I Can’t Believe It’s Not Fake News : Training a Classifier to Distinguish Between Posts in r/news and r/nottheonion

Selecting my subreddits

r/news

18 million subscribers

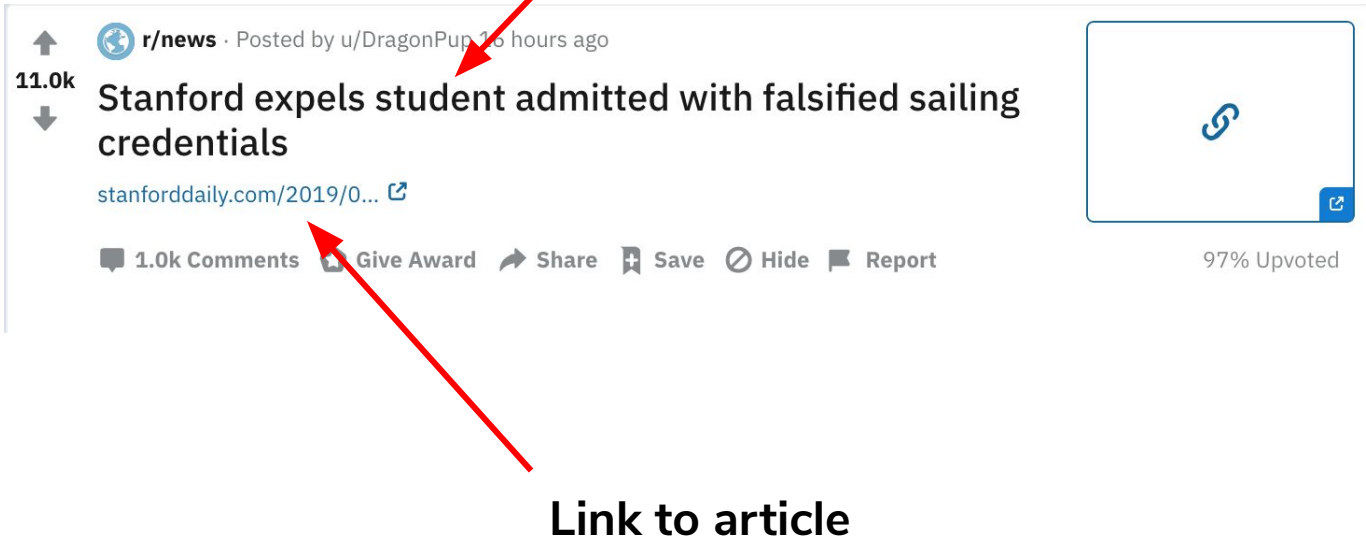
“/r/news is: real news articles, primarily but not exclusively, news relating to the United States and the rest of the World. /r/news isn't: editorials, commercials, political minutiae, shouting, justin bieber updates, kitty pictures. ”

r/nottheonion

15 million readers

“For true stories that are so mind-blowingly ridiculous that you could have sworn they were from The Onion.”

Anatomy of r/news posts



Anatomy of r/nottheonion posts

markste4321 891 points · 3 hours ago

Guess those yoga classes were a waste of time then!

Reply Give Award Share Report Save

Data Collection and Cleaning

Data were imported via Pushshift.io

1000 posts per subreddit

Used post title as primary data

Dropped “removed” posts

Replaced dropped rows with new headlines

... but still not enough!

Modeling

CountVectorizor

TF-IDF

x

Multinomial NB

Logistic Regression

Modeling

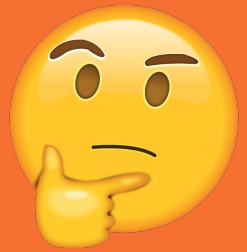
```
from sklearn.model_selection import GridSearchCV

pipe_params = {
    'cvec__stop_words': [None, 'english'],
    'cvec__max_features': [2750, 2800, 3000],
    'cvec__ngram_range': [(1,1), (1,2), (1,3)]
}
gs = GridSearchCV(pipe, param_grid=pipe_params, cv=5)
gs.fit(X_train, y_train);
print(gs.best_score_)
gs.best_params_
```

Evaluation

Overfit to Training Data Across All Models

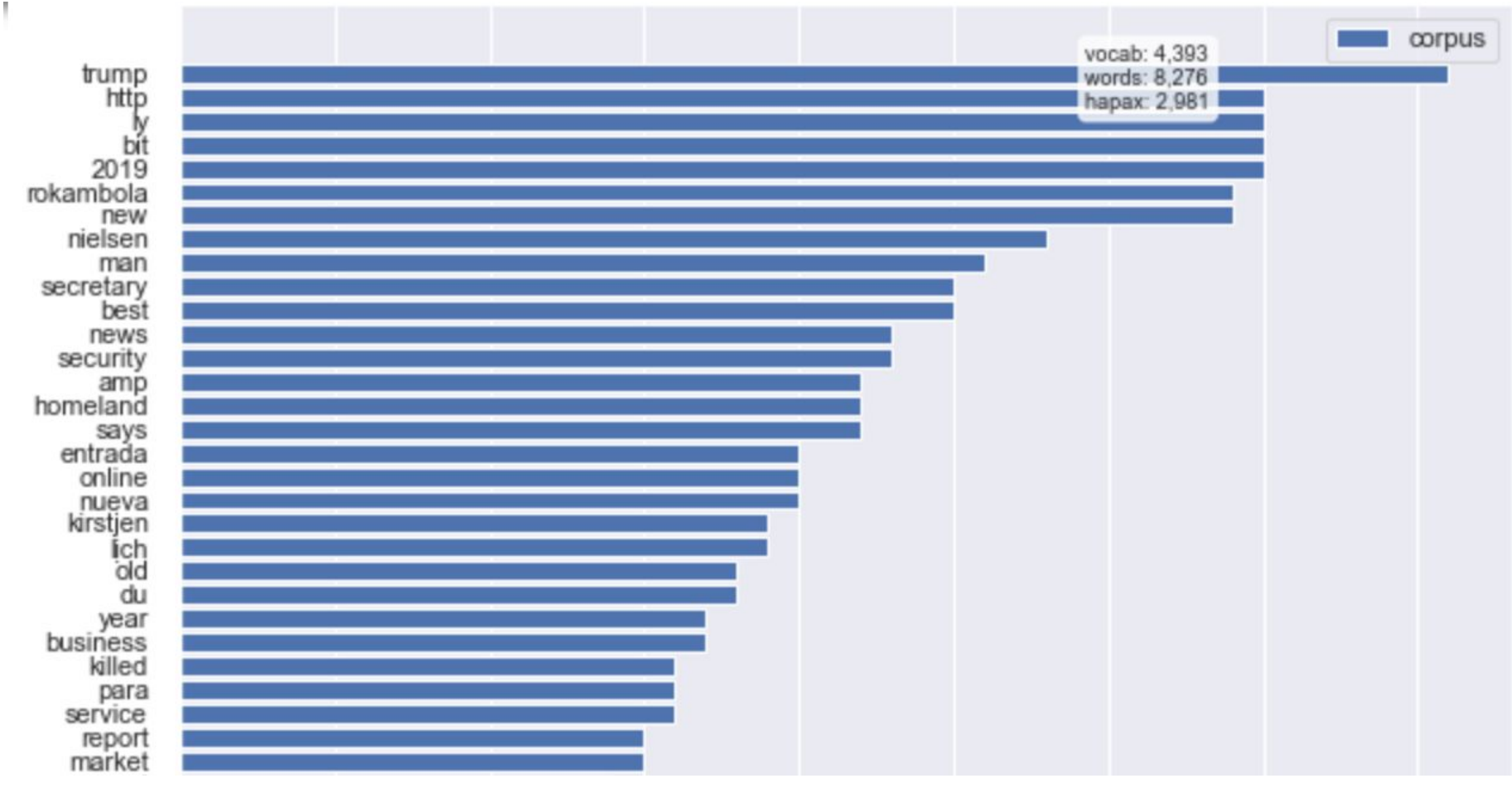
Accuracy score range: 0.76 - 0.79

Highest Scorer was Multinomial NB with the CountVectorizer: 0.79

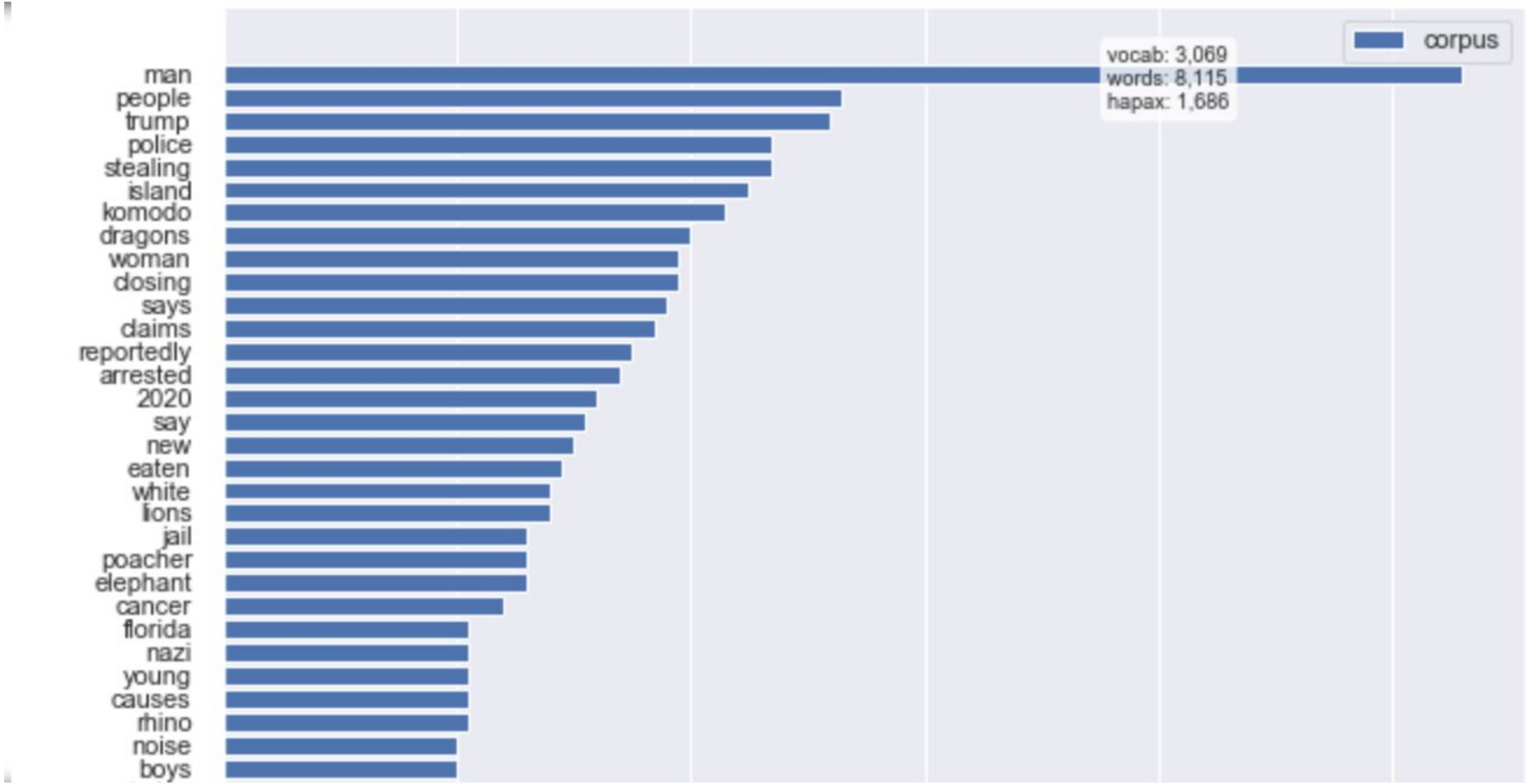


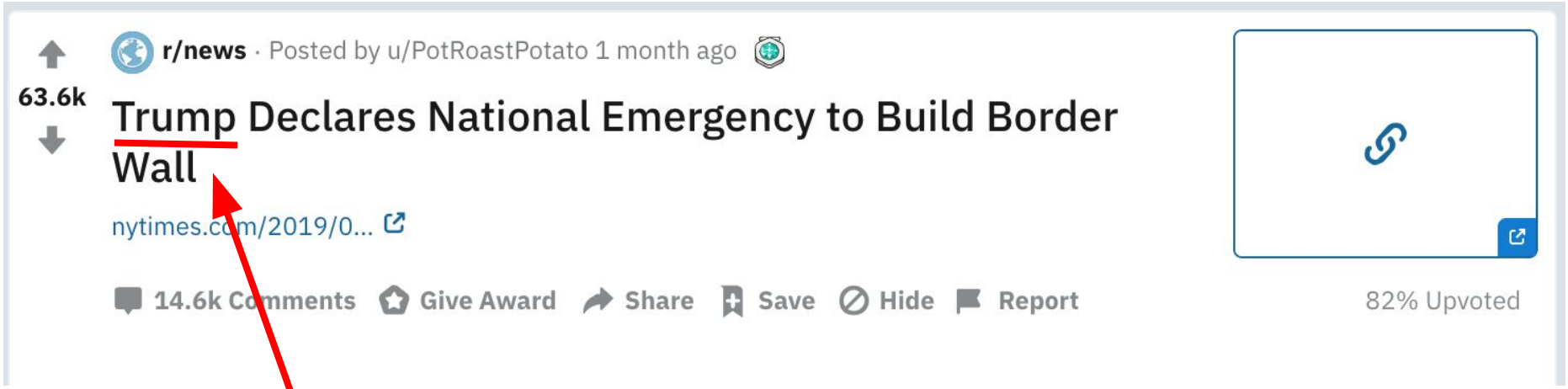
To what can we attribute any distinction between classes?

Word Frequencies for r/news



Word Frequencies for r/NotTheOnion




63.6k
r/news · Posted by u/PotRoastPotato 1 month ago
Trump Declares National Emergency to Build Border Wall
nytimes.com/2019/0...
14.6k Comments
Give Award
Share
Save
Hide
Report
82% Upvoted

Proper Noun

4.4k

r/nottheonion · Posted by u/scgustin 3 years ago

Florida man falls asleep while robbing home

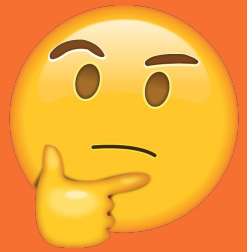
fox43.com/2015/...

/r/all

405 Comments Give Award Share Save Hide Report 93% Upvoted

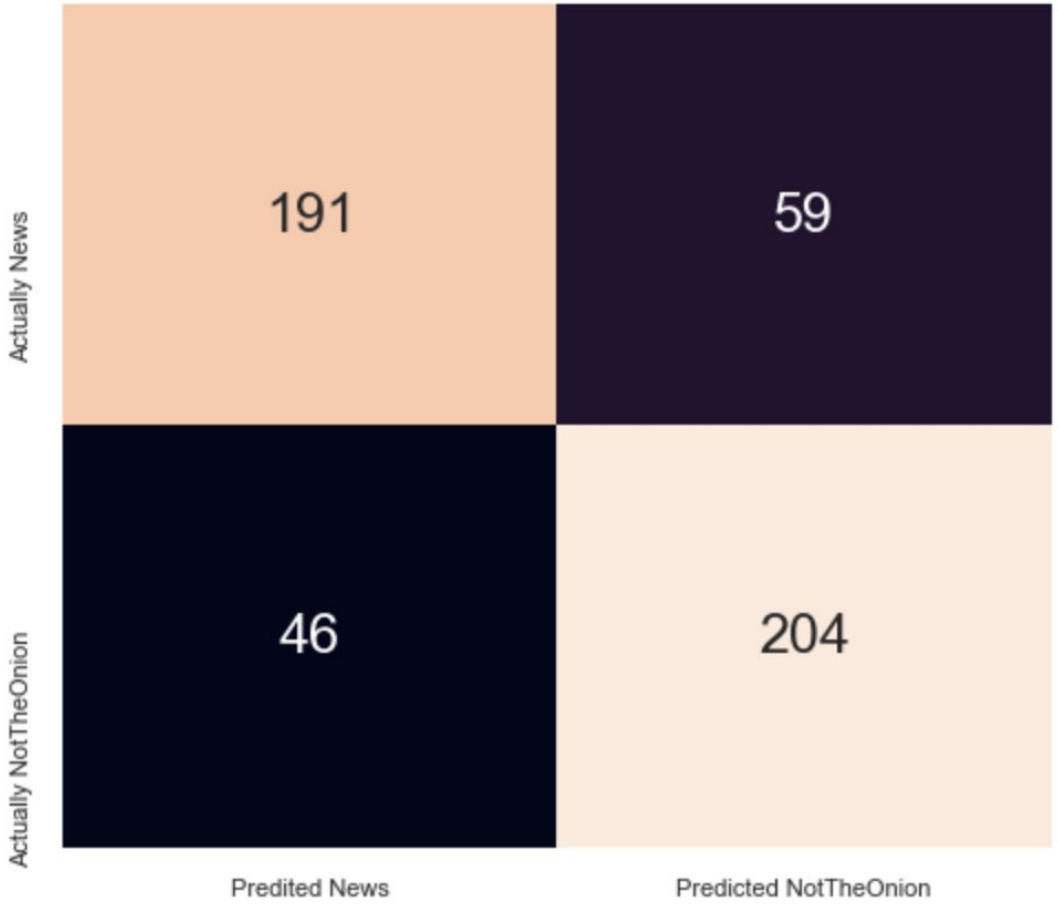
General Noun

Actions



Where was the model struggling?

Confusion Matrix for Multinomial NB



Accuracy
= 79%

669 전세계가 인정하는 '방탄소년단 정국' 몸속 깊이 배인 매너 무한 칭찬
714 오늘(8일) 드디어 공개된 방탄소년단 신곡 '작은 것들을 위한 시' 티저
572 어제자 방콕 콘서트에서 '민트머리+꽃무늬 셔츠'로 역대급 미모 경신한 방탄 뷔
742 방탄소년단x할시, 타이틀곡 '작은 것들을 위한 시' 피처링…작지 않은 엄청난 역대급...
232 방탄소년단 지민, 인성도 으뜸!...함께한 댄서들 일일히 포옹!
330 방탄소년단 지민, 월드투어 마지막 방콕 콘서트에서 열정적인 무대에 관객들 매료~
107 방탄소년단 지민, 갈수록 치솟는 인기 언제까지?
702 방탄소년단 지민, '한국가수 최초' 솔로곡 6000만 더블 돌파 신기록 달성
627 방탄소년단 뷔, 태국공연 피날레 "환상적인 퍼포먼스+무대 향한 열정" 화제
292 “Baba Go Slow Needs To Hurry Up” – American's ...
691 ‘Tháp cơm chiên’ độc đáo chỉ có ở Nhật Bản
645 ‘Morally bankrupt pathological liars’ at Faceb...
159 ‘Eviction’ notices placed on doors of Jewish s...
982 ក្តៅៗ ហុងស៊ុយឆ្នាំកុរ ២០១៩ មិត្តដែលជួយជ...
393 संकल्प पत्र जारी कर बोले मोदी - ये है उनका लक्ष्य
389 राजनाथ सिंह ने बताई 'संकल्प पत्र' की खास बातें
293 मिलिए पहले चरण के POLITICAL धन्नासेठों से
652 मिलिए पहले चरण के POLITICAL धन्नासेठों से
619 भाजपा के 'शत्रु' बने कांग्रेस के मित्र, कहा- स...
631 पूर्व केंद्रीय मंत्री ने #कैप्टन_अमरिंदर पर खड...
404 देश का गौरव आसमान छू रहा है – अमित शाह
364 कौन बनेगा #MP? क्या कहती है होशियारपुर की जनता...
614 कौन बनेगा #MP ? क्या कहती है JALANDHAR की जनता...
376 इन 20 राज्यों के 91 सीटों पर होगा पहले चरण में...

TO DO:

- Remove non-ASCII headlines
- Normalize the corpus (via lemmatization)
- Identify useful stopwords

(Im/Ap)plications

Training a classifier to identify patterns in news headlines can be the first step in identifying “fake news” or building a “clickbait” filter, among other things.

Questions?

You can find me here:

- Twitter : @jonruizruiz
- github.com/jon-ruiz