Sarcasm Detector

Hunter Sapienza

Problem Statement

- What keywords differentiate sarcastic and non-sarcastic article headlines?
- Which classification model performs most accurately in predicting sarcasm within headlines?

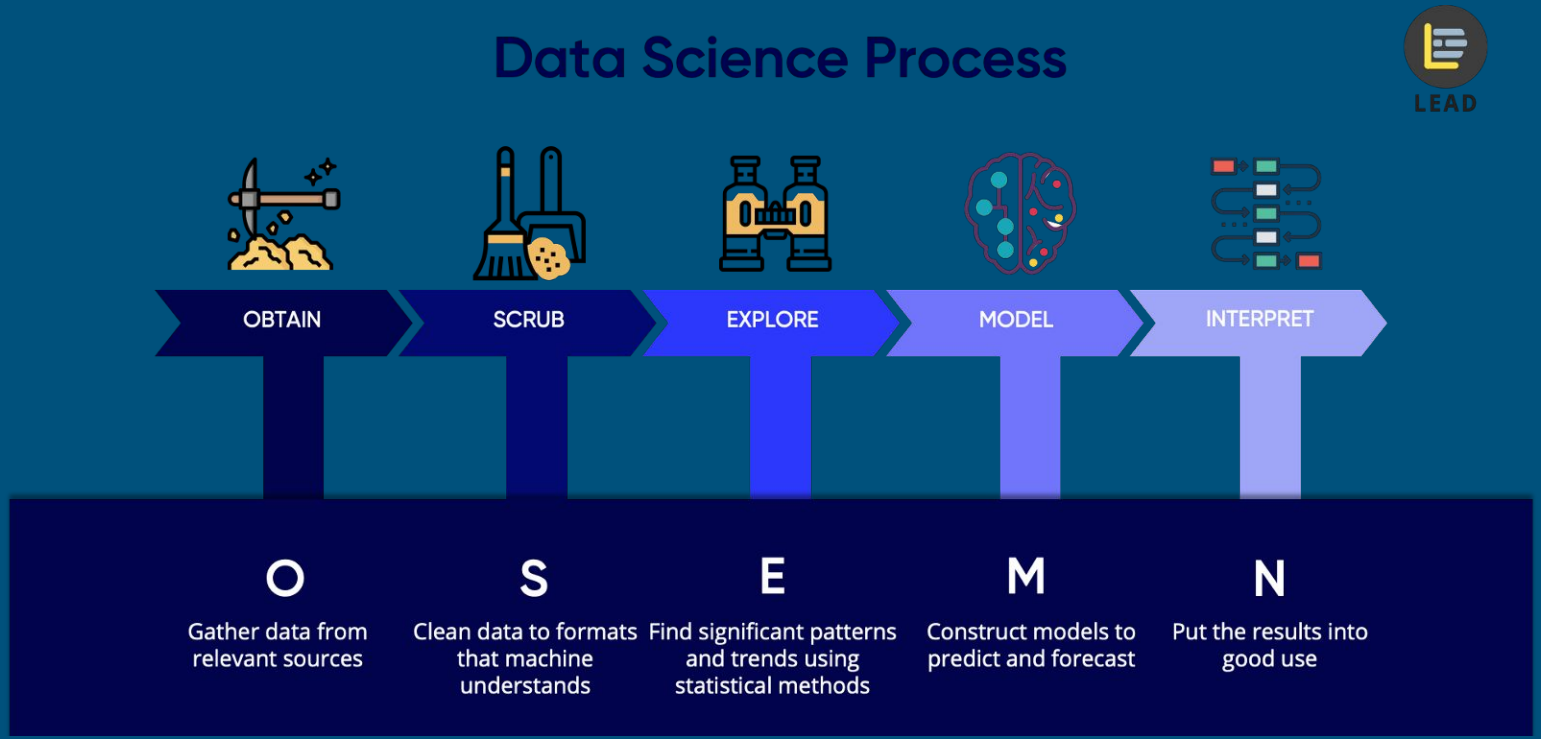
Business Value

How can we predict article content?

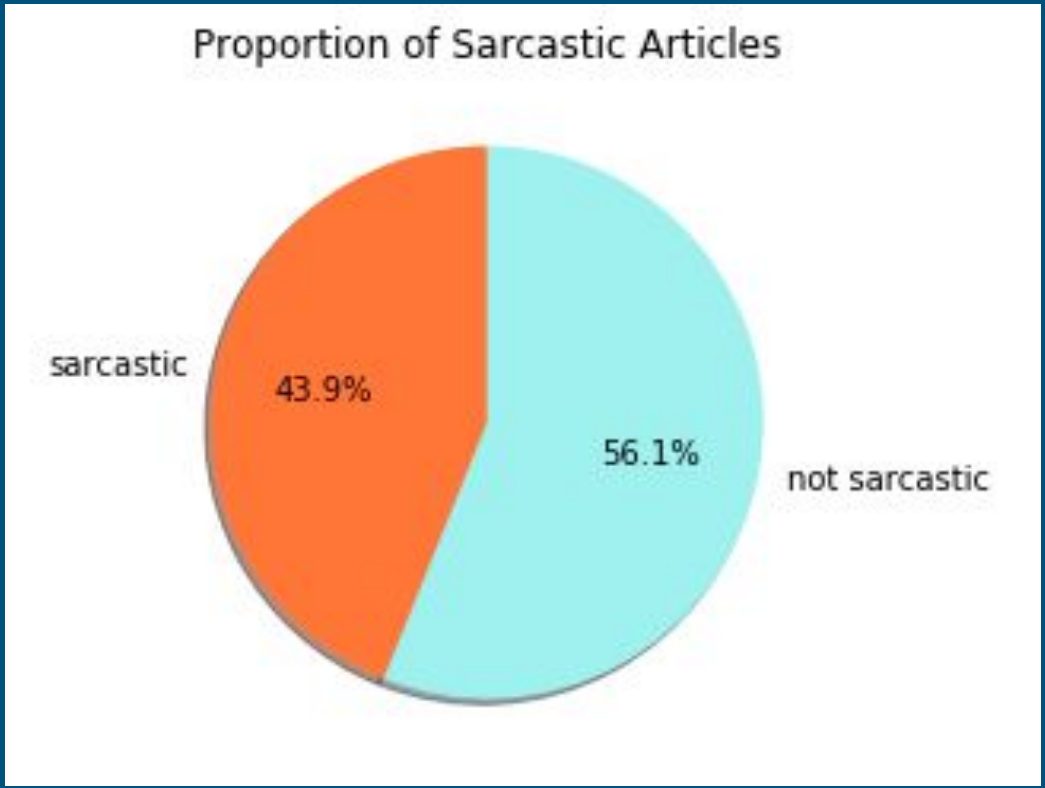
- Differentiate between real news and fake news
- Analyze keywords across current news sources
- Foundation for future NLP work in news and media

News Sources

Data Science Framework



Dataset Composition



Token Summary

	All Headlines	Sarcasic	Non-Sarcastic
Total Tokens	282582	122043	160539
Unique Tokens	29291	19405	19624

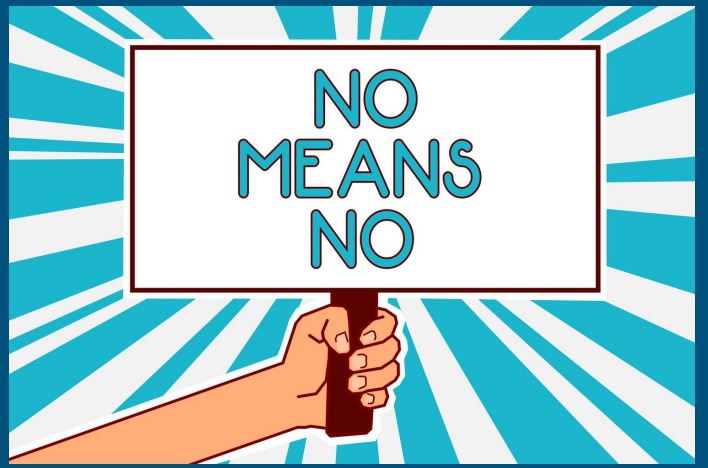
Most Common Words, Bigrams, & Trigrams

Four categories of news content:

- Political themes
- Social issues
- Cultural trends
- Places of interest

Political Themes

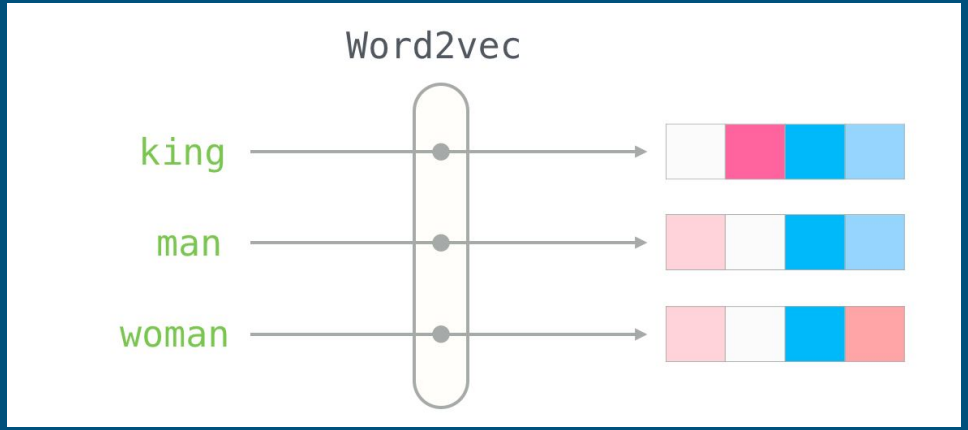
Social Issues

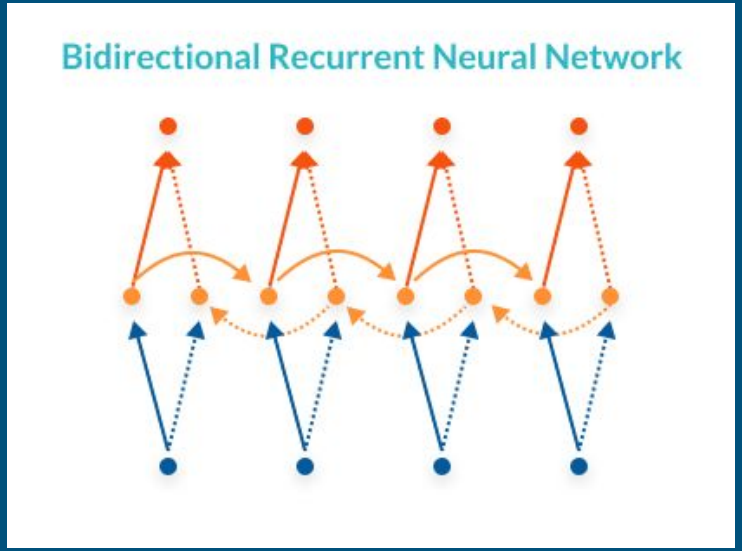


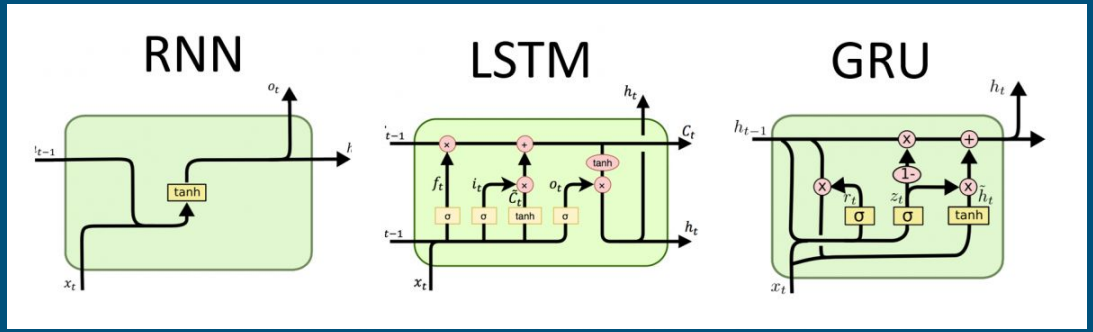
Cultural Trends

Places of Interest

Modeling the Data







Modeling Findings

Layer (type)	Output Shape	Param #
input_3 (InputLayer)	(None, 150)	0
embedding_5 (Embedding)	(None, 150, 128)	3840000
bidirectional_1 (Bidirection	(None, 150, 50)	30800
global_max_pooling1d_5 (Glob	(None, 50)	0
dropout_9 (Dropout)	(None, 50)	0
dense_9 (Dense)	(None, 50)	2550
dropout_10 (Dropout)	(None, 50)	0
dense_10 (Dense)	(None, 1)	51

Total params: 3,873,401
Trainable params: 3,873,401
Non-trainable params: 0

Bidirectional LSTM

```
5342/5342 [==============================]
Test set
  Loss: 0.404
  Accuracy: 0.823
```

Train accuracy: 86%
Validation accuracy: 82%

Test Accuracy: 82.3%

Summary

With over 26,000 article headlines…

From over 280,000 tokens and 29,000 unique tokens…

- Trump, Obama, and Clinton most mentioned people
- References to men 3x as likely as references to women.
- Climate change, health care, sexual assault, and mental health
- Sarcastic bigrams vague, general, and widely relatable.
- New York and North Korea most common places
- Donald Trump appears in nearly all the top lists

Future Work

1. Train models that can differentiate between real news and fake news
2. Detect the presence of bias and subjectivity
3. Models that can predict the objectivity of a particular news article
4. Decrease bias & subjectivity of news sources