Reddit:
Real vs. Parody

Andrew Picart
Data Scientist

Objective and Context

Can headlines for news sites and parody news sites be differentiated?

Construct a model that can predict parody sites and hopefully alert users and content creators.

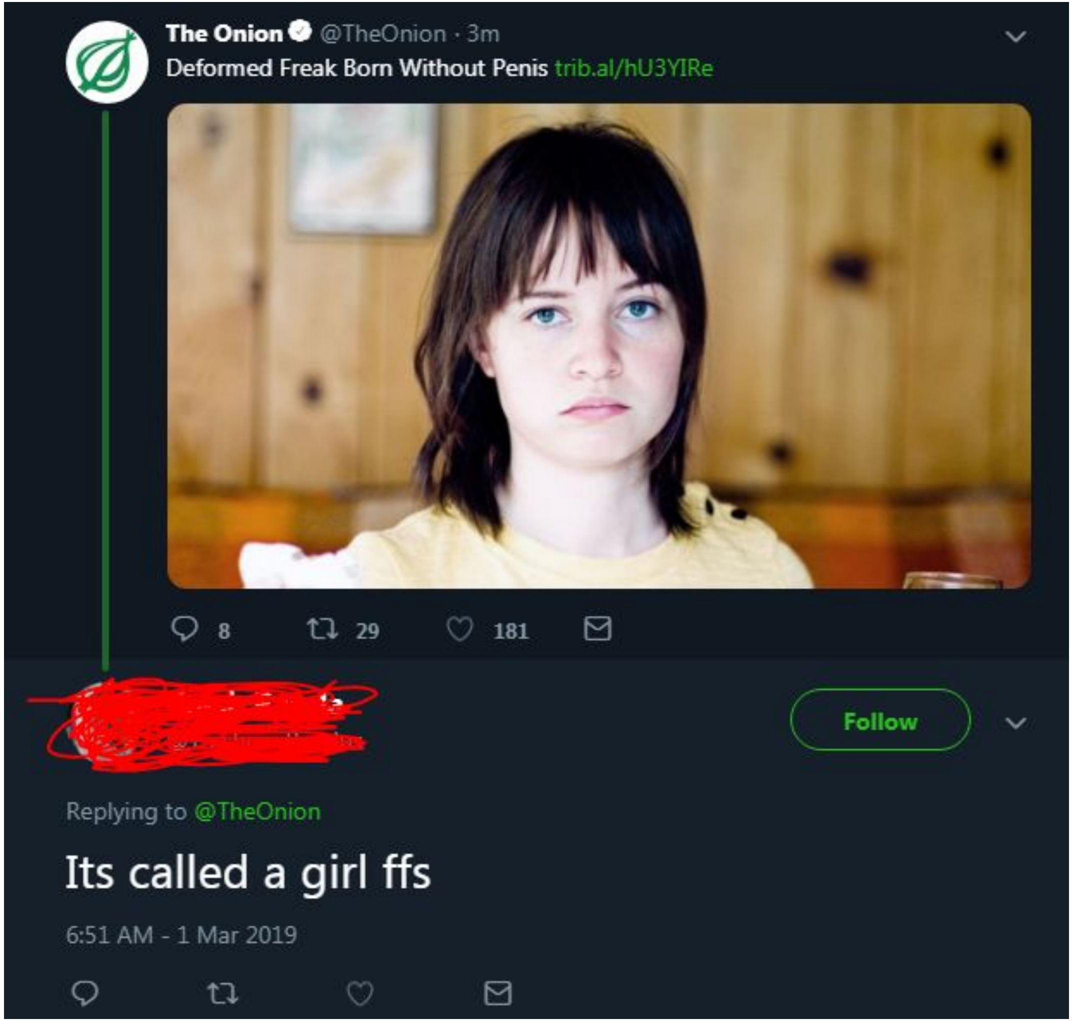
Data is scraped from reddit.com in r/news and r/TheOnion

Top r/theonion posts

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

Heartwarming: When This McDonald’s Employee Had To Walk 20 Miles To Work Because He Couldn’t Afford A Car, The CEO Of McDonald’s Drove Alongside Him To Cheer Him On

‘The Onion’ Proudly Stands With The Media As The Enemy Of The People



The Onion @TheOnion · 3m

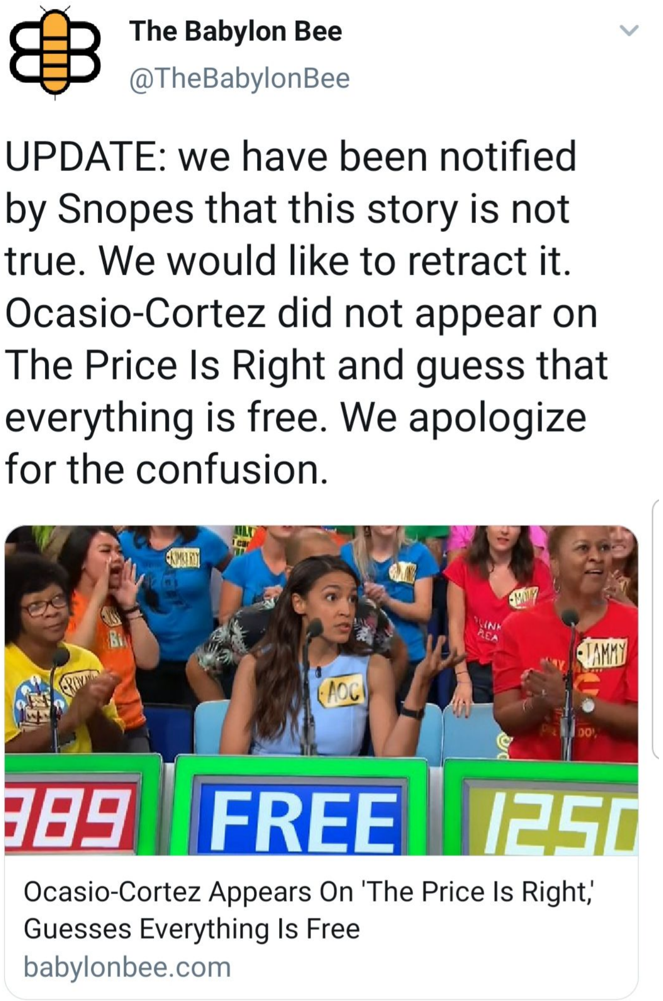
Deformed Freak Born Without Penis trib.al/hU3YIRe

[redacted]

Replying to @TheOnion

Its called a girl ffs

6:51 AM - 1 Mar 2019



The Babylon Bee

@TheBabylonBee

UPDATE: we have been notified by Snopes that this story is not true. We would like to retract it. Ocasio-Cortez did not appear on The Price Is Right and guess that everything is free. We apologize for the confusion.

Ocasio-Cortez Appears On 'The Price Is Right,' Guesses Everything Is Free

babylonbee.com

r/AteTheOnion

Methods

First: Scrape Reddit API

r/news and r/TheOnion

Took posts from subreddit pages: all, new, best(default)

Second: Apply Machine Learning Models

- Multinomial Naive Bayes
- Support Vector Machine
- Random Forest
- k-Nearest Neighbor

Multinomial Naive Bayes

Best Parameters:

```
{'mnb__alpha': 0.1,
 'tfidf__max_df': 0.75,
 'tfidf__ngram_range': (1, 2),
 'tfidf__norm': 'l2',
 'tfidf__stop_words': None}
```

Uses Bayes Theorem

Naively assumes that all features are independent of one another.

Very fast(efficient) and works well in classification problems.

85%

Model accuracy at predicting “r/news” on unseen test data

Confusion Matrix

	Predicted r/TheOnion	Predicted r/news
Actual r/TheOnion	475	59
Actual r/news	78	297

Specificity Rate: .89

TN/(TN + FP)

Words with highest r/news correlation

zumtrel flooby	spectacular	spectacular news	in rpg	in rowboat	speakers white
in road	special meetings	in ringing	in rich	in revealing	in septic
in retrospect	speculate about	speech is	in real	in ring	in search
speech oprah	in same	special flights	in solidarity	in solely	be extra
in solar	in smaller	spazio sells	in sioux	special guest	be fatal

Words with highest r/theonion correlation

his manhattan	trampled to	trampled	in 10	in 2017	In 2018
in 28	in acid	in alabama	in flint	in america	in apparent
in ashley	in austin	in bagel	in barracks	in beating	in blow
trans	in bonuses	trans women	imports	transgender troops	in flordia
in california	trading	in carjacking	trafficking of	trafficking in	in connection

Limitations:

Only dealing with reddit data.

The Onion is very obvious with it’s parody.

Misinformation spreads fast on the internet.