Reddit Natural Language Processing (NLP)

Shengjie Lin

Problem Statement:

- Build a model that classify subreddit post and predict if they are from one of the two subreddit

Subreddits

r/ News

The place for news articles about current events in the United States and the rest of the world.

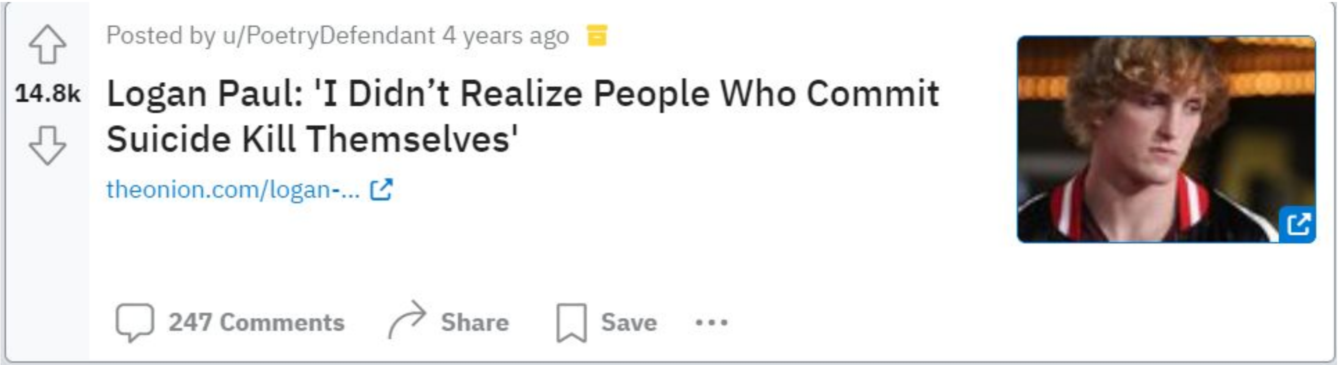
Members: 23.5M

r/ TheOnion

The Onion is an American satirical digital media company and newspaper organization that publishes articles on international, national, and local news.

Members: 162K

Example



Posted by u/PoetryDefendant 4 years ago

14.8k

Logan Paul: 'I Didn’t Realize People Who Commit Suicide Kill Themselves'

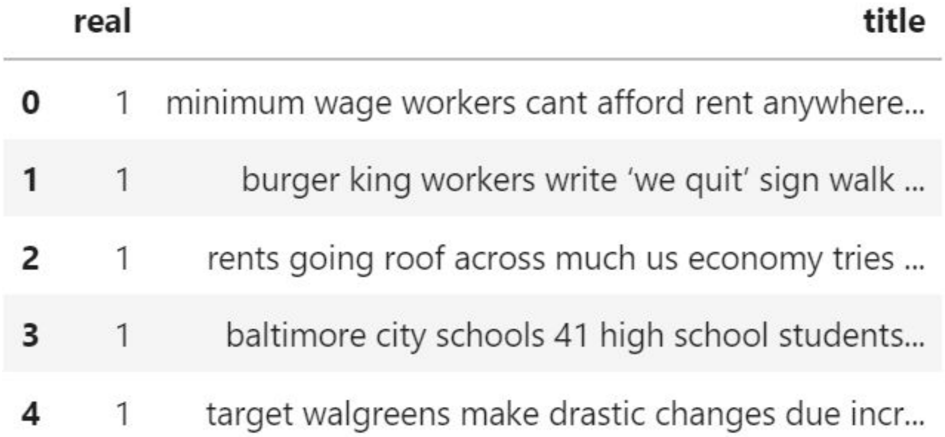
theonion.com/logan-...

Data Collection

- 2500 post from each subreddit
- Time span: 2 years
- Sort: Number of comments

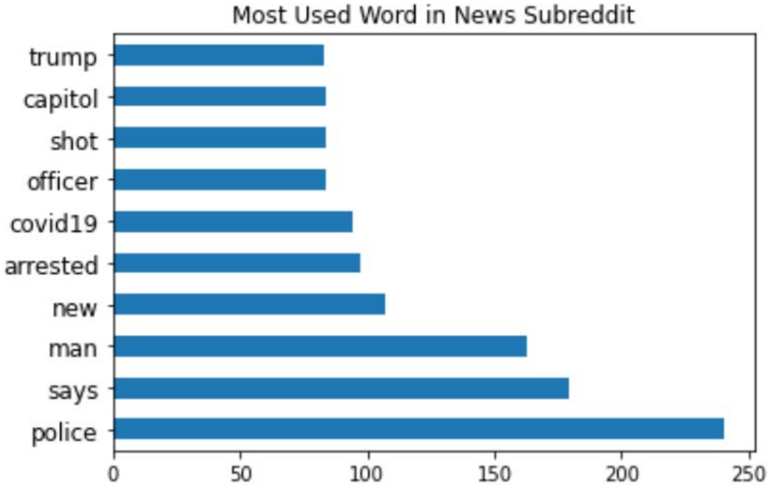
Data Cleaning

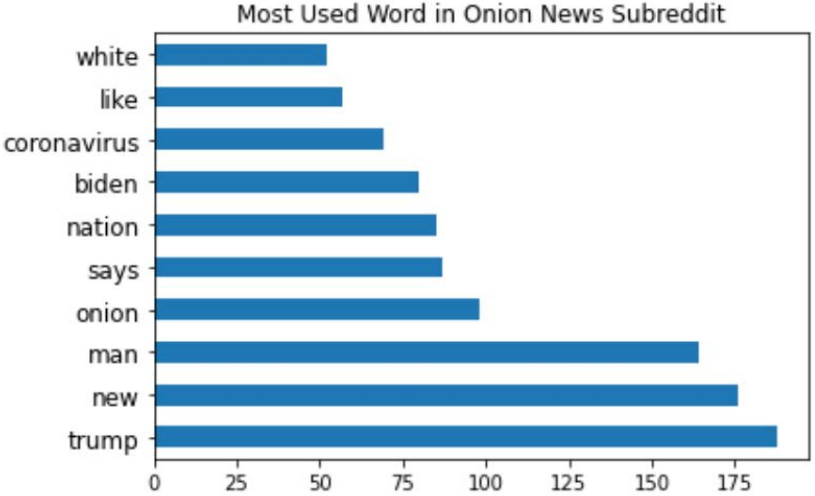
- Remove stop words (nltk.corpus package)
- Change to lowercase
- Remove punctuation
- Remove bad symbol (Ex. /@#$%^&*_~)



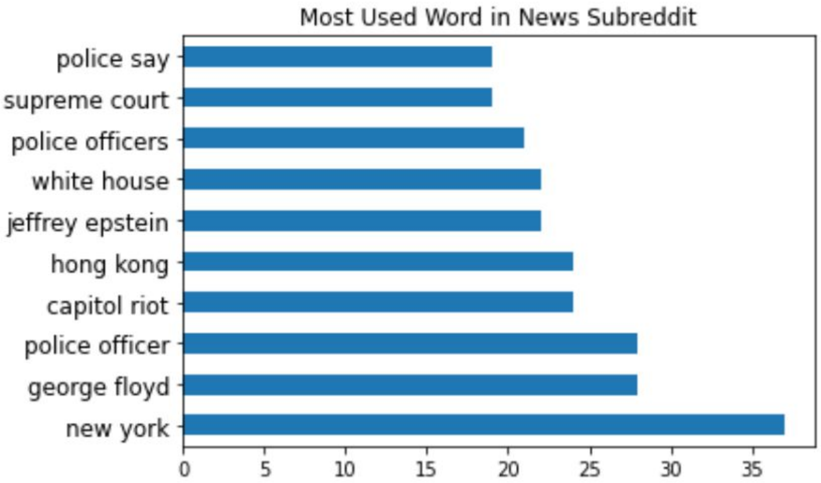
	real	title
0	1	minimum wage workers cant afford rent anywhere...
1	1	burger king workers write ‘we quit’ sign walk ...
2	1	rents going roof across much us economy tries ...
3	1	baltimore city schools 41 high school students...
4	1	target walgreens make drastic changes due incr...

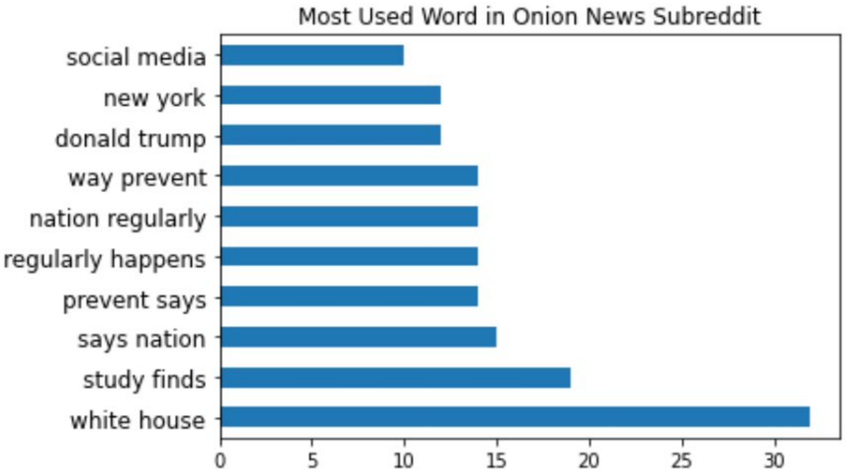
EDA - Top Words





EDA - Top Words 2-gram





Model Used

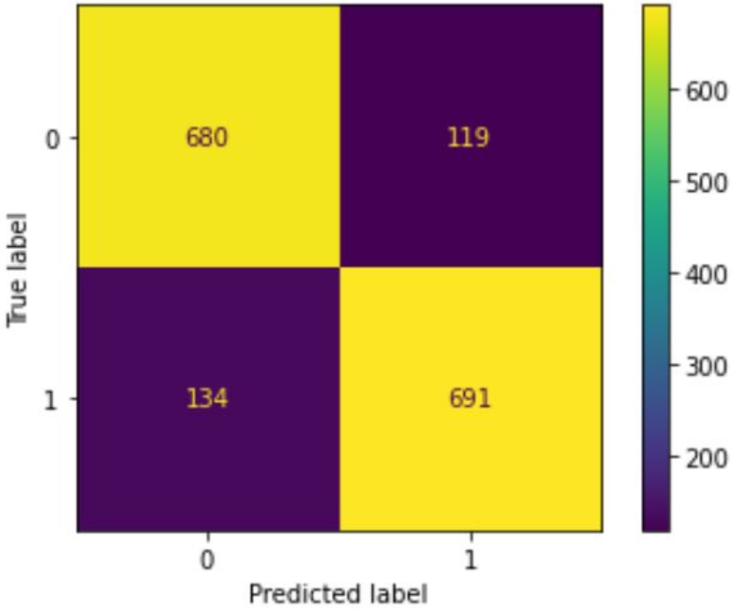
- Bernoulli Naive Bayes
- Random Forest
- AdaBoost

Bernoulli Naive Bayes Model

Train score - 0.845

Test score - 0.844

f1 score - 0.845

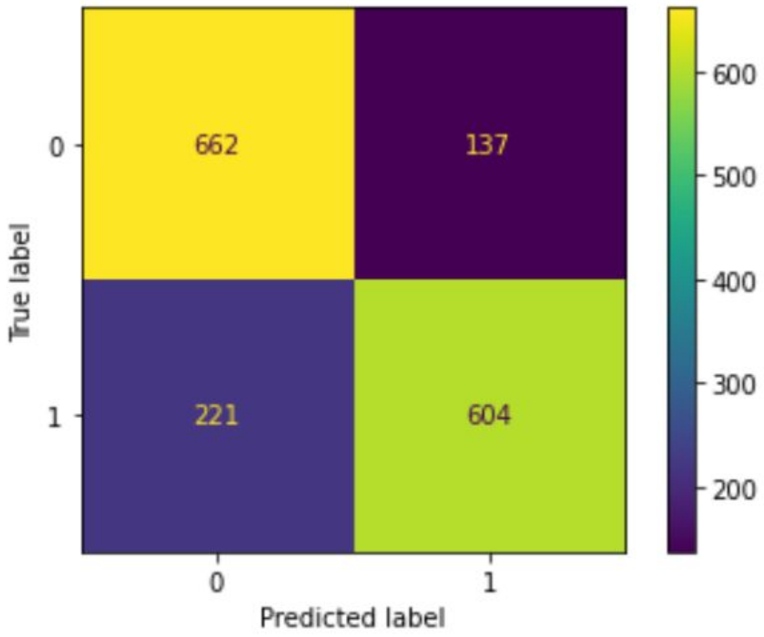


Random Forest Model

Train score - 0.804

Test score - 0.780

f1 score - 0.771

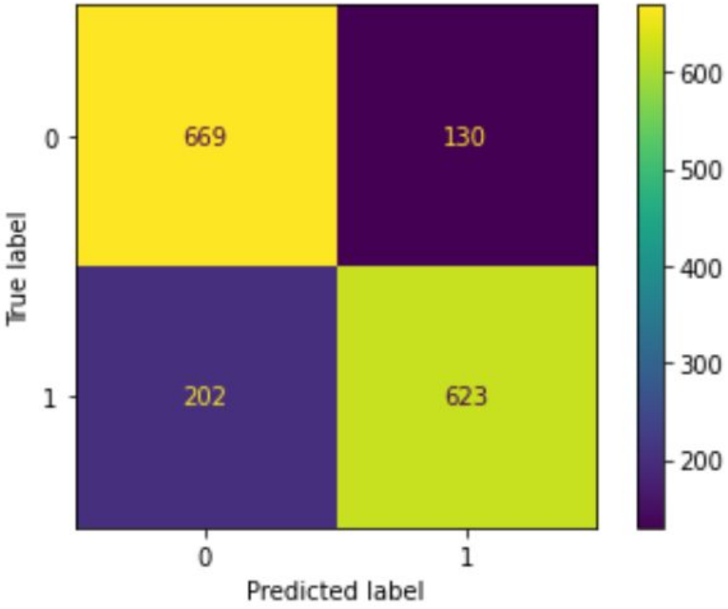


AdaBoost Model

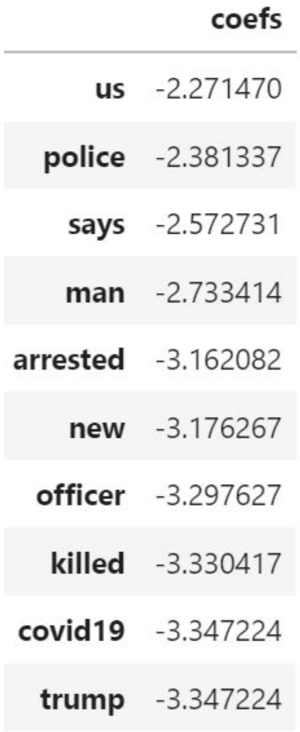
Train score - 0.716

Test score - 0.796

f1 score - 0.790

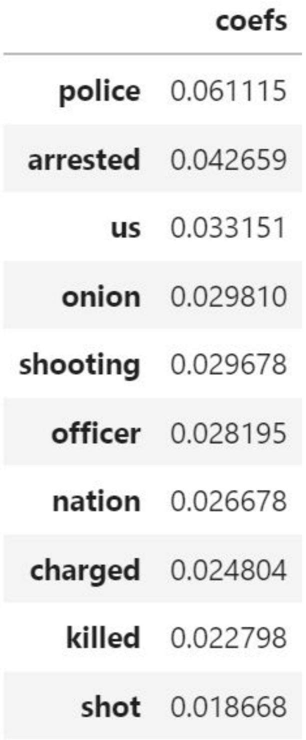


Important Features



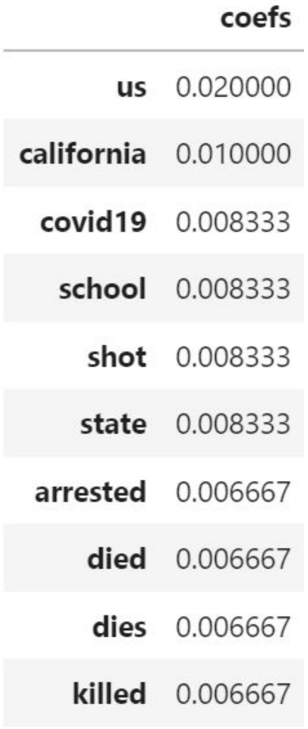
	coefs
us	-2.271470
police	-2.381337
says	-2.572731
man	-2.733414
arrested	-3.162082
new	-3.176267
officer	-3.297627
killed	-3.330417
covid19	-3.347224
trump	-3.347224

Naive Bayes



	coefs
police	0.061115
arrested	0.042659
us	0.033151
onion	0.029810
shooting	0.029678
officer	0.028195
nation	0.026678
charged	0.024804
killed	0.022798
shot	0.018668

Random Forest



	coefs
us	0.020000
california	0.010000
covid19	0.008333
school	0.008333
shot	0.008333
state	0.008333
arrested	0.006667
died	0.006667
dies	0.006667
killed	0.006667

AdaBoost

Conclusion

- Naive Bayes model provide best accuracy score among other, it also has the shortest compute time
- More emphasis on police activity in r/news