Reddit Posts Classification Analysis

Muhamad Tahir

GA Meeting 03/05/2021

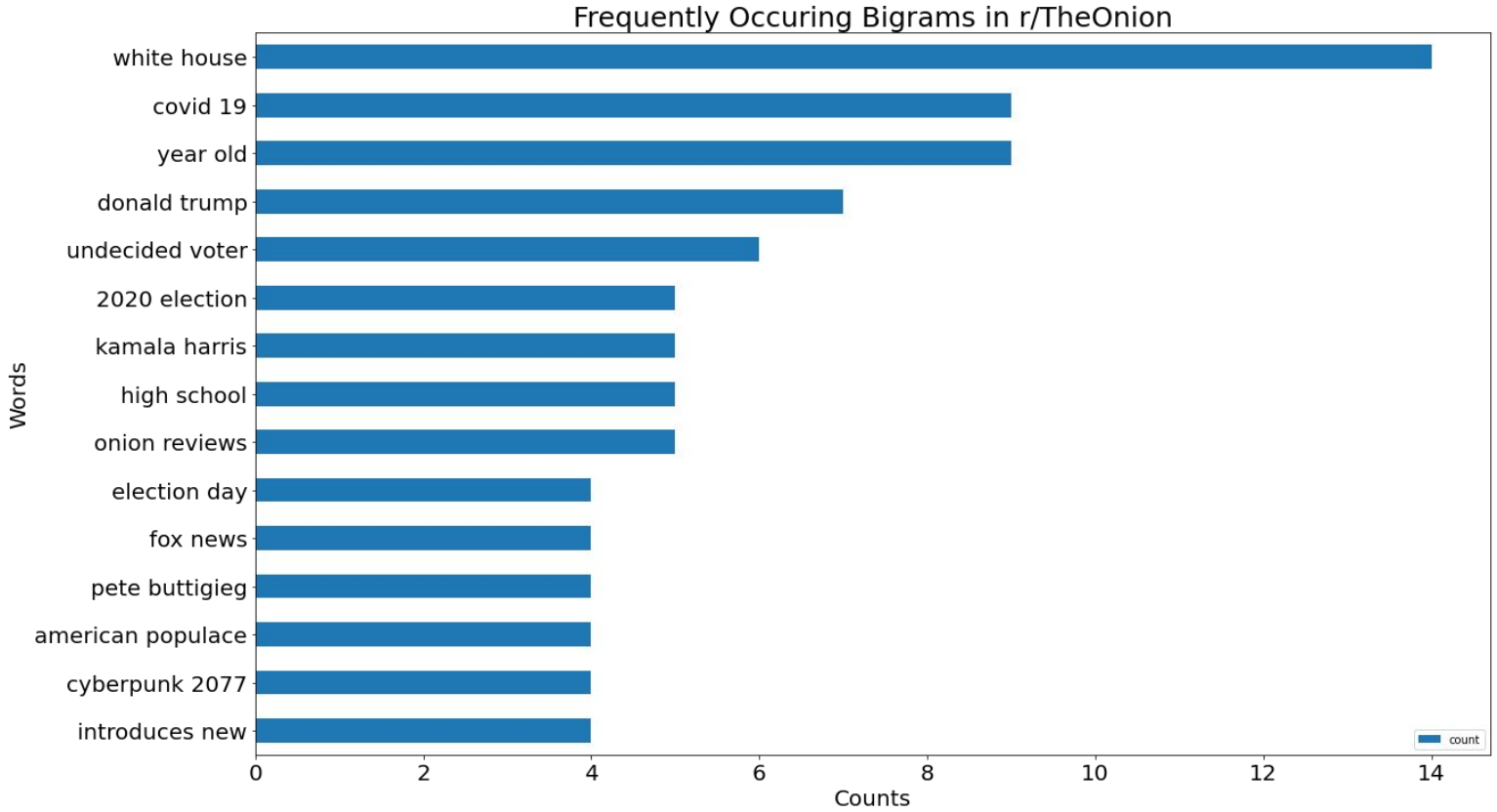
Content:

- Classification Analysis of Reddit Posts
- Cleaning & Exploratory Data Analysis
- Models prediction
- Best Scores Analysis
- Recommendation and Outlook

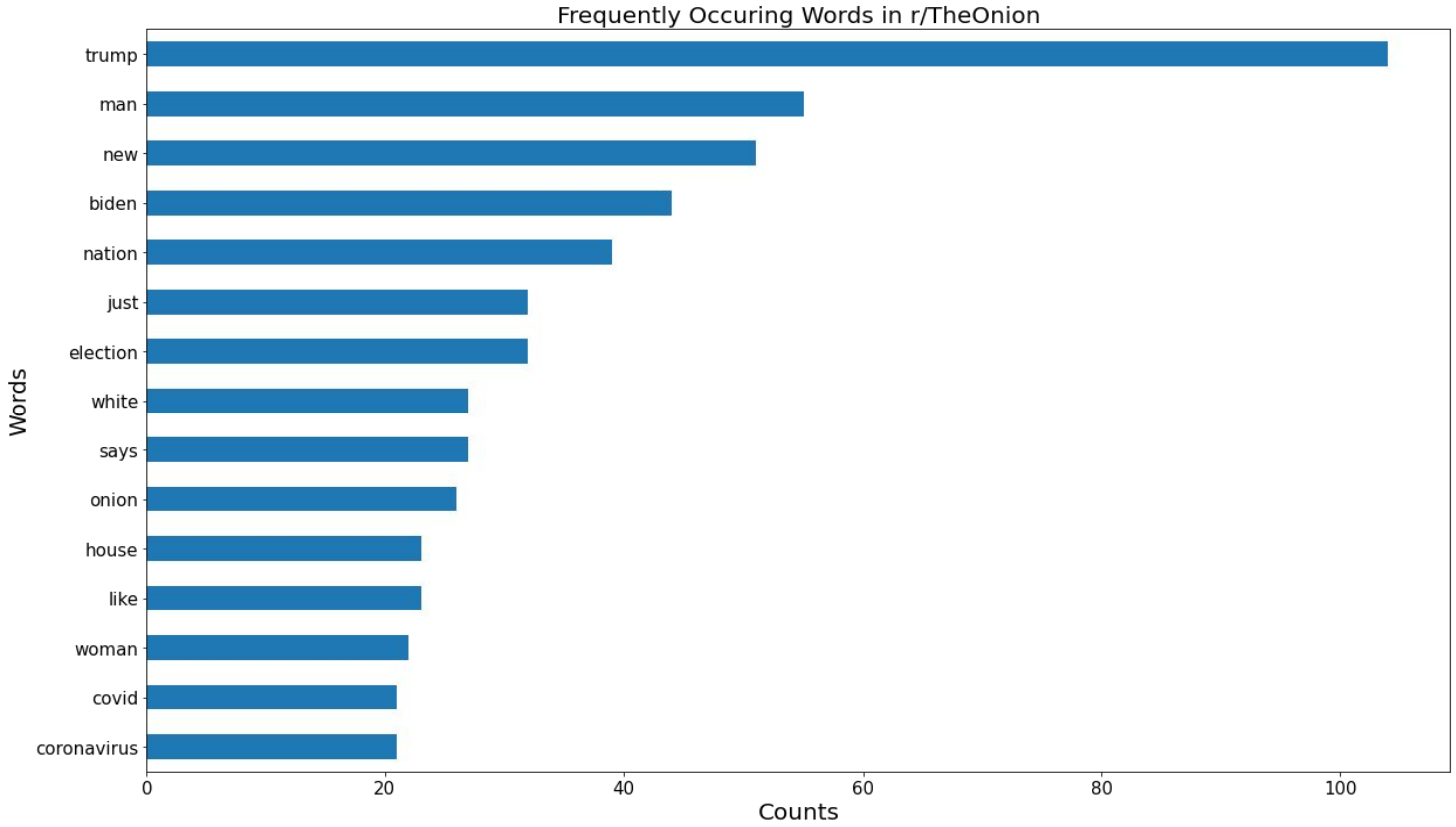
Major Goal:

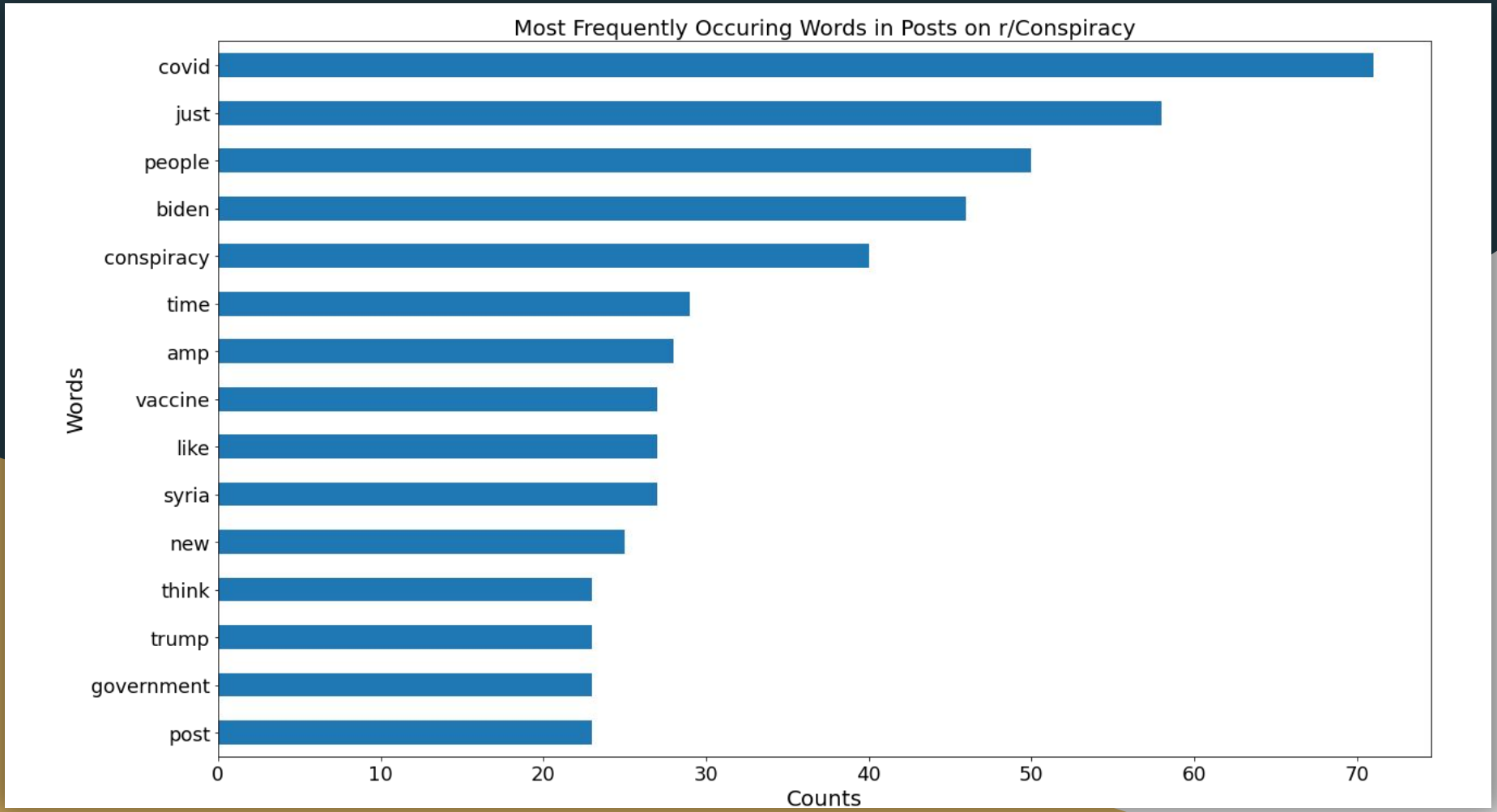
- Explore Two Reddit Posts
- Data cleaned and analyzed
- Models Employed: Logistic Regression, K-Nearest Neighbors, Multinomial Naive Bayes, Random Forest
- Results are Overfitted

Words (stop & Bigram) vs Count

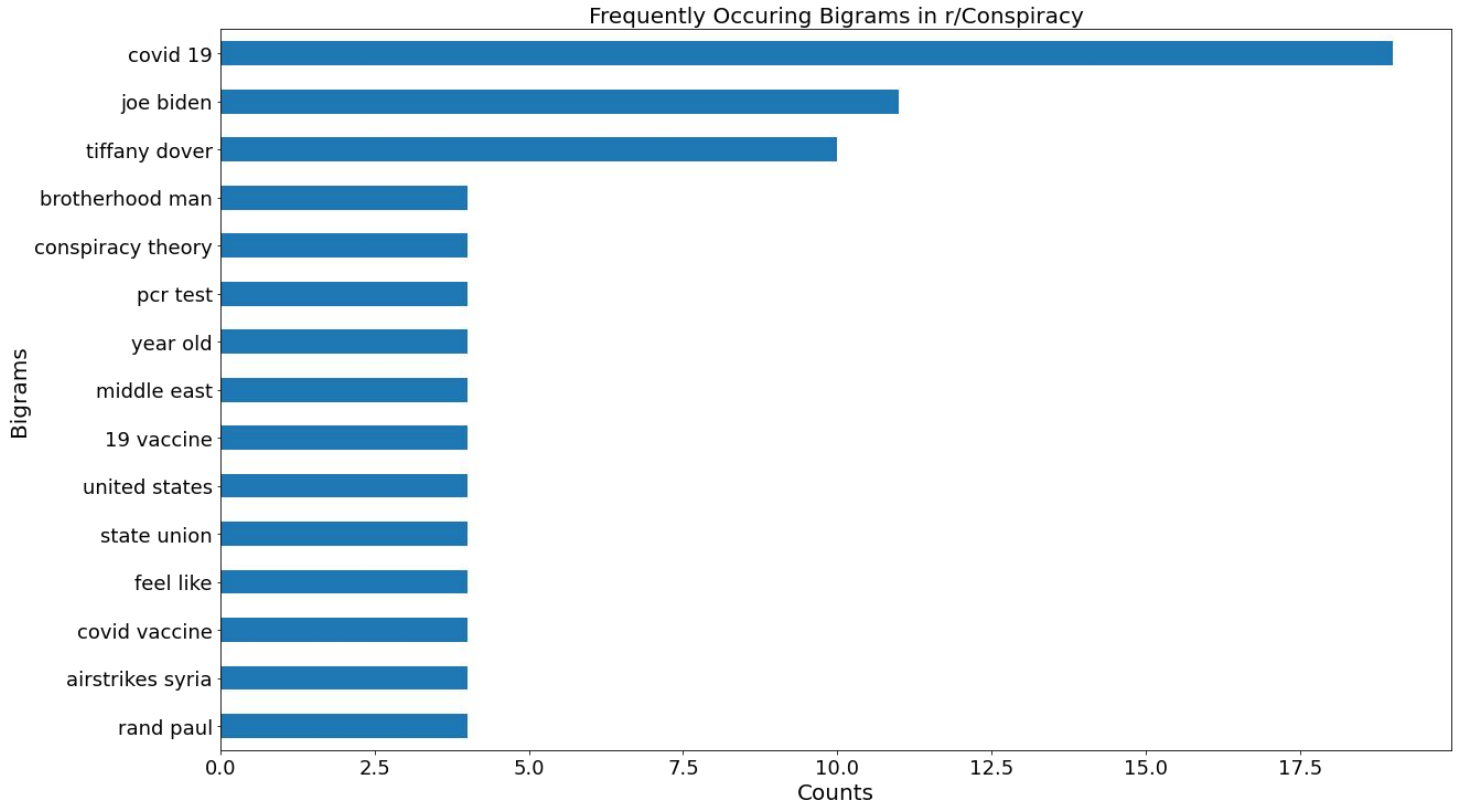


Using Stop Words of “English”




Most Frequently Occuring Words in Posts on r/Conspiracy
covid
just
people
biden
conspiracy
time
amp
vaccine
like
syria
new
think
trump
government
post
Words
0
10
20
30
40
50
60
70
Counts

Stop words & Bigrams



Models and Score

Train and test Scores respectively:

Using CountVectorizer

1- Logistic Regression: 0.94 & 80
2- K-Nearest Neighbors: 0.7 & 0.58
3- Multinomial Naive Bayes: 0.93 & 81
4- Random Forest: 0.99 & 80

Recommendation & Outlook

- Results Overfitted with score: Train 0.93 & Test 0.82
- Difficult to distinguish These Subreddits
- Logistic Regression and Multinomial Naive Bayes: Baseline Scores 0.5
- Boosting or SVM may lead to better result