Social Media

Case Study: Subreddit Political Verbiage

Background

- Political media has become essential to the exchange of political content on platforms the major players being facebook, twitter, and google

Initial Strategic Plan for political candidate support

SERRA
PUBLIC AFFAIRS

Data Collection & Preliminary Cleaning

- Pulled about 3,000 posts from AskPolitics and Conspiracy Subreddits
- Removed duplicate titles, texts removed by moderators, blanks
- Cleaned titles and text with RegEx
- Feature Engineered word counts, punctuation count, and upper/lower case count

Final dataset:

- Ultimately left with approximately 1,000 subreddit posts per subreddit

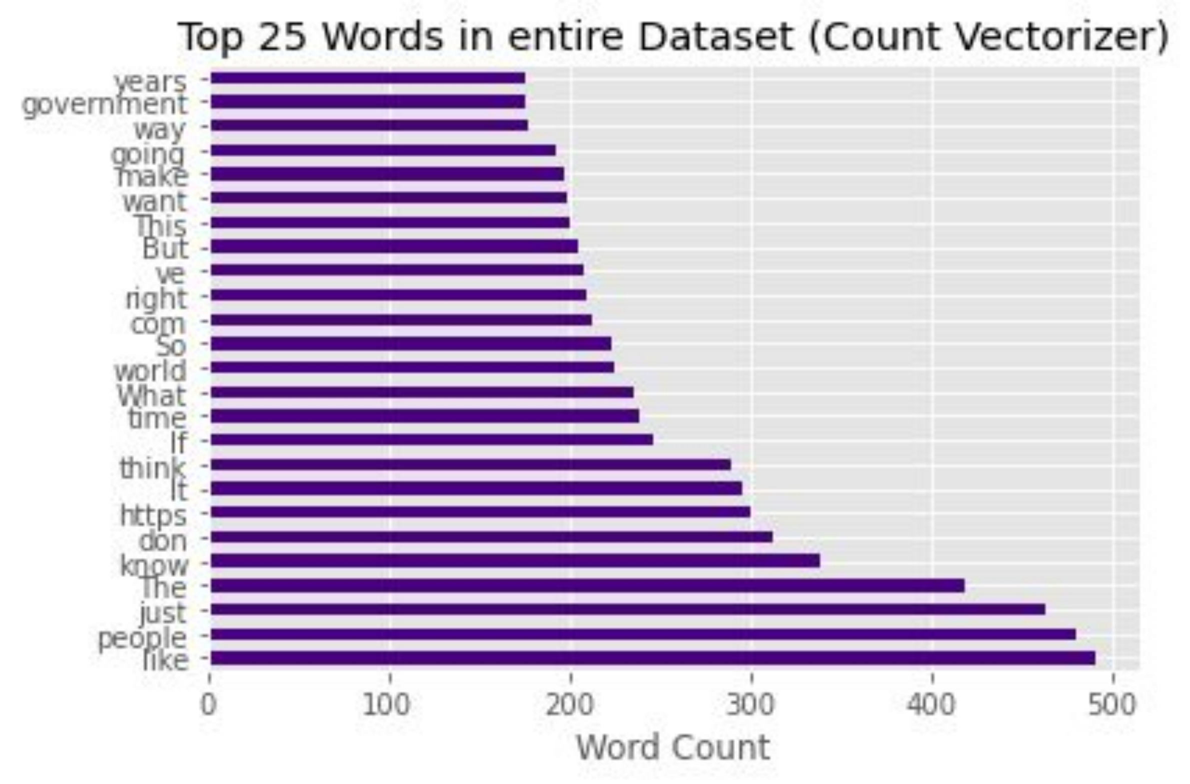
Exploratory Data Analysis

Overall and by subreddits

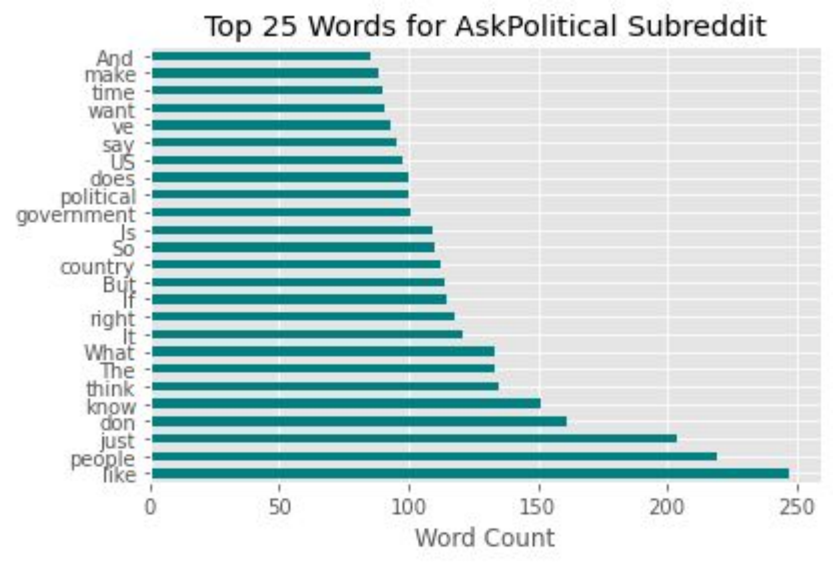
Characteristics explored:

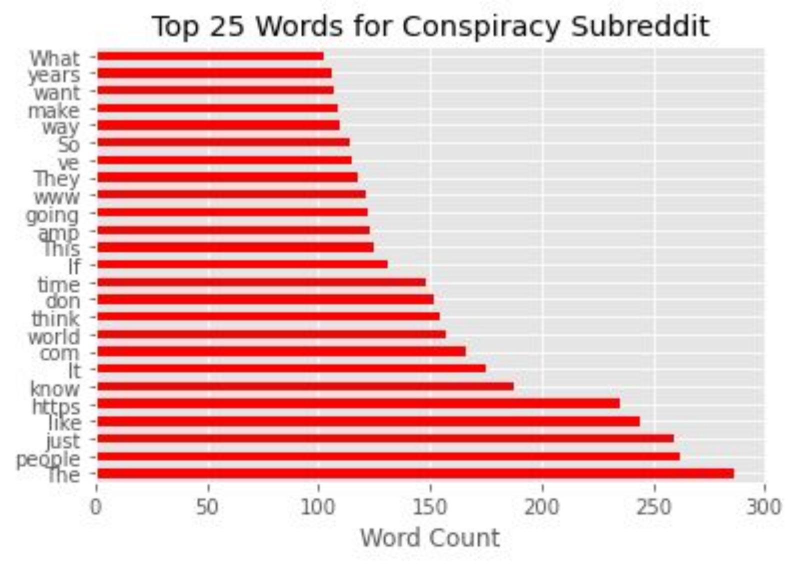
- Top 25 Words
- Word Count
- Punctuation Count
- Correlations among features (including Sentiment Analysis Features)

Top 25 entire Dataset

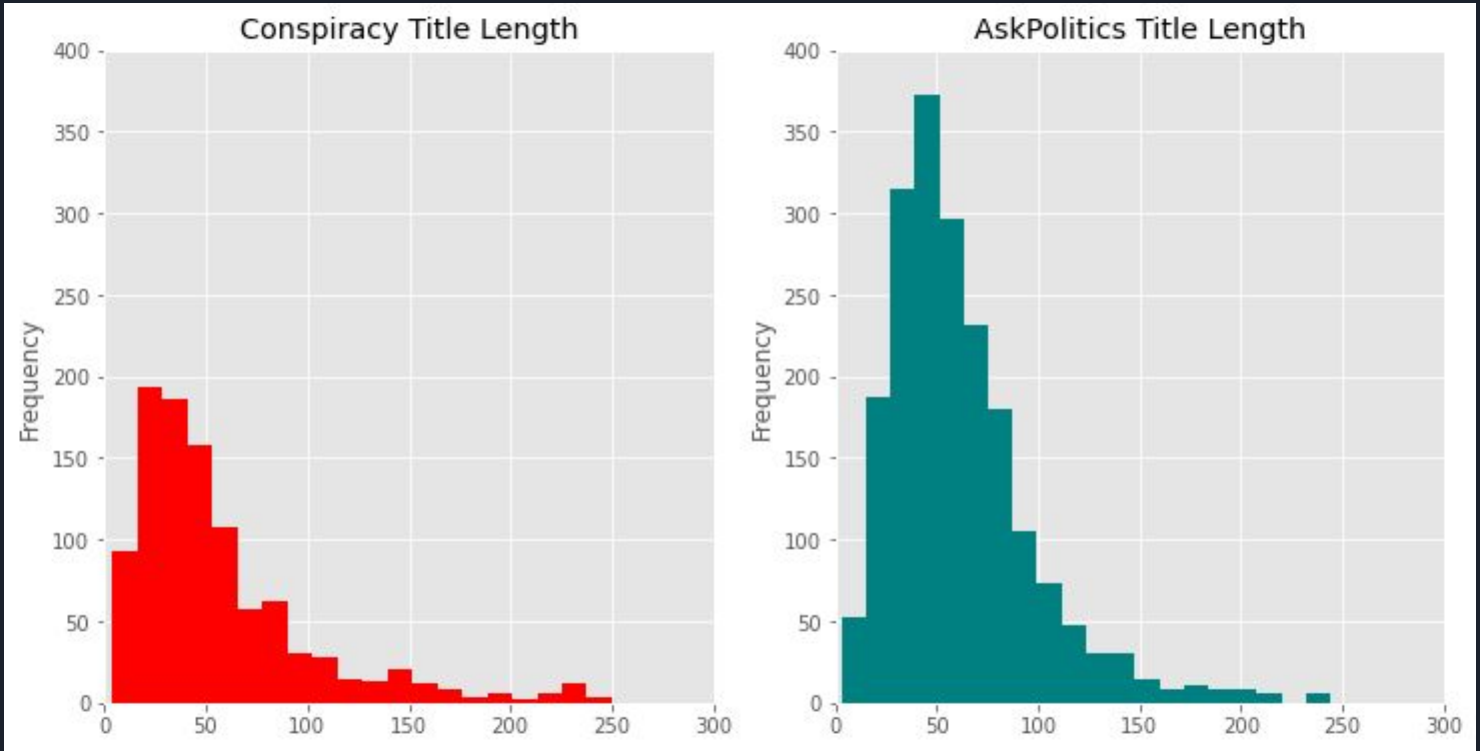


Top 25 per Subreddit

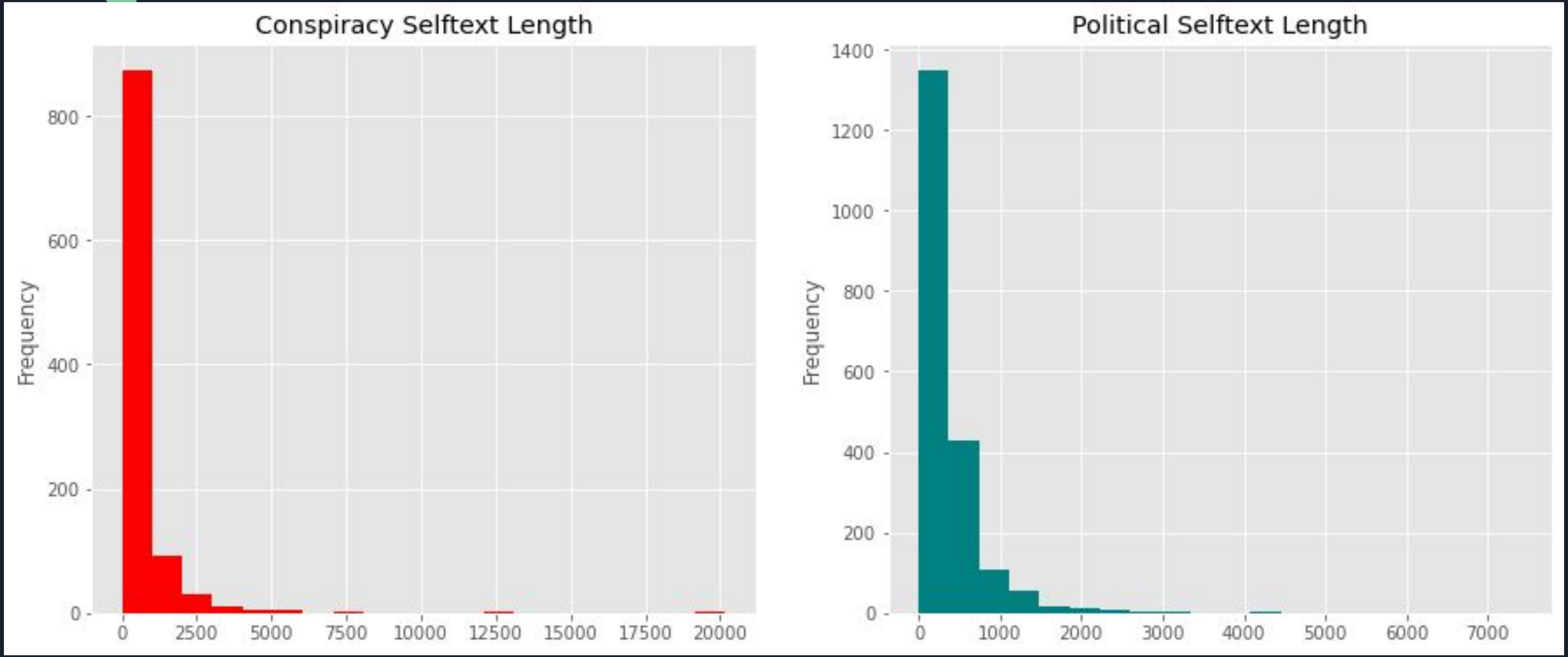

Top 25 Words for AskPolitical Subreddit


Top 25 Words for Conspiracy Subreddit

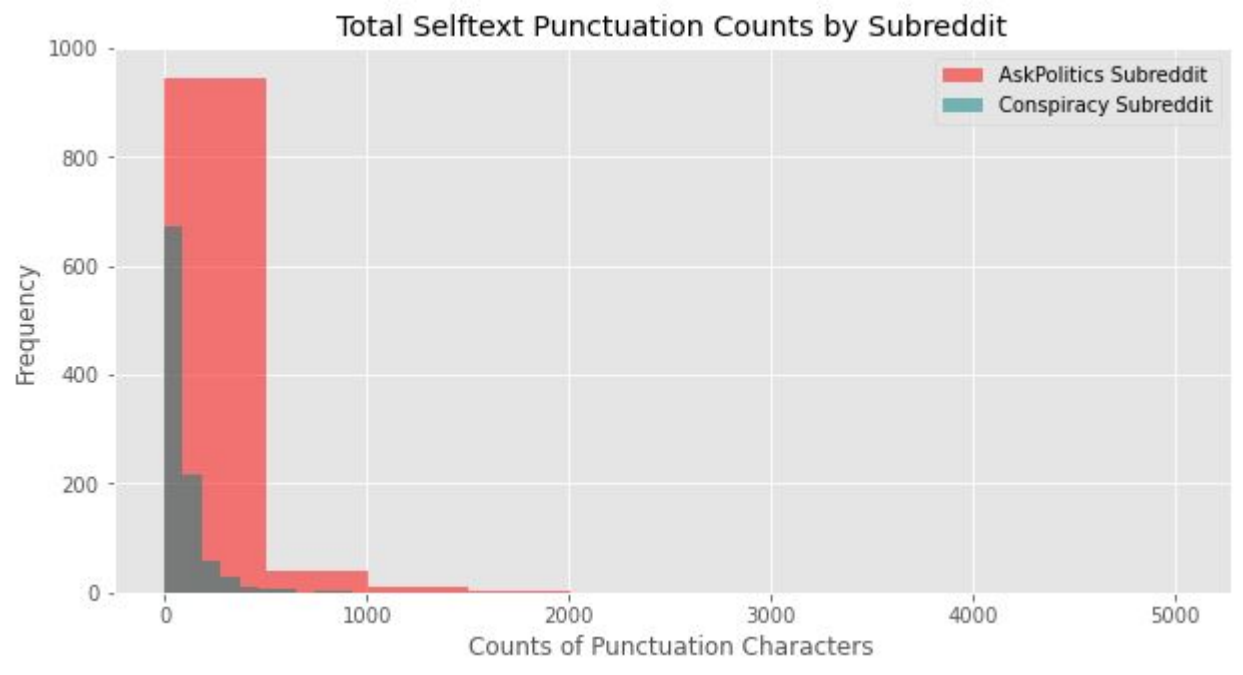
Title Length Distributions per Subreddit

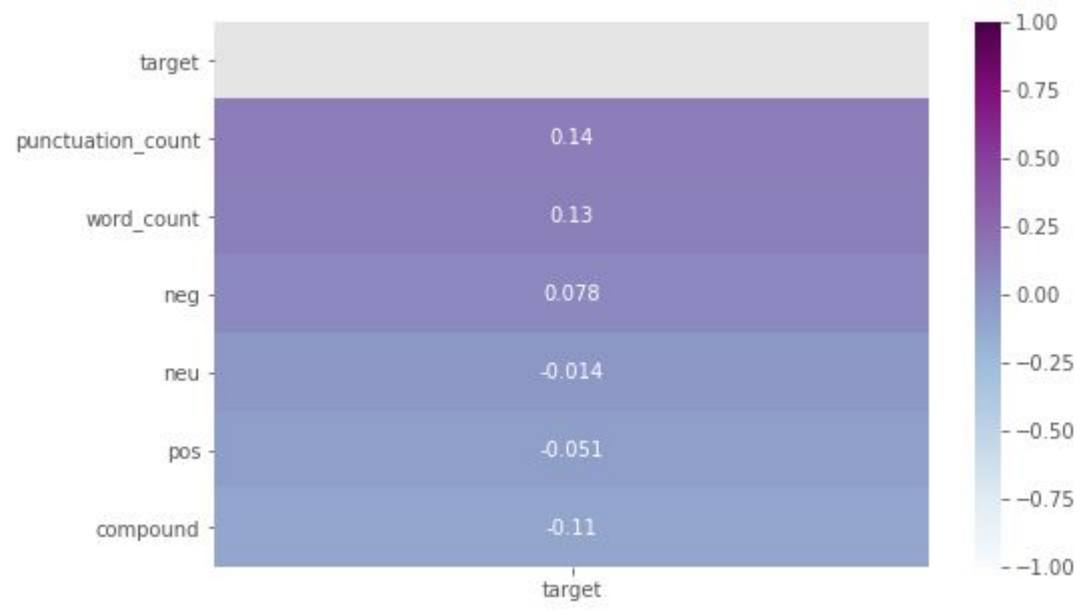


Self Text Length Distributions per Subreddit



Self Text Punctuation Count Distributions per Subreddit




target
punctuation_count
word_count
neg
neu
pos
compound
0.14
0.13
0.078
-0.014
-0.051
-0.11
target
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

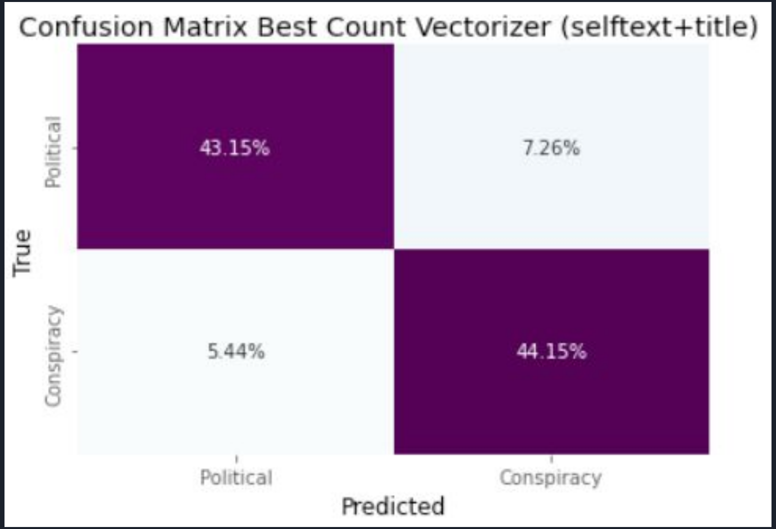
Model Approach

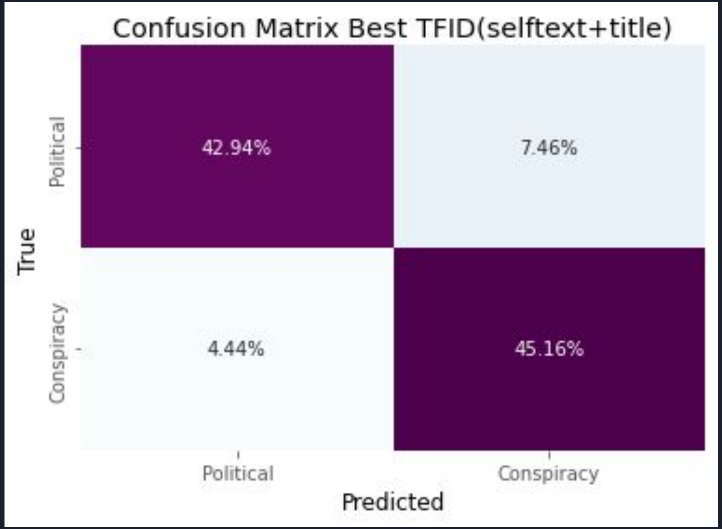
- Logistic Regression prioritized
- Random Forest explored, ultimately not heavily optimized

Misclassifications (TFID vs CV) before Word Cnt + Punc counter

Countvectorizer (First)

TFID (Second)

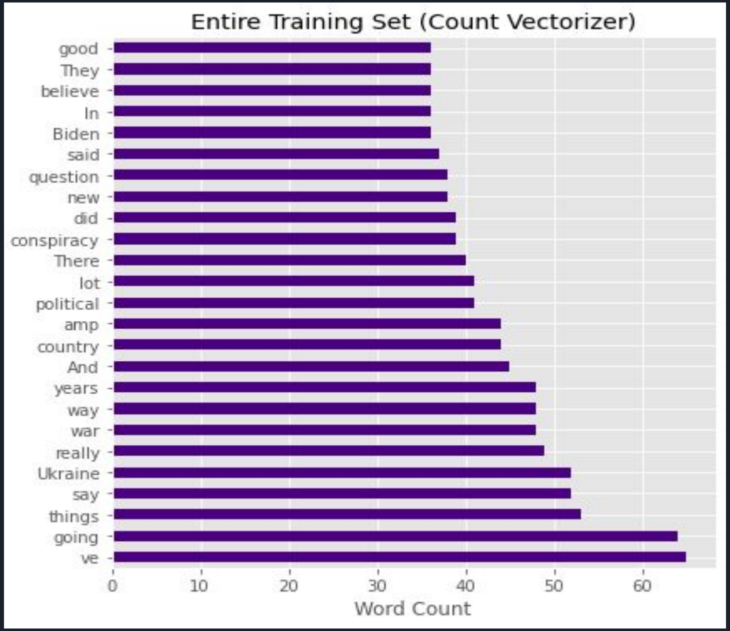




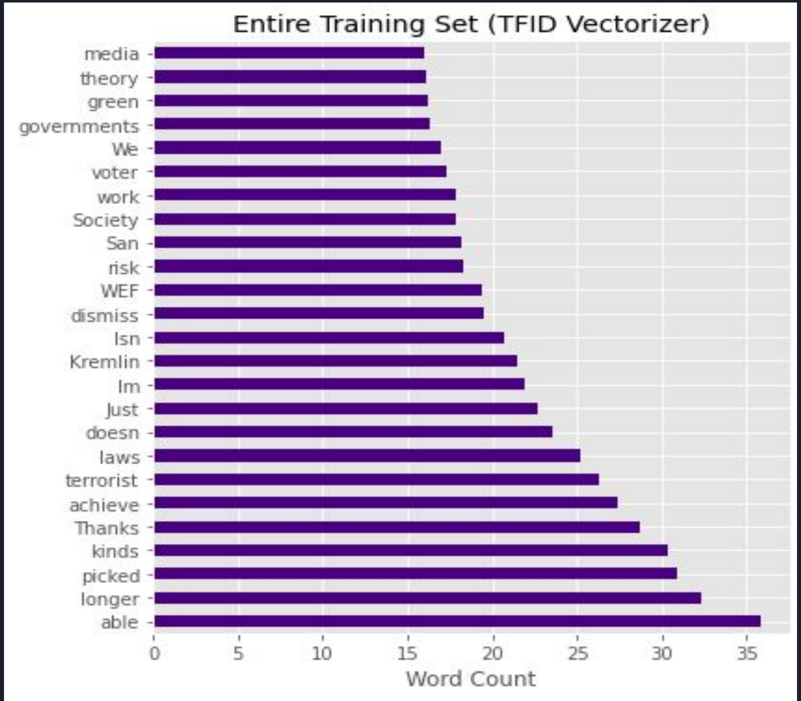
Best Log Ref(count vectorizer)

train 0.9858585858585859

test 0.8729838709677419

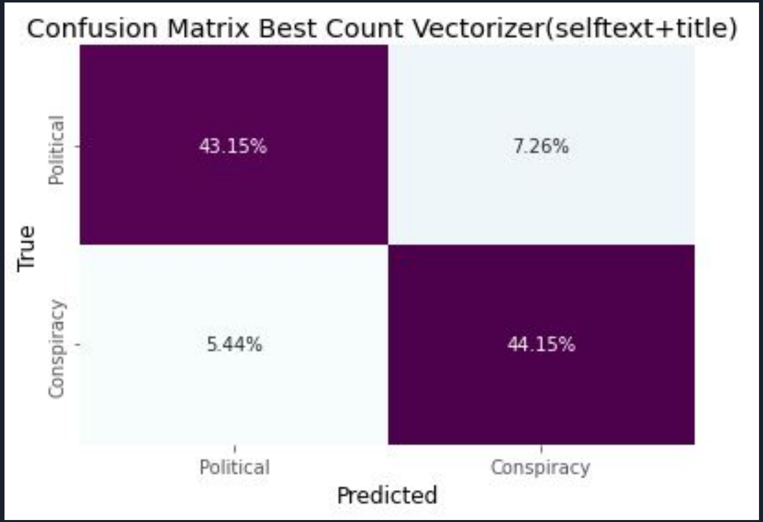


TFID Vectorizer - 0.947(train), 0.88 (test)

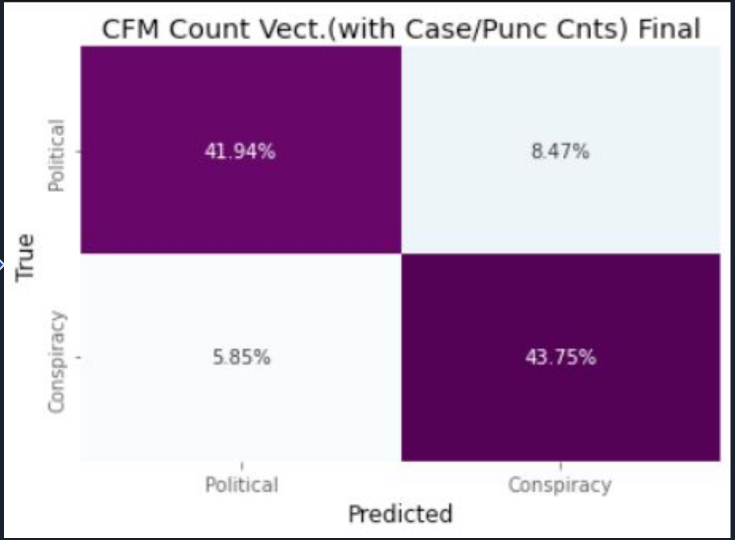


Misclassifications (using best CV) & Word Cnt + Punc counts included

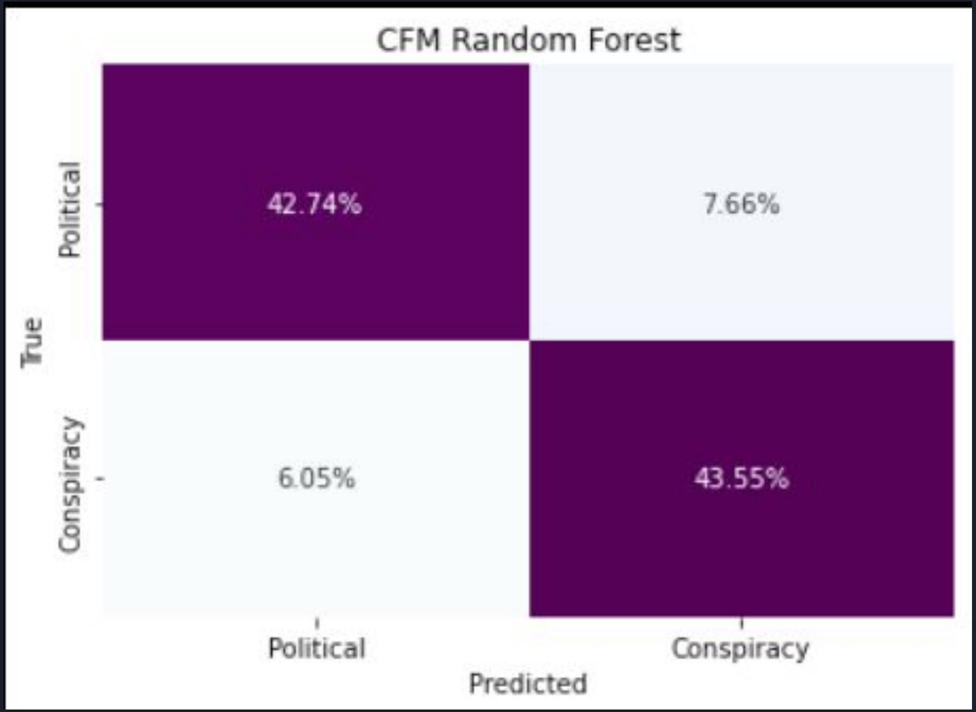
Countvectorizer (before)



Countvectorizer (after)



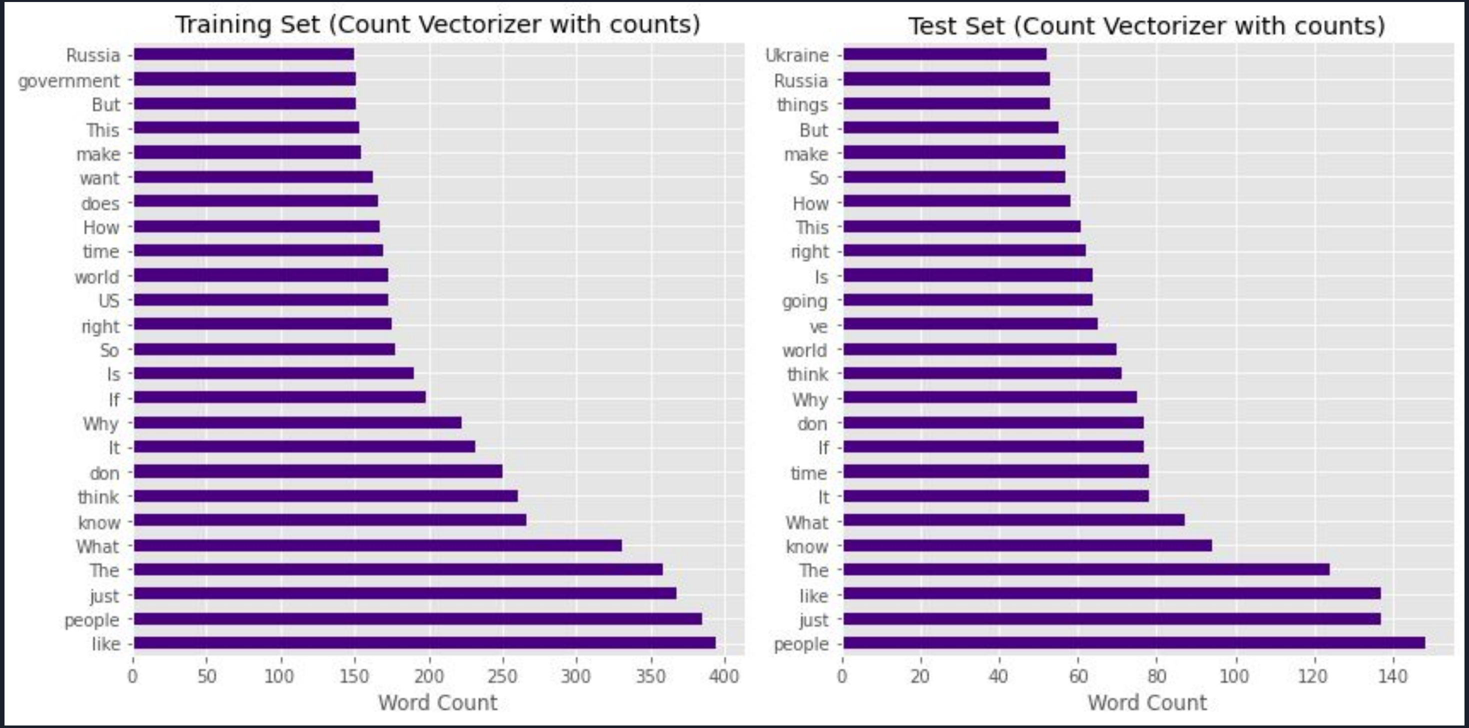
Random Forest



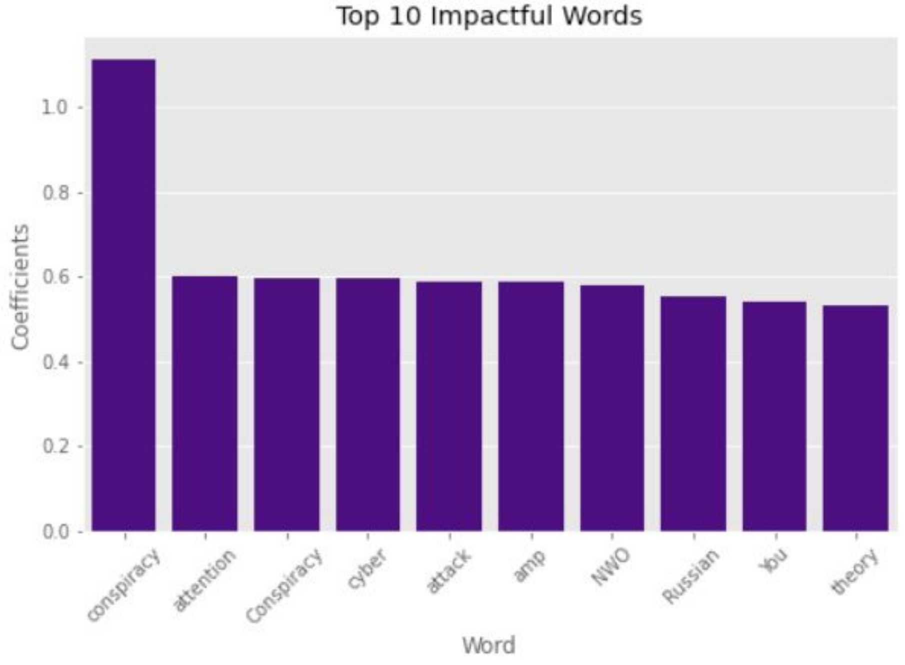
CFM Random Forest

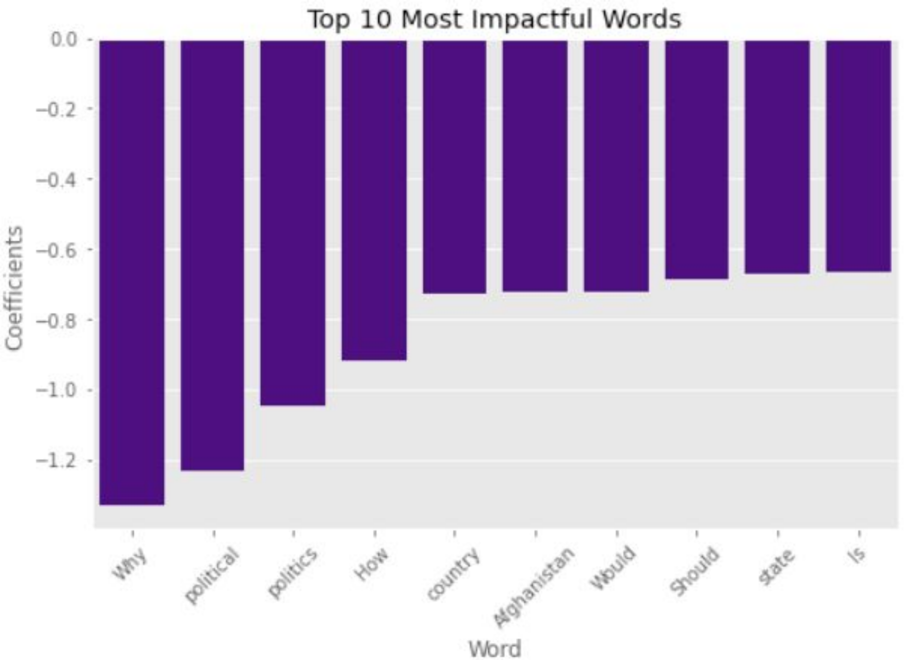
True \ Predicted	Political	Conspiracy
Political	42.74%	7.66%
Conspiracy	6.05%	43.55%

Best Count Vectorization Insights



Best Logistic Regression Insights





Conclusions & Recommendations

- Logistic Regression (without additional features) is best performer
- Coefficients provide insights
- Best text accuracy scores (slightly unbalanced)
- Communications group knows what words to avoid in messaging