NLP PROJECT FOR DISASTER TWEET CLASSIFICATION

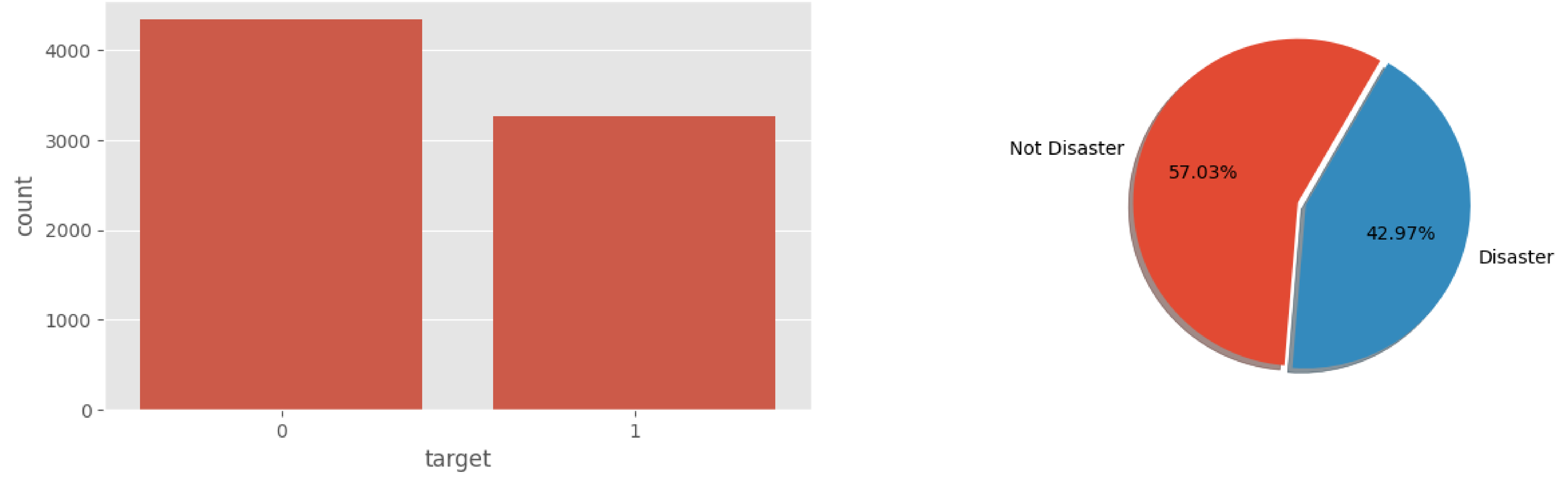
CHALLENGE

THE CHALLENGE IS TO BUILD A MACHINE LEARNING MODEL CAPABLE OF ACCURATELY CLASSIFYING TWEETS AS EITHER RELATED TO REAL DISASTERS OR NOT

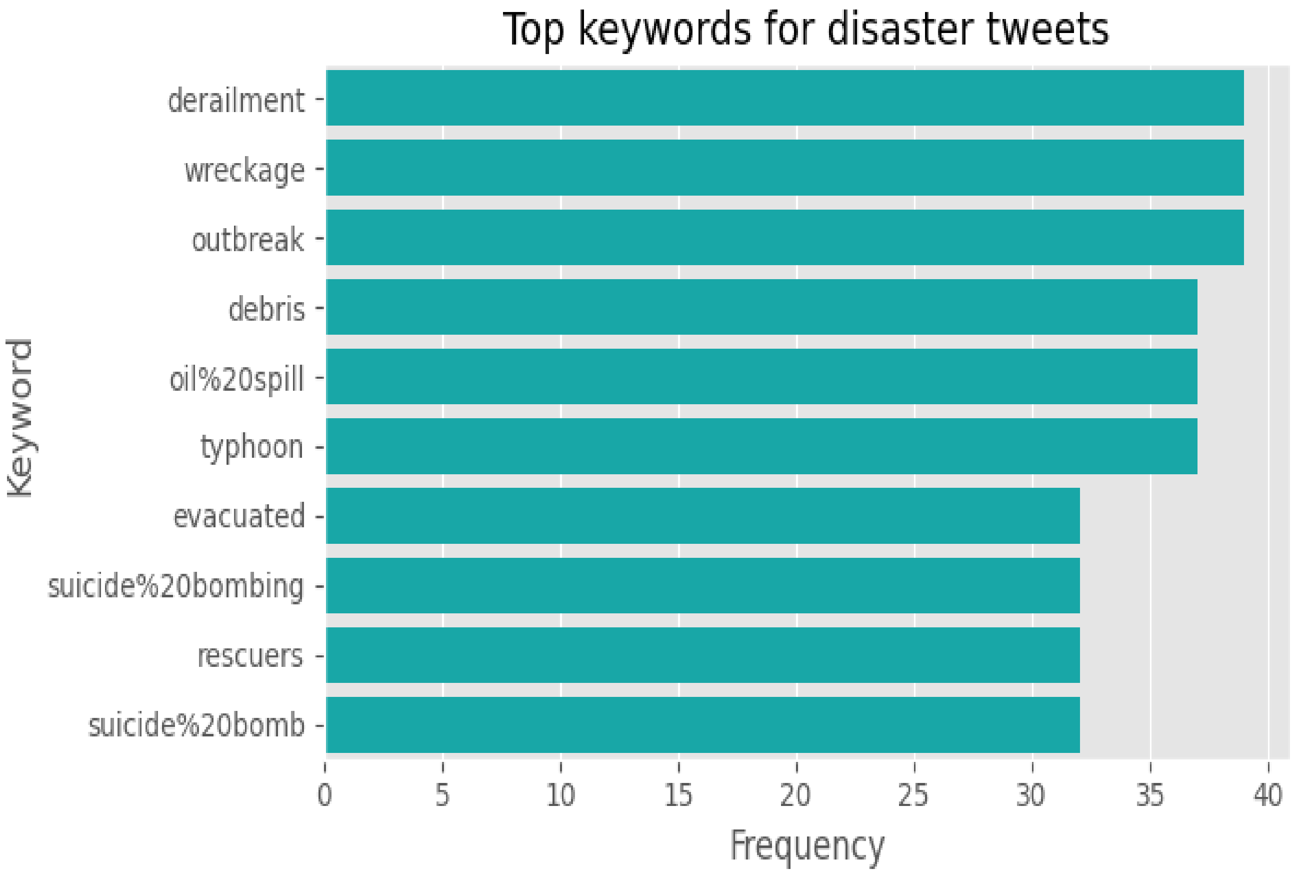
OBJECTIVES

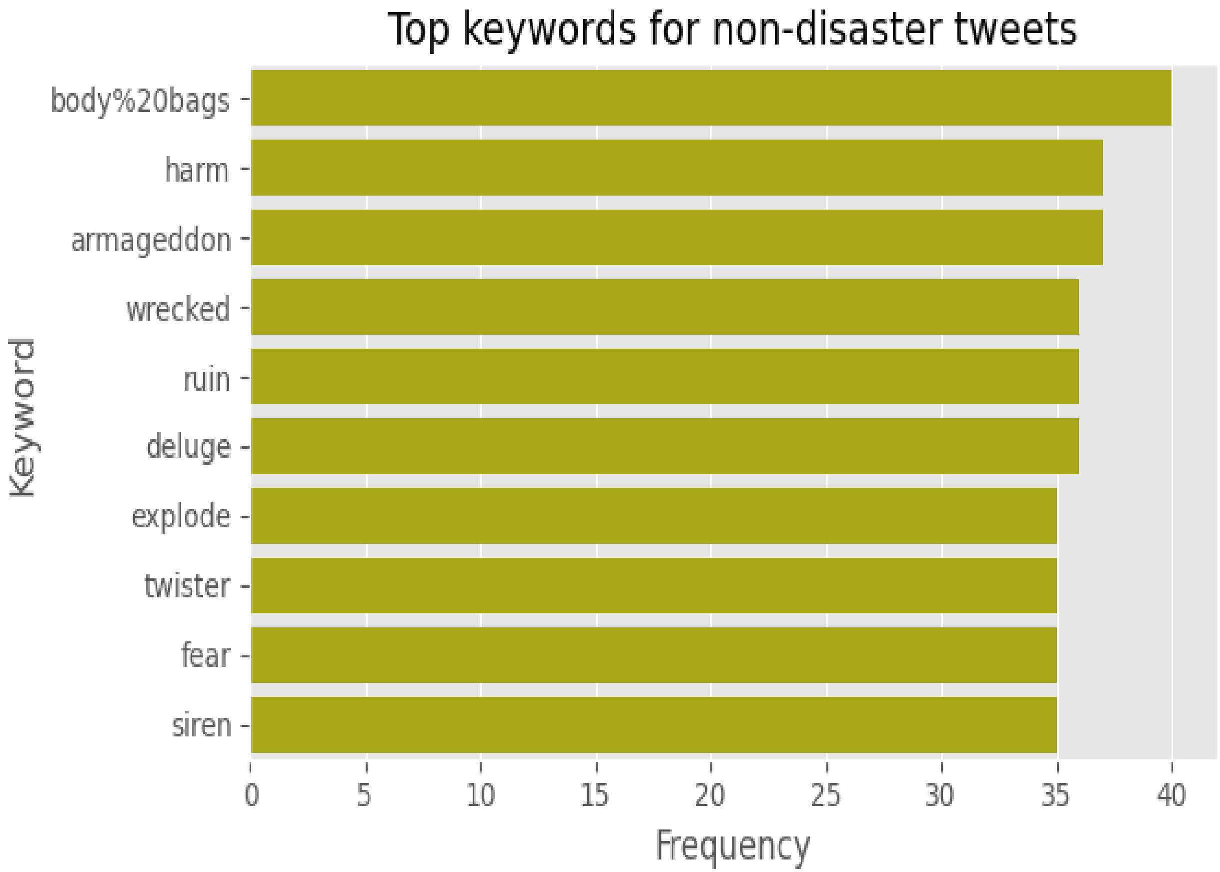
- CLASSIFICATION MODEL DEVELOPMENT
- ACCURACY AND PRECISION
- ROBUSTNESS
- SCALABILITY

Disaster vs Non disaster Tweets

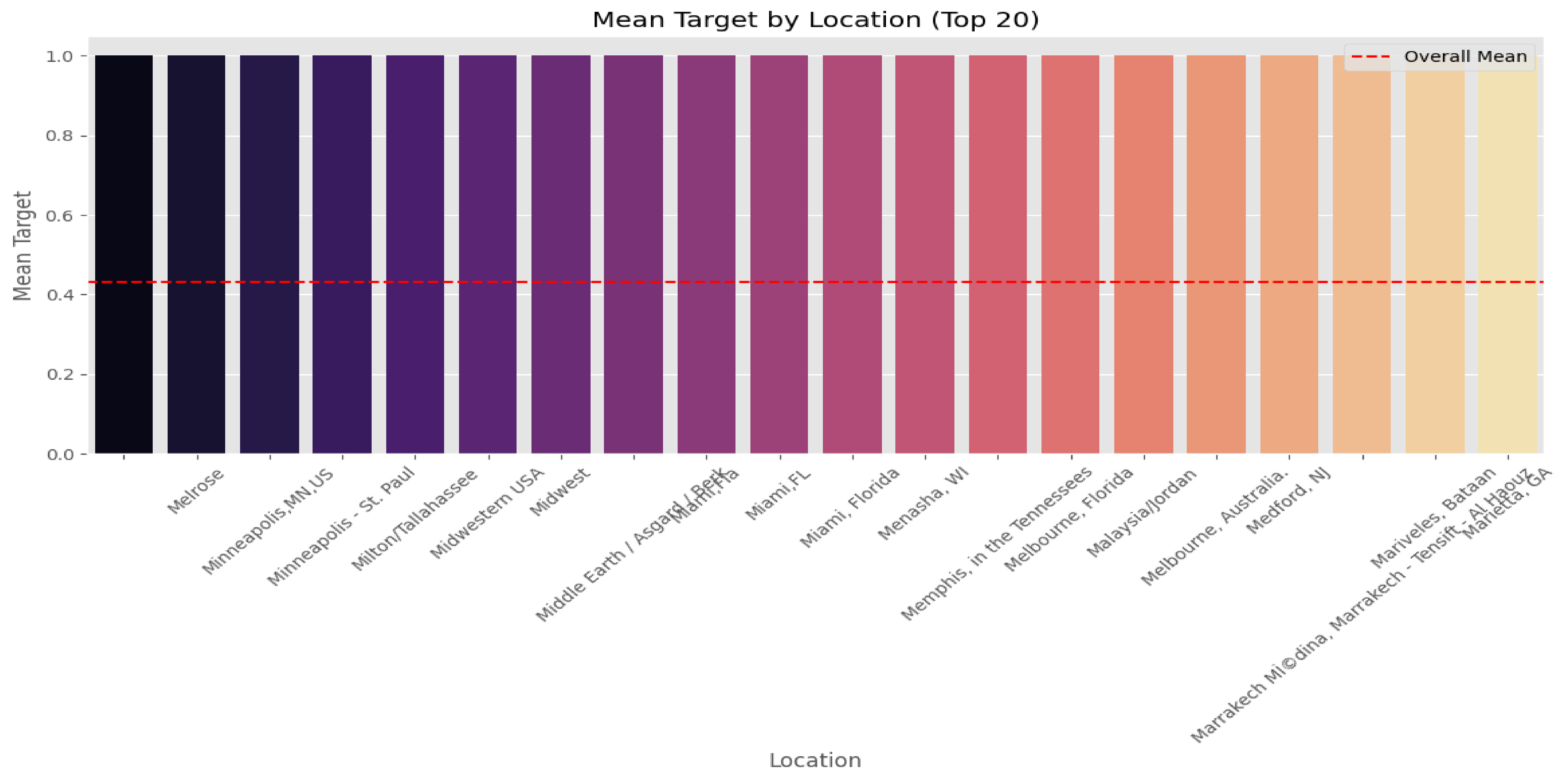


There are more tweets with class 0 (No disaster) 57% than class 1 (disaster tweets) 43%

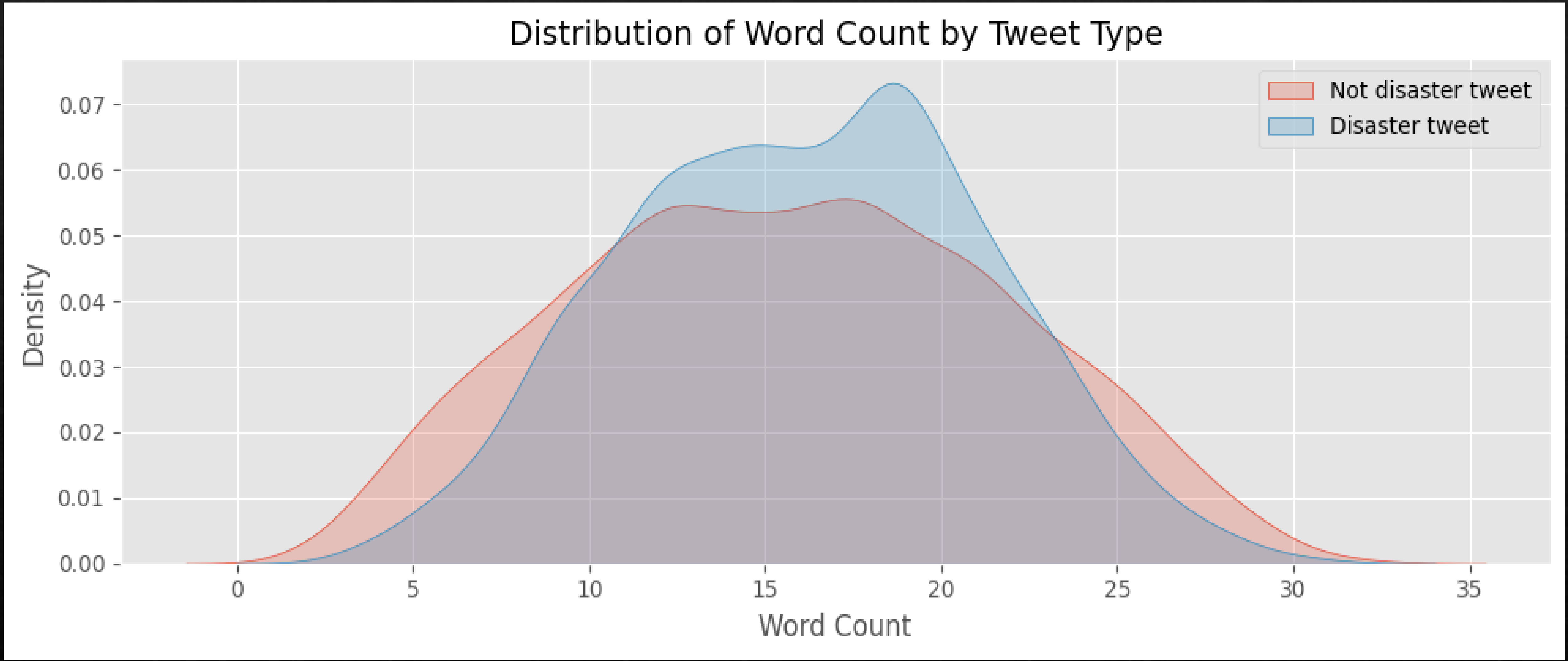




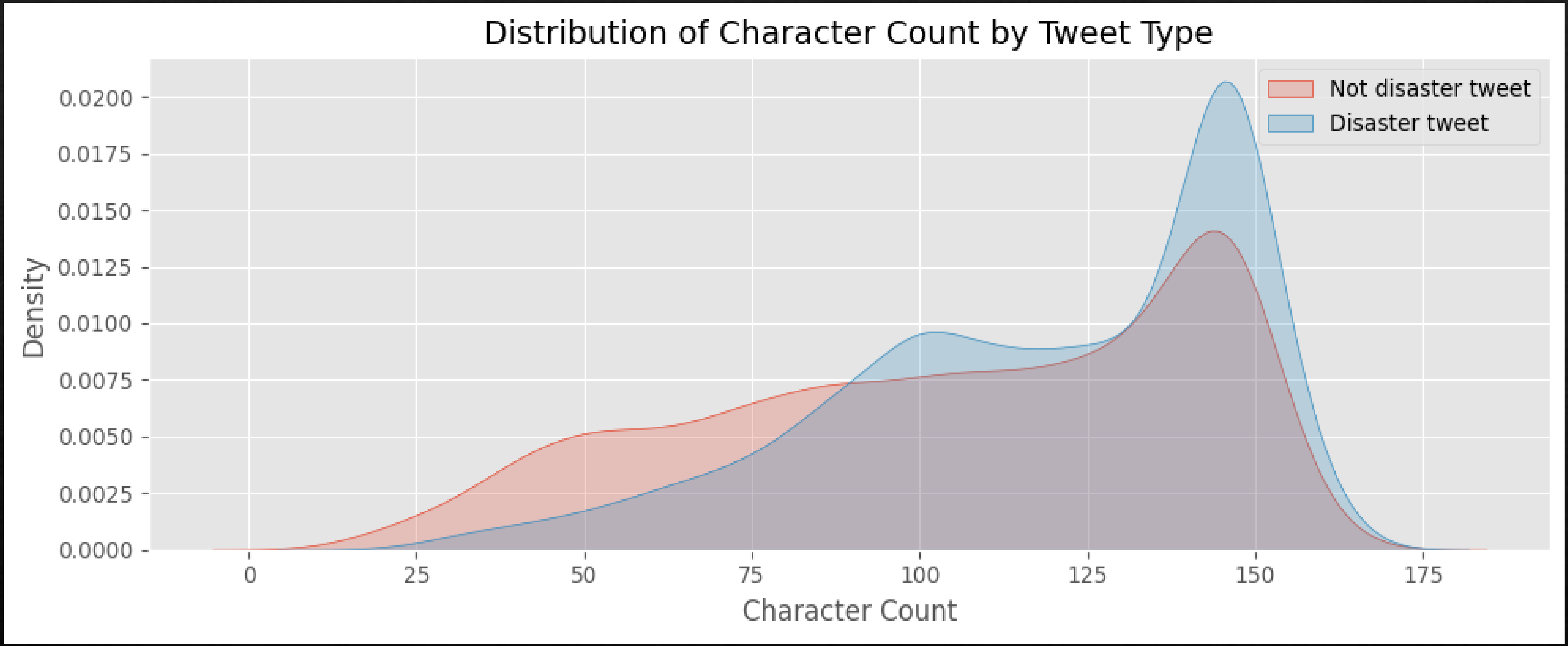
No common top 10 keywords between disaster and non-disaster



Here we have grouped Target and Location to calculate top 20

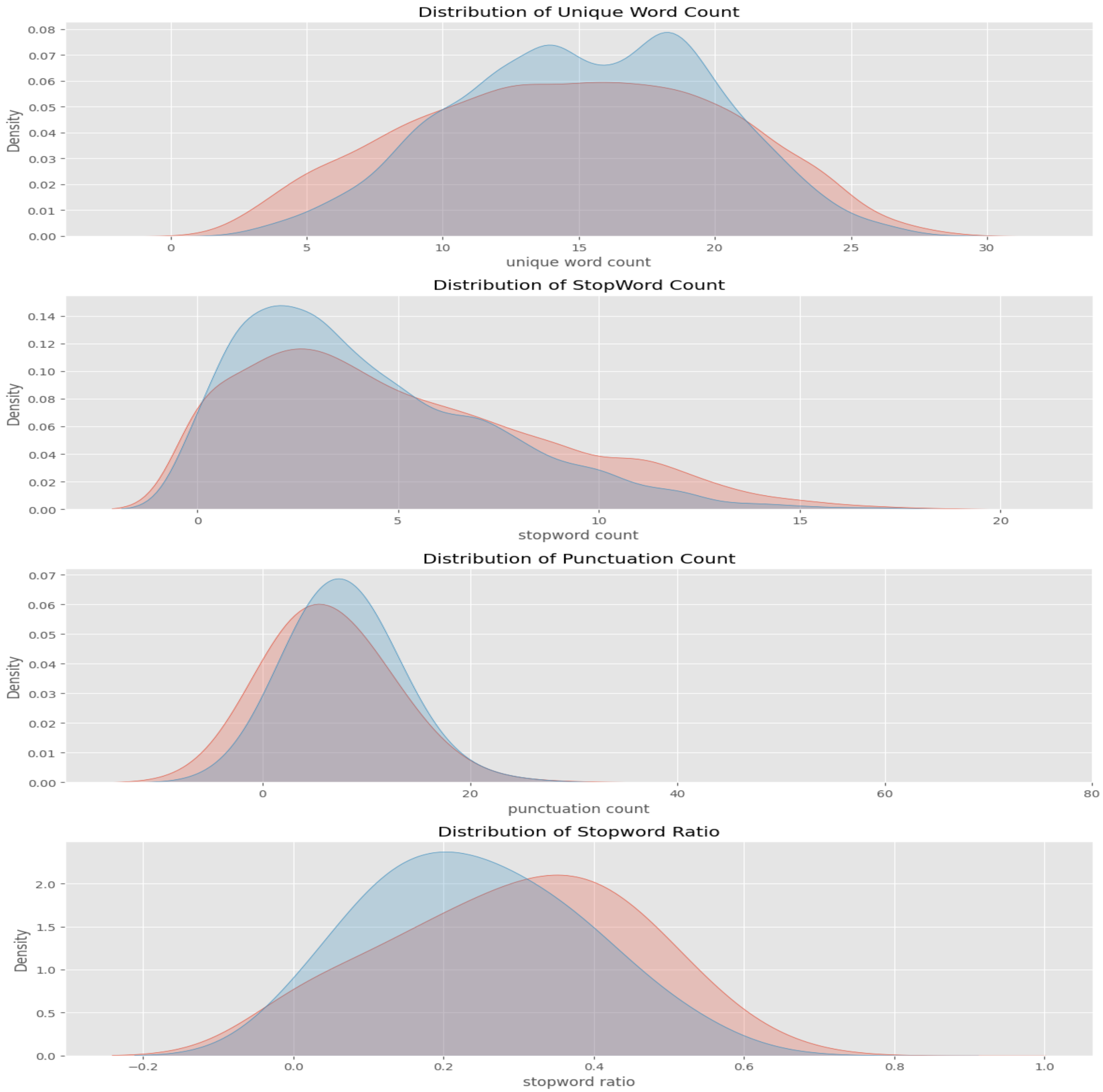


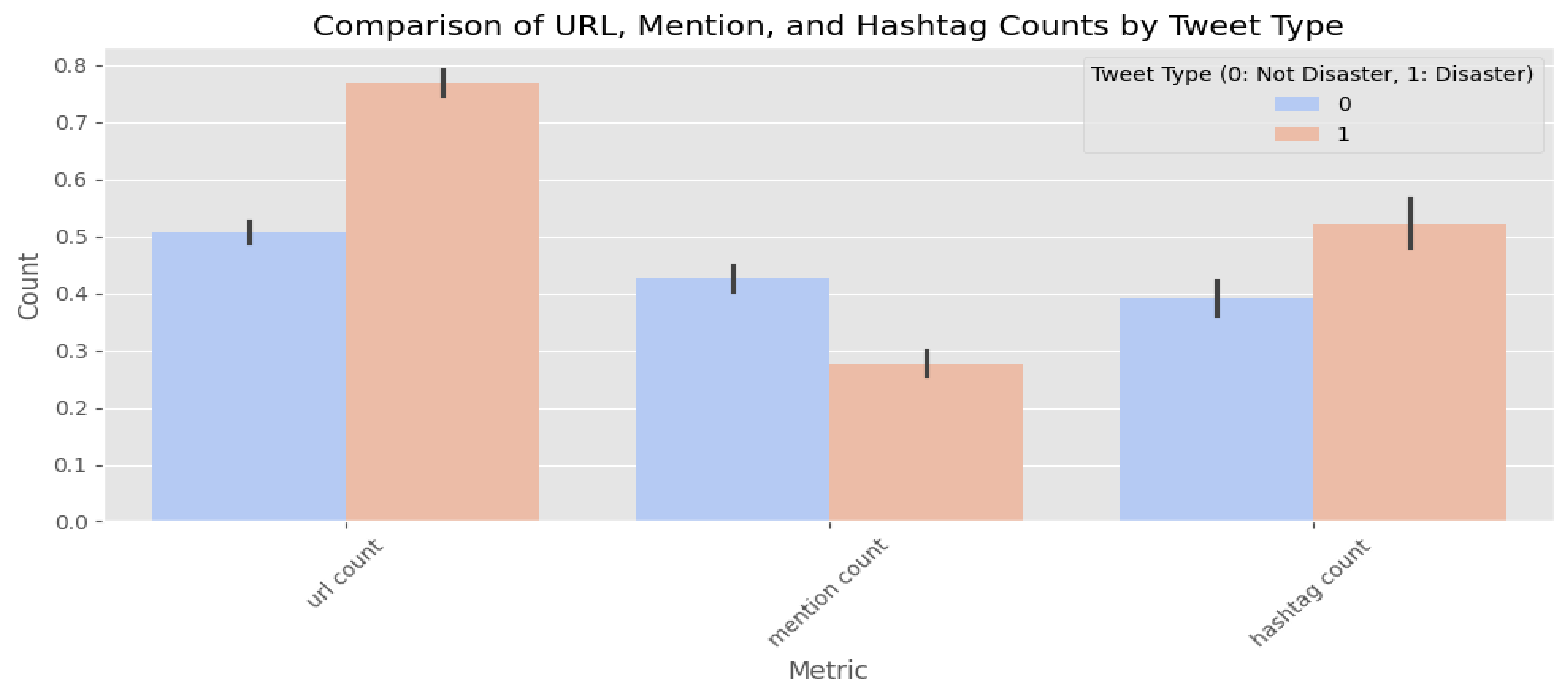
Disaster tweets are more from 15 to 20 word count category as compared to non disaster tweets



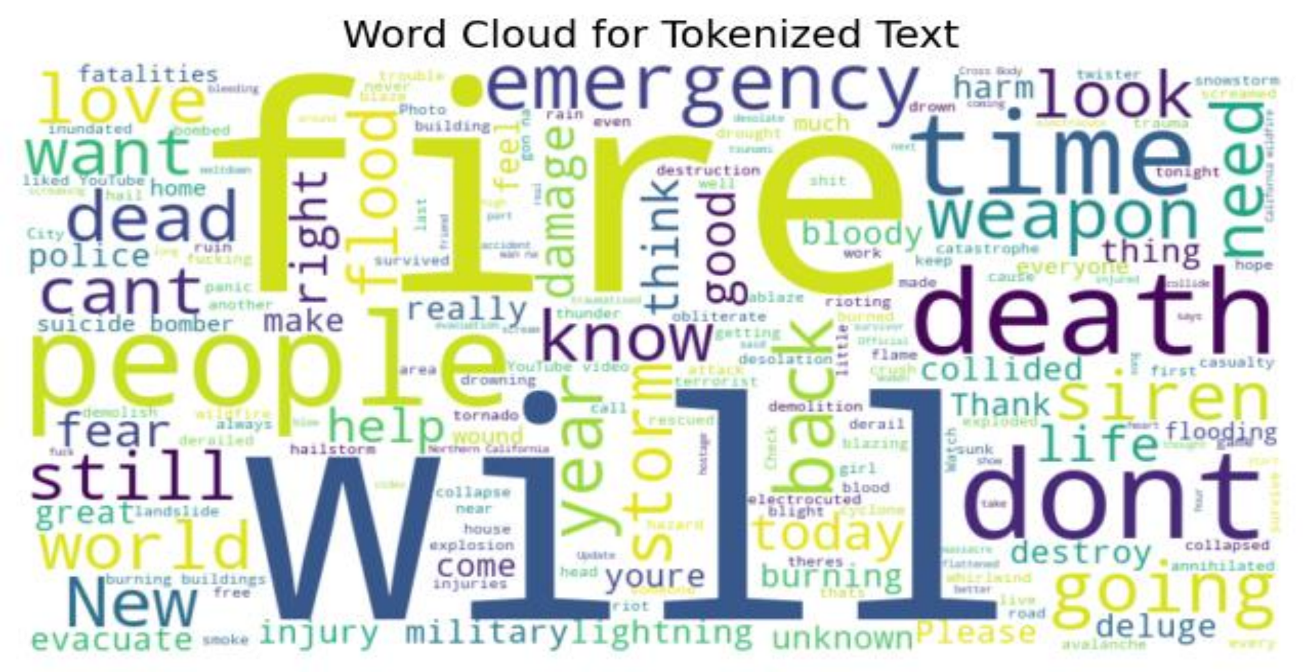
It tells us that very few disaster tweets are less than 50 characters and that the majority of them are more than 125 characters long

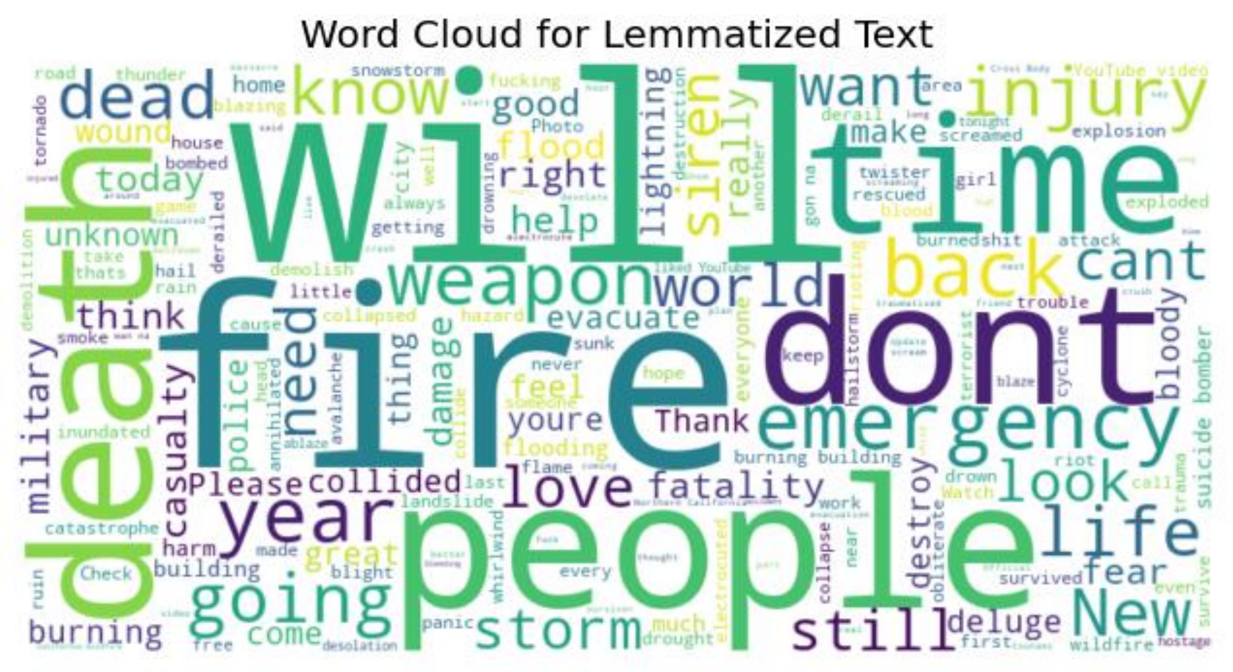
Unique words are much more as compared to Stopwords and Punctuation

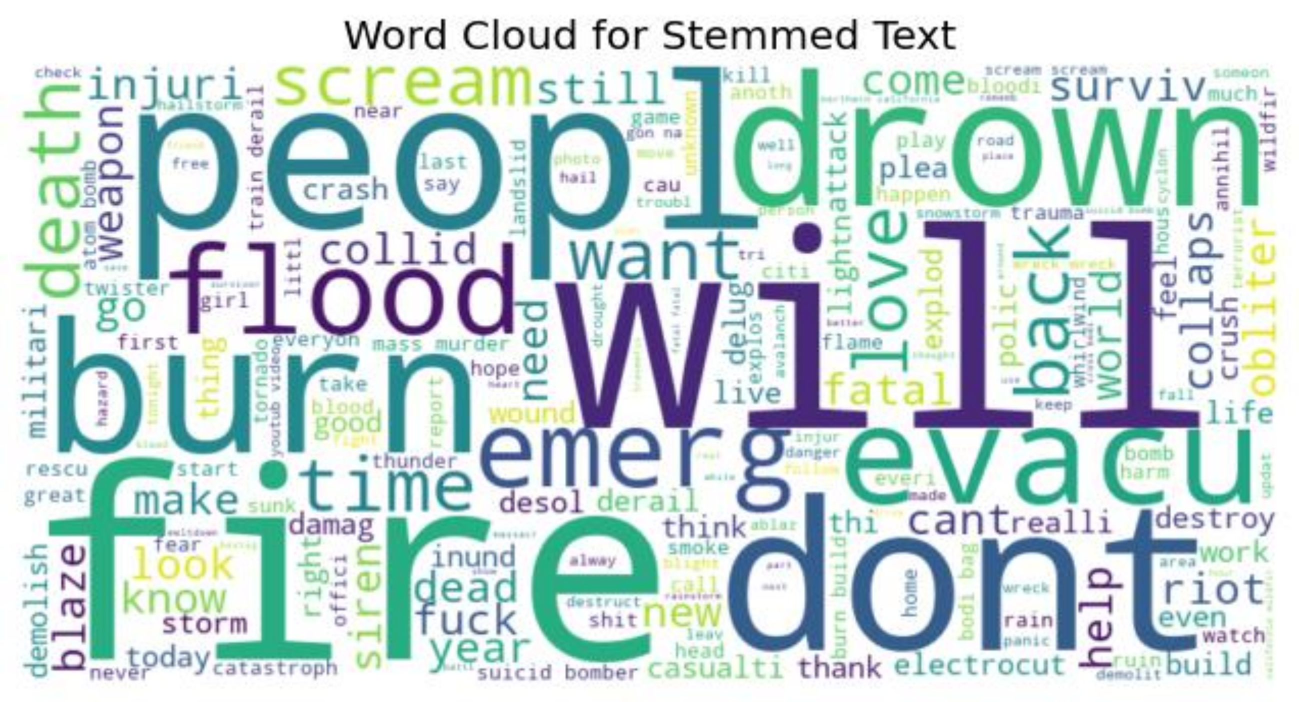




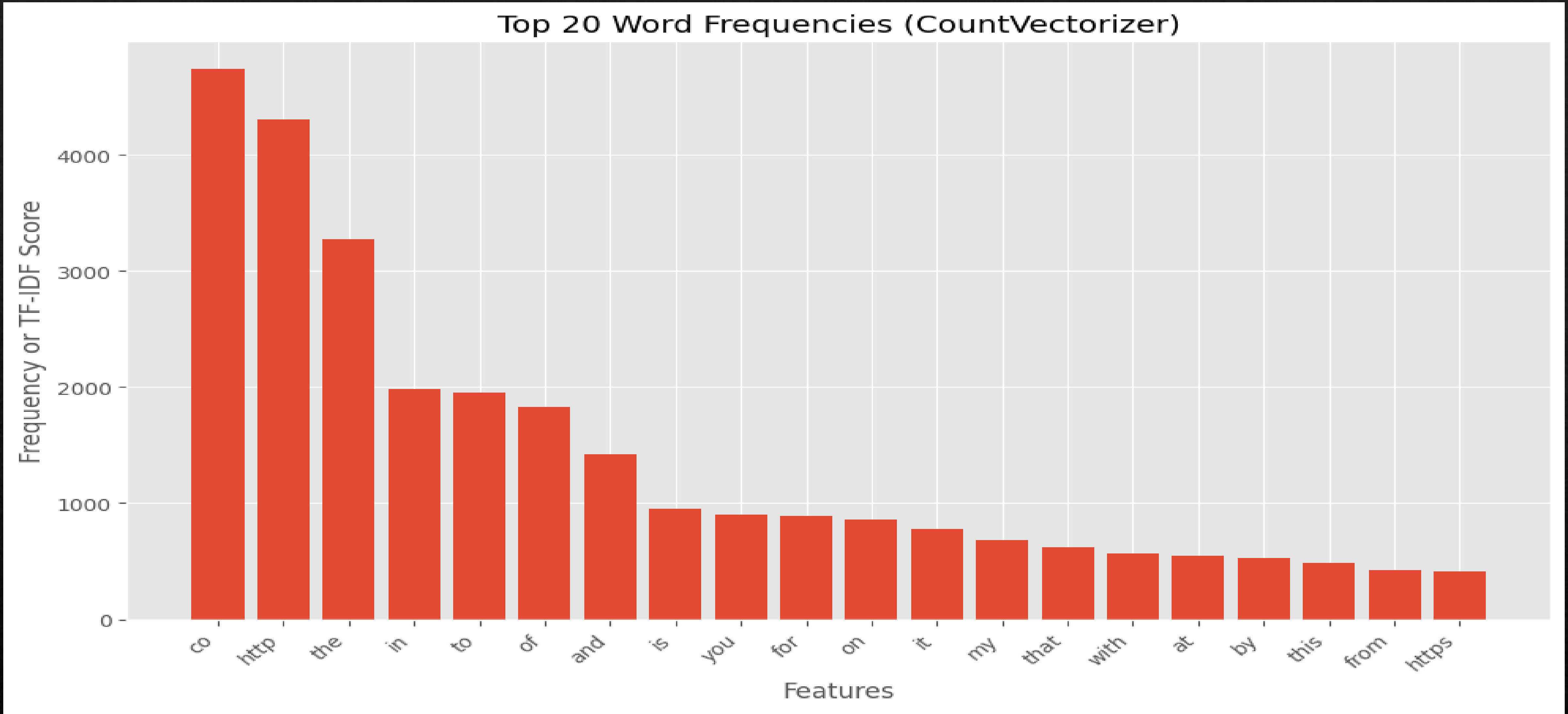
URL counts are maximum followed by Hashtag and Mention count



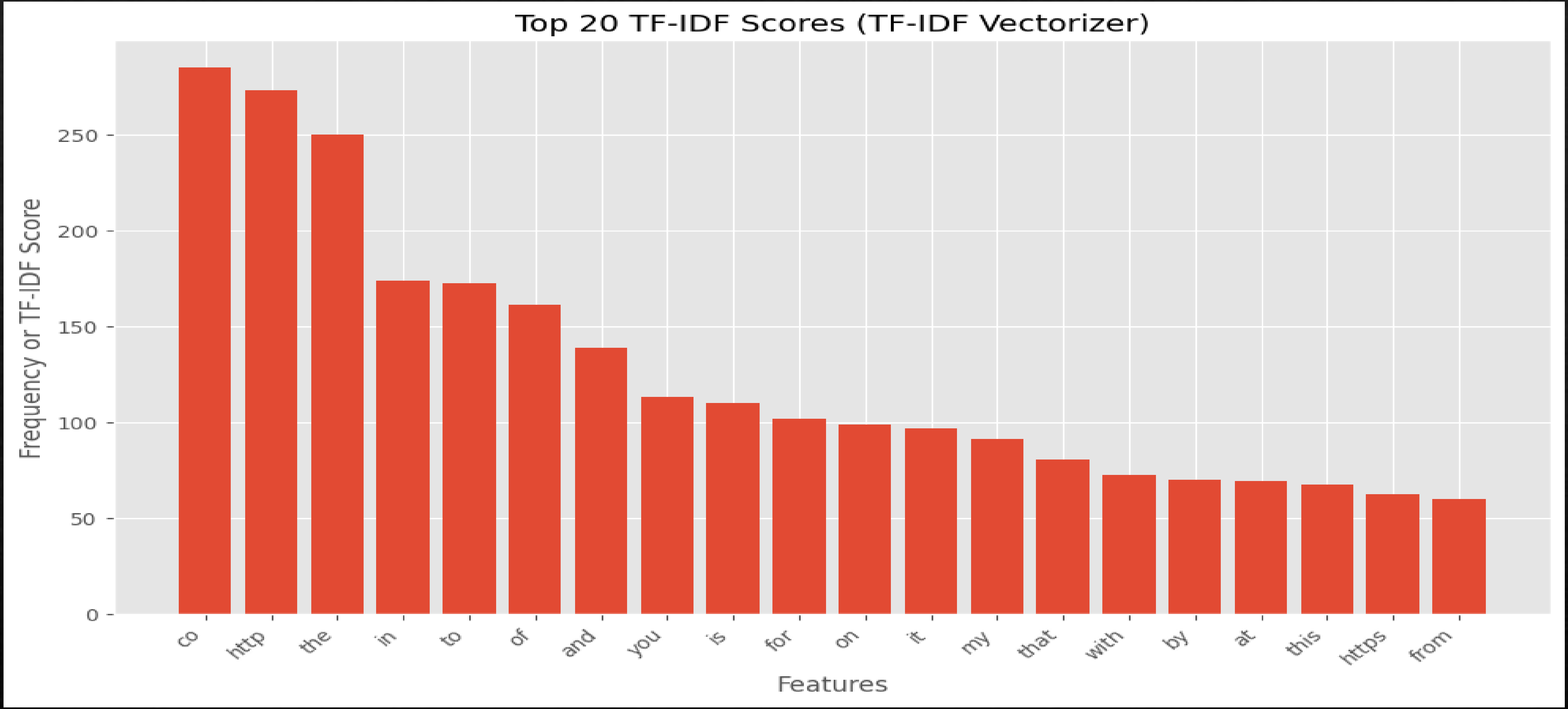




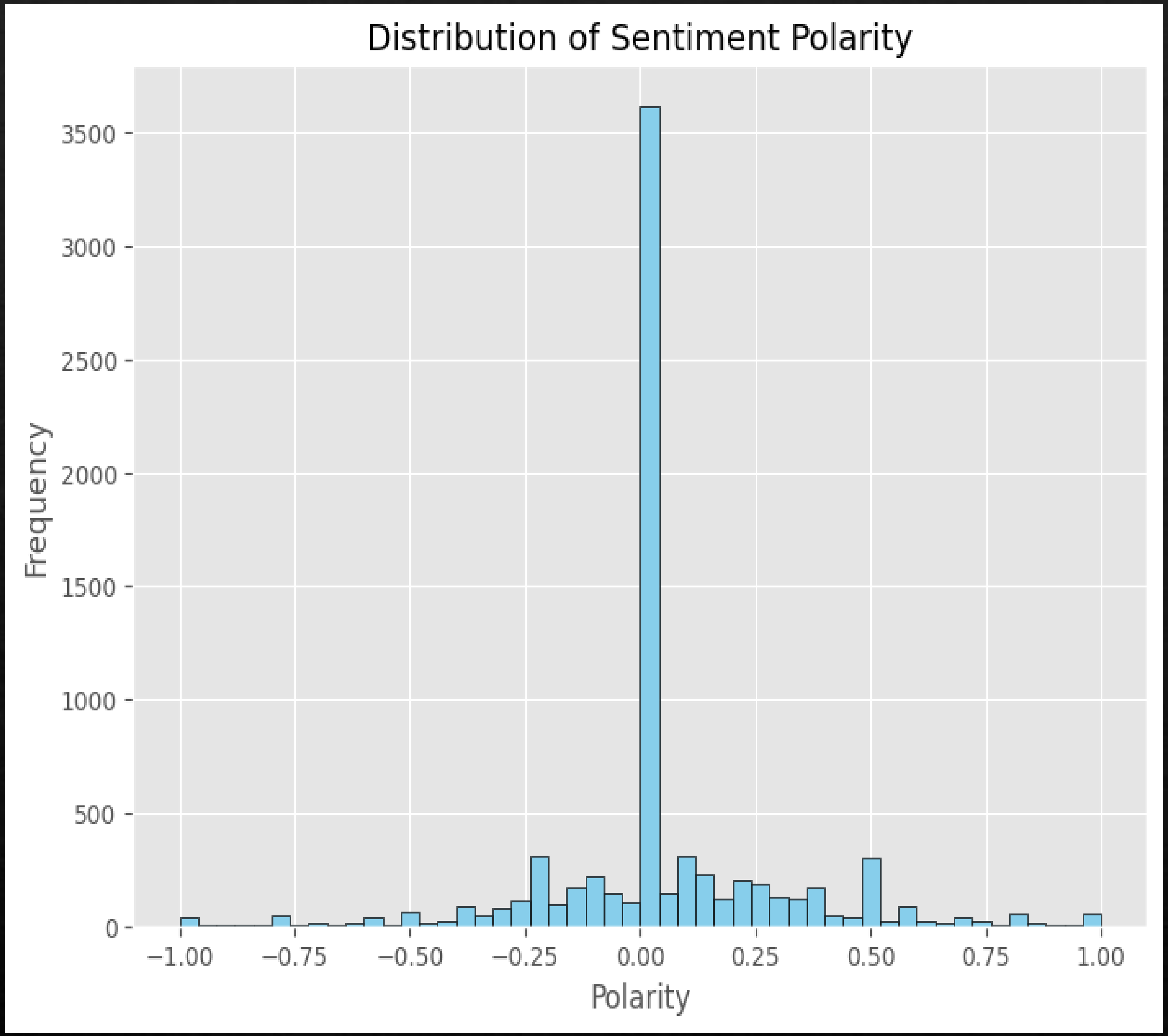
These are the Wordclouds for Tokenized text, Stemmed text and Lemmatized text



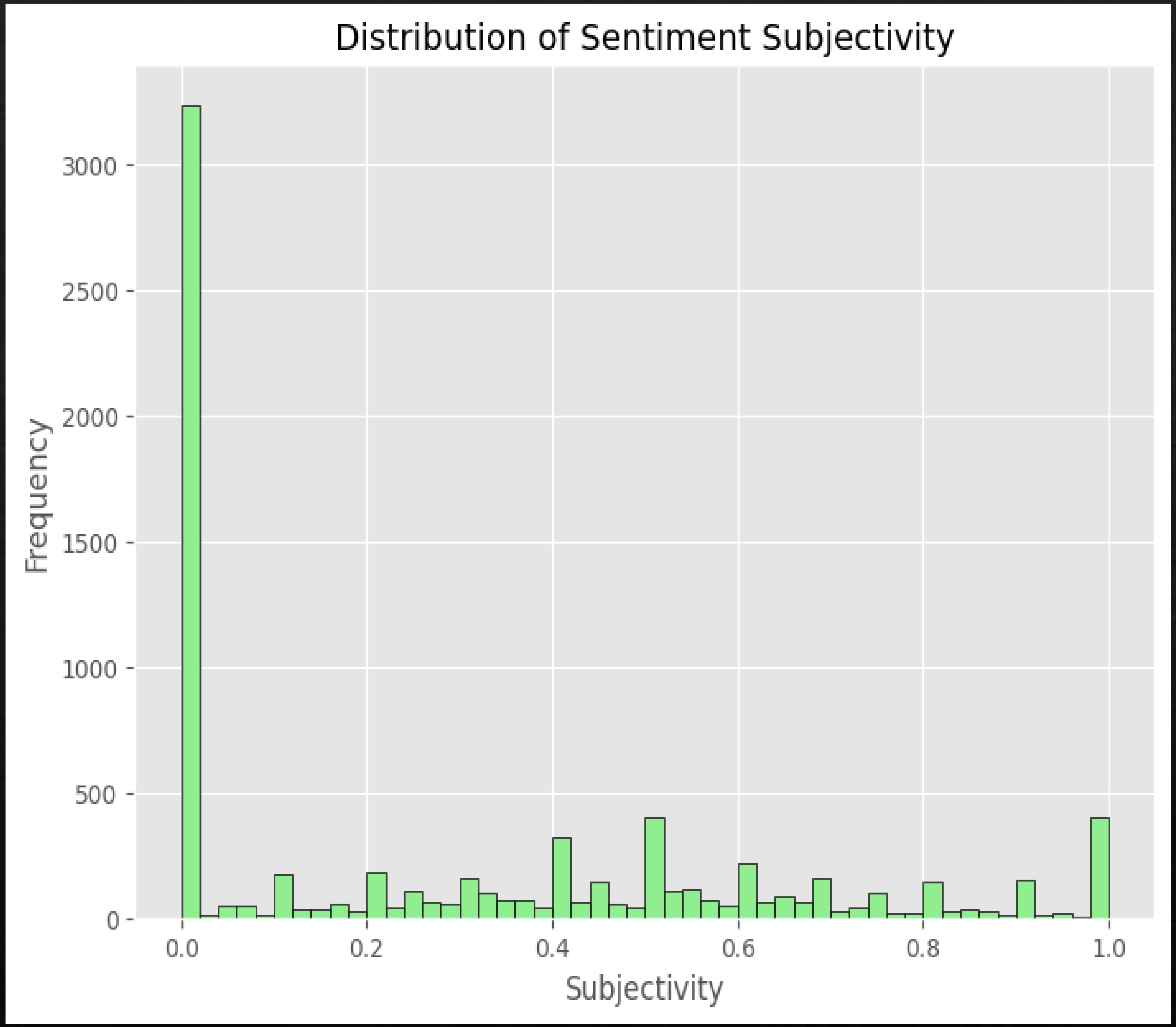
According to Count Vectorizer co and http are the words which are even more than 4000 in numbers followed by the, in, to, of, and



TF-IDF Vectorizer also shows the same result that co, http are maximum followed by the, in, to, of, and

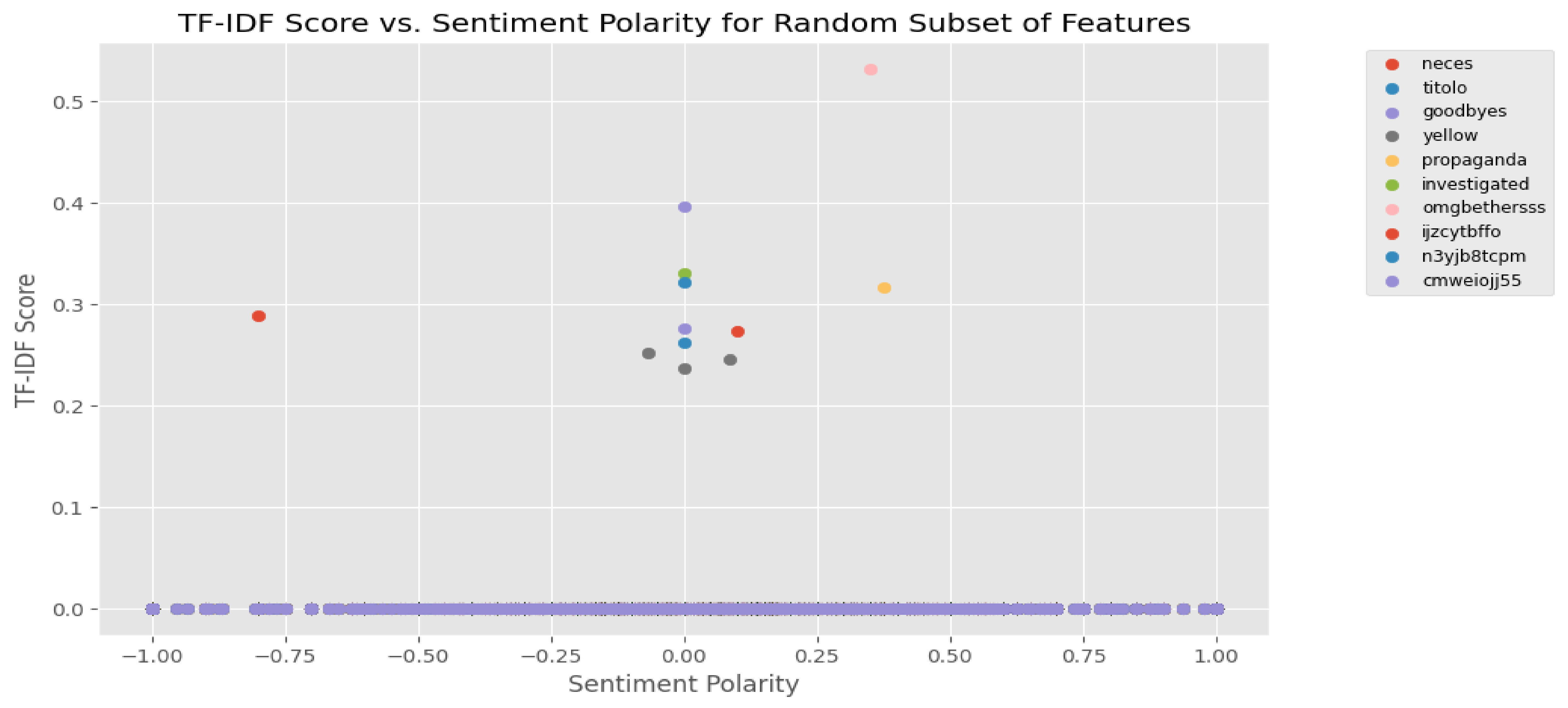


Distribution of Sentiment Polarity

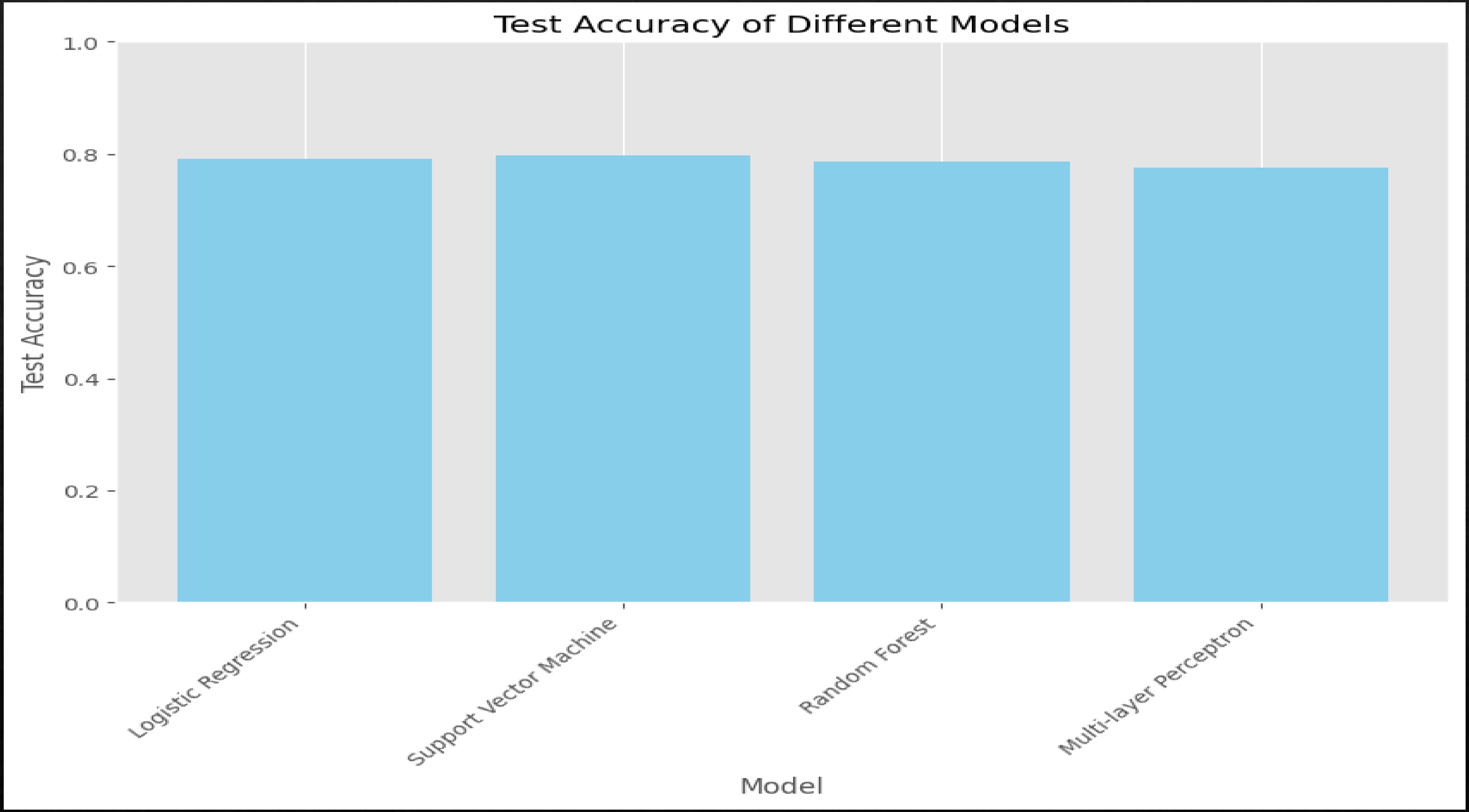


Distribution of Sentiment Subjectivity

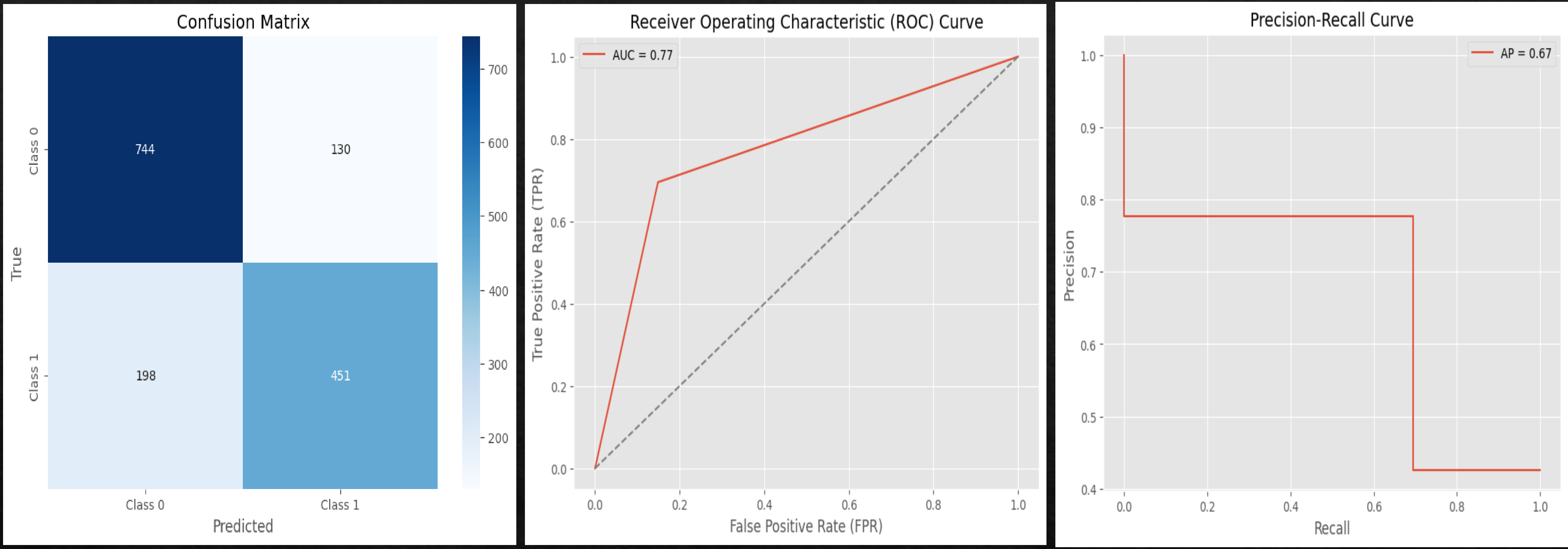
Sentiment Polarity and Sentiment Subjectivity both are maximum at point 0



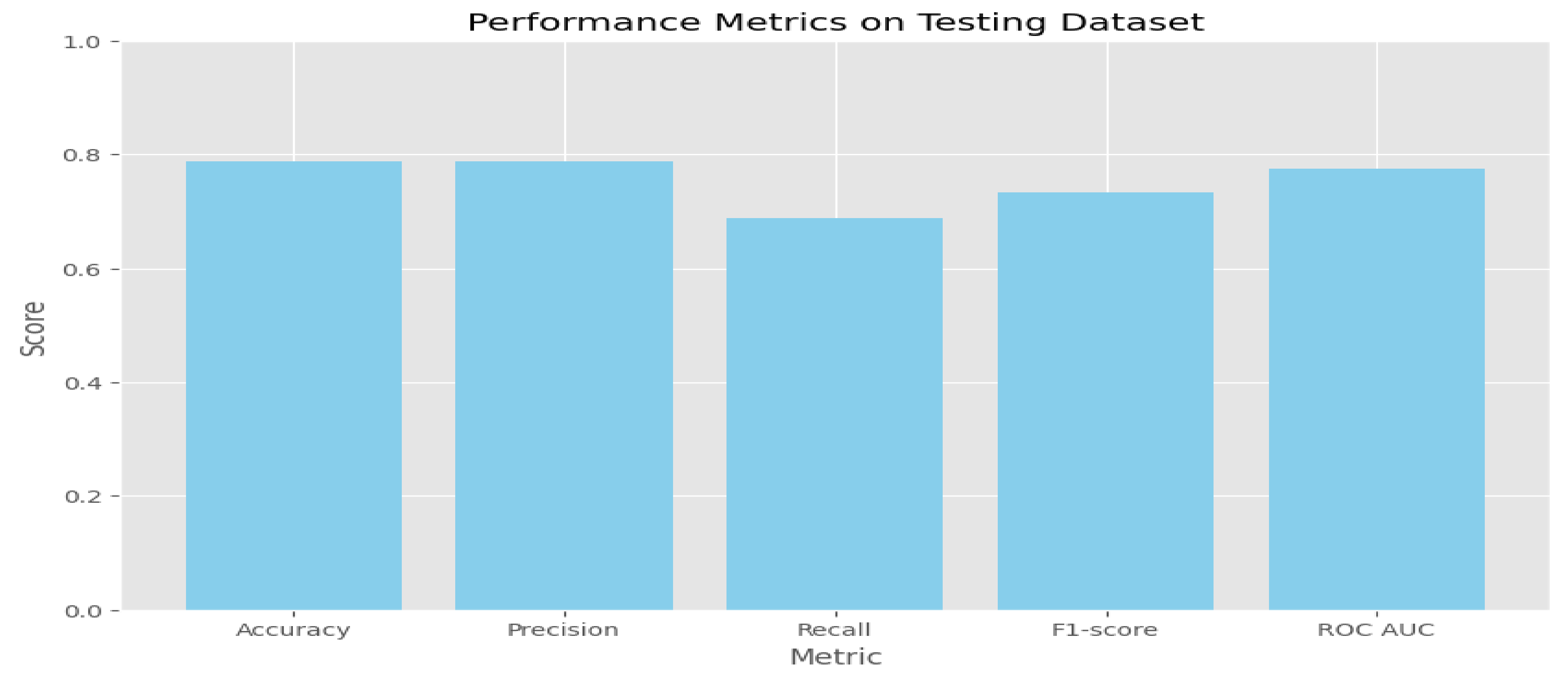
Only 2 or 3 subsets are at points other than 0



The accuracy of all 4 models are almost same but Logistic Regression is the best model in this dataset

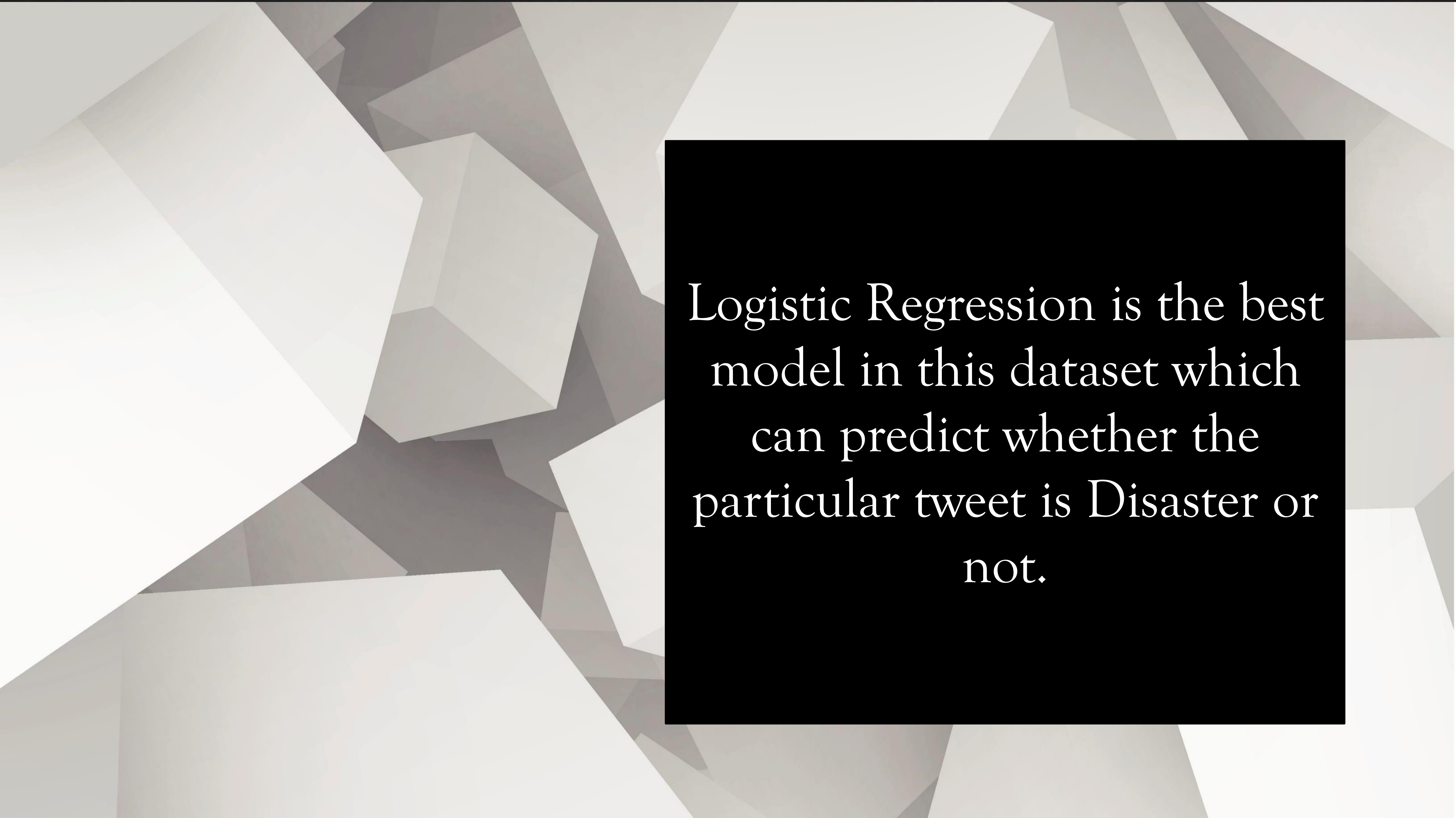


These are the graphs of Confusion Matrix, ROC Curve and Precision-Recall Curve where AUC comes out to be 0.77 and AP is 0.67



```
             RandomForest LogisticRegression SupportVectorMachine MultiLayerPerception


Accuracy          0.78594            0.788575           0.790545           0.754432
 Precision       0.793103             0.769357         0.790493          0.714509
 Recall          0.673344             0.719569         0.691834          0.705701
 F1-score        0.728333             0.743631         0.737880          0.710078
 ROC AUC         0.771454             0.779693         0.777839          0.748159
```



Logistic Regression is the best model in this dataset which can predict whether the particular tweet is Disaster or not.

THANKYOU

By:
NITIMA SAIGAL