Fake News Detection

PRESENTED BY- RIA SINGH

OBJECTIVE

The spread of fake news has become a significant challenge in today's digital world. With the massive volume of news articles published daily, it's becoming harder to distinguish between credible and misleading information. This creates a need for systems that can automatically classify news articles as true or fake, helping to reduce misinformation and protect public trust.

In this assignment, you will develop a Semantic Classification model that uses the Word2Vec method to detect recurring patterns and themes in news articles. Using supervised learning models, the goal is to build a system that classifies news articles as either fake or true.

DATA DICTIONARY

For this assignment, you will work with two datasets, True.csv and Fake.csv. Both datasets contain three columns:

title of the news article

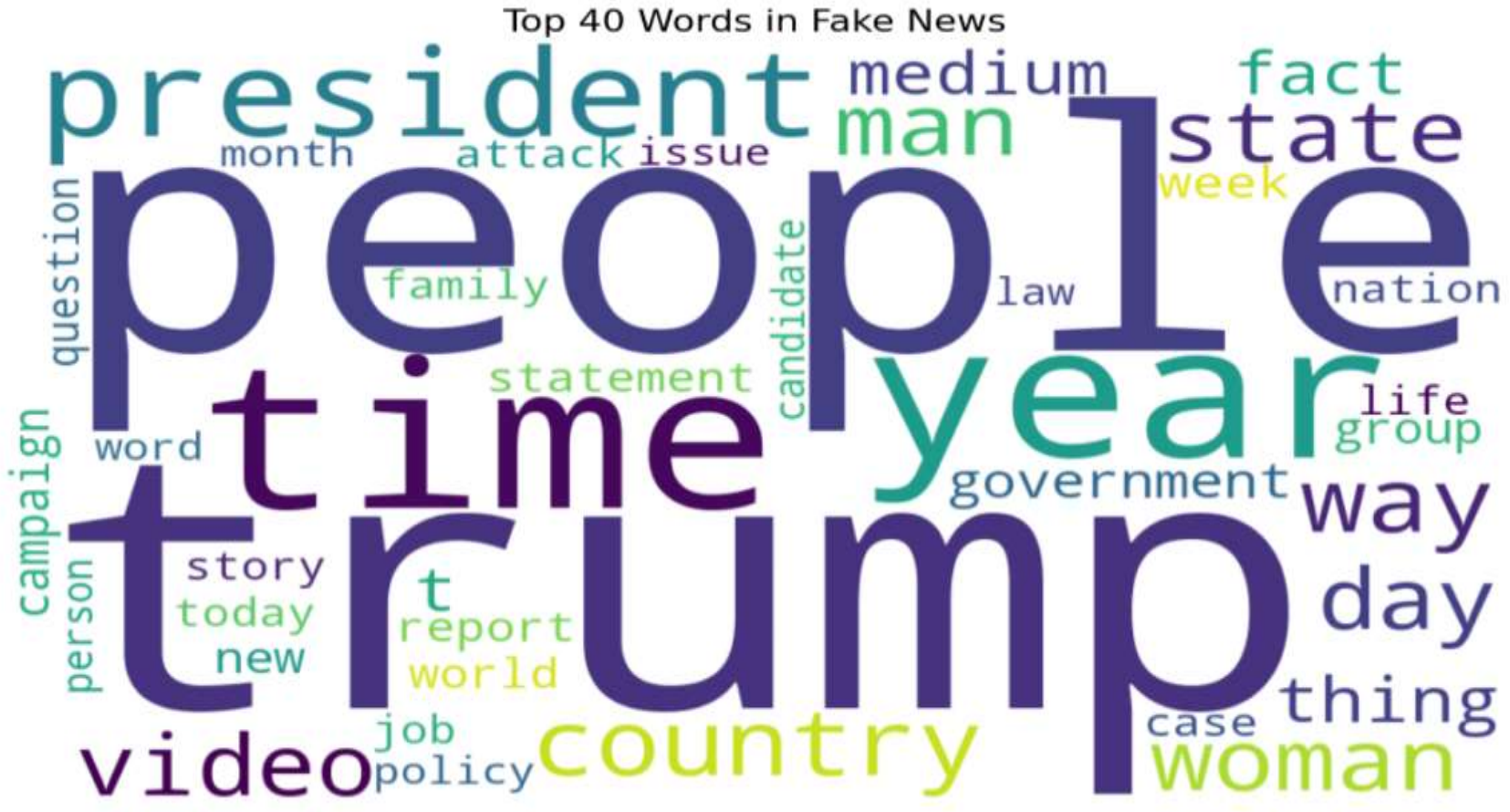
text of the news article

date of article publication

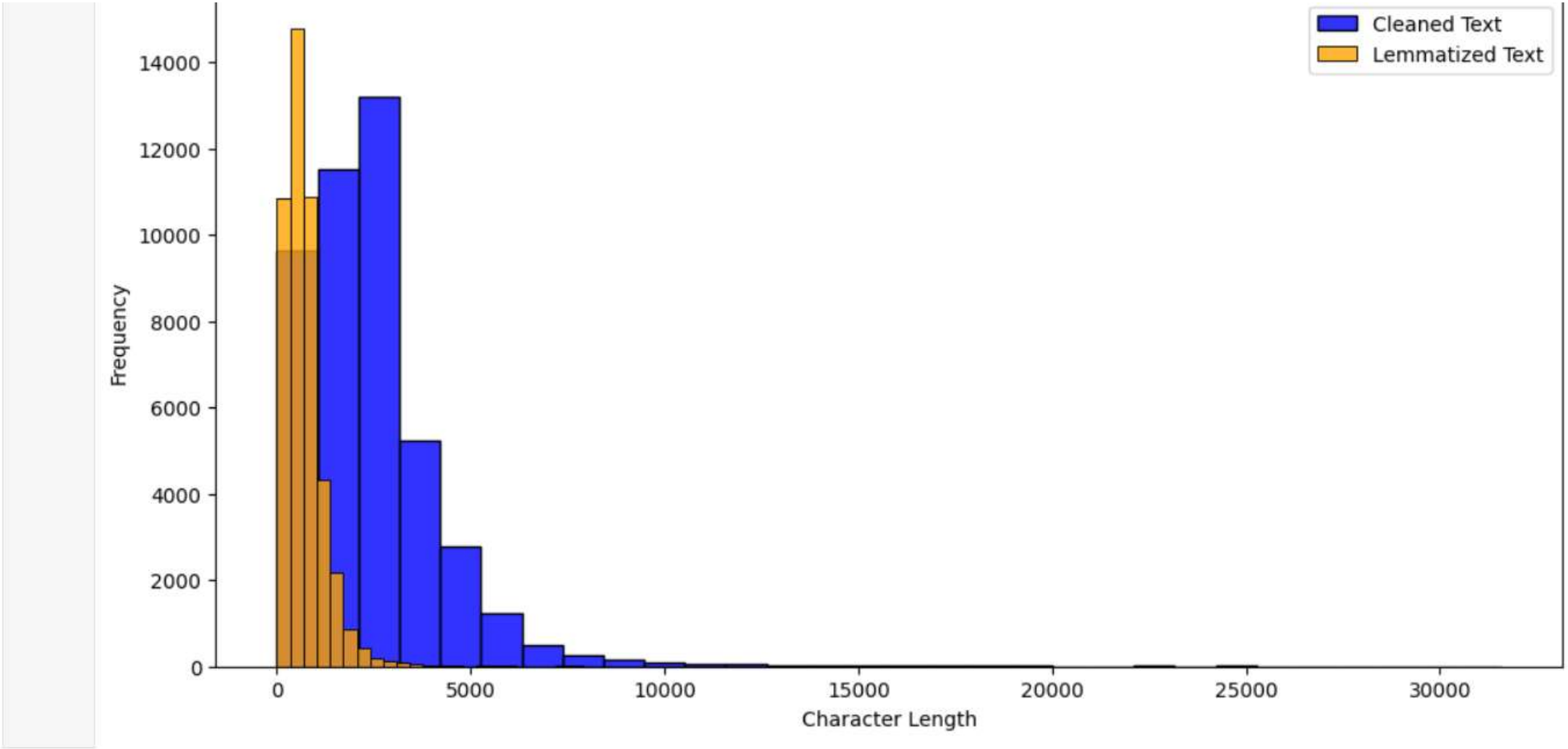
True.csv dataset includes 21,417 true news, while the Fake.csv dataset comprises 23,502 fake news.

TOP 40 WORDS IN TRUE NEWS DATASET

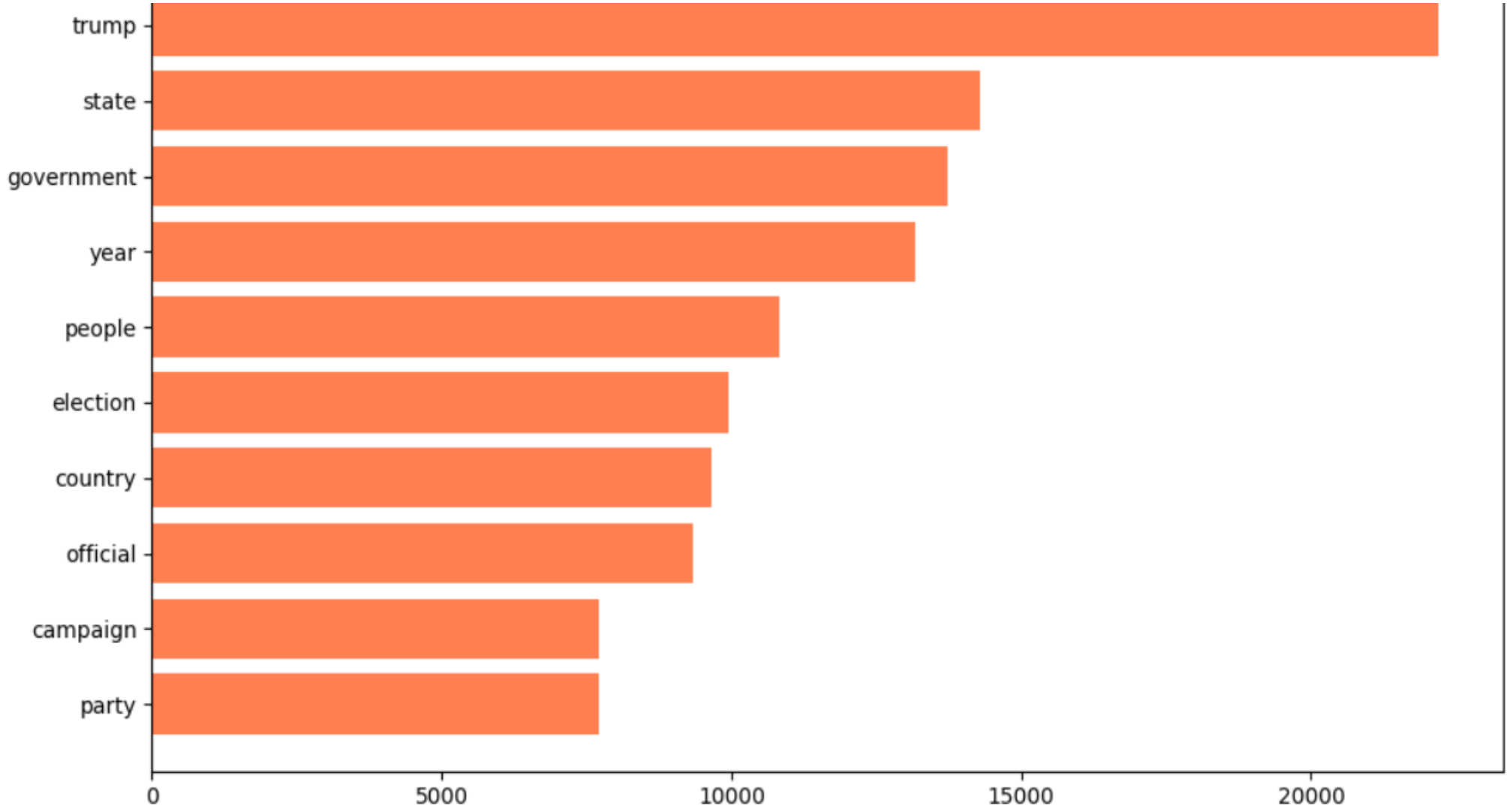
TOP 40 WORDS IN FAKE NEWS DATASET



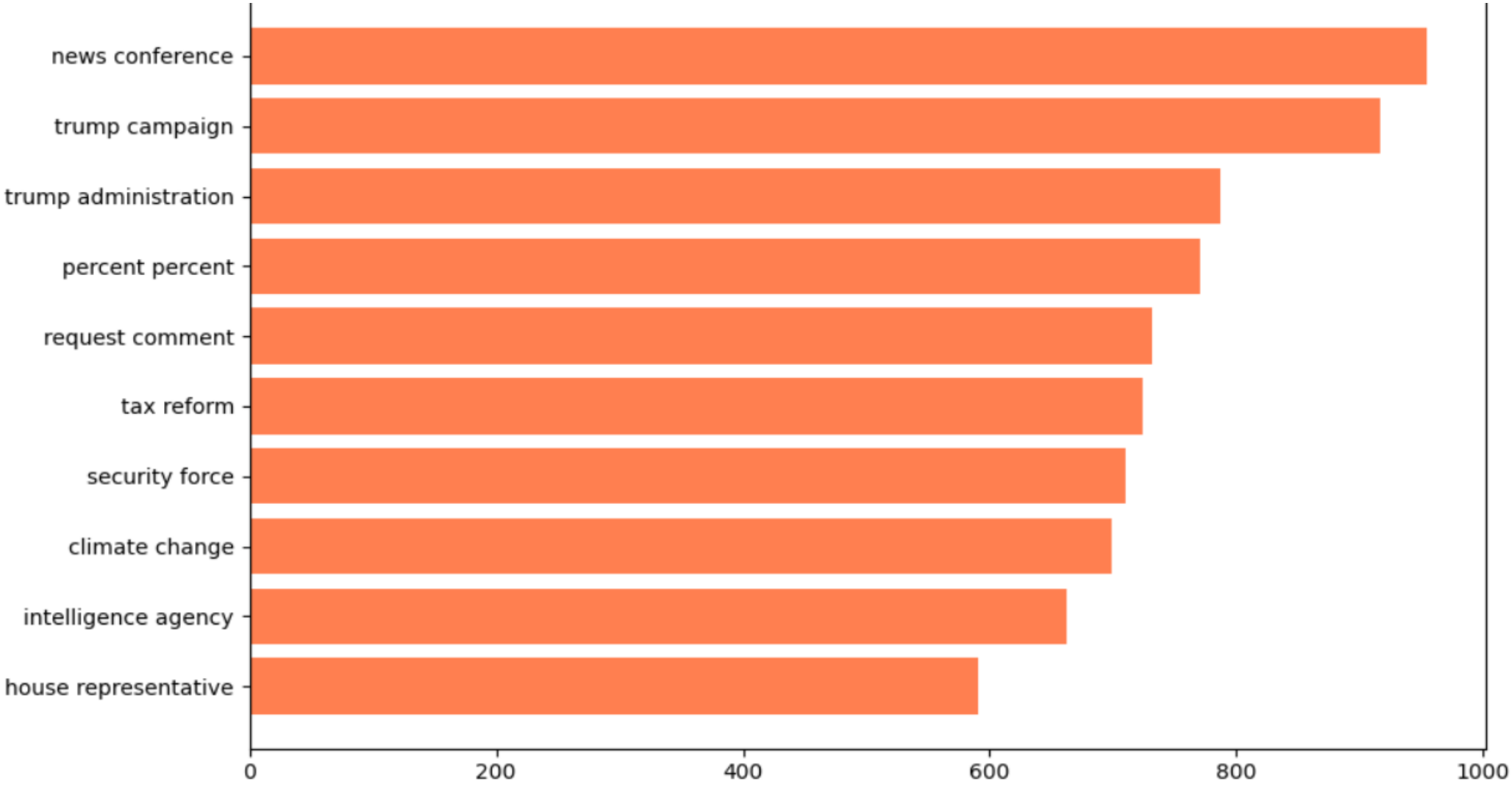
CHARACTER LENGTH FOR CLEANED VS LEMMATIZED NEWS TEXT



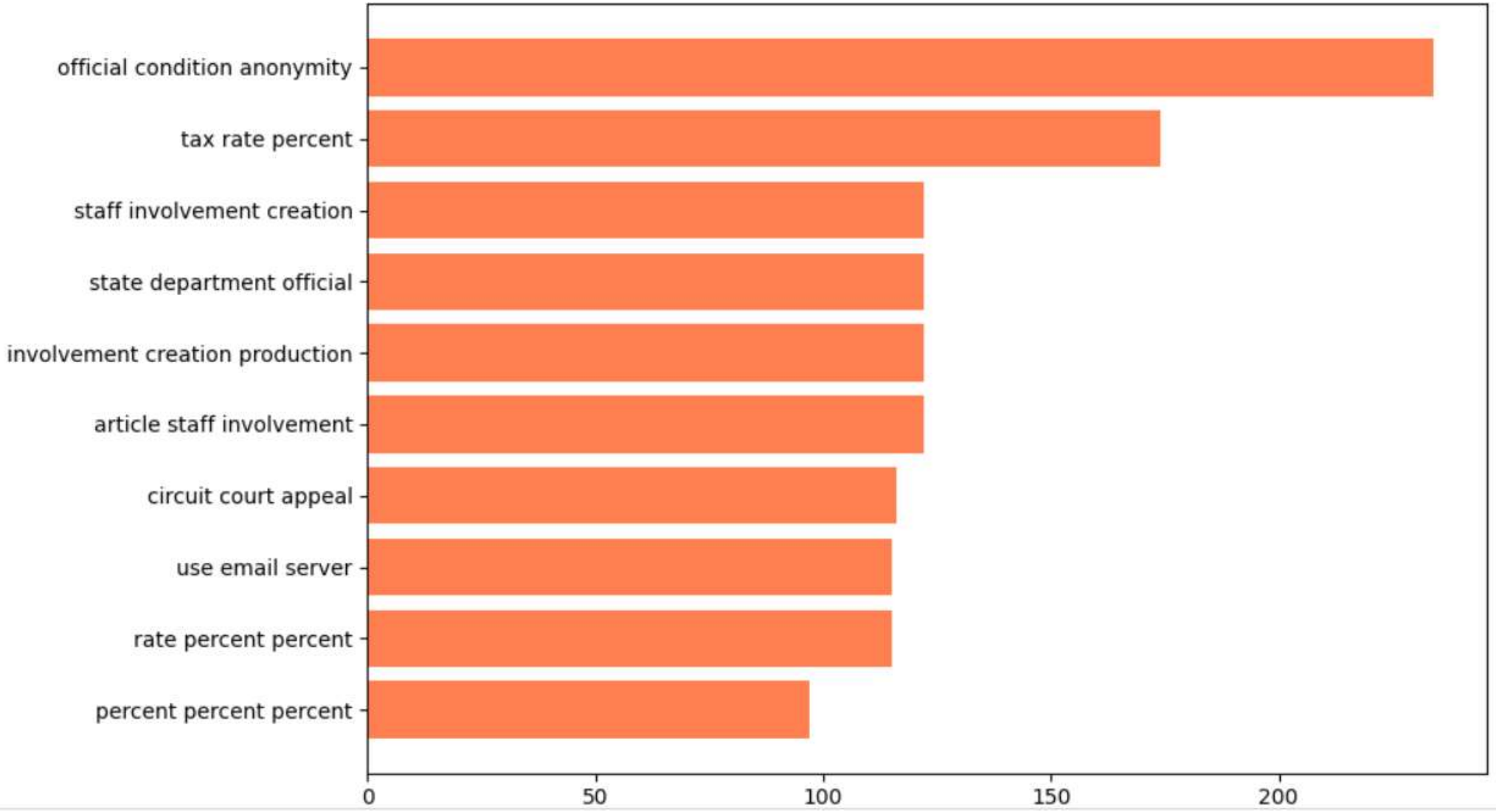
TOP 10 UNIGRAMS IN TRUE NEWS



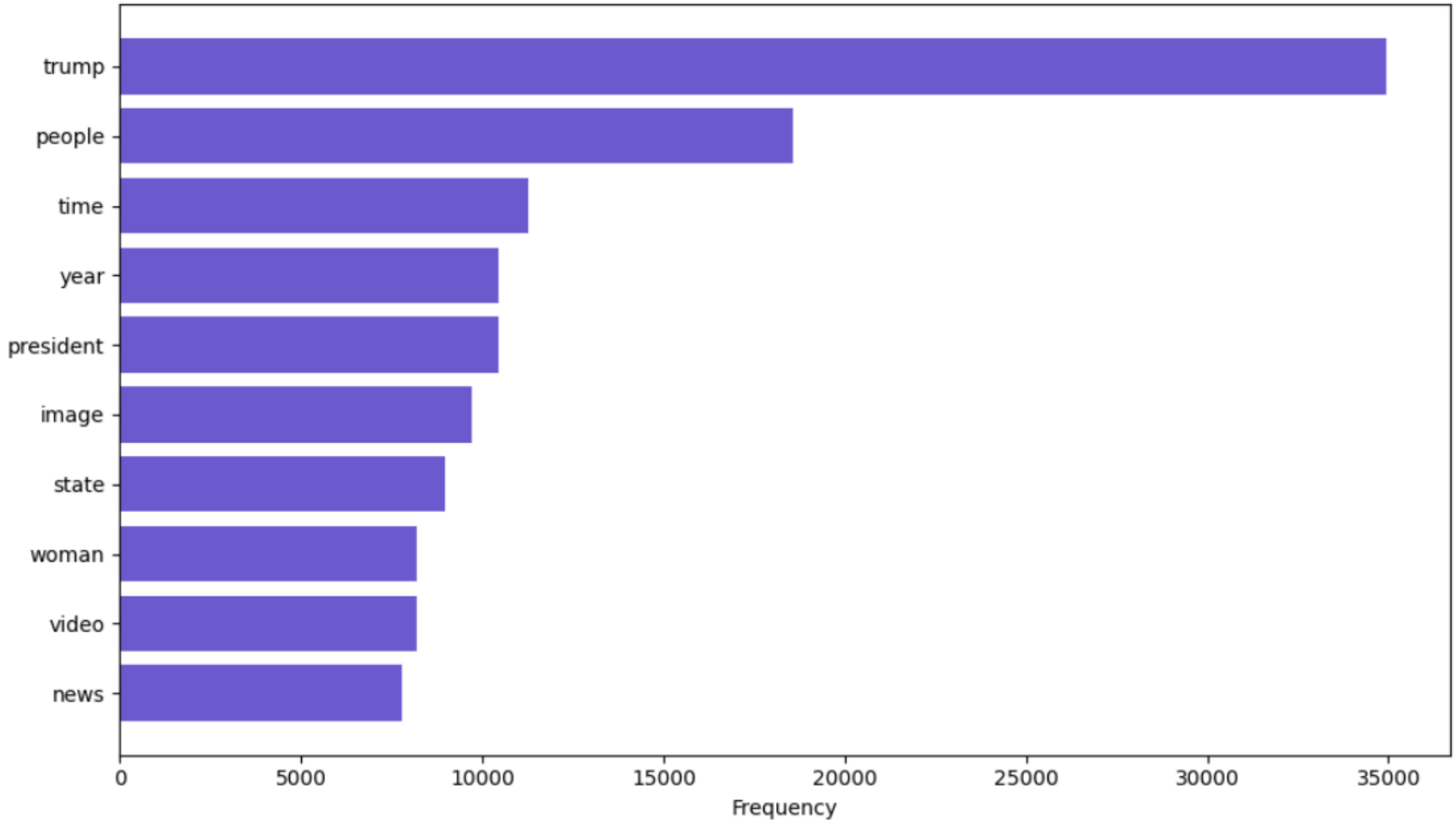
TOP 10 BIGRAMS IN TRUE NEWS



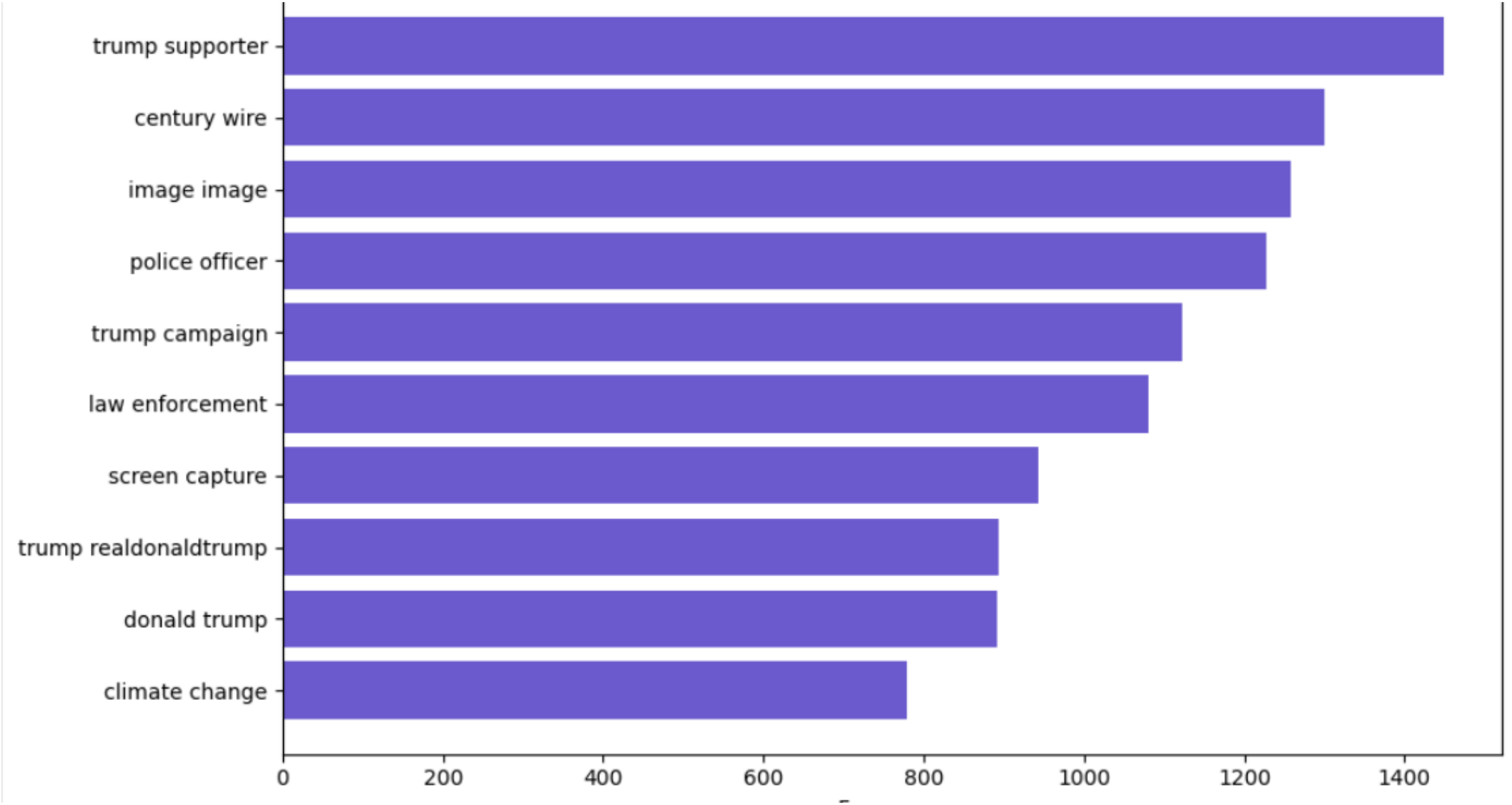
TOP 10 TRIGRAMS IN TRUE NEWS



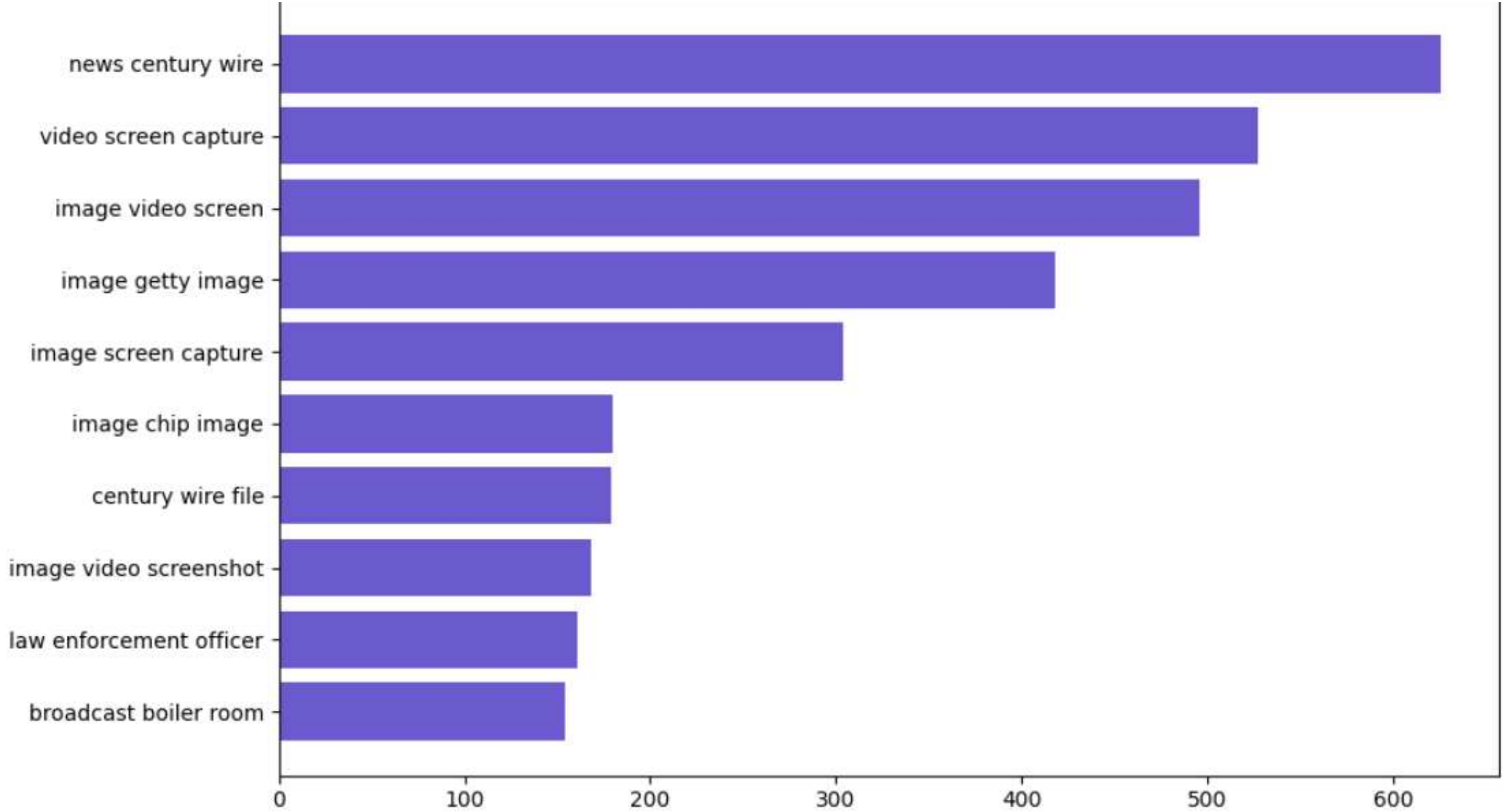
TOP 10 UNIGRAMS IN FAKE NEWS



TOP 10 BIGRAMS IN FAKE NEWS



TOP 10 TRIGRAMS IN FAKE NEWS



MODEL COMPARISON

	LOGISTIC REGRESSION	DECISION TREE	RANDOM FOREST
PRECISION	0.8921	0.8246	0.9048
RECALL	0.8943	0.7900	0.8875
F1 SCORE	0.8932	0.8069	0.8960
ACCURACY	0.8980	0.8197	0.9018

CONCLUSION

True news contain more formal language when compared to the fake news which is less coherent.

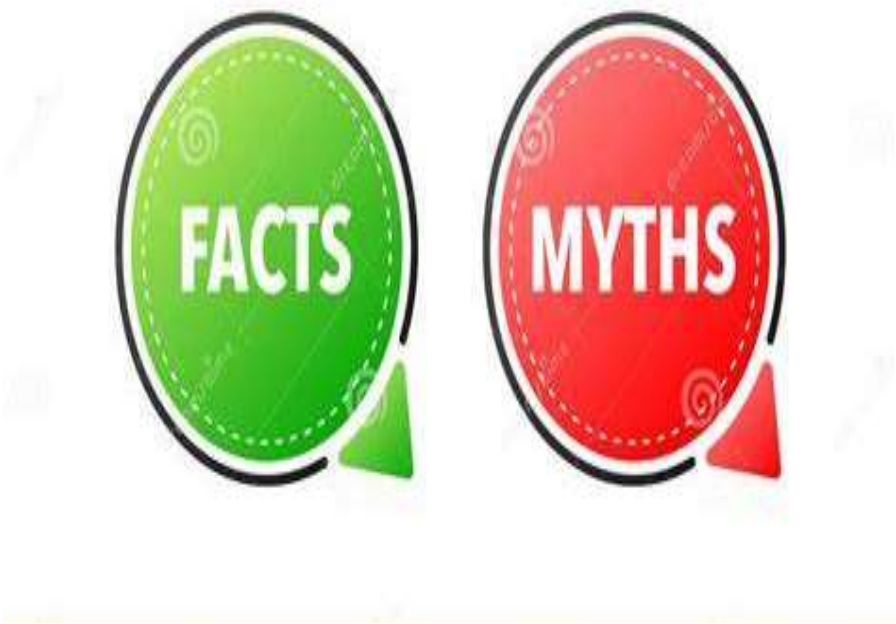
We have used text cleaning, lemmatization, and part-of-speech filtering for semantic classification.

Among the three models, random forest performed the best overall followed by logistic regression.

Lemmatized text is relatively smaller compared to the original news text.

Words such as year, election, government, and trump etc. are prominent which indicates that the news of this data is around election time.

F1 score provides a good balance between false positives and negatives and can be considered a good metric for minimizing spread of wrong information.



Thank you