# Fake News Detection Case Study

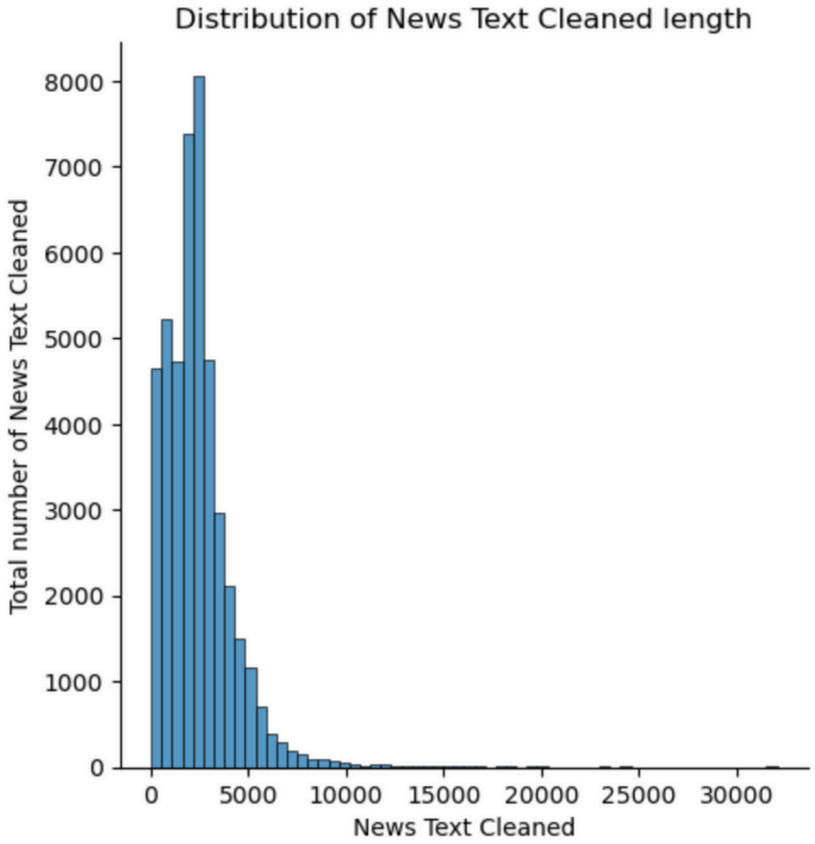
Raghavendran Ramakrishnan

# Problem Statement

- Develop a Semantic Classification Model using Word2Vec method to extract semantic relations from the news text and detect fake news.
- The following tasks would need to be completed: Data Preparation / Text Pre-processing, train validation split, EDA on training data, feature extraction, model training and evaluation.

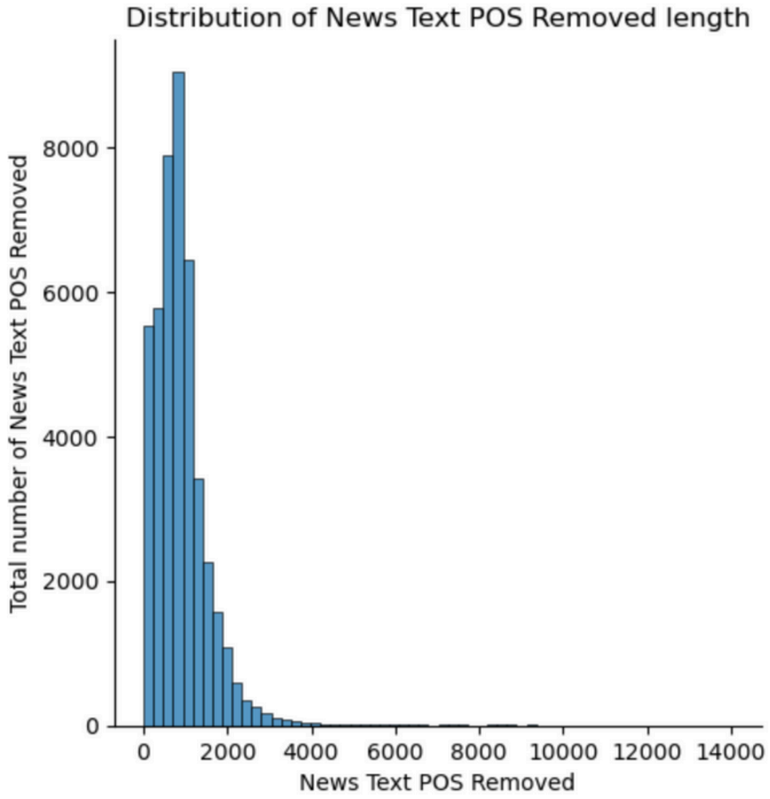
# News Text Cleaned Analysis

- Histogram of news text cleaned and total number of news text cleaned.
- Observation of left skew in the histogram.



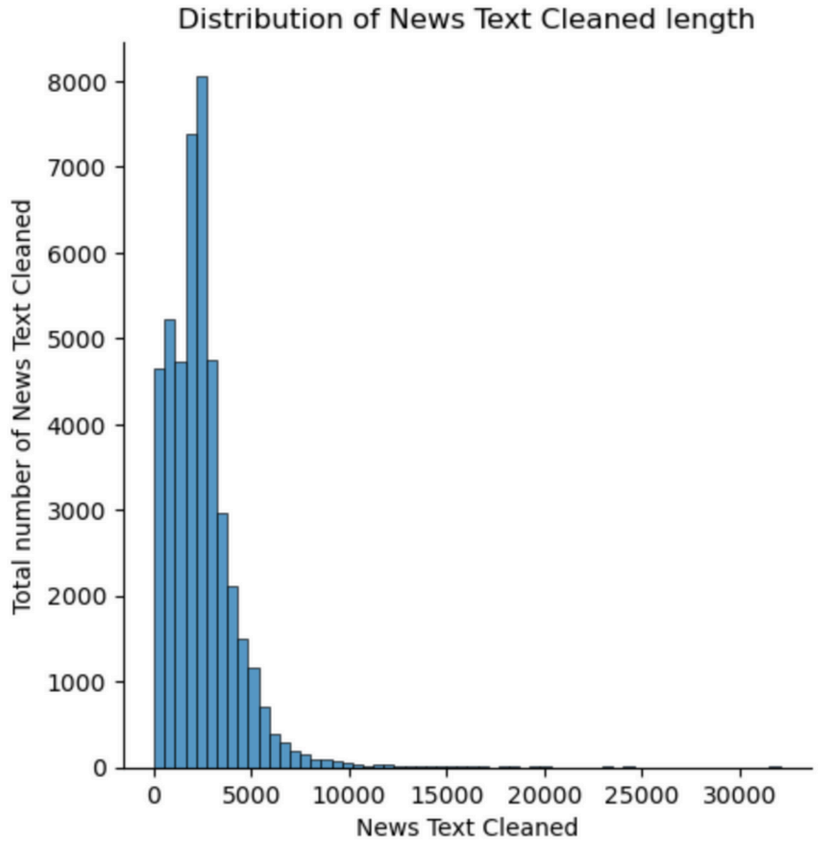
# News Text Lemmatized Analysis

- Histogram of news text lemmatized and total number of news text lemmatized..
- Observation of left skew in the histogram

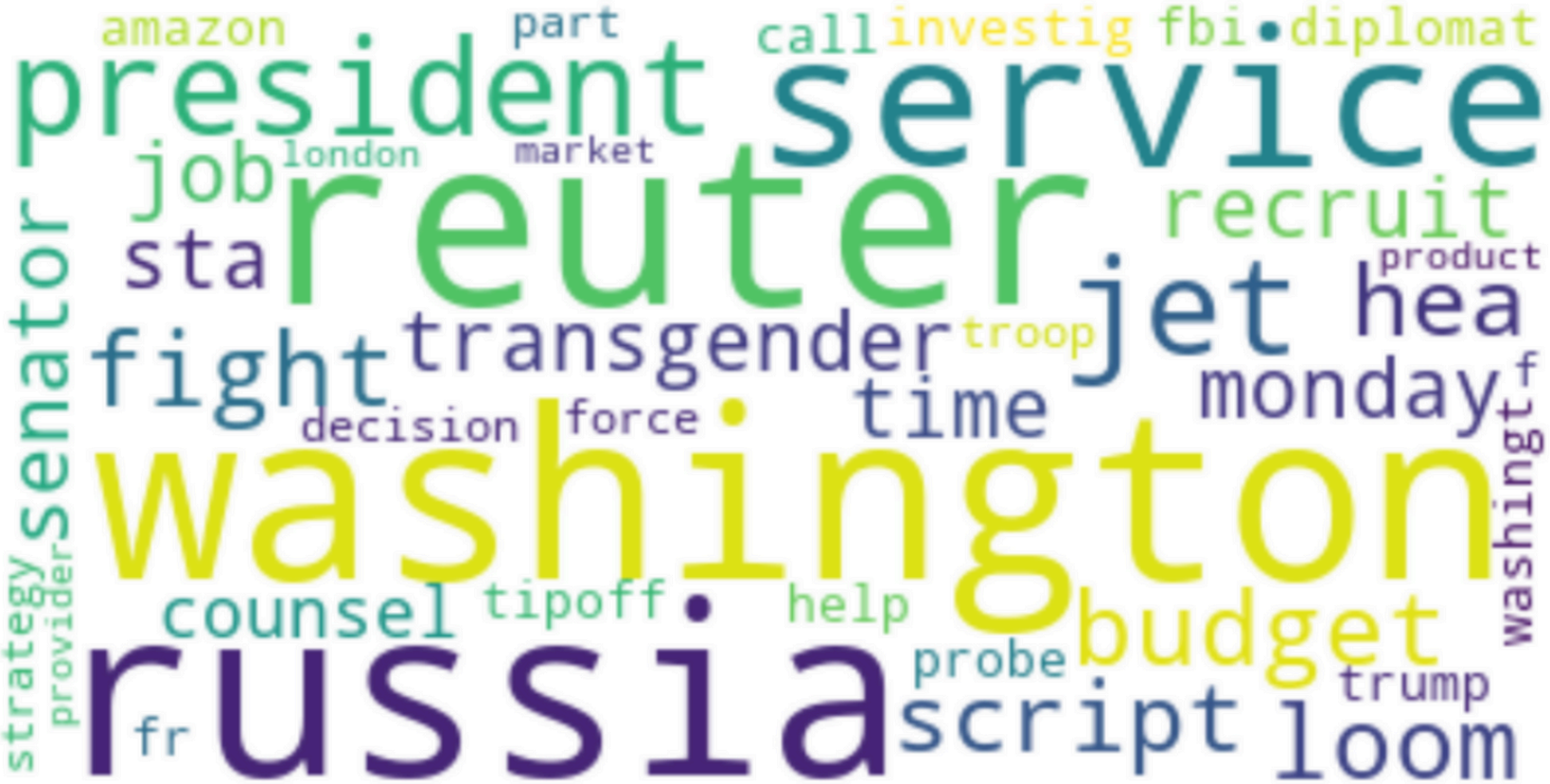


# News Text POS Cleaned Analysis

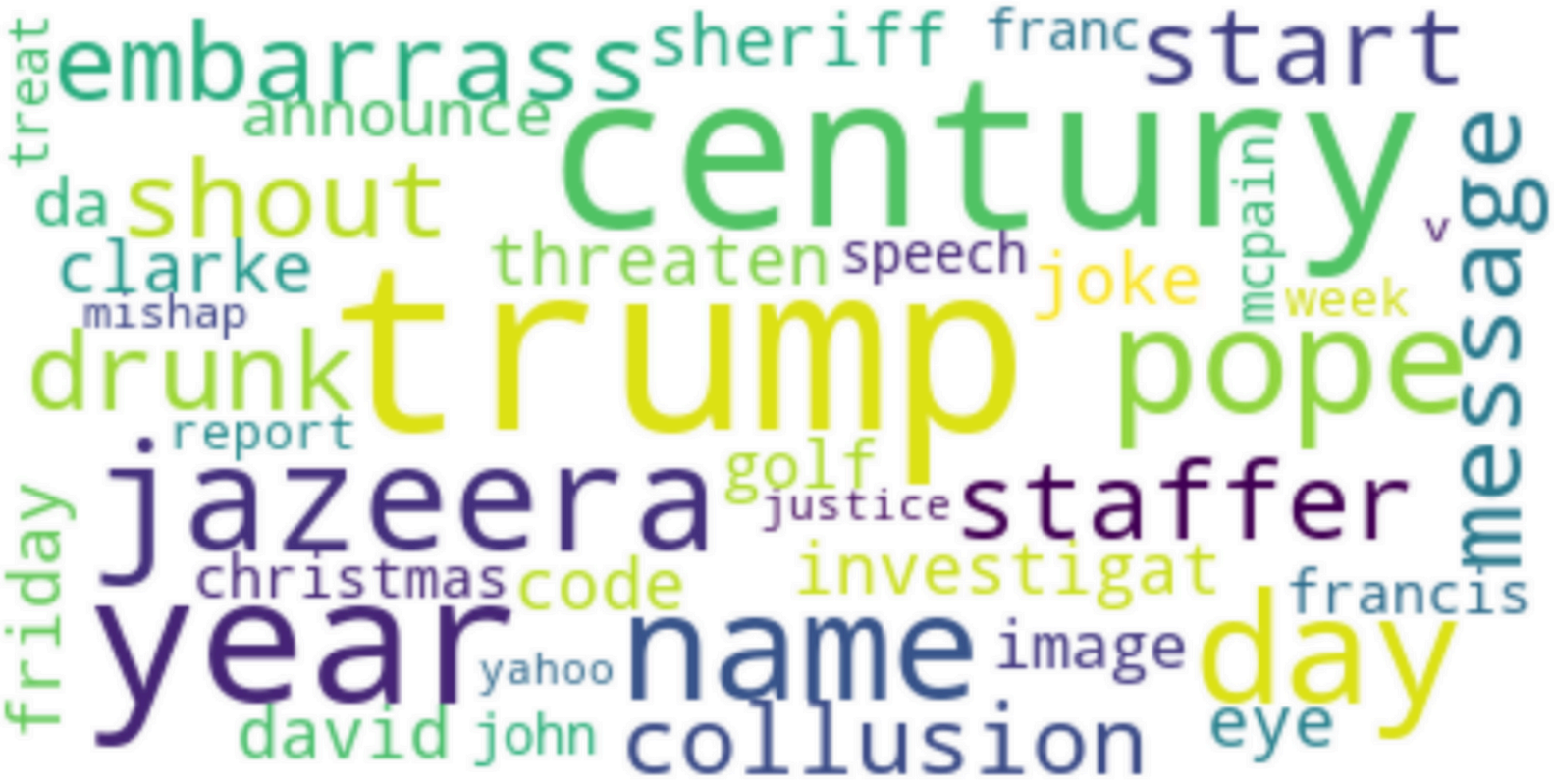
- Histogram of news text cleaned and total number of news text cleaned.
- Observation of left skew in the histogram



# Word Cloud – Real News Top 40 Words

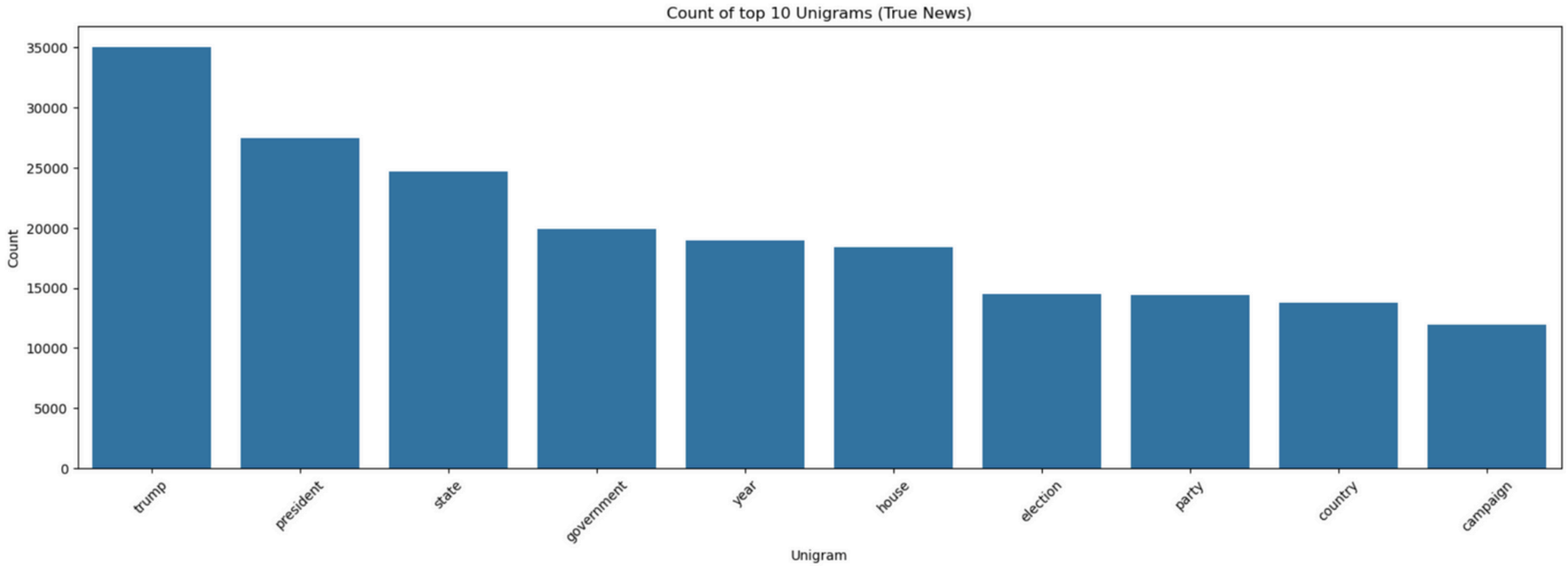


# Word Cloud – Fake News Top 40 Words



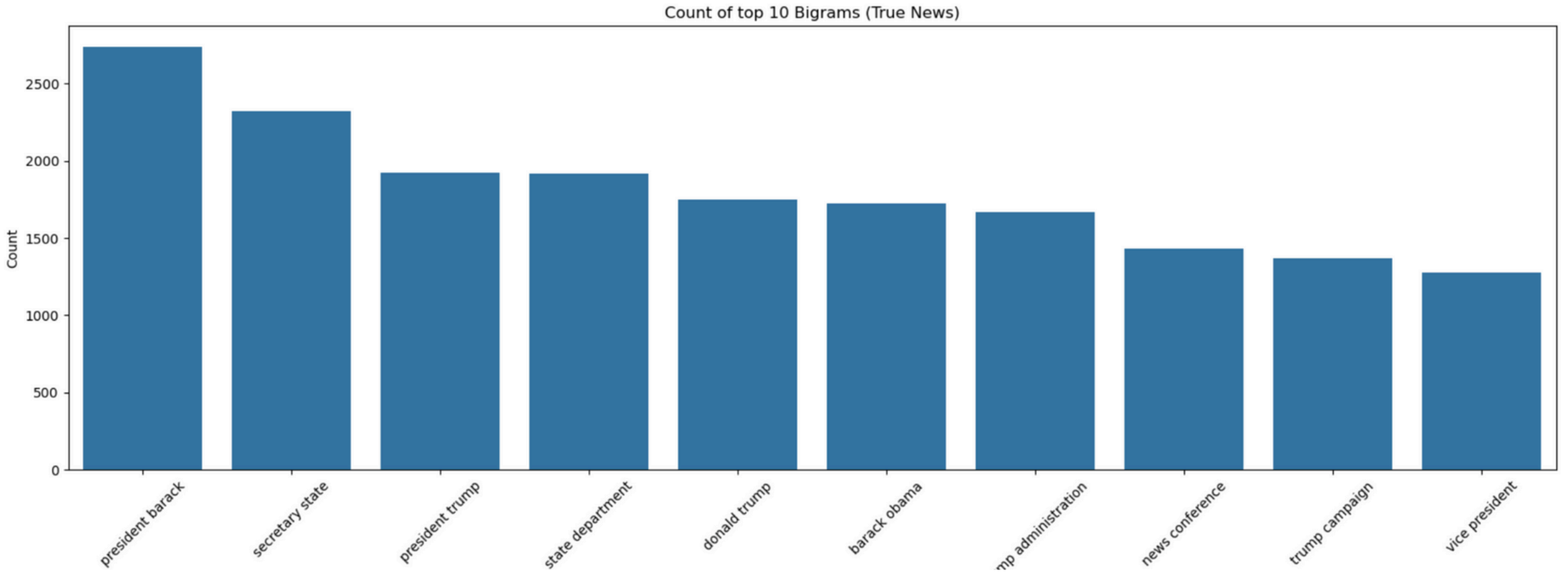
# Top Ten Unigrams (True News)

- Top ten unigrams for true news: trump, president, state, government, year, house, election, party, country and campaign.



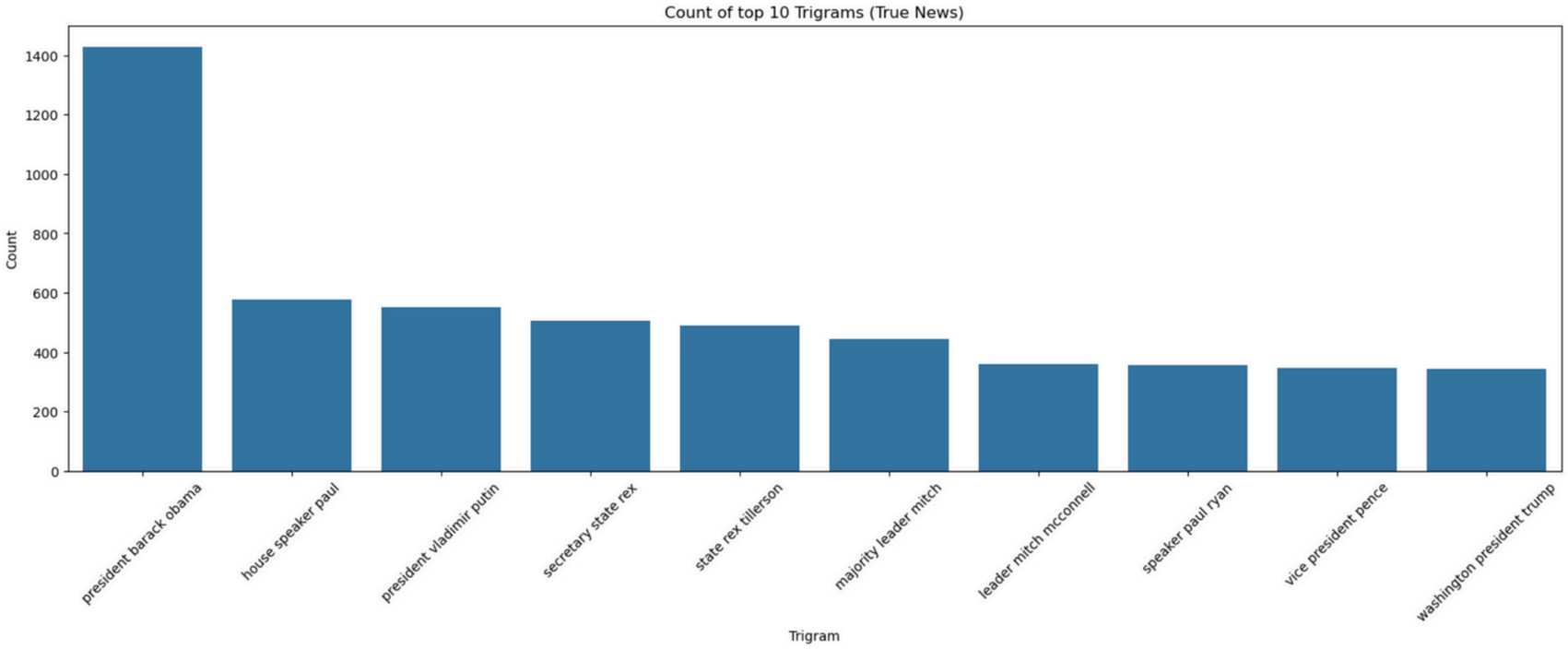
# Top Ten Bigrams (True News)

- Top ten bigrams for true news: president barack, secretary state, president trump, state department, donald trump, barack Obama, news conference, trump campaign and vice president.



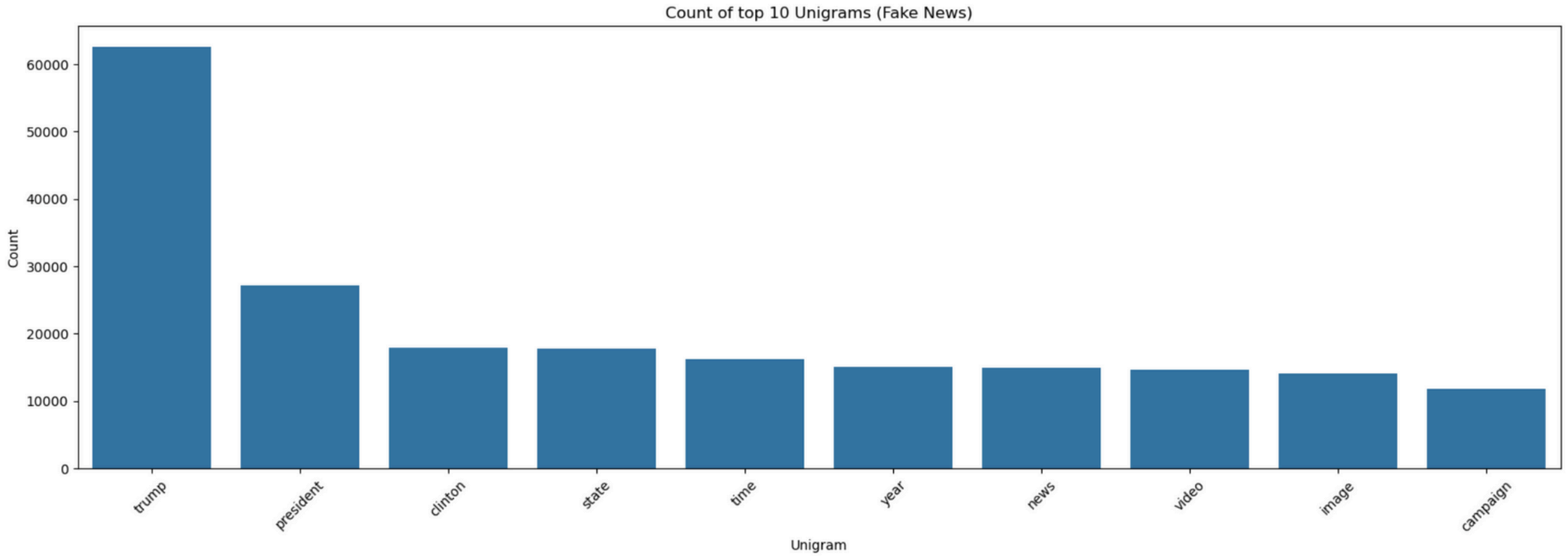
# Top Ten Trigrams (True News)

- Top trigrams: president barack Obama, house speaker paul, president Vladmir putin, secretary state rex, state rex tillerson



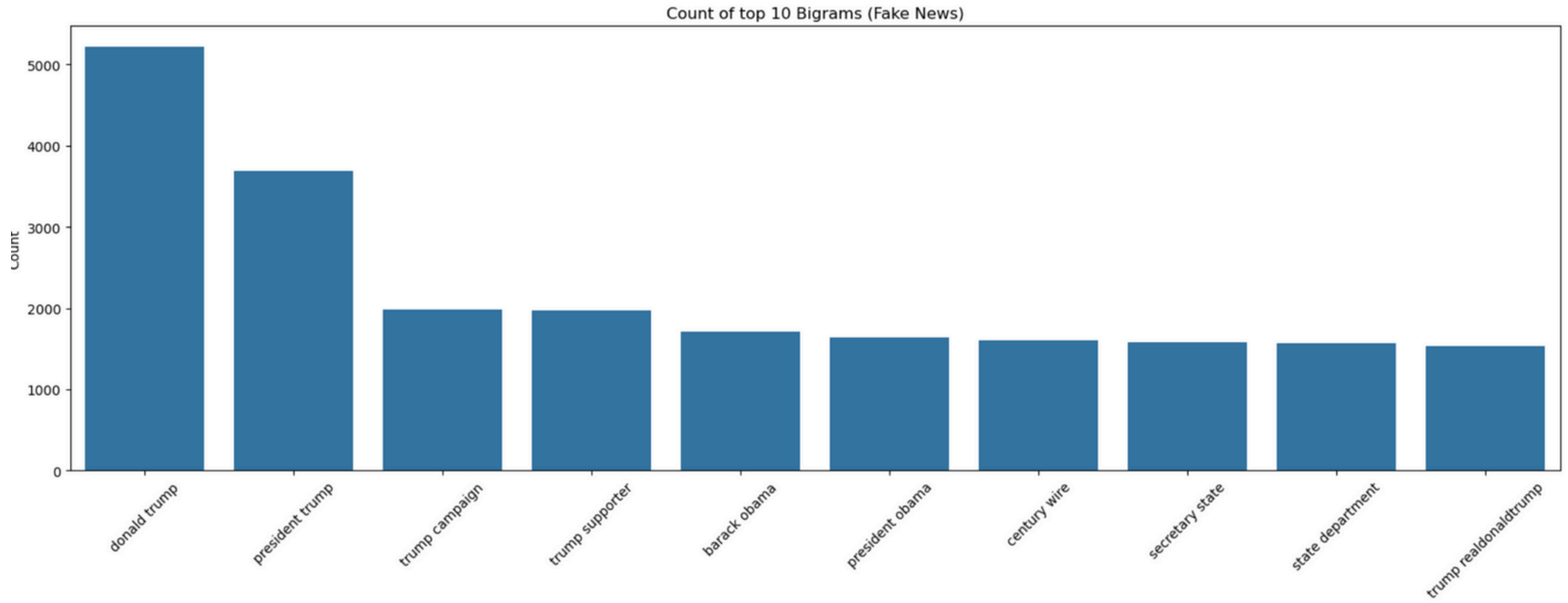
# Top Ten Unigrams (Fake News)

- Top ten unigrams for fake news: trump, clinton, state, time, year, news, video, image, campaign.



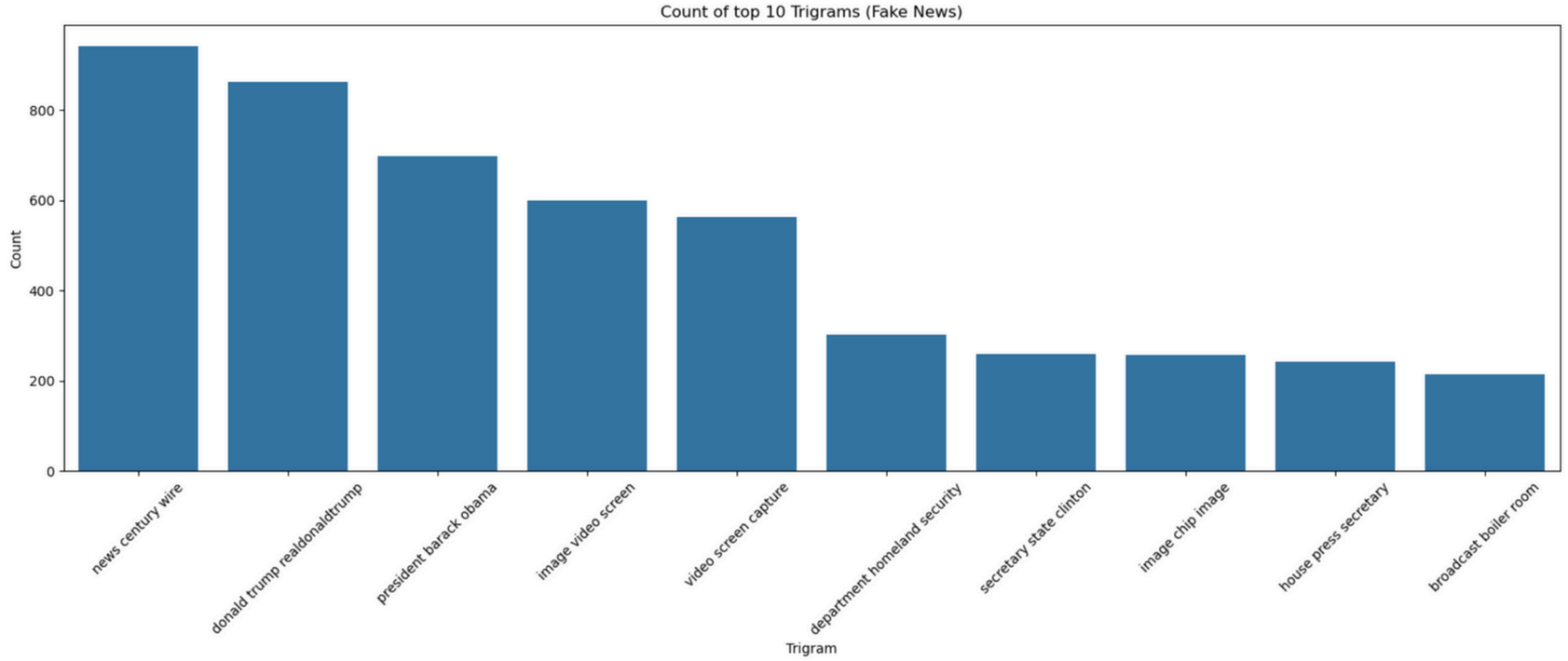
# Top Ten Bigrams (Fake News)

- Top ten bigrams for true news: Donald trump, president trump, trump campaign, trump supporter, president Obama, century wire, secretary state, state department, trump realdonaldtrump.



# Top Ten Trigrams (Fake News)

- Top trigrams: new century wire, Donald trump realdonaldtrump, president barack obama, image video screen.



# Models for False News Classification

- Three Models namely Logistic Regression, Decision Tree Classifier and Random Forest Classifier were created to predict true/false news.
  **Logistic Regression**
- The overall accuracy of the Logistic Regression model is 93%.
- Precision and Recall are 94% and 92% respectively which makes this is a very good model.
  **Decision Tree**
- The overall accuracy of the Decision Tree model is 85%
- Precision and Recall are 84% and 87% respectively
  **Random Forest**
- The overall accuracy of the Random Forest model is 93%
- Precision and Recall are 93 and 94 percent.
  **Conclusion**
  It was found that both the Logistic Regression and Random Forest Models had an overall accuracy and precision around 93 percent.

# Thank You