Satire Detection In News Articles

•••

Alex Walsh, DSIR Data Scientist
2022 - 12 - 23

The Problem

1. Dead man propped up by two other men in attempt to collect pension at post office
2. Man uncooperative after being shot in the head in Grand Rapids, police say
3. U.N. votes to remove Iran from women’s rights council
4. Woke coke: Drug dealers marketing ethically sourced cocaine

The Solution

- r/TheOnion and r/nottheonion tread the line between fiction and reality

- Using the Pushshift API and web scraping, “news” articles can be gathered

- Natural Language Processing techniques can learn the styles of satirical articles

The Process - Aggregation

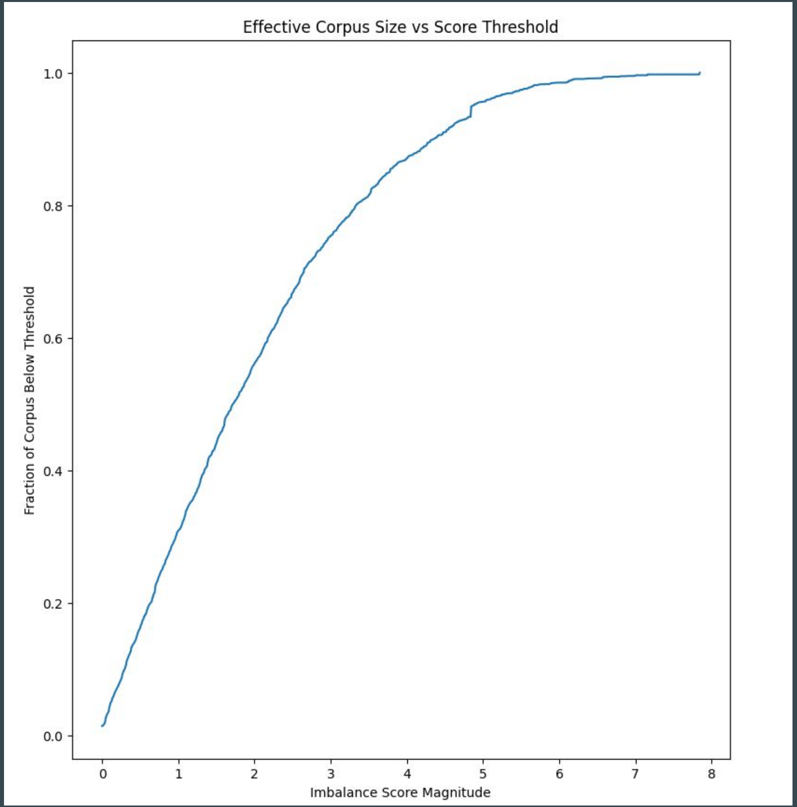
- Posts are gathered from reddit and stored in a SQL database
- A total of over 500,000 posts were gathered
- Of these, about 20,000 posts were chosen for a total of 7.6 million words

the ONION®
America's Finest News Source

Nøt the ONION
REDDIT'S FINEST NEWS SOURCE

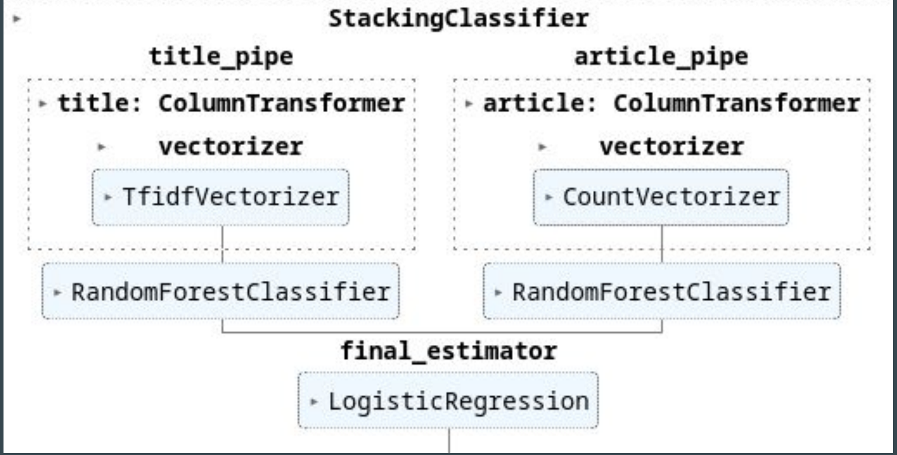
The Process - Cleaning

- There was a sufficiently large amount of data such that posts with missing content could be dropped
- Duplicates were removed with fuzzy matching
- Words were lemmatized to reduce data dimension
- Words were scored based on how imbalanced they were between the two categories



The Process - Modeling

- Articles were split into body and headline
- Each was fed into own model for individual predictions
- These predictions were then fed into a second model for a final prediction

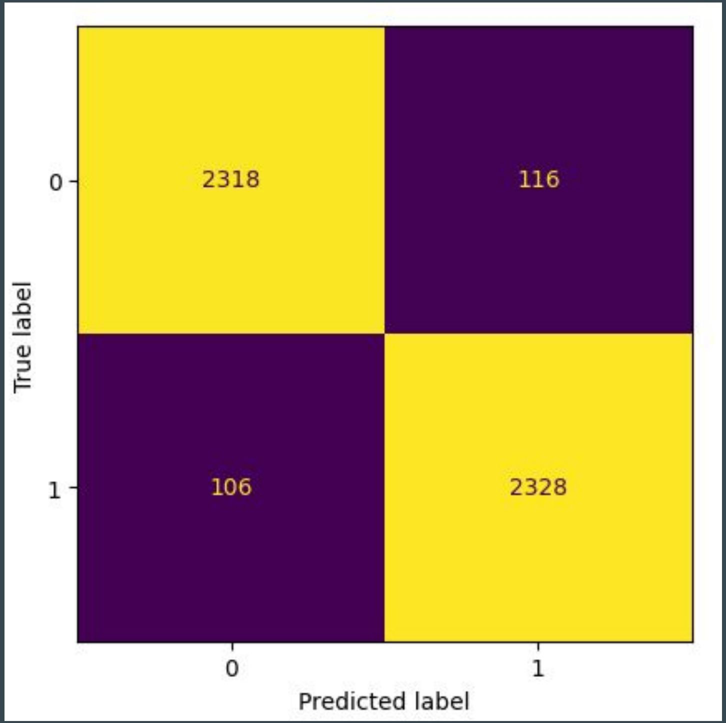


The Results

- The model achieved over 95% accuracy compared to 50% baseline

- Misclassified articles were split evenly between satire and non-satire

- The article body was five times as important for predicting satire compared to the title



Conclusions and Next Steps

- The model was extremely effective at classifying satire, and is viable for deployment following a few additional tests

- The imbalance scoring metric proved effective, but should be improved going forward

- This model can easily be deployed as a browser extension, social media bot, or phone app to allow people to use it

Summary

- Recognizing satire can be difficult
- Machine learning techniques are a viable option to fix this, but they require complex modeling strategies
- Article bodies are more important than their titles