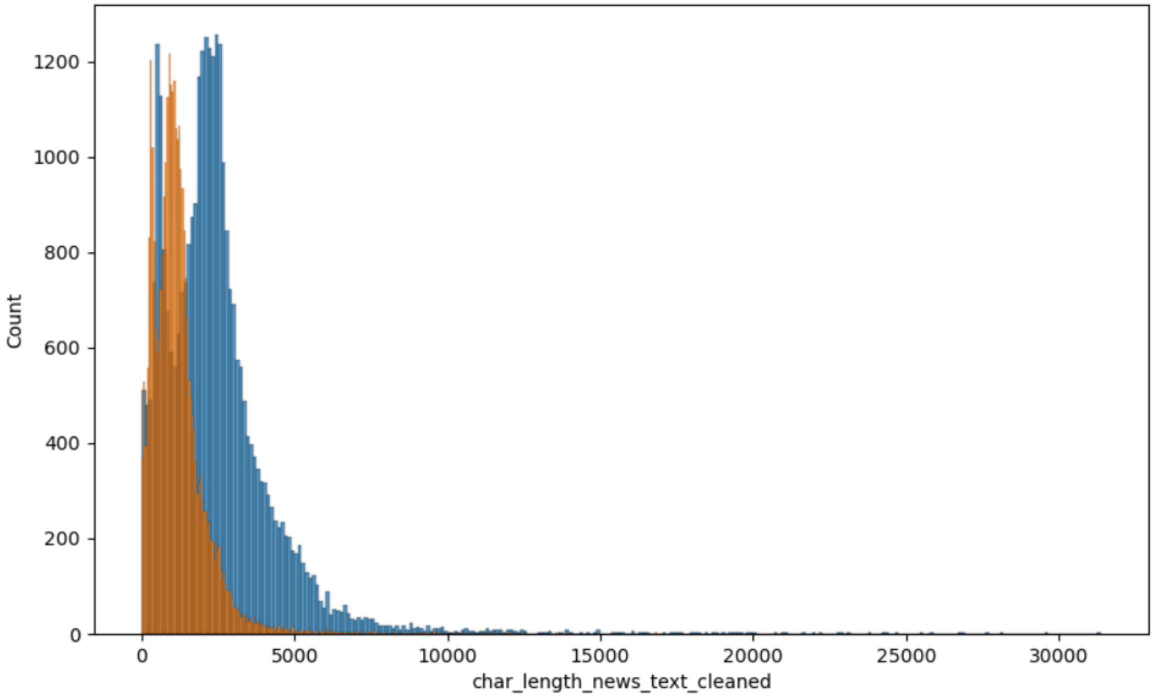
Length of Original, cleaned and Lemmatized texts

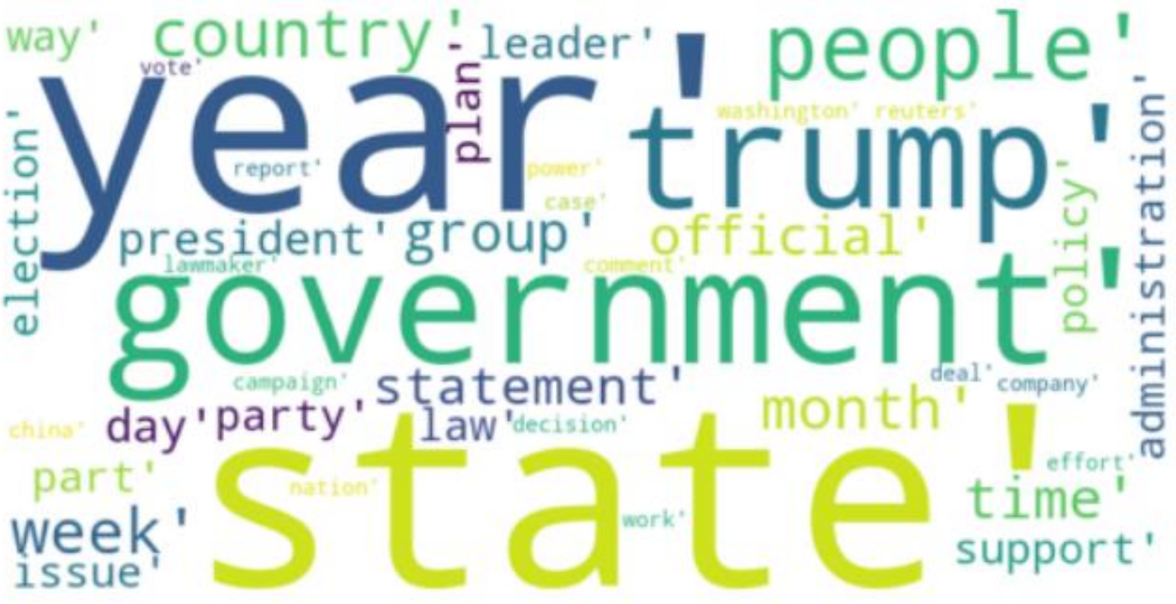
- **Original Text** – No change to the Data provided
- **Cleaned text** – Applied Below operations
  - ○ Converted to Lower case
  - ○ Removed square brackets and content within
  - ○ Removed anything that is not a-z, A-Z, 0–9, or underscore
  - ○ Removed URLs
  - ○ Removed HTML tags
  - ○ Removed punctuation
  - ○ Words that contain numbers in them e.g.: win123, abc2xyz, rule6969
- **Processed text**
  - ○ Removed Stop words
  - ○ Applied WordNetLemmatizer to keep single word variable
  - ○ Applied word_tokenize to Tokenise words
  - ○ Kept only 'NN' and 'NNS' POS tagged words

Graph of length of each Processed text in Train data set

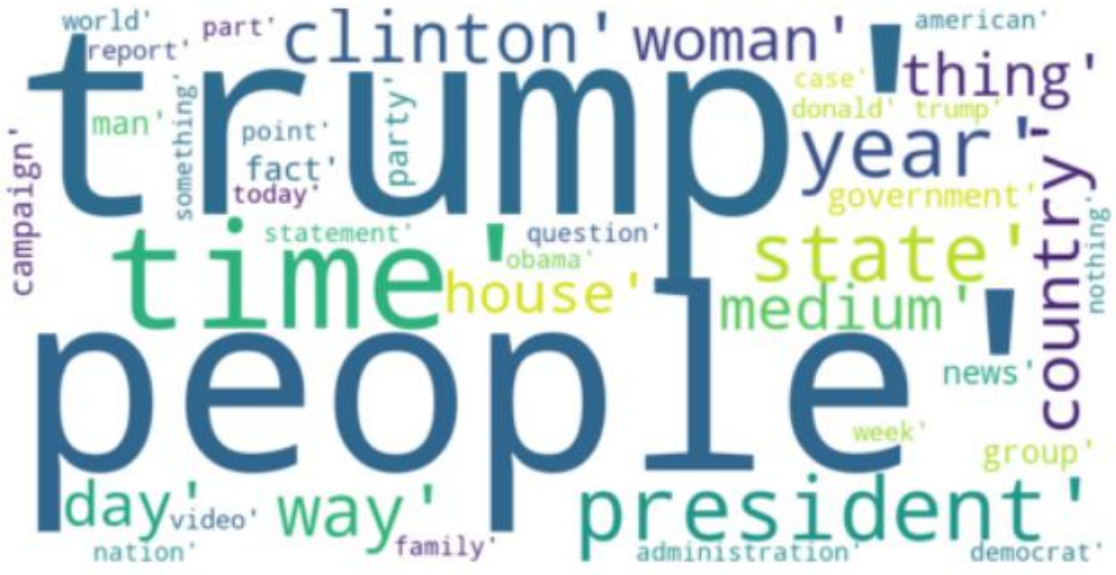


Train Data set

Top 40 words as per ranking in True dataset

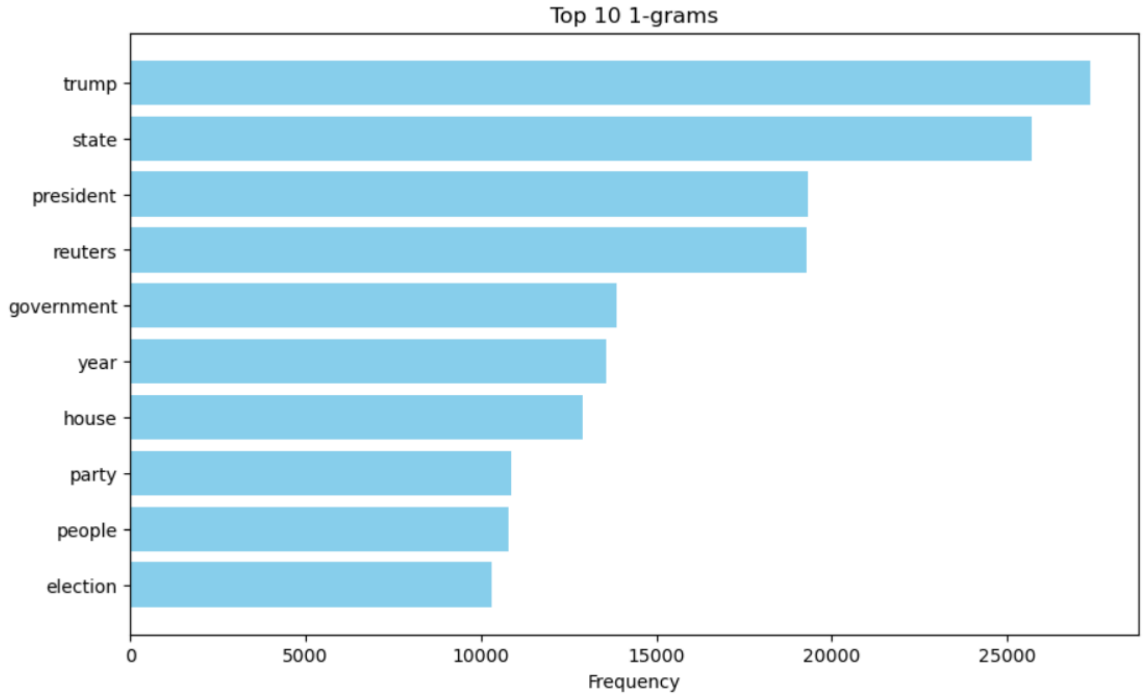


Top 40 words as per ranking in Fake dataset

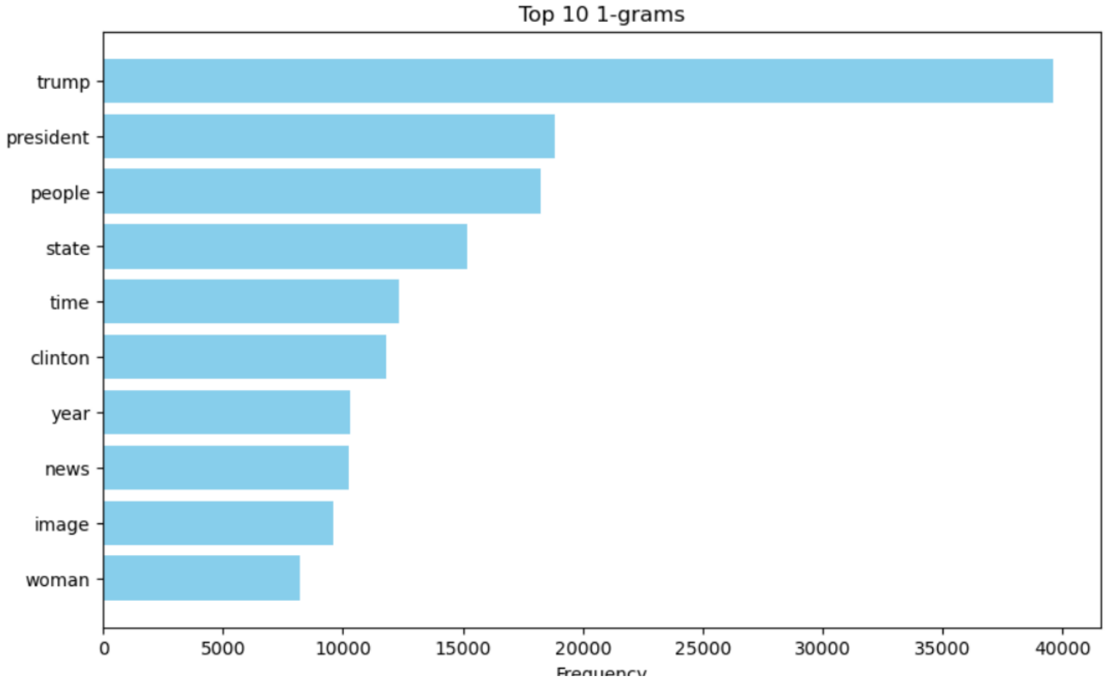


Top 10 Unigram in True data and Fake data set from Train data set

True Data set

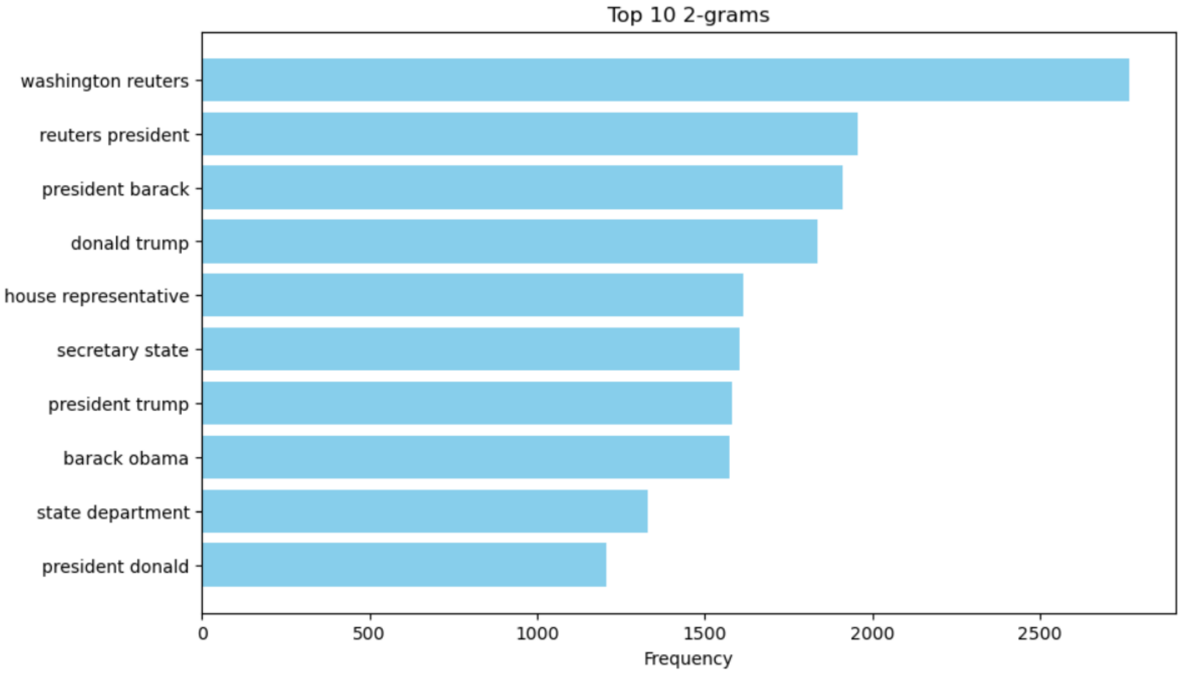


Fake Data set

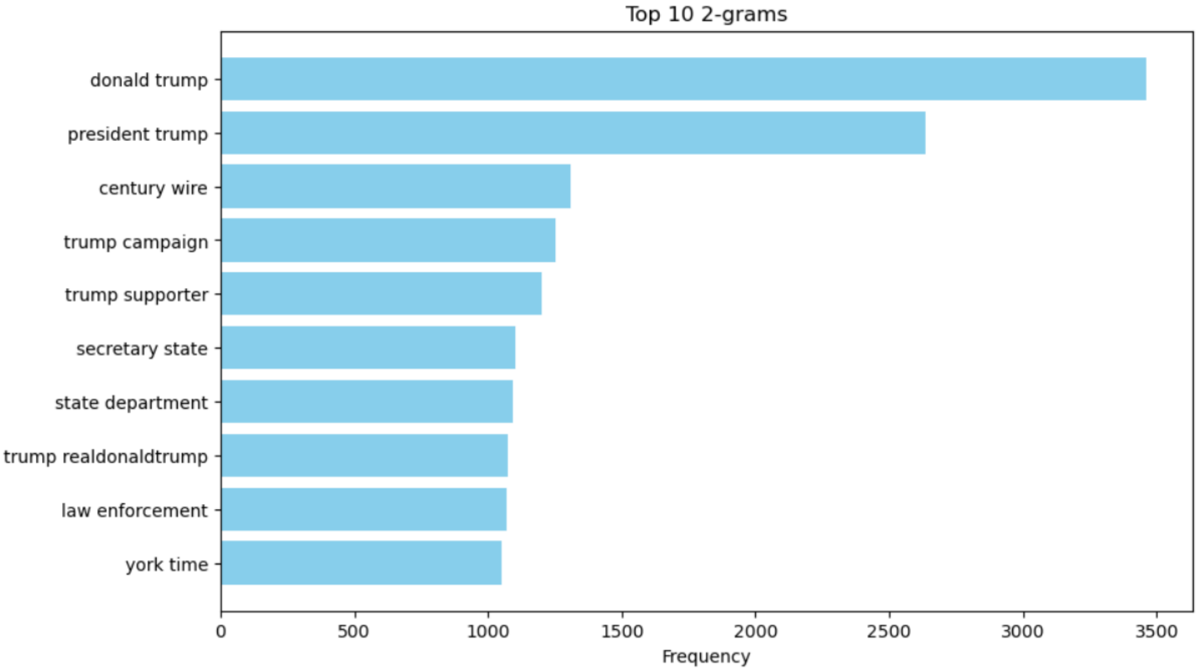


Top 10 Bigram in True data and Fake data set from Train data set

True Data set

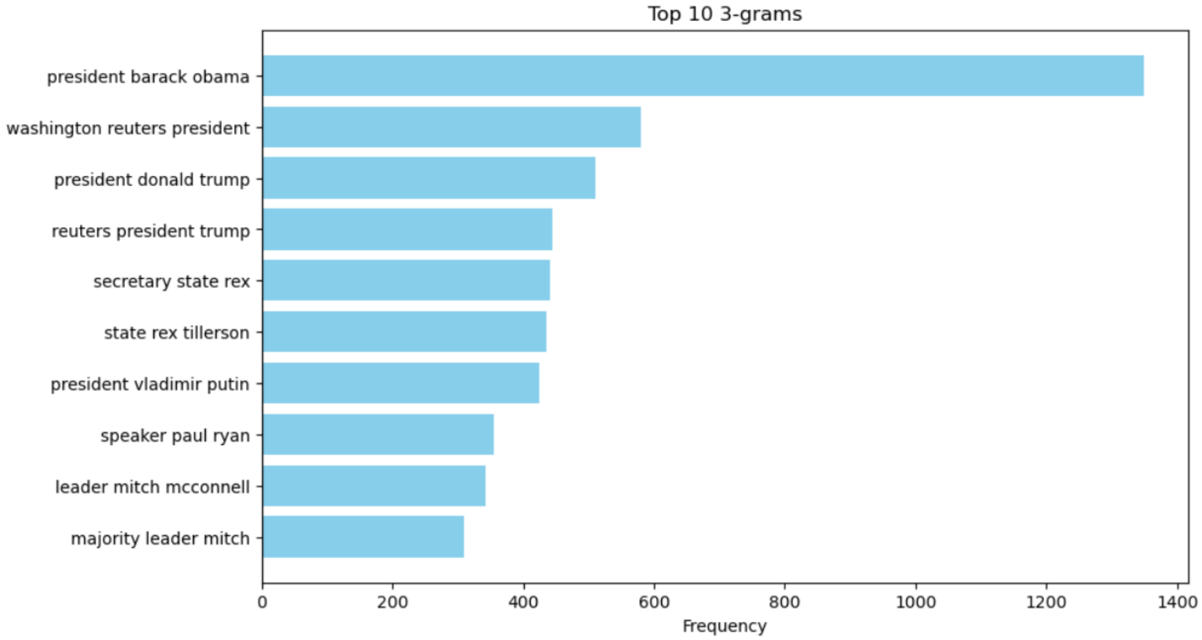


Fake Data set

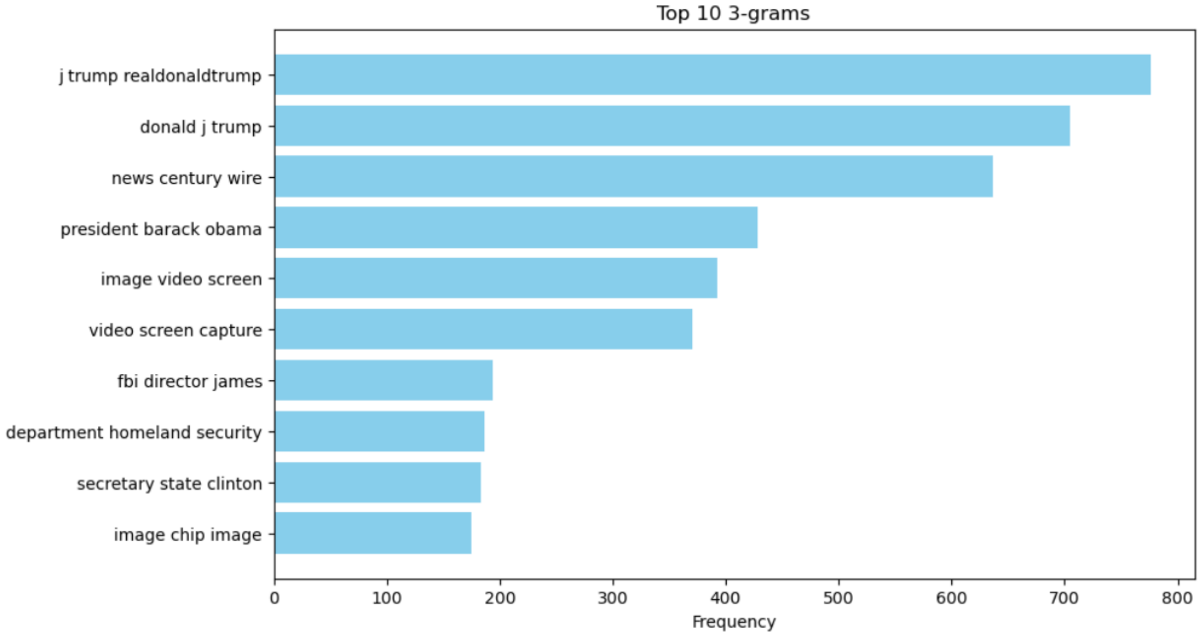


Top 10 Trigram in True data and Fake data set from Train data set

True Data set



False Data set



Conclusion

- The accuracy of decision tree model is less than Random Forest and Logistic regression model by approximately 10%.
- With the help of Logistic regression and Random Forest we are able to detect the fake new with an accuracy of approximately 92%
- After lemmatization, the data reduced significantly which helped us with faster model training.
- Using Unigram, Bigram and Trigram we are able to capture different word combination and identify the most frequently occurring phrase.