Fake News Detection - Semantic Processing

Authors:
Raja Ravi Sekar A
Rajiv Gaba

Table of Contents **4**

Objective

The objective of this assignment is to develop a Semantic Classification model for Fake News Detection

True.csv dataset includes 21,417 true news, while the Fake.csv dataset comprises 23,502 fake news.

Data Preparation

- Data Understanding, Add new column, Merge DataFrames, Handle the null values, Merge the relevant columns, and drop the rest from the DataFrame
- The data is merged, and the null values of the merged dataset

Null values in each column of the merged:

title	21
text	21
date	42
label	0

Text Preprocessing

The new DataFrame named ‘processed’ is created, and the data is converted to lowercase, removing text in square brackets, punctuation, and words with numbers

POS Tagging and Lemmatization

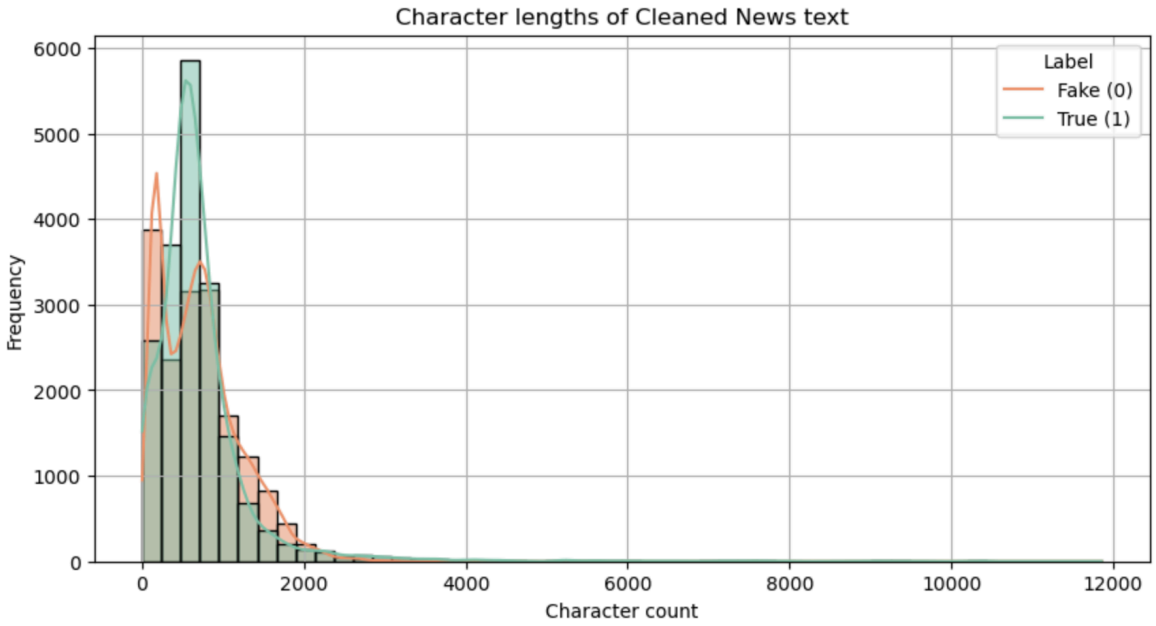
- Created the function for POS tagging and lemmatization, filtering stopwords, and keeping only NN and NNS tags using Spacy and the cleaned data is saved as csv file

Train Validation Split

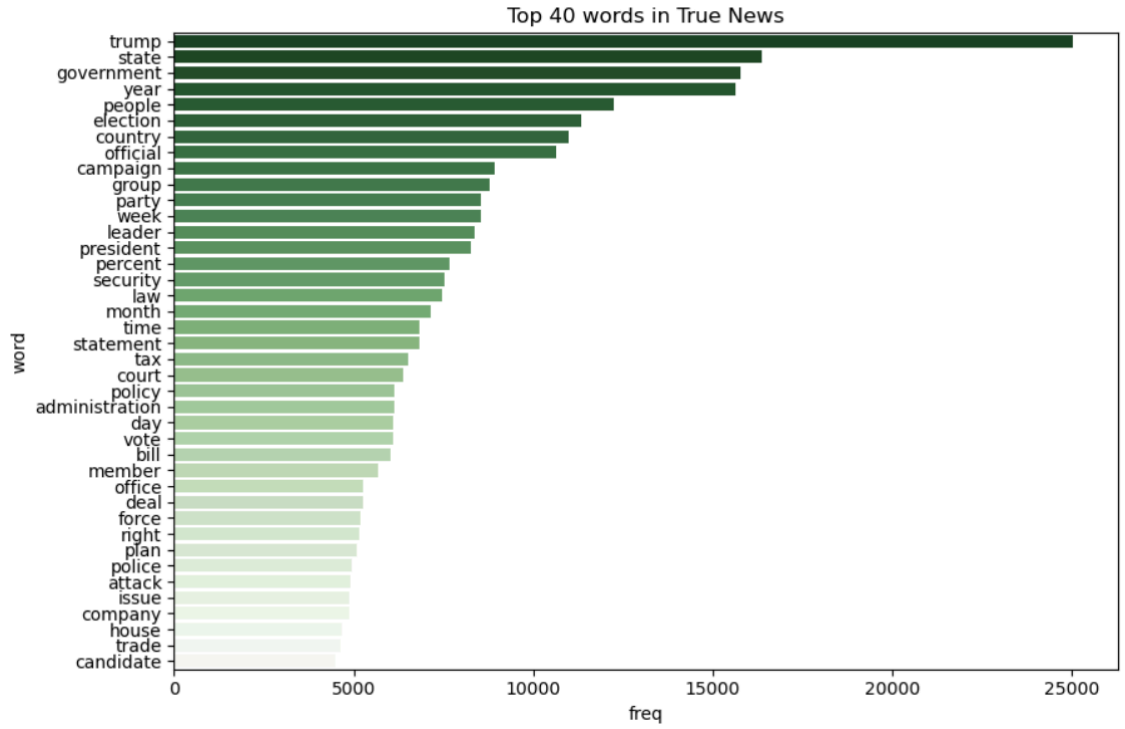
- The data is split into Train and Validation

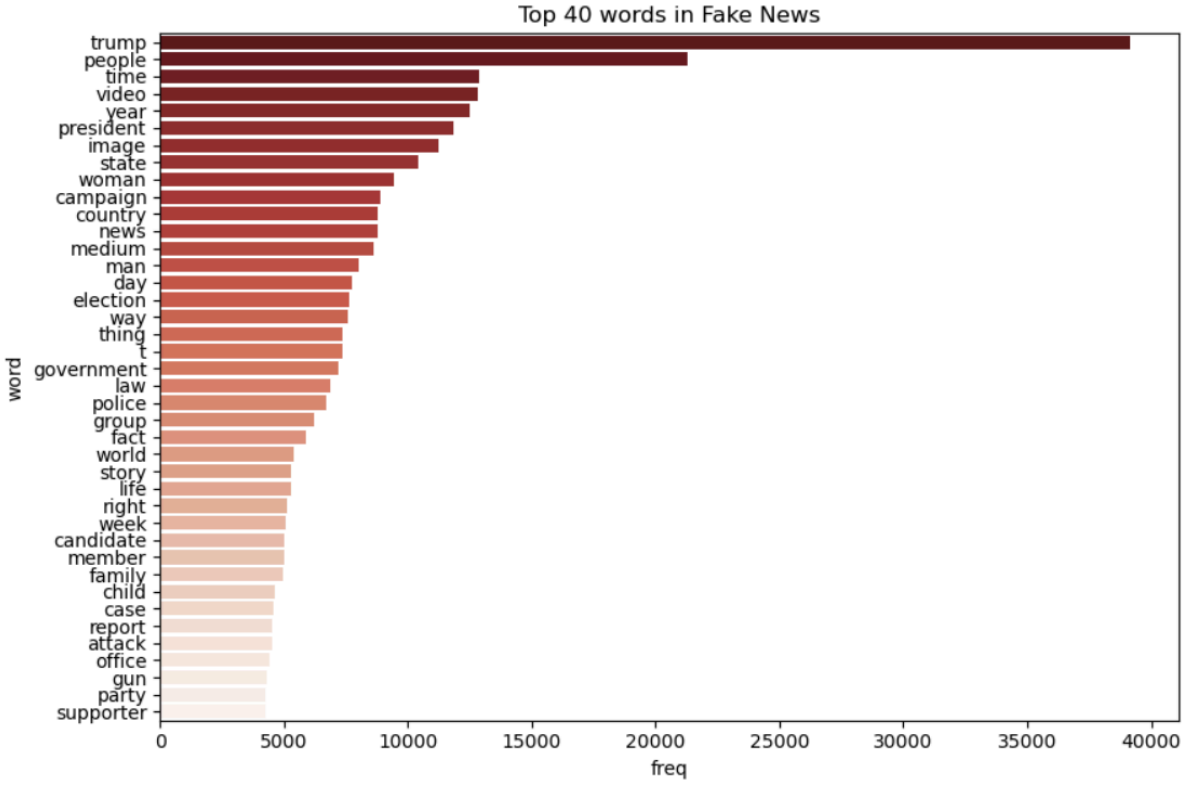
Exploratory Data Analysis on Training Data

Character length graphs for cleaned news text and lemmatized news text with POS tags removed



Top 40 words by frequency among true and fake news in the Training data after processing the text

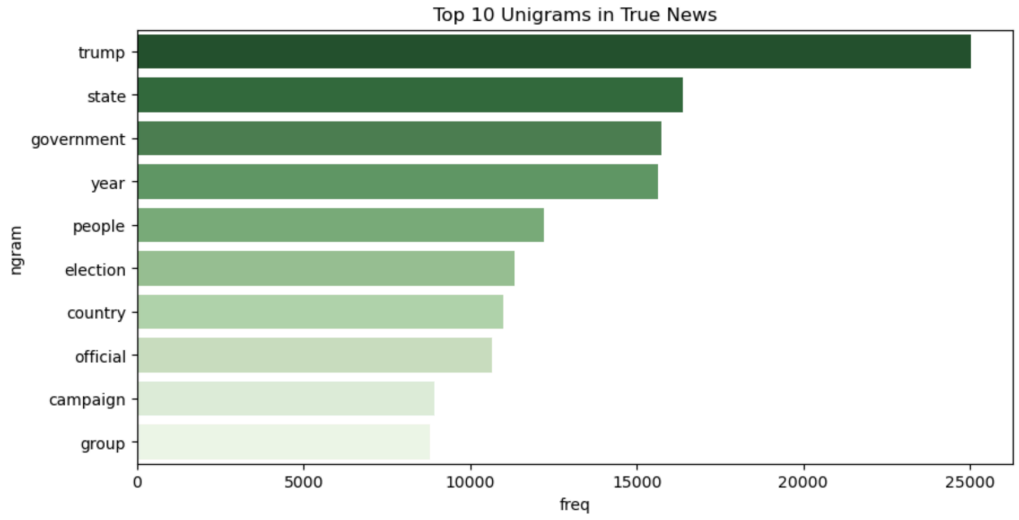


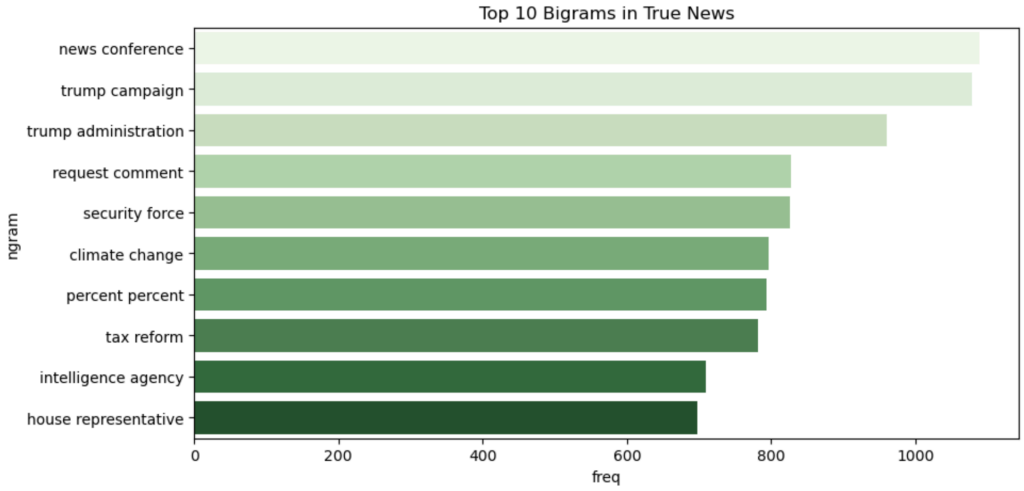


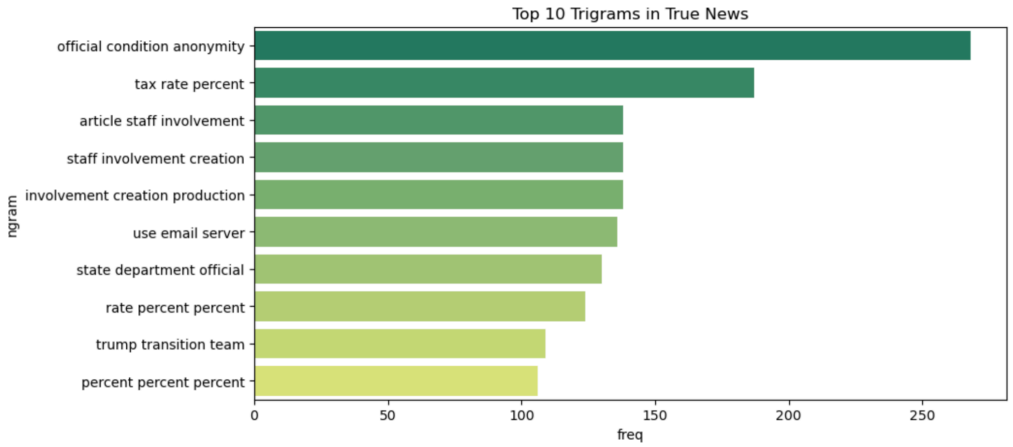
WordCloud for True News

WordCloud for Fake News

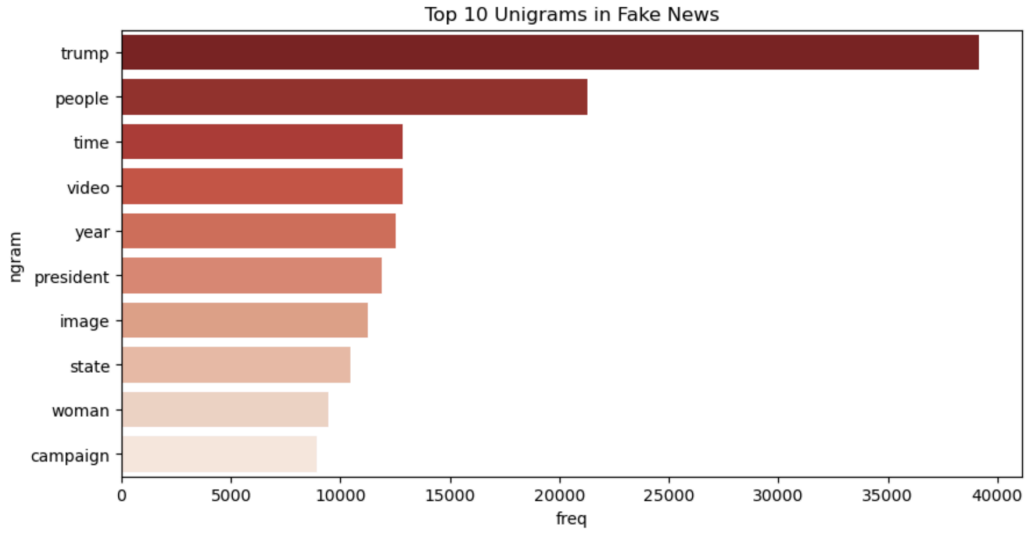
Top unigrams, bigrams, and trigrams by frequency in true news after processing the text

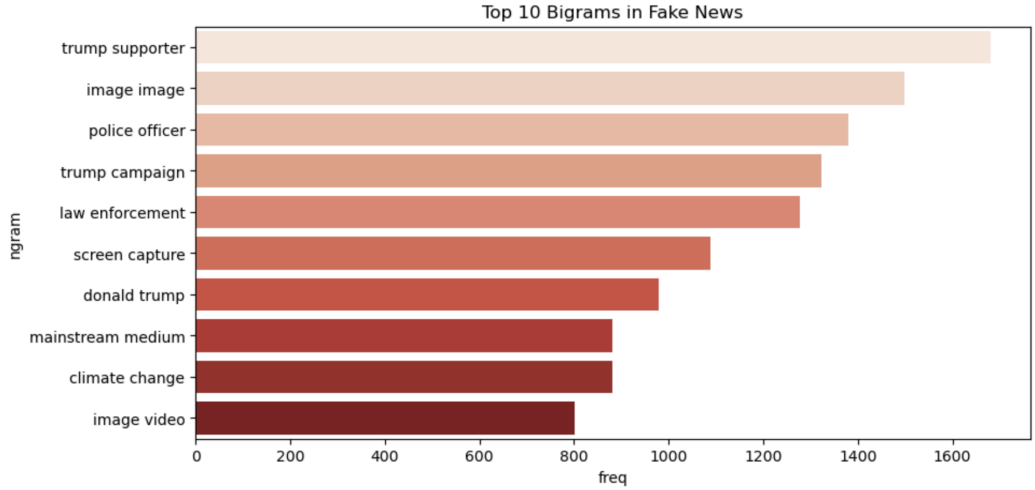


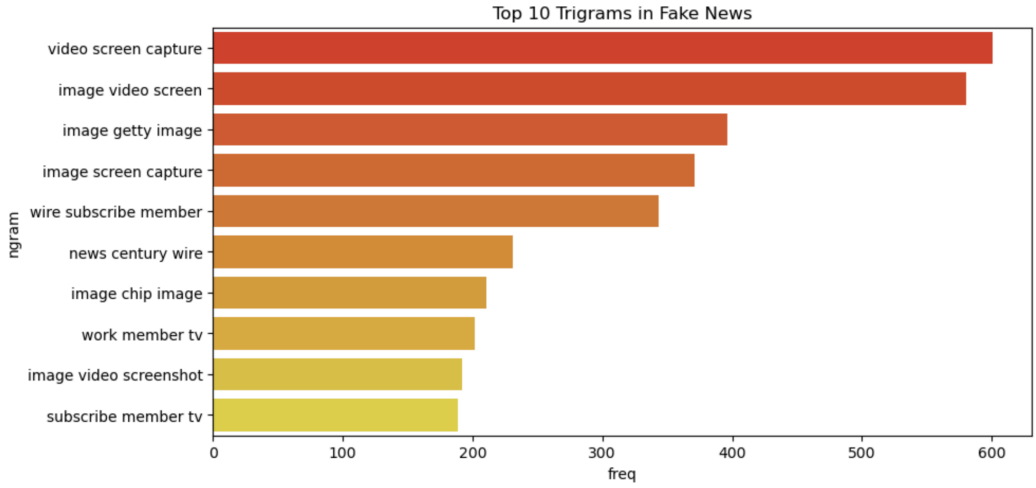




Top unigrams, bigrams and trigrams by frequency in fake news after processing the text

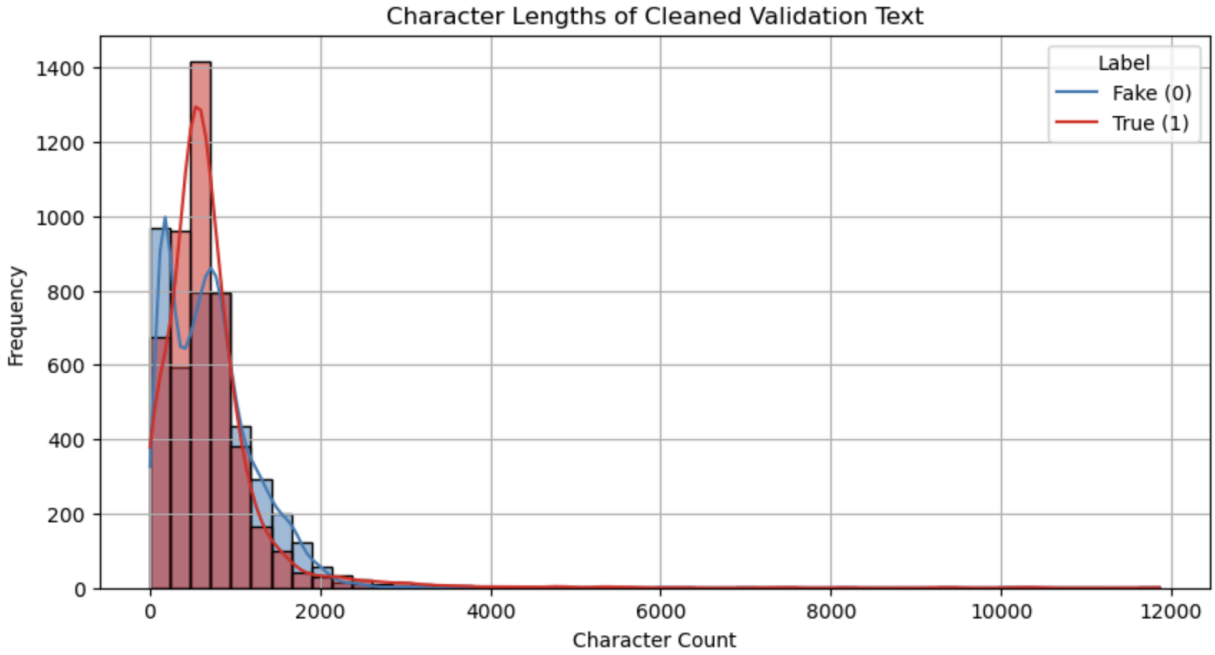




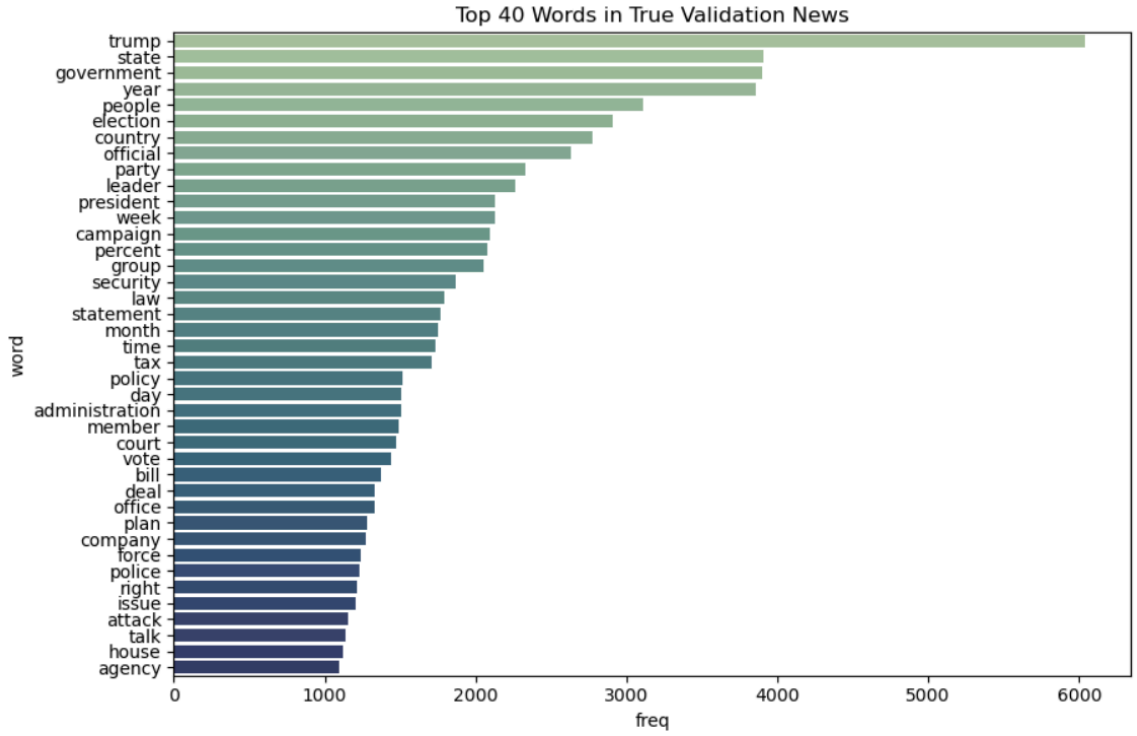


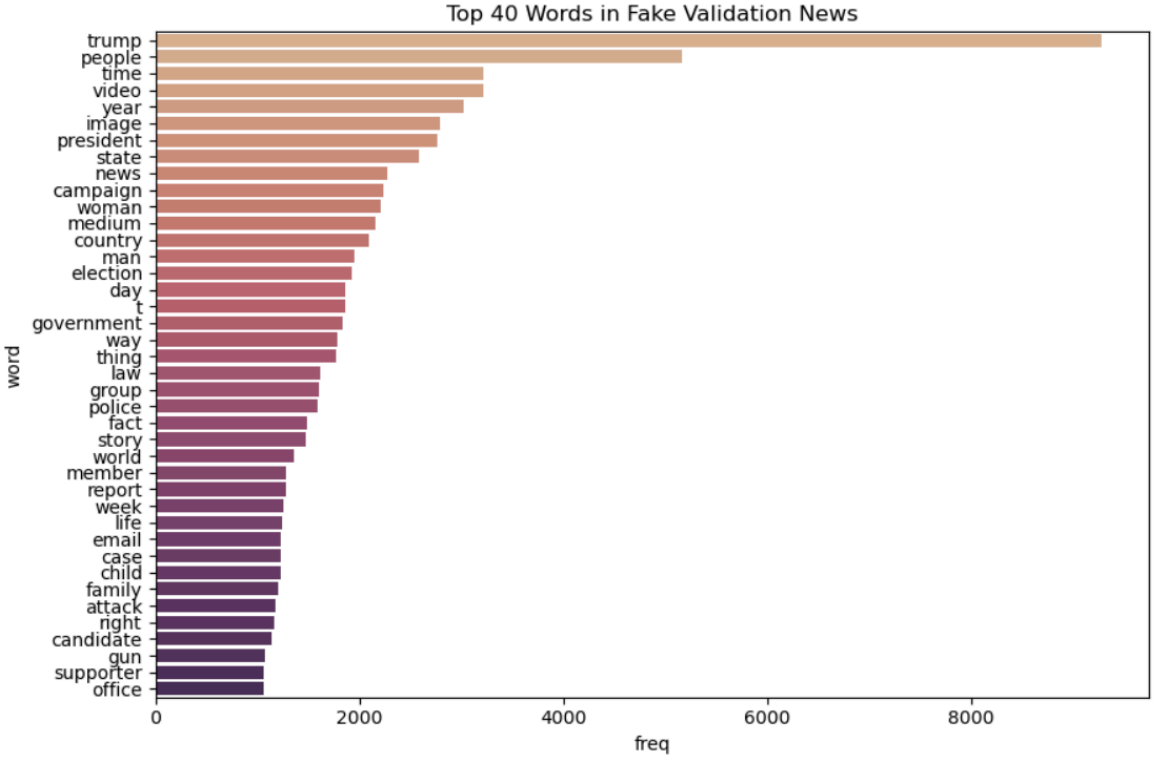
Exploratory Data Analysis on Validation Data

Character lengths of cleaned news text and lemmatized news text with POS tags removed

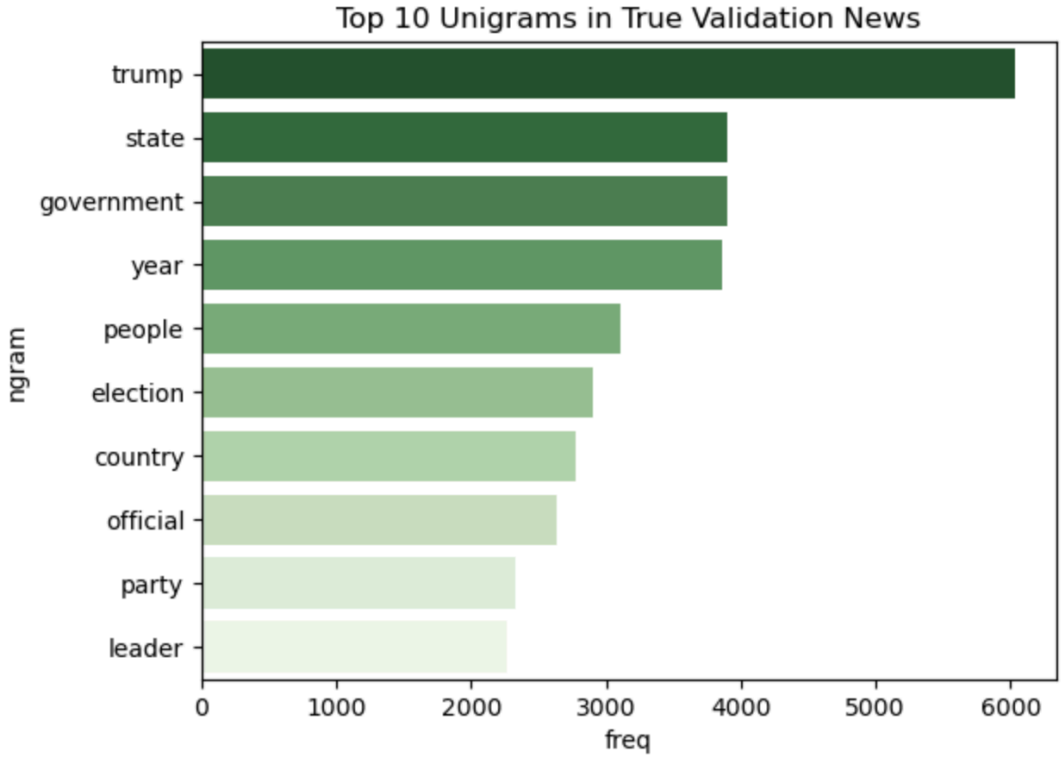


Top 40 words by frequency among true and fake news after processing the text

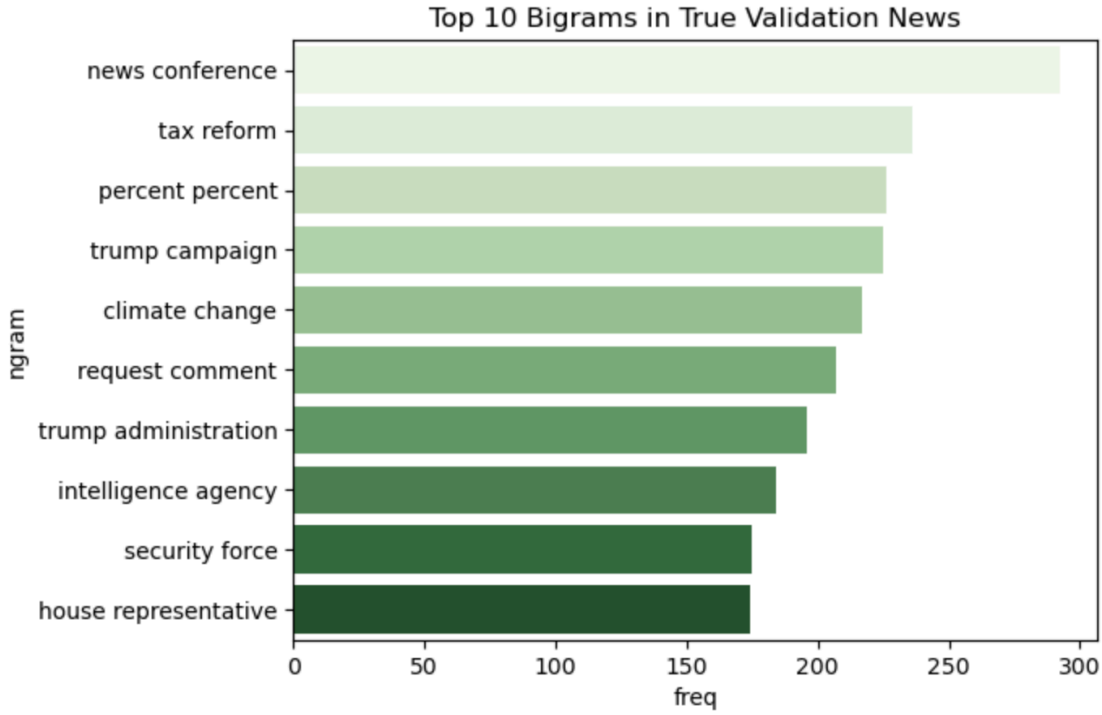




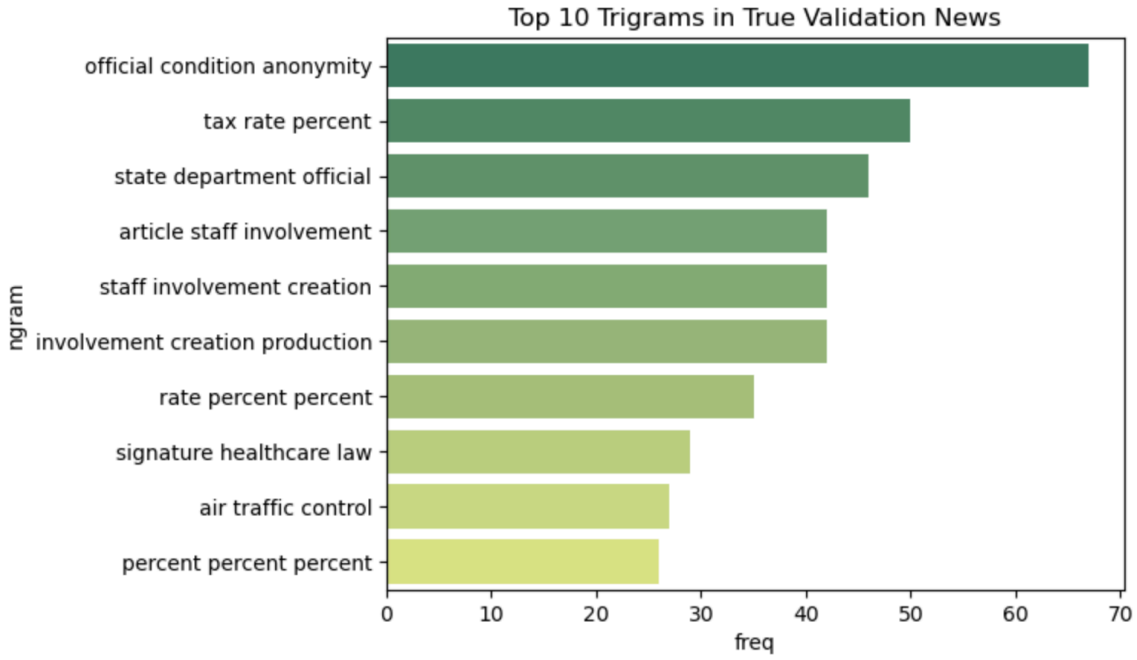
Top 10 unigrams by frequency in true news



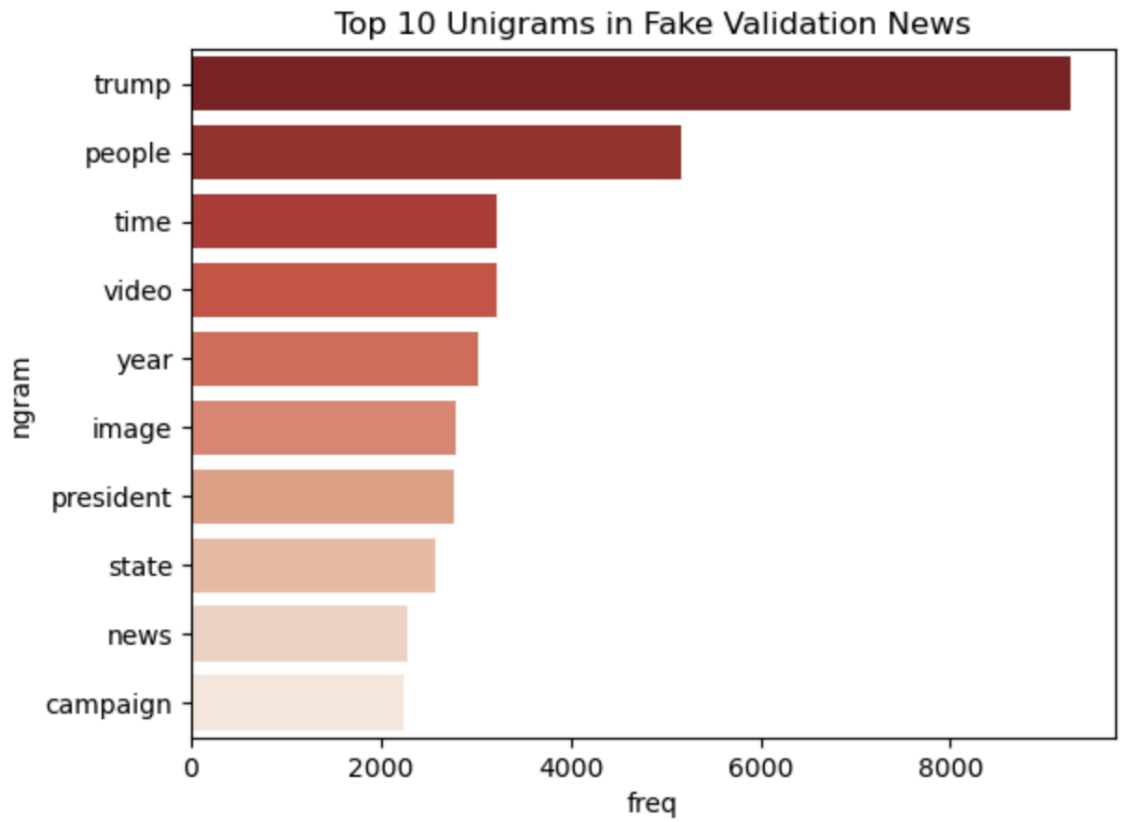
Top 10 bigrams by frequency in true news



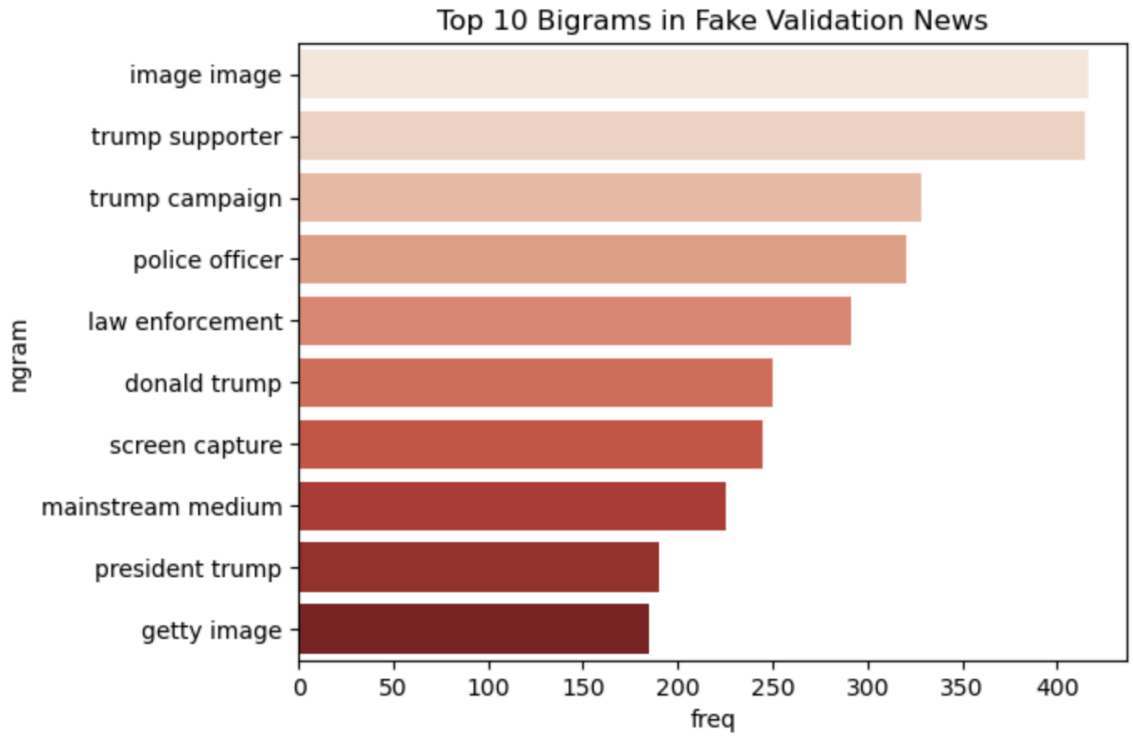
Top 10 trigrams by frequency in true news



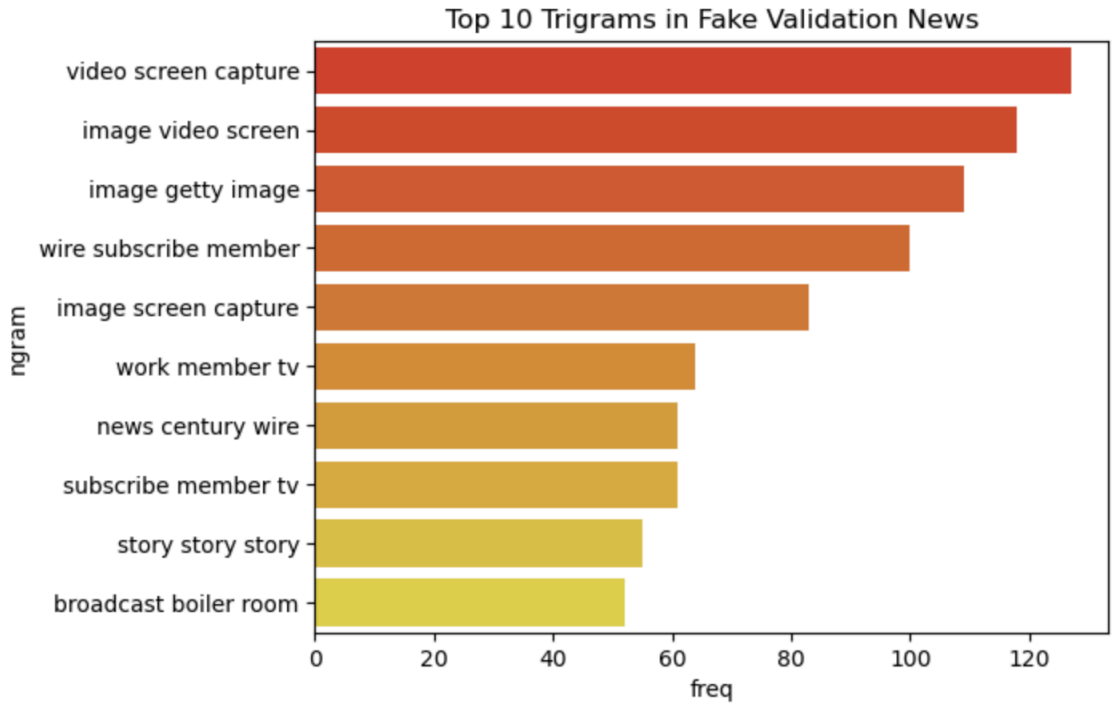
Top 10 unigrams by frequency in fake news



Top 10 bigrams by frequency in fake news

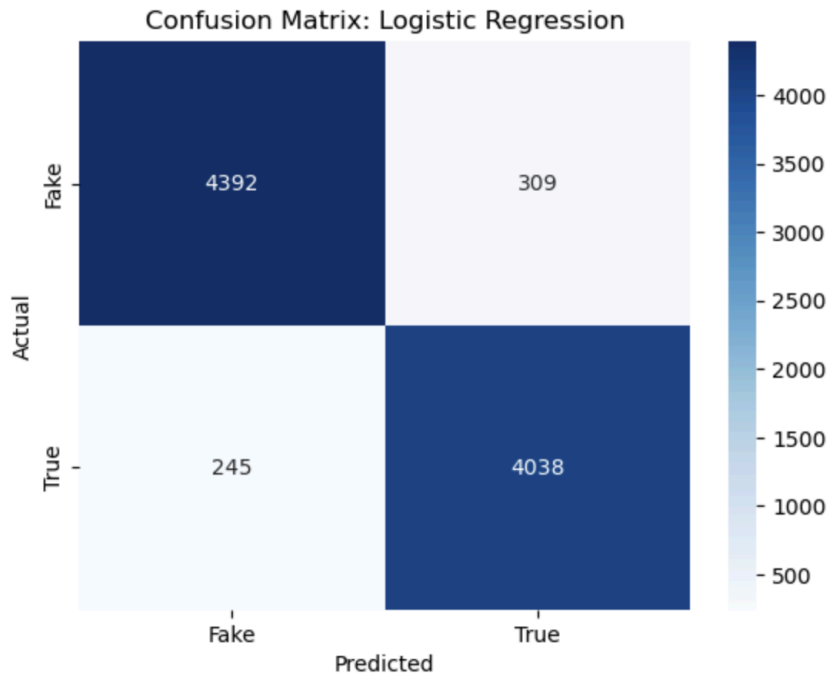


Top 10 trigrams by frequency in fake news

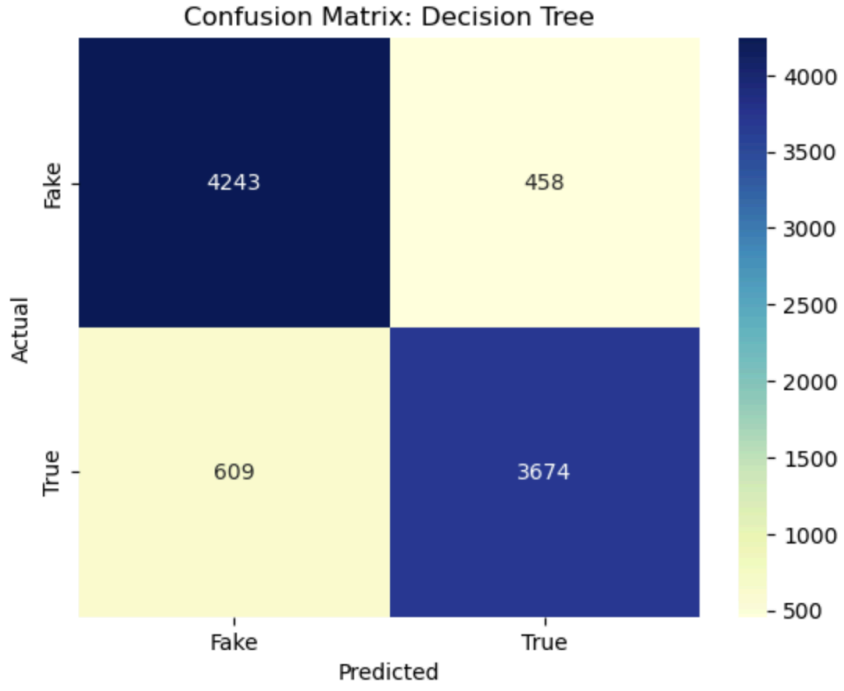


Feature Extraction (Word2Vec) and Model Training

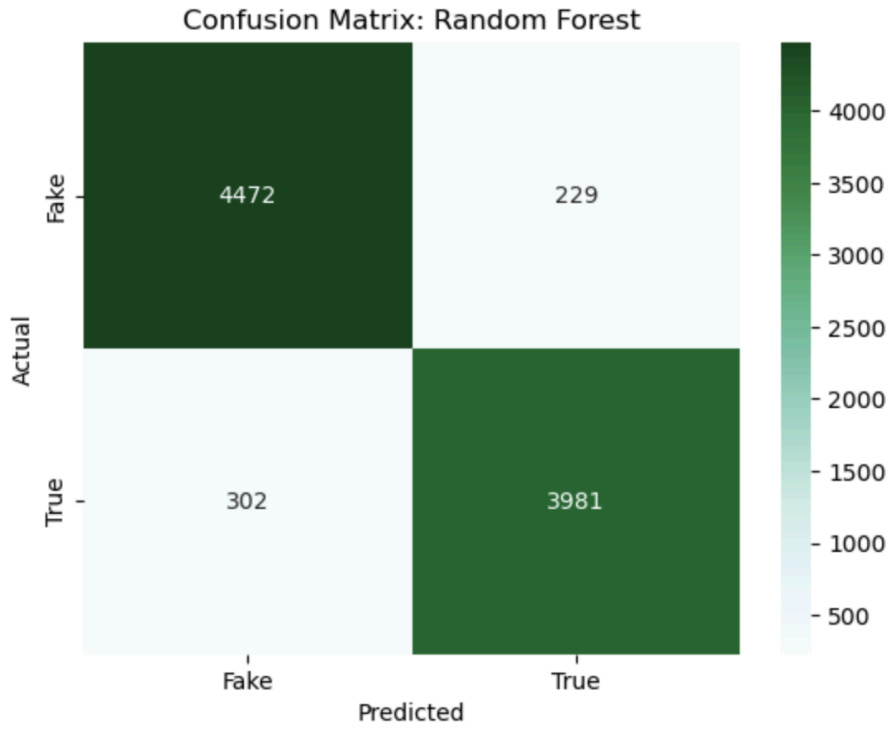
Confusion Matrix of Logistic Regression Model



Confusion Matrix of Decision Tree Model



Confusion Matrix of Random Forest Model



Conclusion

The aim is to distinguish between true and fake news using semantic understanding.

- Preprocessed and cleaned text
- Extracted meaning-rich word representations using Word2Vec
- Built and evaluated three supervised models on semantic document vectors

Traditional text classification often fails to detect semantic subtleties. Using Word2Vec, the model captures semantic relationships between words. This solution is generalizable and robust for real-world misinformation detection systems.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9417	0.9332	0.9454	0.9392
Decision Tree	0.8898	0.9021	0.8625	0.8818
Random Forest	0.9390	0.9450	0.9260	0.9354

Best Model: Logistic Regression, Random Forest slightly outperformed in precision, but the Logistic Regression provided: Highest overall F1 Score (0.9392) and the Best Recall (0.9454) — which is critical in this context

The model will catch false news even at the cost of a few false alarms

Hence, Logistic Regression was chosen as the most balanced and reliable classifier.

- End of Report -