Benjamin Karlsberg https://github.com/benkarlsberg/fake-news-detector

FAKE NEWS DETECTOR

What is "Fake" News?

In a broad sense, "fake" news contains information that is false or exaggerated beyond objective facts. Fake news articles tend to contain language that is opinionated and biased towards one viewpoint in order to promote a political agenda or to generate views for advertising.

How the Detector Works

By honing in on the language that tends to be used in "real" and "fake" news articles, the detector is able to calculate the probability of the article being from a factual source.

*Note: this does not imply that the detector verifies actual claims in the articles. It primarily suspects the article of being biased or not.

Where Did It Learn From?

The data that was used to train the model contains over 70,000 labeled articles that were sourced from "Politifact.com" (80%), "The New York Times" (10%), and "The Onion Magazine (10%)." 47% of the articles were labeled as "fake" and 53% as "true."

Generating Features:

- Lemmatization (Root Words)
- Bigrams/Trigrams
- Remove "stop words"

TF-IDF Vectorizer:

- 20,000 Feature Limit (Terms)
- Gives weight to each feature
- Based on importance within entire corpus of articles

Random Forest Classifier:

- Determines feature importances by splitting
- Gini Impurity: Decides split quality by incorrect classification likelihood

Model Selection:

The optimal model was decided by performing Cross Validation on the training set for several different models, including: Logistic Regression, Gradient Boosted Classification, Naive Bayes Classification, and Random Forest Classification. The Random Forest Classifier performed the best in terms of accuracy (total correct) and precision (fewest false positive results for real news being classified as fake).

Results On Final Holdout Set:

After an 80-20% train-test split on all of the data, the Random Forest Classifier performed well with an accuracy score of 93.5%, precision score of 96.1%, and recall score of 89.7%.

Technologies Used:

Numpy, Pandas, Matplotlib, Scikit-Learn, Jupyter Lab, Flask (Web Application), Amazon Web Services (EC2)