

# /r/TheOnion VS /r/nottheonion


Distinguishing absurdity from...absurdity?  
Nicholas Jin





# Satire/Sarcasm detection is hard

Therefore we should expect this problem to be  
hard, right?



# Quiz time!

Women Wanted For Stealing McDonald's Toilet Paper

Geno Smith Reportedly Fires Agent After Fall To N.Y. Jets

'Mine Is Bigger': Trump Dares Kim Jong-Un To Compare Nuclear Buttons

PETA Sues Photographer On Monkey's Behalf To Give Monkey Copyright

'No Pants Subway Ride' Kicks Off With Train Commuters Around The World Stripping Off

Magazine Behind UVA Rape Hoax Begs Obama To Do Something About Fake News

Melania Trump To Be Protected By Special All-Female Japanese Police Force

Tennessee Trying To Re-Ban Gay Marriage

Pa.'s Worst Toll Evader Named Stiff To Pay Up: \$128K In All

Man Turns Dead Cat Into Helicopter

## Part 2

Day Without A Woman Sees Thousands Leave Work

Mute, Terrified Rubio Awakes To Find Self Unable To Vocalize Any Unscripted Sentiment

Entirety Of Hollywood Film Industry Replaced With 40,000 Christopher Plummerts

Cackling Mitch McConnell Reveals To Stunned Democrats Hes Been Working Undercover For Republican Party This Whole Time

Sometimes It Feels Like Im In Prison Too, But Then I Go Home

United 93 Director Announces Remastered Edition Digitally Removing WTC From Film

Washington Post Offers Non-Subscribers 10 Free Articles To Fact Check Per Month

Q Forced To Resign From Department Of Agriculture For Improper Filing Of Expense Reports

Nations Gay Straw Men March On Washington For Right To Marry Animals

Old Lady At Parade Flapping Little American Flag Like A Motherfucker

### CountVectorizer models

Model	Training time (s)	Train Accuracy	Test Accuracy	/r/theonion corpus accuracy	/r/nottheonion corpus accuracy
LogReg	43	.984	.800	.791	.805
...with h-opt	15	.969	.801	.793	.805
MultinomNB	.1	.954	.812	.811	.806
...with h-opt	10	.944	.812	.805	.803
ComplementNB	.1	.954	.812	.810	.807
...with h-opt	11	.955	.812	.800	.799
BernoulliNB	.1	.955	.813	.801	.816
...with h-opt	11	.962	.810	.795	.811

### Tf-idf models

Model	Training time (s)	Train Accuracy	Test Accuracy	/r/theonion corpus accuracy	/r/nottheonion corpus accuracy
LogReg	26	.976	.799	.810	.795
...with h-opt	13	.913	.783	.797	.783
MultinomNB	.1	.955	.805	.826	.792
...with h-opt	11	.999	.819	.824	.789
ComplementNB	.1	.955	.807	.823	.797
...with h-opt	11	.940	.815	.812	.799
BernoulliNB	.1	.956	.807	.813	.812
...with h-opt	11	.904	.806	.781	.815
SVC	281*	.997	.797	.804	.812
...with h-opt	382*	.997	.796	.804	.812

\*SVCs trained extremely quickly, but evaluating them on the holdouts took so long that I included that time.

# Logistic Regression Coefs

EDA *still* wasn't thorough enough!

Garbage in, garbage out!

odds multiplier		odds multiplier	
around	3.571741	don	0.116656
nation	3.587011	police	0.167605
mueller	3.650384	arrested	0.177086
back	3.688191	accused	0.198700
pros	3.808212	re	0.223864
hes	3.852983	jail	0.237137
cons	4.305074	china	0.237268
how	4.435756	texas	0.255145
introduces	4.456310	stolen	0.264024
just	4.514982	sues	0.267916
cant	4.524034	us	0.281179
doesnt	4.609962	because	0.288155
shit	4.660664	onion	0.290433
wont	4.770151	toilet	0.291731
area	4.833877	bans	0.292858
fucking	5.511220	ends	0.300187
trumps	5.741289	his	0.303455
announces	6.354899	pizza	0.304854
nations	6.365406	storm	0.314677
tips	7.374050	florida	0.318325

## Future work:

Word embeddings as lower-dimensional features.