# PREDICTING FAKE NEWS

Dana Hackel

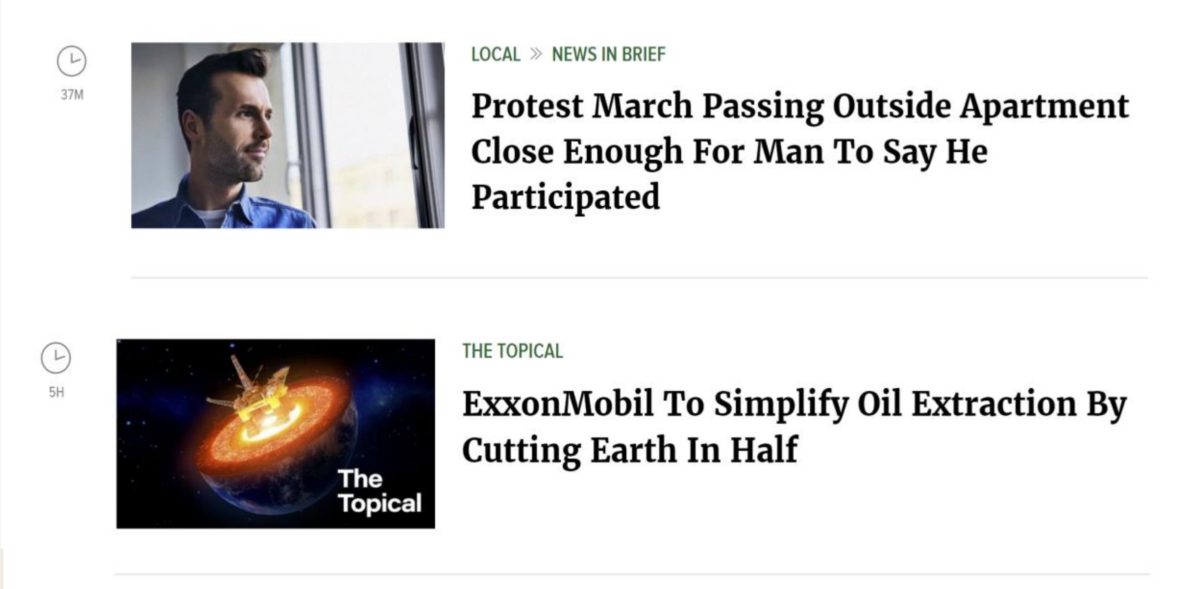
8/28/2020

# Presentation Agenda

- Problem Statement
- Data Collection/ Cleaning
- Data Exploration
- Classification Models
- Future Directions

# Problem Statement

- With the rise of social media and click-bait articles comes the rise of pseudo-news
- I will use subreddit posts from ‘r/theonion’ and ‘r/news’ to create classification models (logistic regression and support vector model) assessed on accuracy. The models will be used to predict if a post is from The Onion or actual News in order to help users differentiate news from satire



37M

LOCAL » NEWS IN BRIEF

**Protest March Passing Outside Apartment Close Enough For Man To Say He Participated**

5H

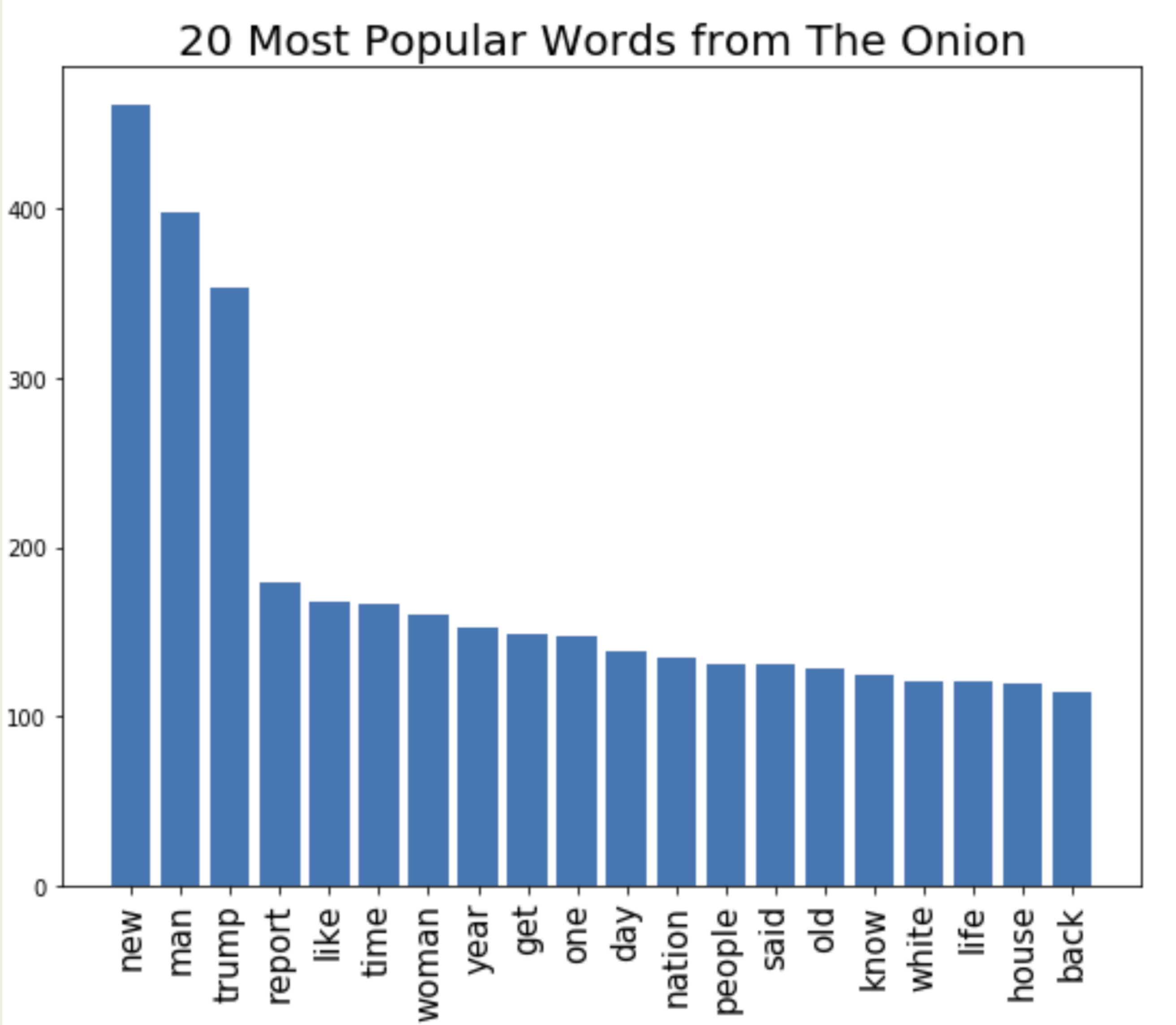
THE TOPICAL

**ExxonMobil To Simplify Oil Extraction By Cutting Earth In Half**

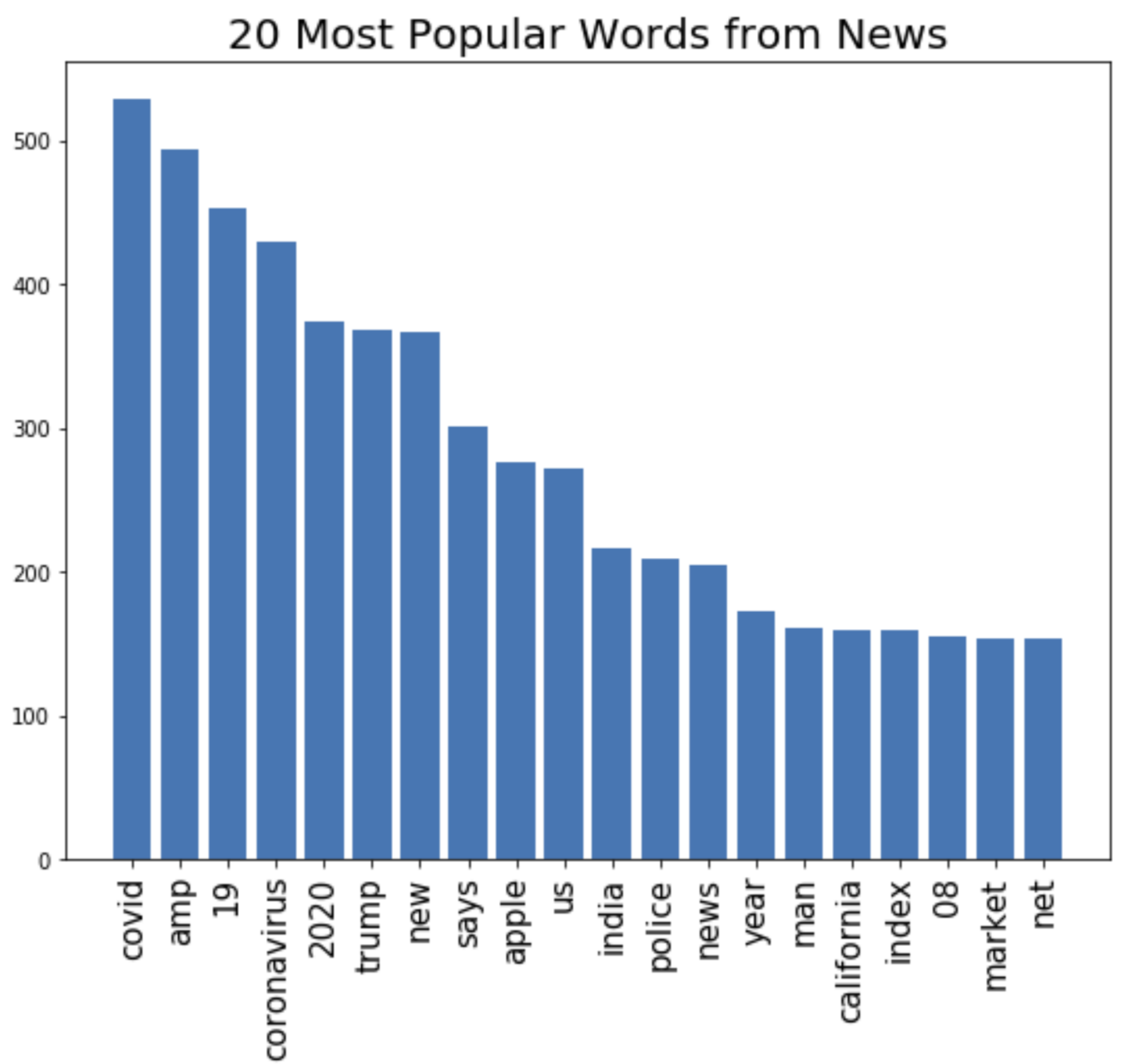
# Data Collection/ Cleaning

- Used Pushshift API to scrape posts from both subreddits
  - *100 posts at a time, every 3 seconds*
  - *7,178 from The Onion*
  - *10,000 from News*
- Removed posts which had a duplicate title
- No ‘selftext’ for any post
  - *Null value*
  - *[deleted]*

# Exploring the Data



20 Most Popular Words from The Onion



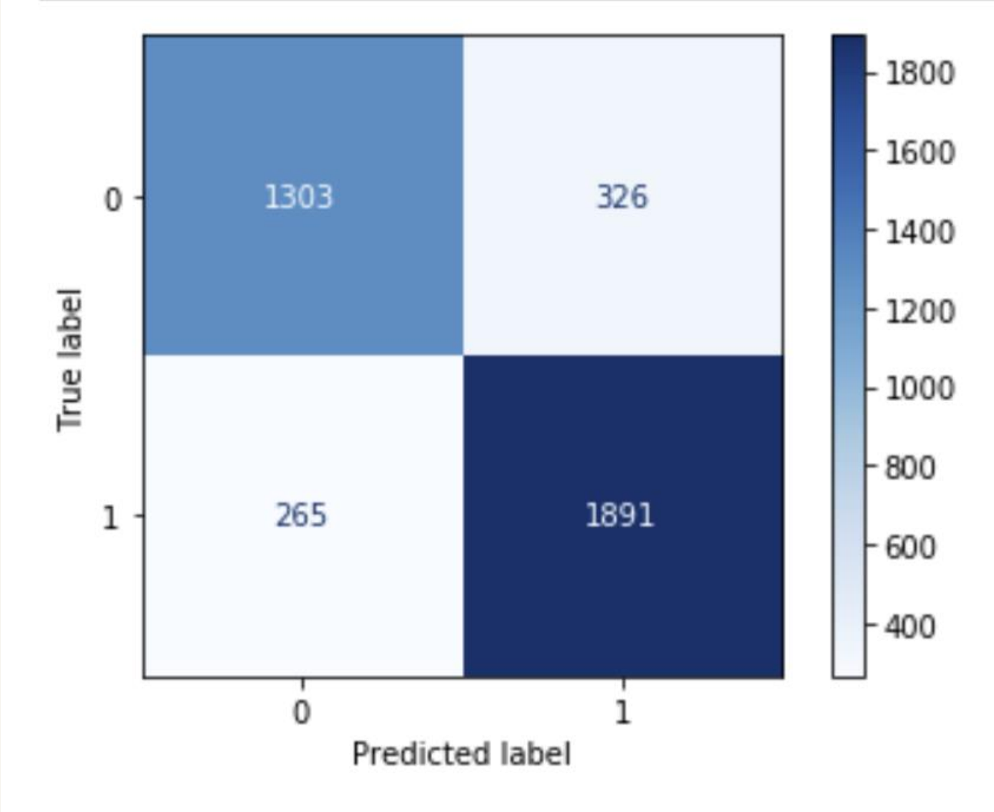
20 Most Popular Words from News

# Logistic Regression Model

- Used to model the probability of a certain class (i.e. ‘The Onion’ or ‘News’)

- Baseline: 43% Onion Posts, 57% News

- Variation: 9%
- Accuracy: 84.4%
- Precision: 85.3%



0 : ‘The Onion’
1 : ‘News’

# Mis-predictions (from Logistic Regression Model)

## Predicted News, actually The Onion (False Positive)

- Benefits Of Open Office Not Extended To CEO
- Most Anticipated TV Shows Of 2019
- Tips For Taking Care of House Plants

## Predicted The Onion, actually News (False Negative)

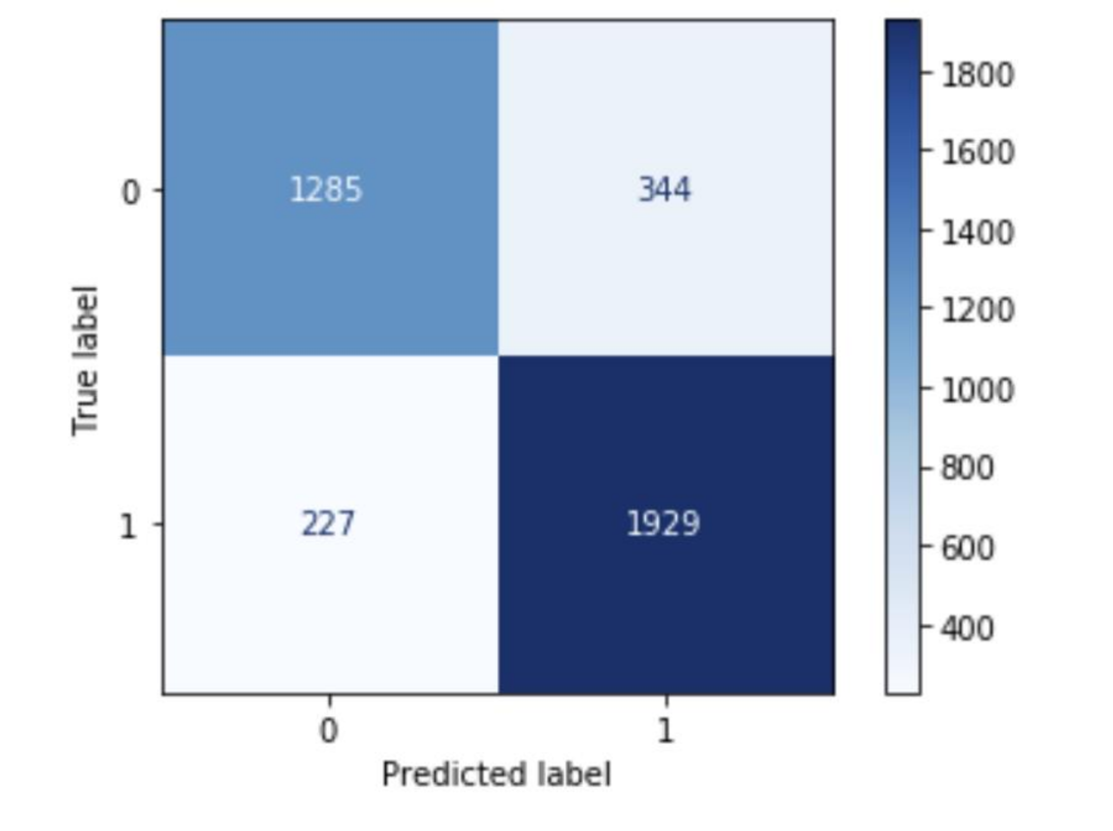
- Falwell's use of yacht comes under scrutiny
- Nike keeps on betting on Zero Waste
- Scarred by my Own Hands

# Support Vector Classification

- Finds a hyperplane for the best way to divide the two classes when plotted

- Baseline: 43% Onion Posts, 57% News

- Variation: 11%
- Accuracy: 85.0%
- Precision: 84.9%



0 : ‘The Onion’
1 : ‘News’

# Other Models

- KNearestNeighbors: 65.0%
- Random Forest: 57.4%

# Future Directions

- Stronger processor for SVM grid search (or a lot of free time)

- Filter future Reddit posts
  - *Help moderators determine ‘fakes’ for either subreddit?*
- ‘Fake News’ filter on Chrome?

# QUESTIONS?

# References

- https://www.reddit.com/r/TheOnion/
- https://www.reddit.com/r/news/
- https://github.com/pushshift/api
- https://blog.aylien.com/support-vector-machines-for-dummies-a-simple-explanation/#:~:text=Support%20vectors%20are%20the%20data,elements%20of%20a%20data%20set.
- Lessons!