Identifying Fake News using NLP

Creating a model from r/TheOnion and r/Politics

Outline

Executive Summary

Fake News is a growing concern and there are significant efforts to identify fake articles and curtail them.

Reddit is a free public forum where anyone with an account can submit and link posts.

- r/TheOnion is a group dedicated to TheOnion articles.
- r/Politics is a group dedicated for discussing current world politics.

Approach

1. Webscraping

- Reddit API
- 2 sets of 500 posts

2. Preprocessing

- Removing http links
- Removing blank space
- Lemmatizing and Stemming

3. Modelling

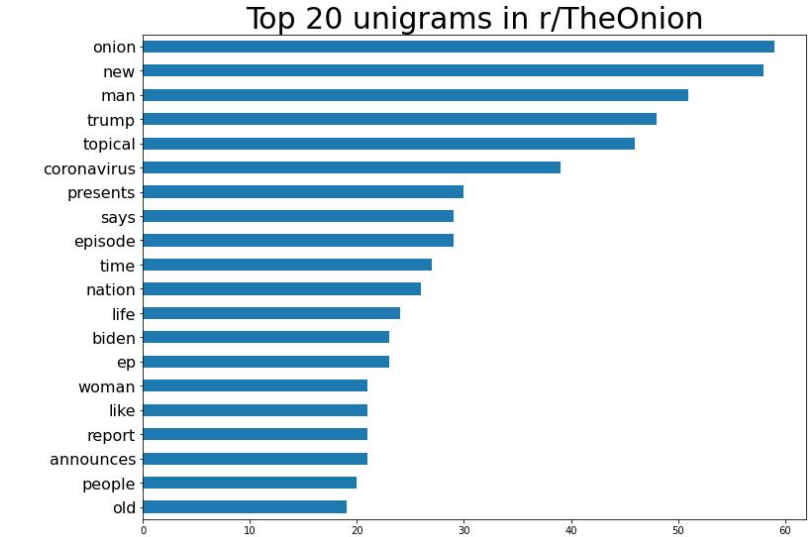
- Count Vectorizer / TFIDF
- Logistic Regression, Multinomial NB, Gaussian NB, Decision Tree, Bagging Classifier, K-Nearest Neighbours, Random Forest, Extra Trees, Support Vector Classifer
- Gridsearch

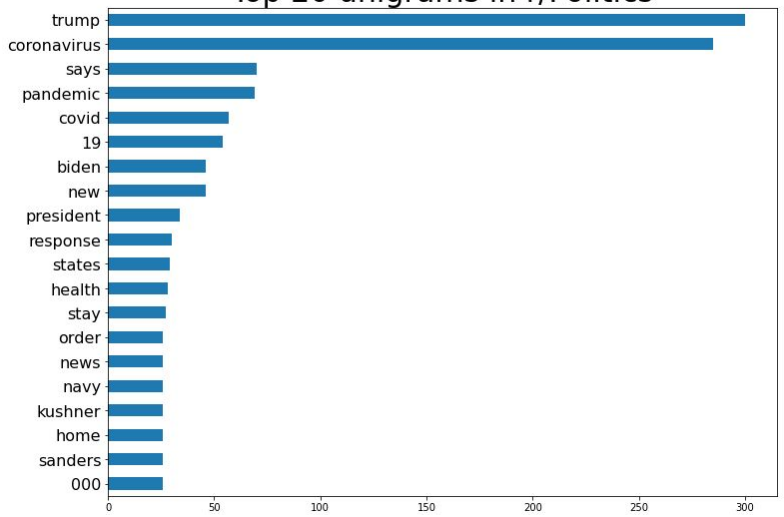
Observations

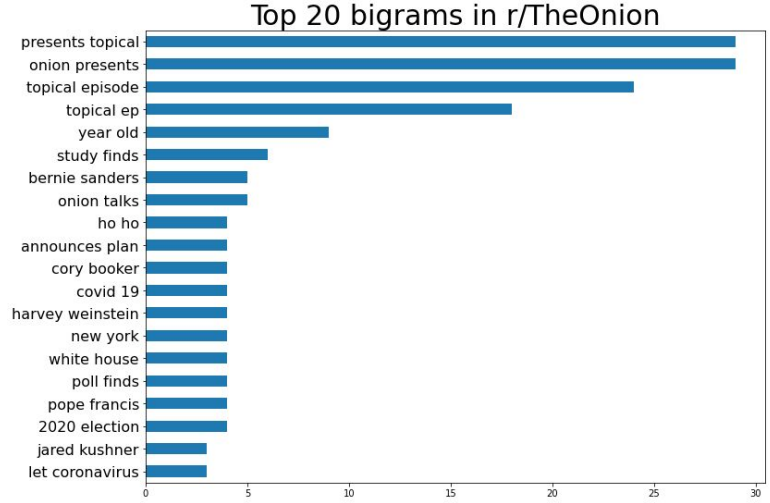
1. **Word Count**
   - Comparing 1&2 n-grams

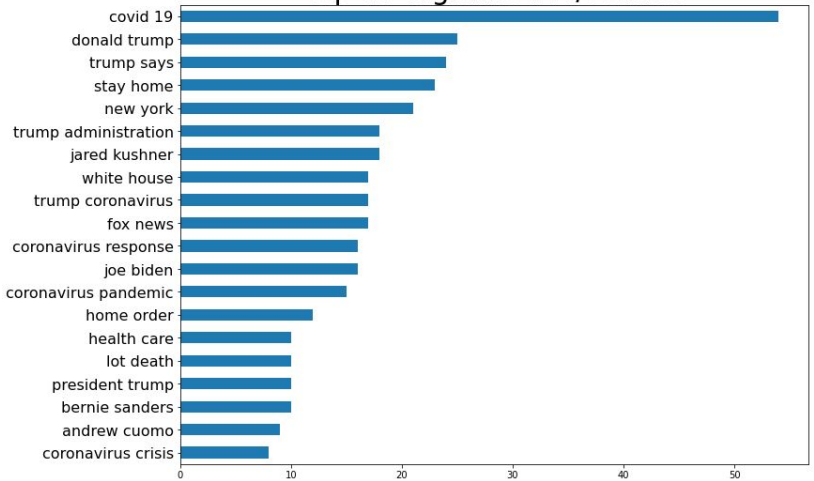
2. **Accuracy**
   - Highest Score
   - Confusion Matrix

3. **ROC/ AUC curve**


Top 20 unigrams in r/TheOnion
onion
new
man
trump
topical
coronavirus
presents
says
episode
time
nation
life
biden
ep
woman
like
report
announces
people
old
0
10
20
30
40
50
60

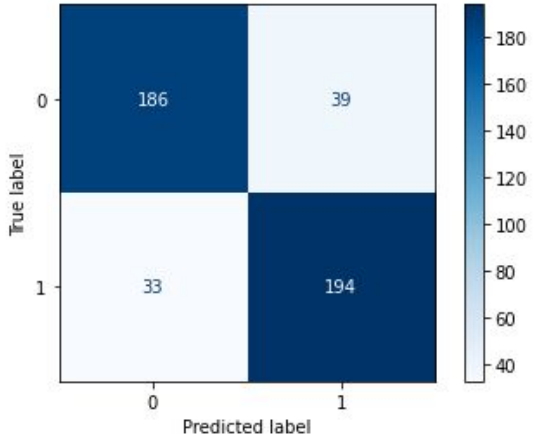

Top 20 unigrams in r/Politics
trump
coronavirus
says
pandemic
covid
19
biden
new
president
response
states
health
stay
order
news
navy
kushner
home
sanders
000
0
50
100
150
200
250
300


Top 20 bigrams in r/TheOnion
presents topical
onion presents
topical episode
topical ep
year old
study finds
bernie sanders
onion talks
ho ho
announces plan
cory booker
covid 19
harvey weinstein
new york
white house
poll finds
pope francis
2020 election
jared kushner
let coronavirus
0
5
10
15
20
25
30

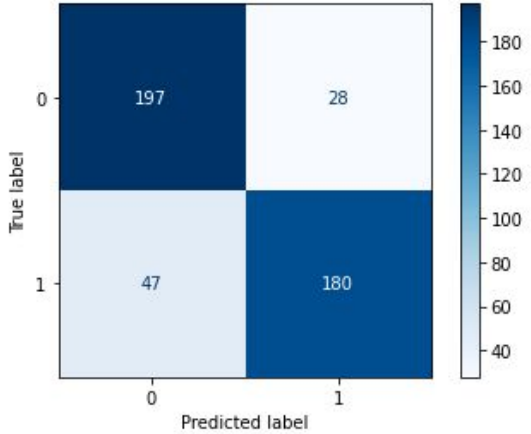

Top 20 bigrams in r/Politics
covid 19
donald trump
trump says
stay home
new york
trump administration
jared kushner
white house
trump coronavirus
fox news
coronavirus response
joe biden
coronavirus pandemic
home order
health care
lot death
president trump
bernie sanders
andrew cuomo
coronavirus crisis
0
10
20
30
40
50

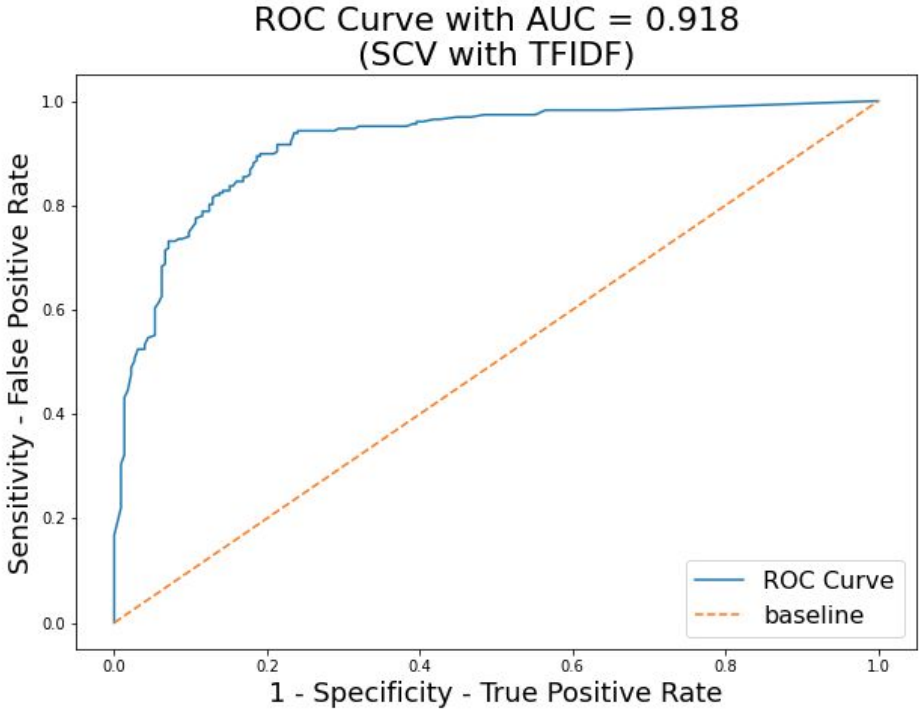
Model	Training Score	Test Score	ROC AUC Score
Logistic Regression	0.9845	0.8407	0.9119
SVC	0.9963	0.8407	0.9178

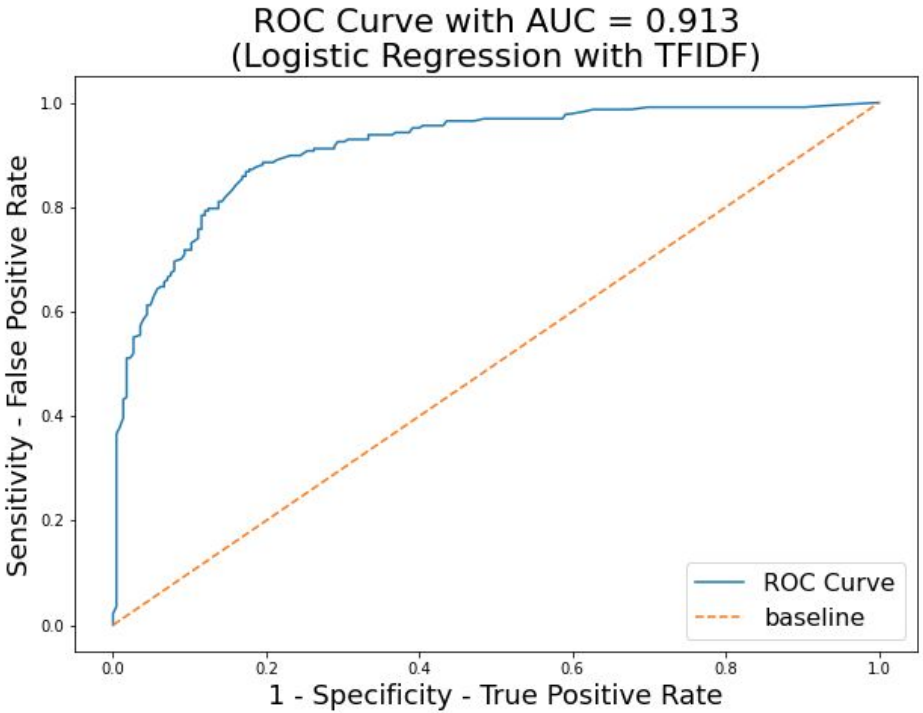
SVC:



Logistic Regression:




ROC Curve with AUC = 0.918
(SCV with TFIDF)
Sensitivity - False Positive Rate
1 - Specificity - True Positive Rate
ROC Curve
baseline
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve with AUC = 0.913
(Logistic Regression with TFIDF)
Sensitivity - False Positive Rate
1 - Specificity - True Positive Rate
ROC Curve
baseline
0.0
0.2
0.4
0.6
0.8
1.0

Conclusions

- 17 models trained
- Highest Accuracy Score: 0.8407 | Highest AUC: 0.9178
- From these metrics, SVC can be considered as the best model to identify Fake News.

Improvements:

- Images, emojis, or any other non alphanumeric character.
- Context analysis
- Feature Engineering
- Stopwords

Questions?