Detecting Fake News Using Machine Learning

Arti Ravi Garg, Maybel Herrera, Rahul Gurujapu

What is fake news?

- False or misleading information presented as news.
- Often used to manipulate opinions, spread misinformation, or generate revenue.

Why it matters?

- Affects public perception and decision-making.
- Undermines trust in credible journalism.
- Has significant societal, political, and economic implications.

Goal

With AI rapidly advancing, fake news and AI-generated videos and images have become increasingly common, making it difficult to distinguish between reality and deception. Whether it's a heartwarming animal rescue, a violent attack, or a political statement, AI-generated content can blur the line between fact and fiction. As technology evolves, the ability to detect fake news becomes more crucial than ever. Without reliable detection methods, misinformation can manipulate public perception, influence important decisions, and erode trust in media. Developing AI-driven tools to combat fake news is essential to maintaining an informed and responsible society.

About the Data

Kaggle Fake News
https://www.kaggle.com/datasets/ruchi798/source-based-news-classification

- news_articles.csv

10 column 2047 Instances :

Column names are:

Author	title
text	language
Type	label
hasImage	date
time	timezone

Data clean up and transformation included :

-Remove columns that would not be needed for the analysis.

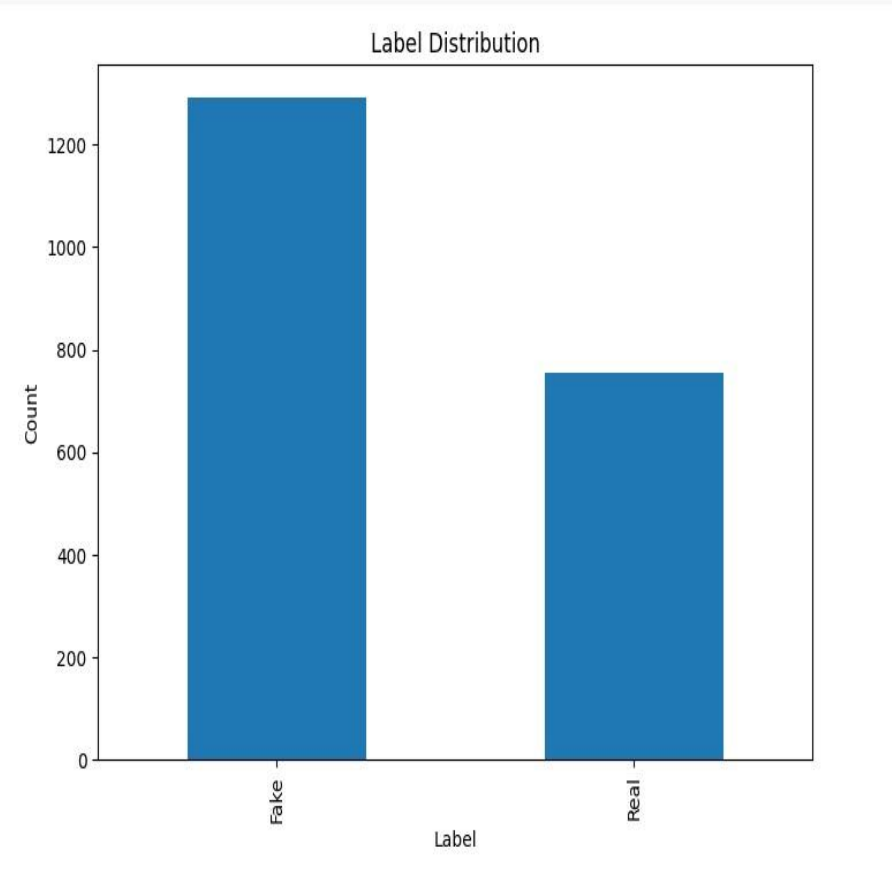
-Remove rows with null values

-Transformed published column: Split the column in to 3 separate columns: date, time and timezone.

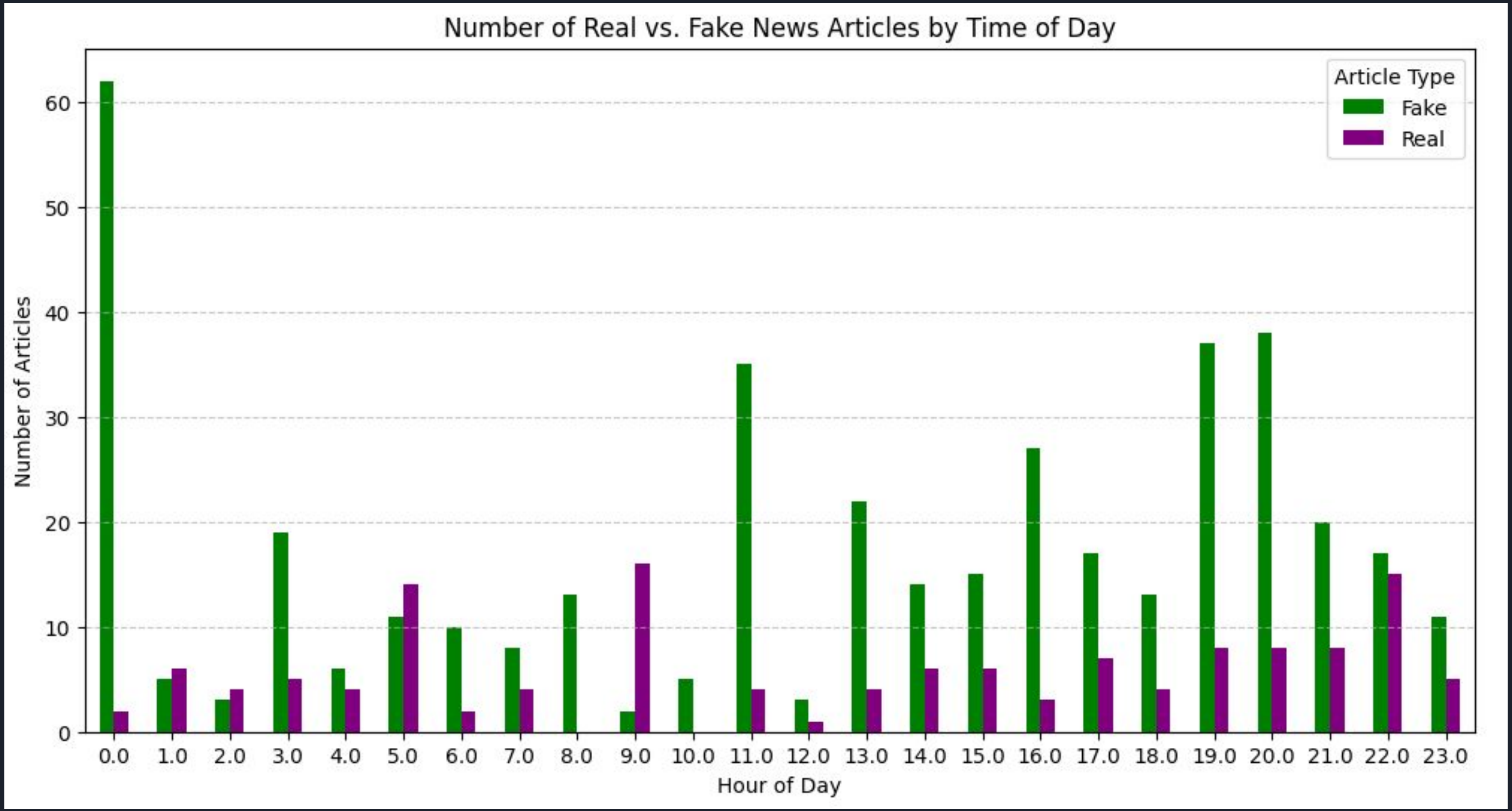
Initial Data exploration

Label Column

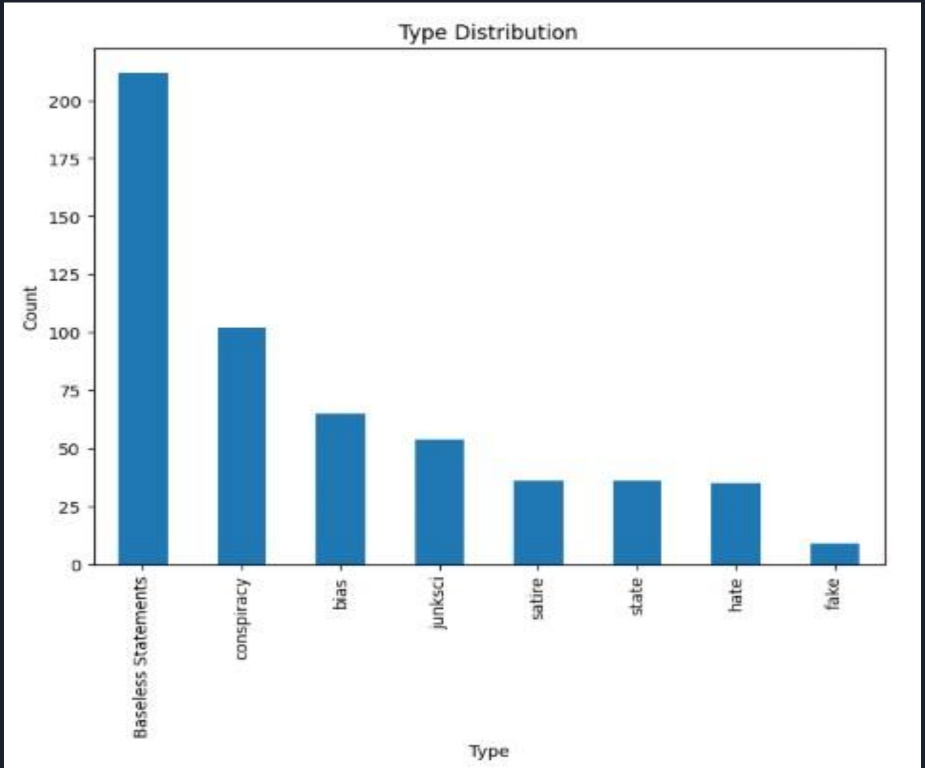
Data set has about 1200 Fake news
800 Real news
Challenge- class imbalance with the data set having more fake news. This might be a problem when training the model.

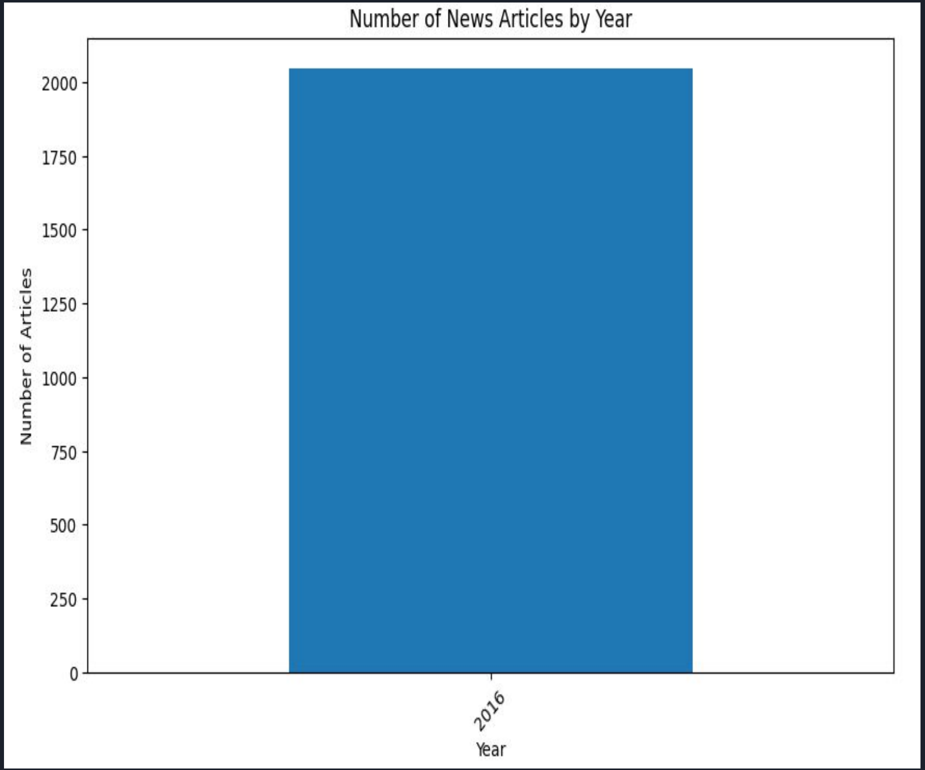


Peak of articles are published in the night , specifically at midnight. Second peak happens close to noon and third peak is in the evening.



News articles were all written in 2016 and the two most common themes were baseless statements and conspiracy.





Pre-processing

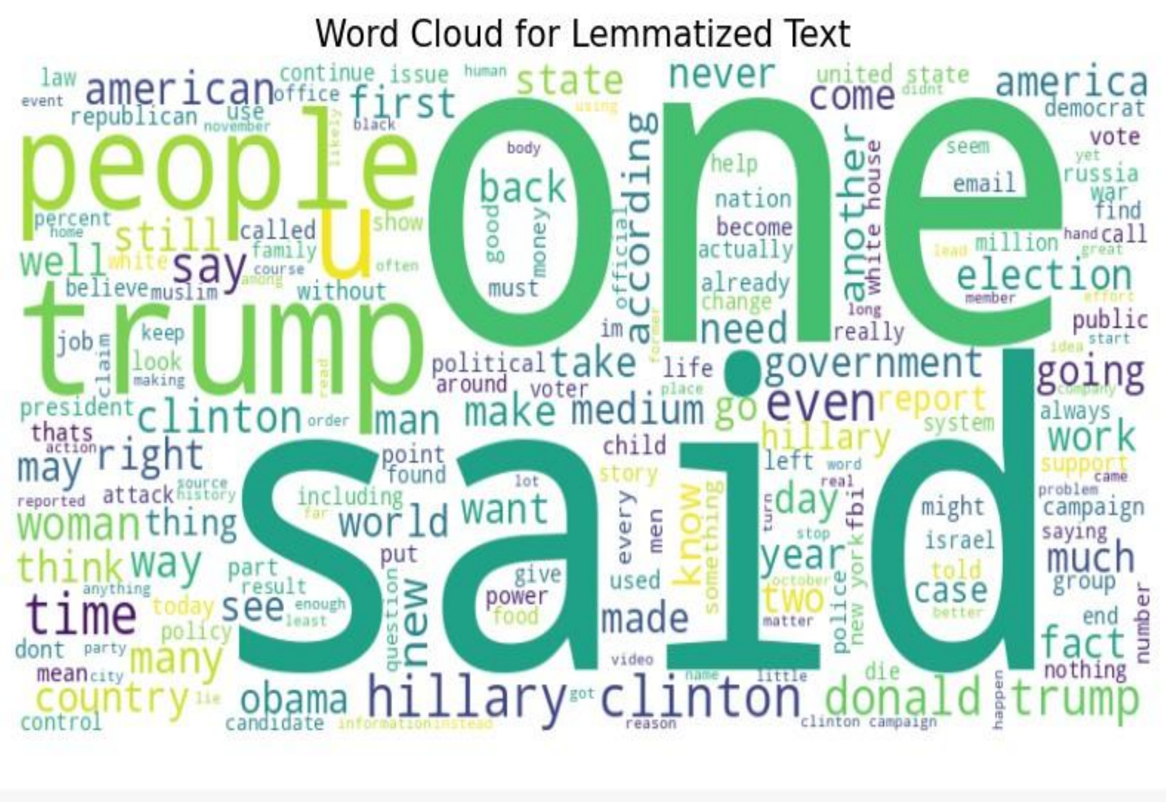
- Removed special characters
- Convert to lowercase
- Tokenized
- Removal of stopwords
- Converted words to their base form using lemmatization

Word Clouds

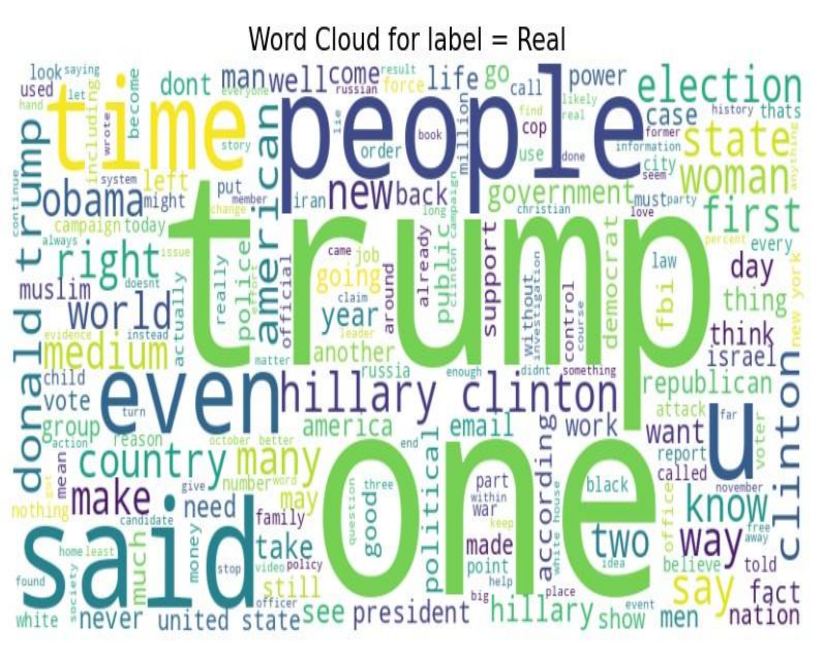
Key Terms in 2016 News Coverage

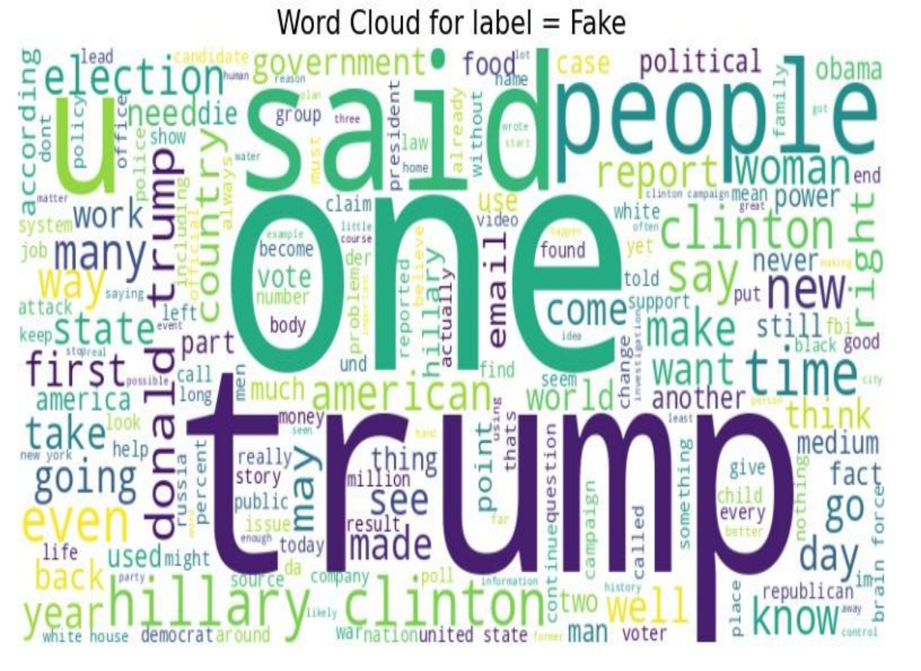
The news articles were published in 2016 . This word cloud reflect the most frequent terms. Which reflect the focus of the news articles to be political. The most common terms are :
-One
-Said
-people
-trump

The word ‘one’ and ‘said’ indicates a lot of singular statements.


Word Cloud for Lemmatized Text

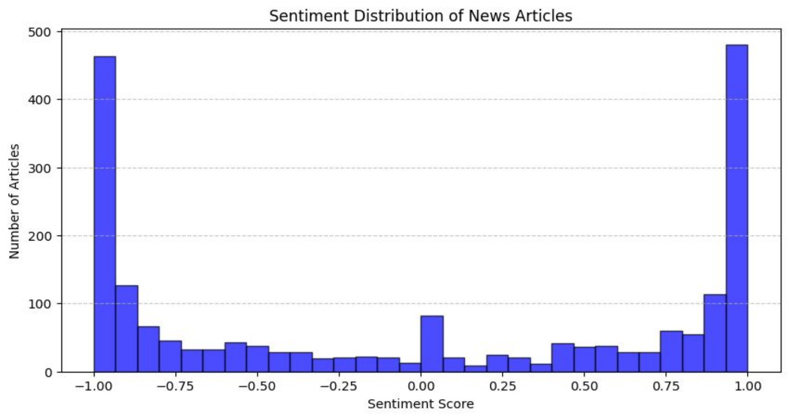
Comparing Language in Real vs. Fake News



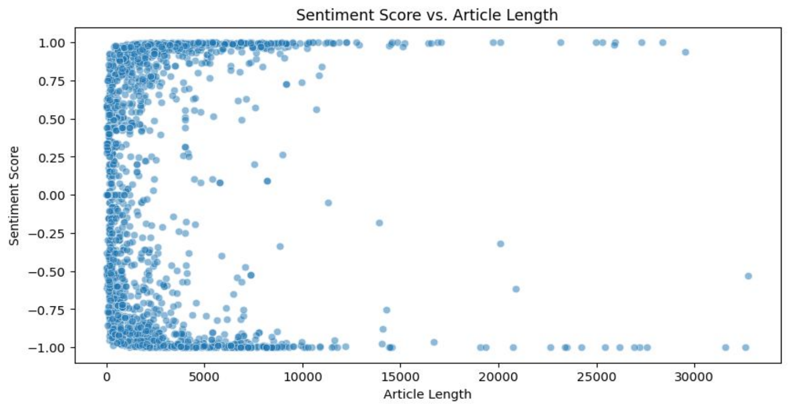


Both real and fake news articles prominently feature words like 'Trump,' 'one,' 'said,' and 'people.'

Analyzing Sentiment



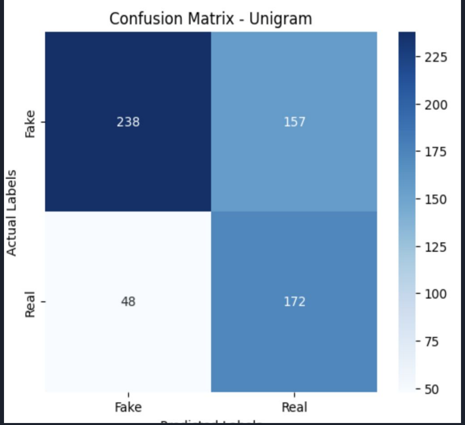
Sentiment Distribution of News Articles: A histogram showing the distribution of sentiment scores in news articles. The majority of articles have extreme sentiment scores (-1 or 1), with fewer articles having neutral or moderate sentiment.

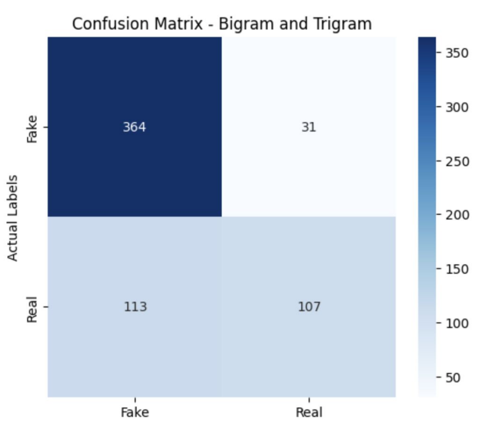


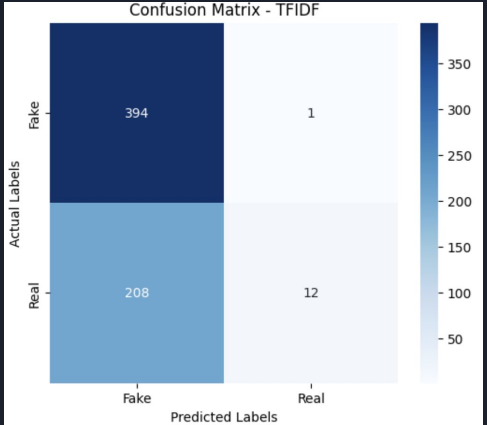
Sentiment Score vs. Article Length: A scatter plot showing the relationship between sentiment scores and article length. The points are concentrated at extreme sentiment scores (-1 and 1), with varying article lengths.

Naive Bayes

Naive Bayes: MultiNomial







Accuracy:67%

Accuracy:77%

Accuracy: 66%

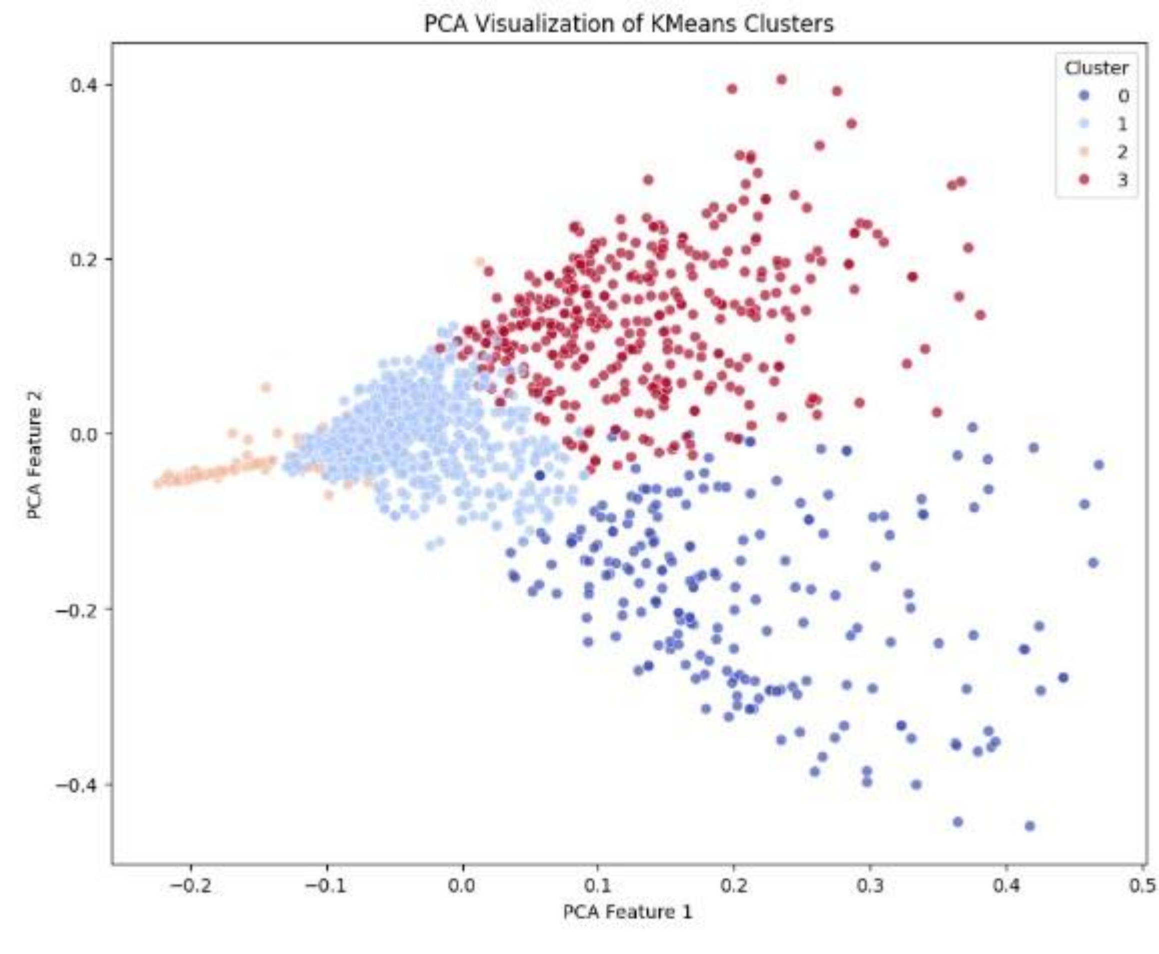
k-Means Clusters

k-Means Clusters

- This visualization represents the results of K-Means clustering using PCA for dimensionality reduction.

-Each color represents a different cluster.The clusters indicate natural groupings within the dataset based on textual similarities.

-Optimal Clusters: A group of 4 clusters was identified as the best fit since there is very little overlap.

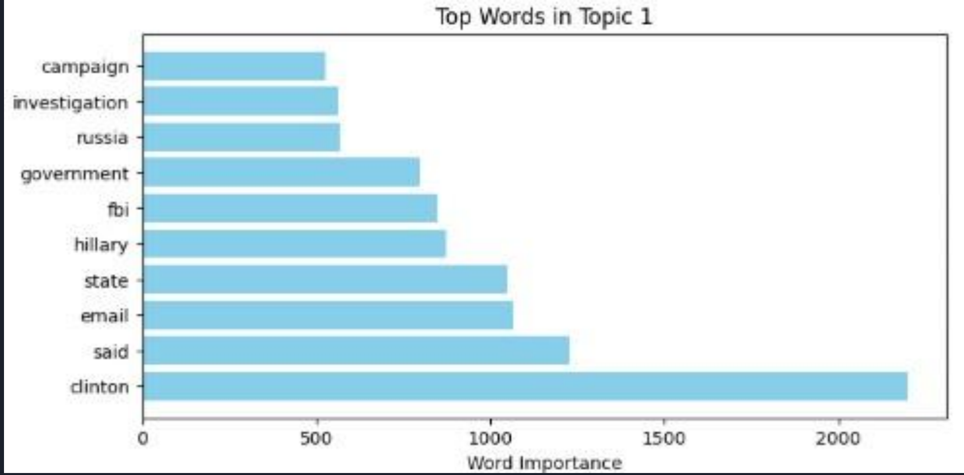


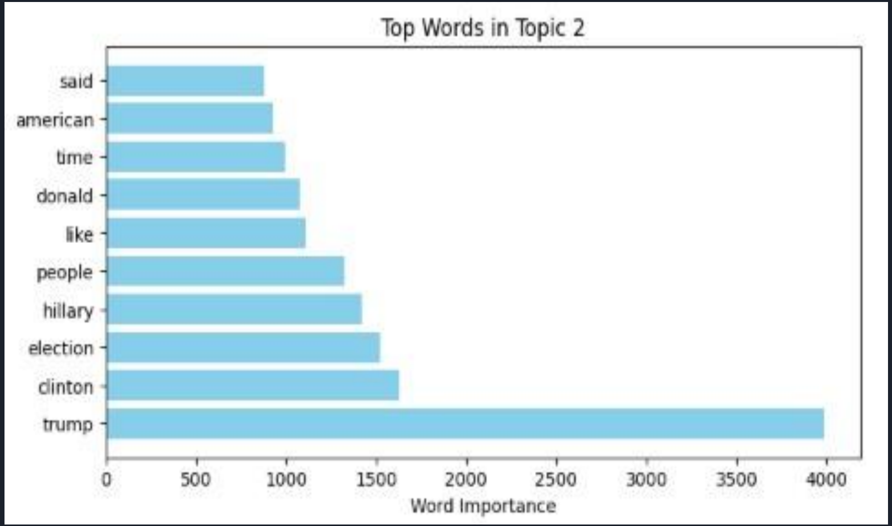
Topic Modeling using Scikit-learn LDA

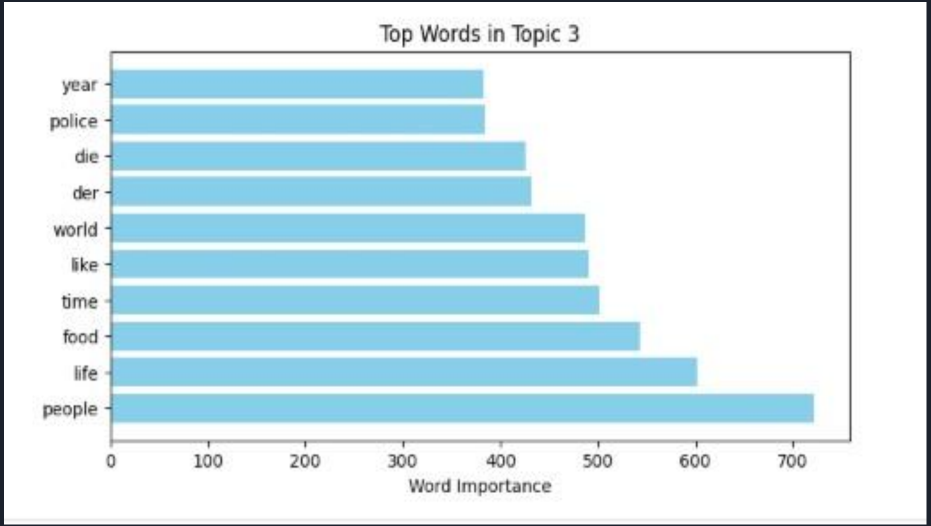
Topic 1 Label : Hillary Clinton Email Investigation & 2016 Election Controversy

Topic 2 label : 2016 U.S. Presidential Election & Political Communication

Topic 3 Label : Global Issues: Society, Life, and Law Enforcement

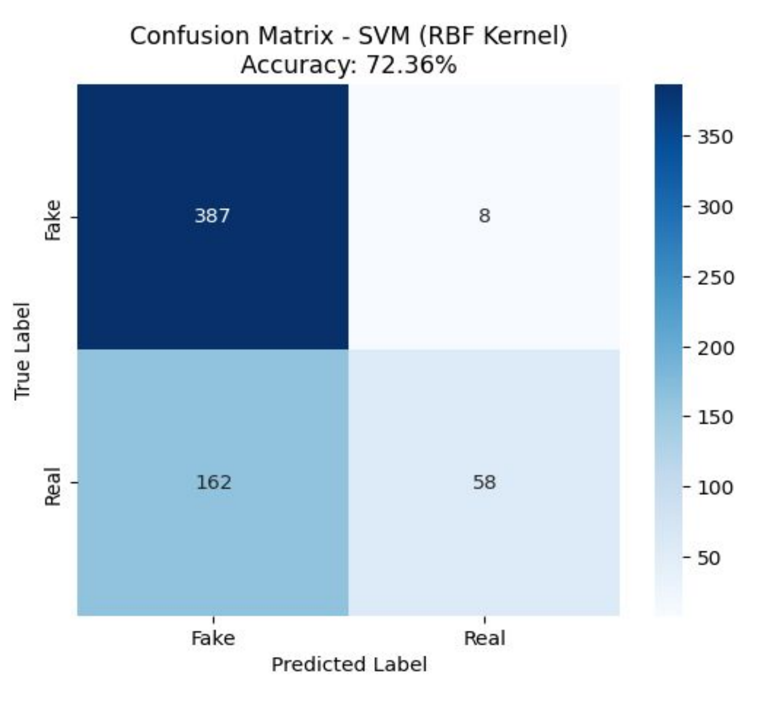






SVM Model

One of our top performing models was the SVM (sector vector machines)using Boolean features and RBF kernel



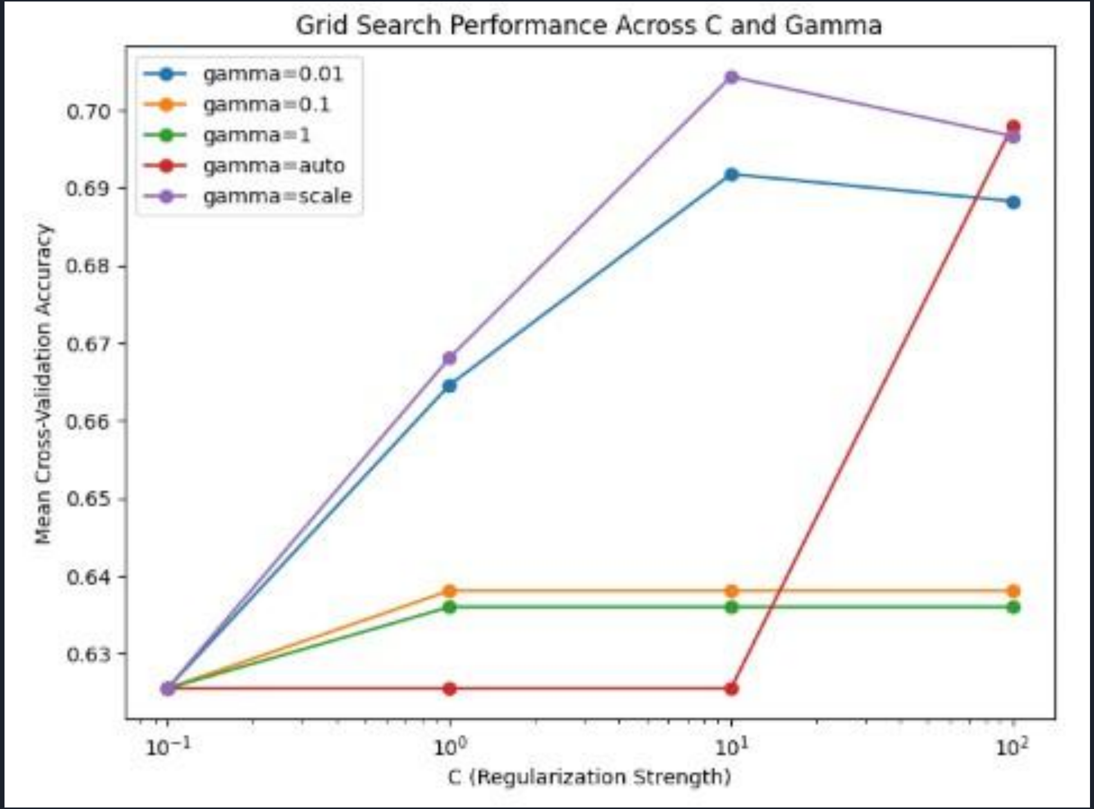
Grid search Performance graph

The graph shows how accuracy changes with different C and gamma values.

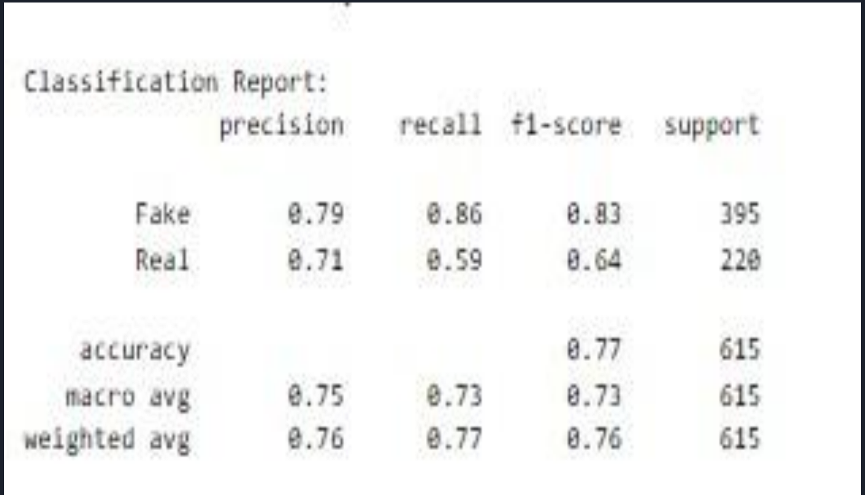
5-Fold Cross-Validation was used to evaluate model performance across different hyperparameters.

Best Hyperparameters Found:
Regularization Strength (C): 10
Gamma: Scale

This combination achieved the highest cross-validation accuracy.



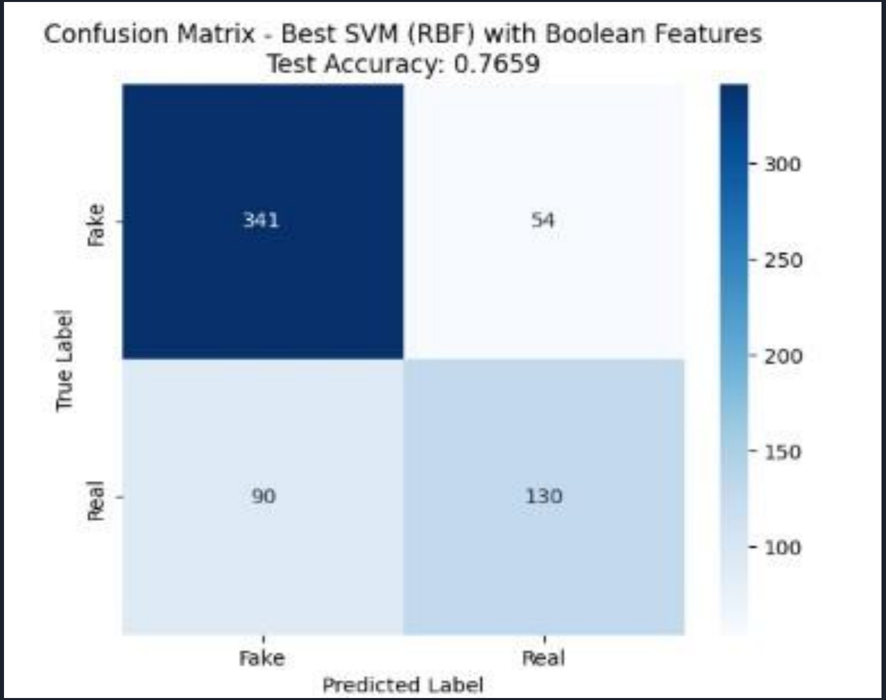
SVM Model with RBF Kennel



```
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
```



Confusion Matrix - Best SVM (RBF) with Boolean Features
Test Accuracy: 0.7659

Questions?