



# Headline Satire Detector



## ABOUT:

Classifying news articles through headline analysis using NLP and deep learning methods



## TOOLS:



Keras



learn



BeautifulSoup



## TIMELINE:

5 weeks



/gabrielatanumihardja



/gabrielaeaton

# All Generalizations are Wrong, Including This One...

---

- **Satire & Natural Language Processing:**
  - Can ML models detect the subtle nuances?
  - Satire can be difficult to differentiate by headlines



# Can You Tell?

---

- **Sample headlines:**

- The Lamp Was a Clue to a Life I Didn't Know My Mother Had
- DNC Pours All Campaign Funding Into New York, California  
To Win Popular Vote By Even Greater Margin Than 2016



# Can You Tell?

---

- **Sample headlines:**

- The Lamp Was a Clue to a Life I Didn't Know My Mother Had
- **DNC Pours All Campaign Funding Into New York, California  
To Win Popular Vote By Even Greater Margin Than 2016**



# Project Framework

1



## Data acquisition and cleaning

Scrape headlines from four online news sources

2



## Train baseline models and evaluate

Logit, SVM, Naïve Bayes, KNN, Random Forest

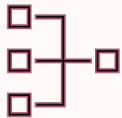
3



## Adjust data

Resample by year and sources

4



## Train advanced models and deep learning models

AdaBoost, XGBoost, BERT model

# Results + Predicting Sources

---

- **Validation accuracies ranging from 63% - 90%**
  - **BERT model** - 90% validation accuracy
- **Predicting source of article:**
  - **BERT model** - 80% validation accuracy
  - **The Beaverton** headlines often misclassified as belonging to **the Onion** (20%)

# Predictive Words

---

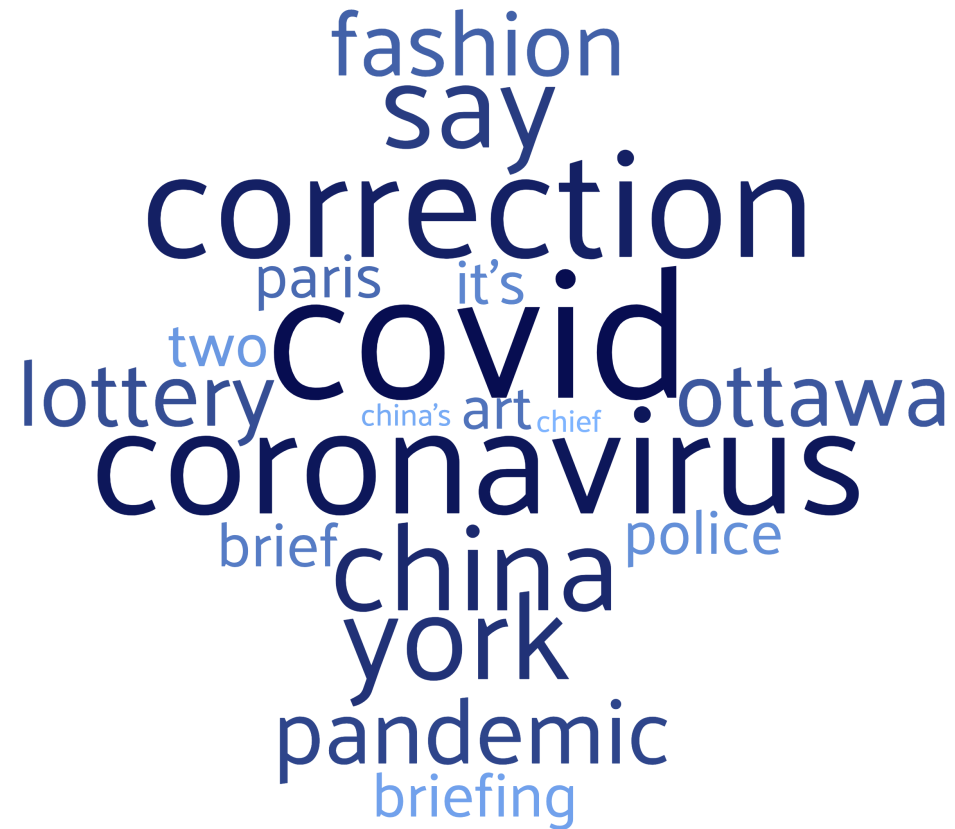
## Satire



A word cloud for the 'Satire' category. The words are primarily in shades of red and orange. The most prominent words are 'local', 'man', 'area', 'nation', and 'harper'. Other visible words include 'guy', 'single', 'nfl', 'week', 'fan', 'realize', 'year-old', 'friend', 'tim', 'entire', 's\*\*t', 'clearly', 'reveals', and 'k'.

guy  
local  
man  
area  
nation  
harper  
year-old  
single  
nfl  
week  
fan  
realize  
friend  
tim  
entire  
s\*\*t  
clearly  
reveals  
k

## Legitimate



A word cloud for the 'Legitimate' category. The words are primarily in shades of blue. The most prominent words are 'covid', 'coronavirus', 'china', 'york', 'pandemic', 'correction', 'fashion', 'say', 'ottawa', 'police', 'briefing', 'lottery', 'two', 'paris', 'it's', 'china's', 'art', 'chief', 'brief', and 'pandemic'.

fashion  
say  
correction  
covid  
coronavirus  
china  
york  
pandemic  
ottawa  
police  
briefing  
lottery  
two  
paris  
it's  
china's  
art  
chief  
brief

# Serve Up! What's Next?

×

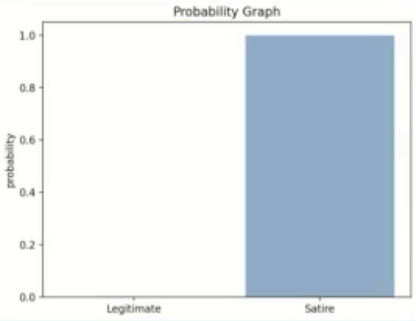
What would you like us to predict?

Satire ▾

This model was trained to predict satirical qualities of headlines from the Beaverton, the Onion, the Globe and Mail, and the New York Times.

So, is it a satire?

Probability Graph



Category	Probability
Legitimate	0.0
Satire	1.0

☰

## TWO AND TWO MAKES FIVE


-- George Orwell, 1984

---

Enter a headline below

You People Made Me Give Up My Peanut Farm Before I Got To Be President

We predict that this is a **satirical news** with **100.0%** probability!



The people heard it, and approved the doctrine, and immediately practiced the contrary

— Benjamin Franklin, The Way to Wealth



# Thank you!

---



[linkedin.com/in/gabrielatanumihardja/](https://linkedin.com/in/gabrielatanumihardja/)



[github.com/gabrielaeton](https://github.com/gabrielaeton)