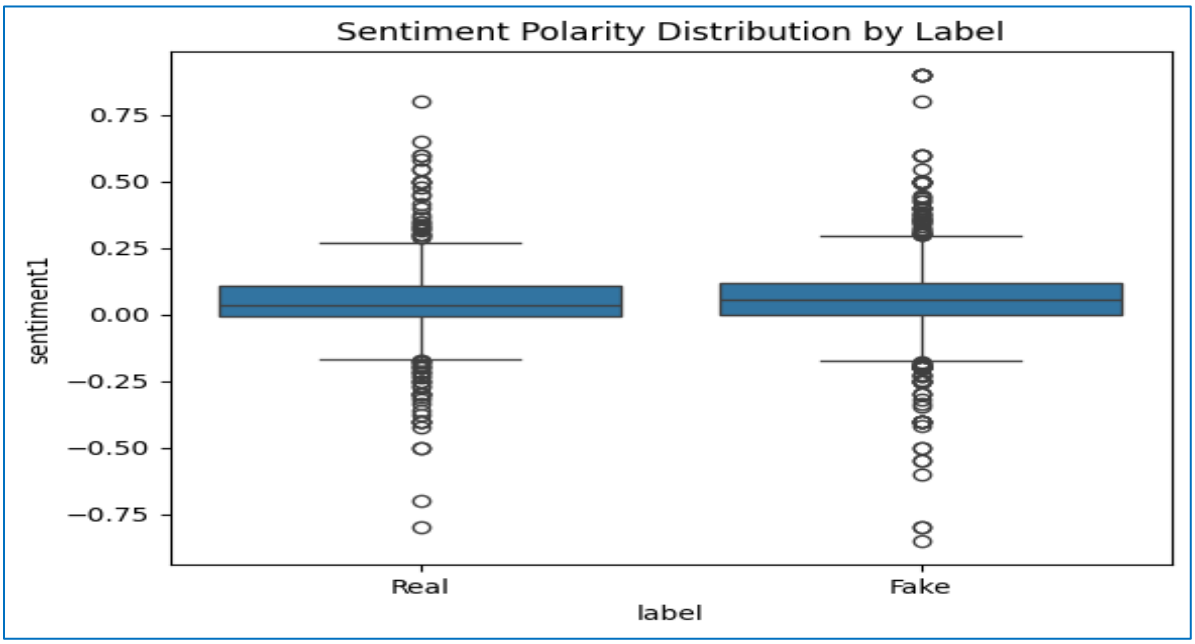
The top 10 words are:

people, life, food, time, like, world, der, die, police, year

This topic likely clusters articles discussing societal issues, police actions, food security, life experiences, and possibly international topics. The possible topic theme in these clistes of news articles could be Society, Life, and Law Enforcement.

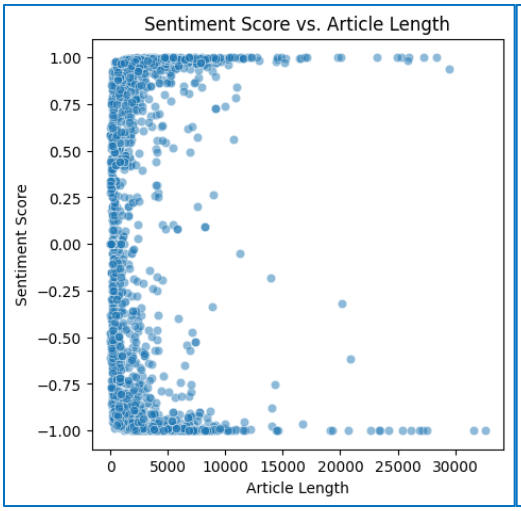
Sentiment Analysis

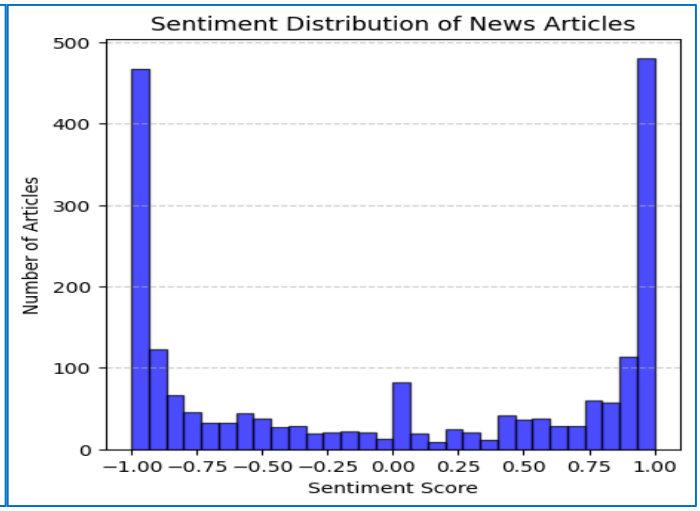
We Sentiment analysis for the study by measuring the polarity of each article's text using the TextBlob library. This box plot visualizes the distribution of sentiment polarity for both **Real** and **Fake** news labels. The spread and the position of the whiskers indicate a similar variability in sentiment for both categories. This similarity in sentiment distribution for both real and fake news suggests that sentiment alone may not be a strong discriminator between the two categories.



We also examined the sentiment scores in relation to the article length. The analysis shows that most articles fall within the 0-10,000-word range, with sentiment scores scattered between -1 and

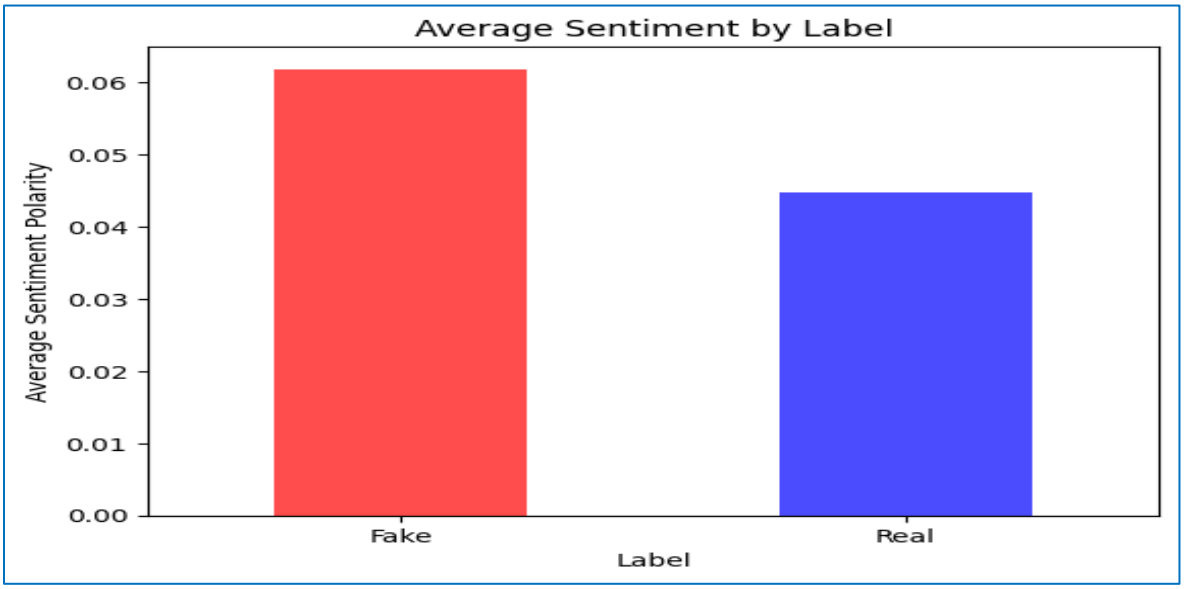
+1. Interestingly, articles with fewer than 5,000 words have sentiment scores spread across the entire range from -1 to +1. In contrast, articles with a length between 5,000 and 10,000 words tend to exhibit stronger sentiment, either closer to -1 or +1.





We also observed that most articles tend to have extreme sentiment scores, either -1 or +1.

We also analyzed the sentiment distribution between fake and real news. The results show that fake news tends to have a higher sentiment score of around **0.06**, while real news has a slightly lower sentiment score of approximately **0.045**. The slightly higher positivity in fake news could imply that fake news may use more emotionally charged language to engage readers. This difference in sentiment could be a potential indicator to help distinguish between real and fake news.



Vectorization

Prior to fitting machine learning models, articles were transformed into numeric representations. We applied three vectorization techniques to create different models.

1. Unigram vectorization focused on the frequency of individual words with minimum document frequency 5

2. while n-gram vectorization considered pairs or triplets of adjacent words to better capture local context with minimum document frequency 1 and

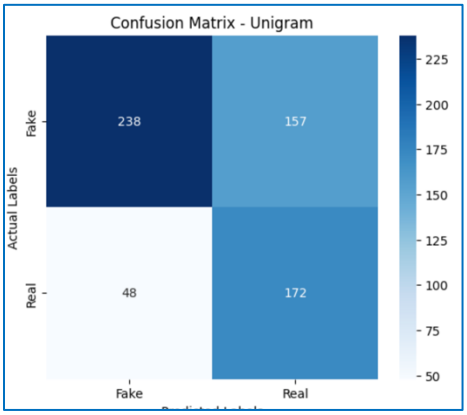
3. TF-IDF, on the other hand, emphasized terms that were common in individual articles yet infrequent across the entire dataset, thus highlighting discriminative words with document frequency 1.

Model Development and Evaluation

The data was split into training and testing sets to create our models. We took 70% of our data as train data and 30% data as our test data. Two main algorithms were explored: Multinomial Naive Bayes (MNB) and Support Vector Machines (SVM). MNB was trained using unigram, n-gram, and TF-IDF features, and its performance was evaluated using confusion matrices and classification reports. These visualizations clarified how effectively the classifier identified real and fake articles.

Multinomial Naive Bayes-

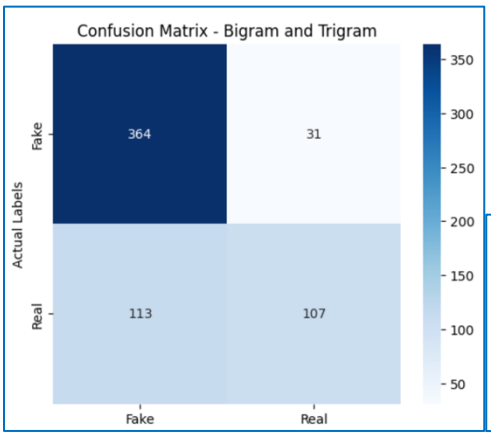
1. **<u>Unigram vectorized MNB-</u>**



	precision	recall	f1-score	support
Fake	0.83	0.60	0.70	395
Real	0.52	0.78	0.63	220
accuracy			0.67	615
macro avg	0.68	0.69	0.66	615
weighted avg	0.72	0.67	0.67	615

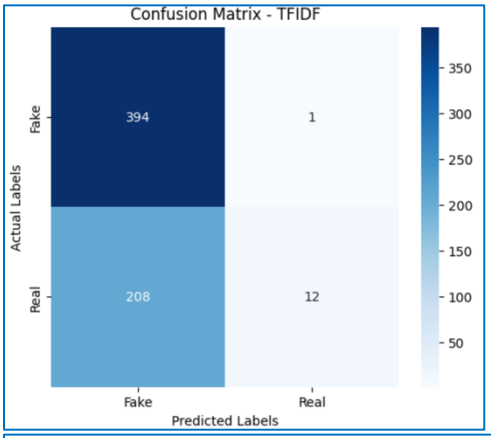
The classification report shows that the model performs better at identifying fake news, with a precision of 0.83 and a recall of 0.60, leading to an F1-score of 0.70. For real news, the model has a lower precision of 0.52 but a higher recall of 0.78, resulting in an F1-score of 0.63. The overall accuracy is 67%.

2. **<u>n-gram Vectorized MNB-</u>**



	precision	recall	f1-score	support
Fake	0.76	0.92	0.83	395
Real	0.78	0.49	0.60	220
accuracy			0.77	615
macro avg	0.77	0.70	0.72	615
weighted avg	0.77	0.77	0.75	615

The classification report shows that the model performs well with fake news, achieving a precision of 0.76, a high recall of 0.92, and an F1-score of 0.83. For real news, the precision is slightly higher at 0.78, but recall drops to 0.49, resulting in a lower F1-score of 0.60. The overall accuracy is 77%. The model is highly effective at identifying fake news but struggles with correctly predicting real news.



	precision	recall	f1-score	support
Fake	0.65	1.00	0.79	395
Real	0.92	0.05	0.10	220
accuracy			0.66	615
macro avg	0.79	0.53	0.45	615
weighted avg	0.75	0.66	0.54	615

The classification report indicates that the model performs well in detecting "Fake" news with a high precision of 0.65 and perfect recall of 1.00, meaning it correctly identifies nearly all fake news articles. However, its performance on "Real" news is poor, with a low recall of 0.05, meaning it misses most of the actual "Real" articles. The overall accuracy is 66%, This suggests that the model needs improvement in detecting "Real" news.

Compare all Three models of Multinomial Naive bayes- The **N-gram model** provides the best balance between accuracy and performance, with good recall for "Fake" and decent F1-score for both classes. The **Unigram model** performs moderately but is outperformed by the N-gram and TF-IDF models. The **TF-IDF model** struggles with a low recall for "Real" news despite its high precision.

Cross Validationon of all three MNB Model for check overfitting-

1. Unigram Vector:

- Individual words (unigrams) are used as features.
- The cross-validation scores vary between 0.63 to 0.72.
- Average Accuracy: 0.6818 (≈ 68.18%).

2. N-gram Vector:

- This considers sequences of N words instead of single words.
- Scores range between 0.62 to 0.70.
- Average Accuracy: 0.6667 (≈ 66.67%).

3. TF-IDF Vector:

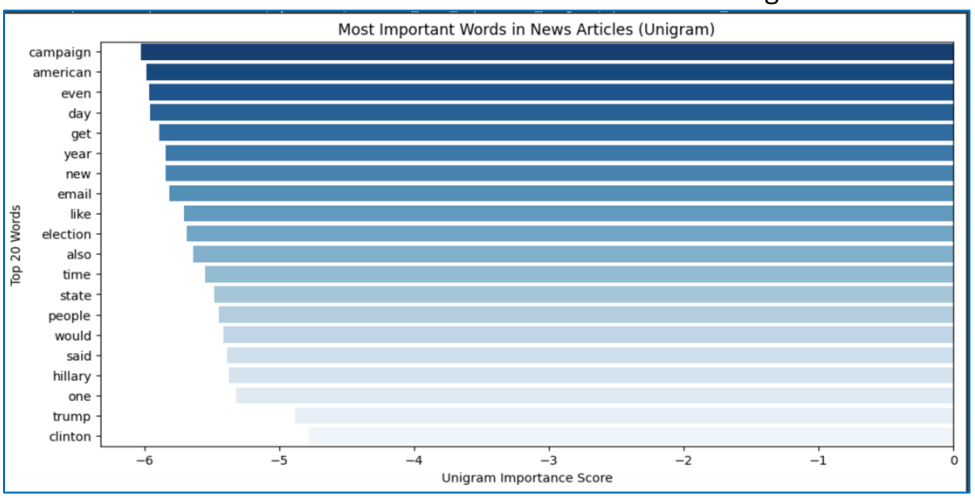
- This technique weighs words based on importance using Term Frequency-Inverse Document Frequency (TF-IDF).
- Scores range between 0.63 to 0.68.
- Average Accuracy: 0.6549 (≈ 65.49%).

We can see still unigram and TF-IDF model are getting approximately same result, but N-gram model is showing only 66% accuracy whereas the original n-gram model was showing 77% accuracy which suggest there is overfitting in this model. The model was highly accurate on the training set but performed significantly worse on unseen data. This suggests it memorized specific word sequences rather than understanding general language patterns.

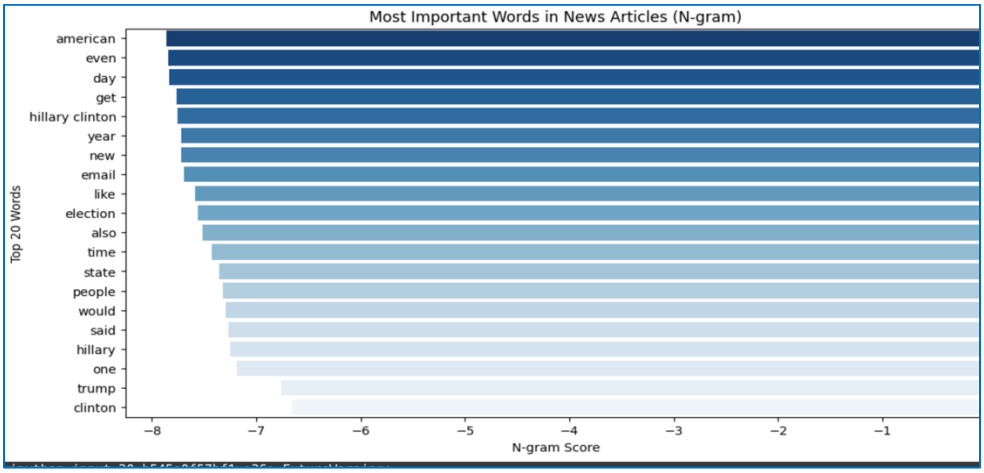
Top 20 important feature for each model-

The presence of terms like **'trump,' 'hillary,' 'clinton,'** and **'campaign'** in all models shows that political topics are central in both real and fake news, with these words helping the model to

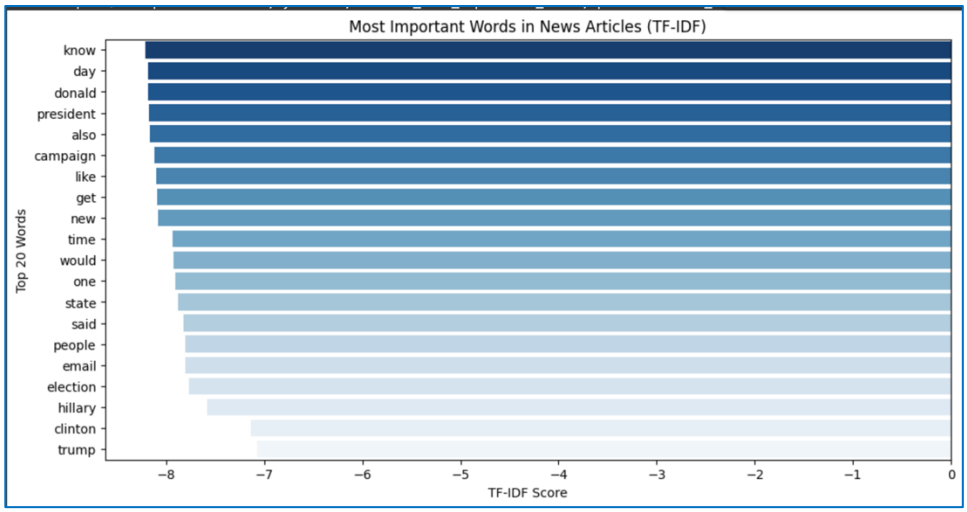
differentiate between the two based on their context and usage.



The top words here are similar to the unigram chart, suggesting consistent relevance in the dataset. The scores are more negative, possibly indicating a different weighting or scaling approach.



Words like "know," "day," "Donald," and "president" are significant here, indicating they are distinct but important in certain articles.

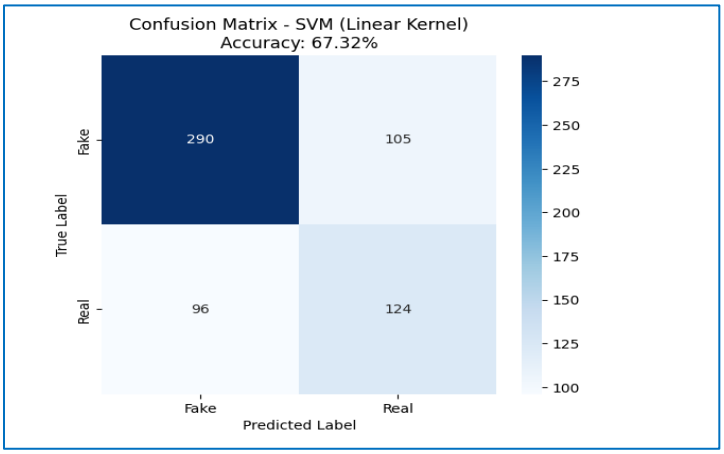


SVM Models-

SVM experiments included comparisons between linear and RBF kernels, as well as the incorporation of different vectorization features. The results, illustrated in confusion matrices and accuracy plots, demonstrated that certain configurations (such as RBF kernel combined with Boolean features) provided higher accuracy than simpler approaches.

First experiment SVM with term frequency features and linear kernel

The first SVM model train was ran with a linear kernel and term frequency feature. This model displayed an overall accuracy of 67.32%. Classifying 290 true fake reviews and 124 true real reviews.

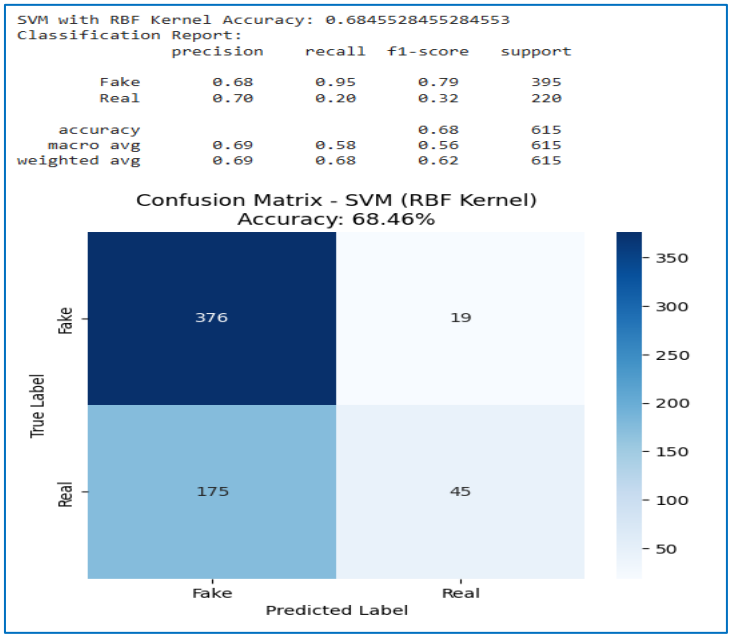


The classification report shows that fake class performed better: Higher precision (75%) and recall (73%).

```
Classification Report:
              precision    recall  f1-score   support

        Fake       0.75      0.73      0.74       395
        Real       0.54      0.56      0.55       220

    accuracy                           0.67       615
   macro avg       0.65      0.65      0.65       615
weighted avg       0.68      0.67      0.67       615
```

Second SVM Experiment was changed to have RBF kernel and improved the accuracy slightly to 68.46%

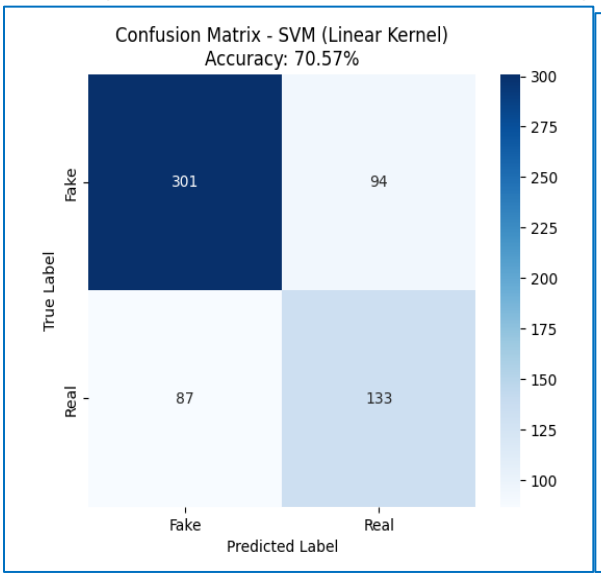
- Fake Precision (68%) vs. Fake Recall (95%)
  The model catches almost all Fake reviews (high recall) but at the cost of some misclassifications (moderate precision).
- Real Precision (70%) vs. Real Recall (20%)
  The model is very cautious when predicting Real reviews, which explains why Real recall is so low.

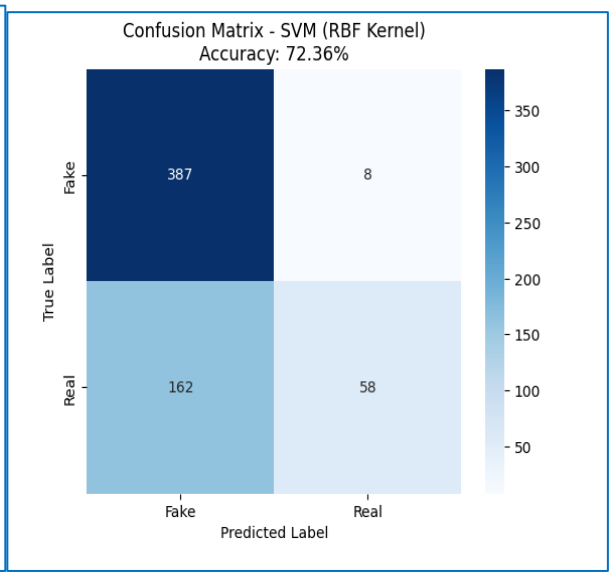


SVM with Boolean features

The next two SVM Models were ran with boolean features to test if there was any improvement with this type of vectorization method. Same two kernels were used “Liner” and “RBF “The model’s ran with Boolean features showed improvement with both kernels. The linear kernel had an overall

accuracy of 70.57% while RBF had an accuracy of 72.36%.





The precision scores were better on the model with RBF Kernel which had 70% for fake reviews and 88% for real reviews. Meaning that when the model predicts a review as real, it's correct 88% of the time. The model is highly selective and only predicts "Real" when it is very sure. As a result, it rarely predicts "Real" (hence, recall is low), but when it does, it's mostly correct (high precision).

SVM with Linear Kernel (Boolean Features) Accuracy: 0.7056910569105691
Classification Report:

	precision	recall	f1-score	support
Fake	0.78	0.76	0.77	395
Real	0.59	0.60	0.60	220
accuracy			0.71	615
macro avg	0.68	0.68	0.68	615
weighted avg	0.71	0.71	0.71	615

SVM with RBF Kernel (Boolean Features) Accuracy: 0.7235772357723578
Classification Report:

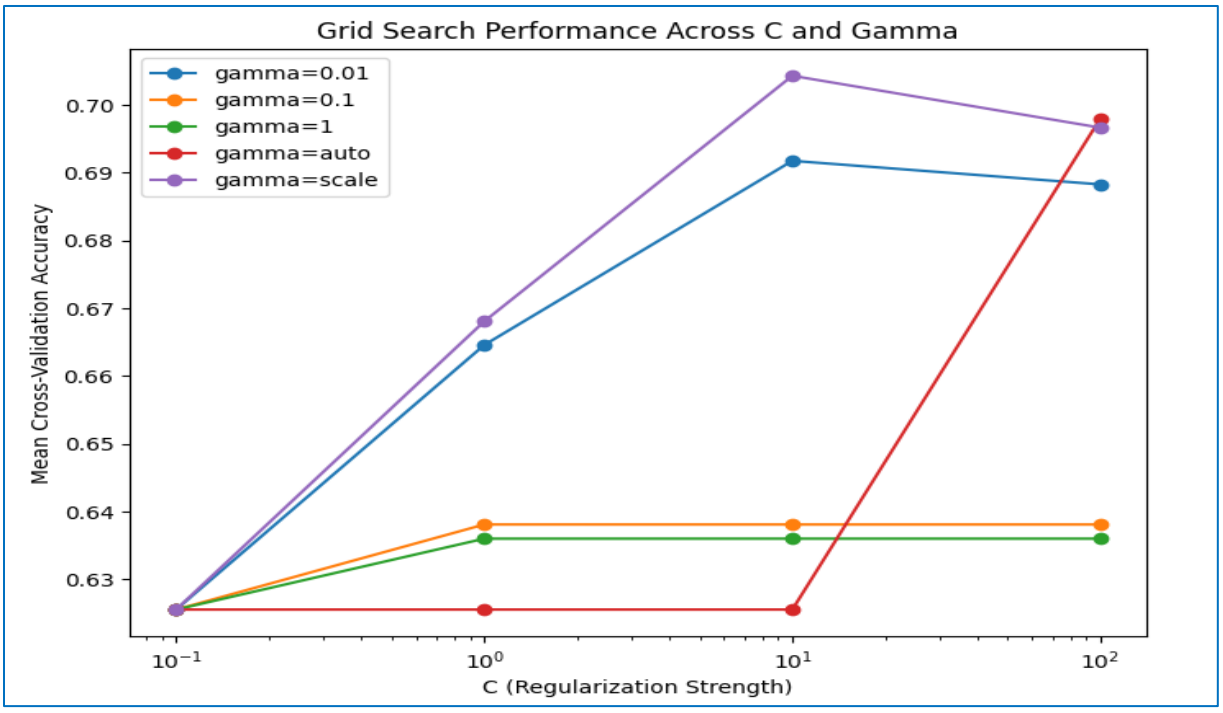
	precision	recall	f1-score	support
Fake	0.70	0.98	0.82	395
Real	0.88	0.26	0.41	220
accuracy			0.72	615
macro avg	0.79	0.62	0.61	615
weighted avg	0.77	0.72	0.67	615

To further refine performance, hyperparameter tuning was conducted via GridSearchCV, systematically searching combinations of parameters to discover optimal settings. This final set of tuned parameters led to notable improvements in classification, underscoring the significance of careful model optimization. These hyperparameters were added to the best SVM Model which was Boolean features with an RBF Kernel.

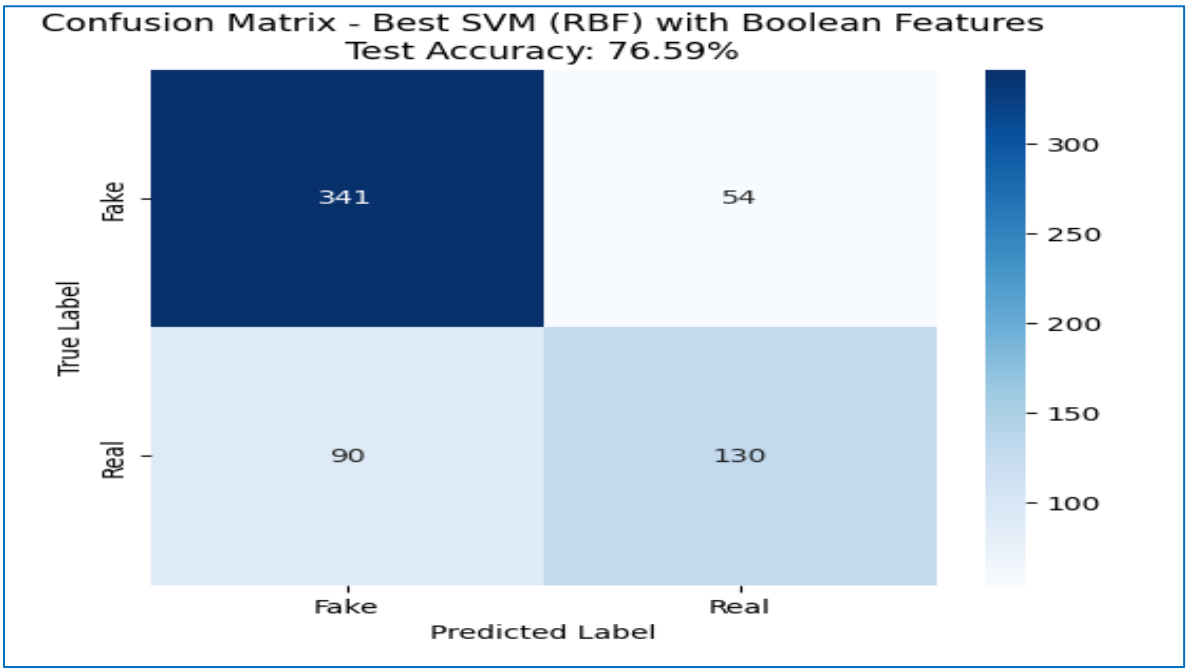
The hyperparameter grid for tuning was defined as follows:

Regularization strengths C if (0.1, 1, 10, 100) since the kernel being used is RBF kernel gamma options needed to be added. We defined the gamma options as scale, auto, 0.01, 0.1, and 1.

The grid search was performed using a 5- cross validation. The best hyperparameter combination found was a C strength of 10 and gamma set to scale.



The model correctly classifies 76.58% of all reviews. This is the best accuracy achieved so far, showing improvement.



The classification report results

- Precision: 79% → When the model predicts "Fake," it is correct 79% of the time.
- Recall: 86% → The model correctly identifies 86% of all actual Fake reviews.
- Precision: 71% → When the model predicts "Real," it is correct 71% of the time.
- Recall: 59% → The model correctly identifies only 59% of Real reviews (better than before but still needs improvement).

```
Best Model Test Accuracy: 0.7658536585365854

Classification Report:
              precision    recall  f1-score   support

        Fake       0.79      0.86      0.83       395
        Real       0.71      0.59      0.64       220

    accuracy                           0.77       615
   macro avg       0.75      0.73      0.73       615
weighted avg       0.76      0.77      0.76       615
```

Conclusion

In conclusion, we applied fake news detection with the assistance of various machine learning techniques and text mining practices. We have utilized a large set of data preprocessing, exploratory data analysis (EDA), vectorization techniques, and various machine learning algorithms to identify how efficient it is to distinguish between real news and fake news. In addition, we also performed topic modeling for understanding the context and theme of our data better.

We trained MNB and SVM machine learning models to identify fake and real news. We ran several models with different vectorization features to test which combination of features provided optimal results. The results showed that the best multinomial model using n-gram features had an overall score of 77% accuracy and the best SVM model was found using grid search defining hyperparameters and 5 cross validations with boolean features on a RBM kernel work best with approximately 77% of accuracy. Both models proved to be just as equally accurate, but based on precision and recall values the SVM seem to outperform the multinomial navies bayes model. In addition, we tested with 5 cross validations on the multinomial and the N-gram multinomial did not pass the overfitting test. Therefore, we consider SVM as our best Model.

The real challenge in this model was the data set, there was an imbalance between the quantity of real v fake news and the strong focus on the 2016 U.S. Presidential Election, which limits the model's ability to generalize to non-political or more diverse fake news contexts. An additional challenge the similarity in vocabulary between real and fake news suggests that more nuanced feature extraction techniques may be needed.

In the future, the model could be improved and obtain better results if it Incorporates more diverse and balanced datasets, which can help generalize the model for broader applications. We can also try Techniques like transformer-based models (BERT) to better capture contextual nuances. We can also train the model to not only learn the linguistic patterns but also the published time pattern of when real and fake news is published.

In summary, while our models demonstrated moderate success in detecting fake news, the complexity of misinformation suggests a need for more advanced, context-aware approaches. As fake news continues to evolve, developing more sophisticated detection methods is crucial for maintaining information integrity in digital media.