Using Data Science to Identify Fake News

galvanize

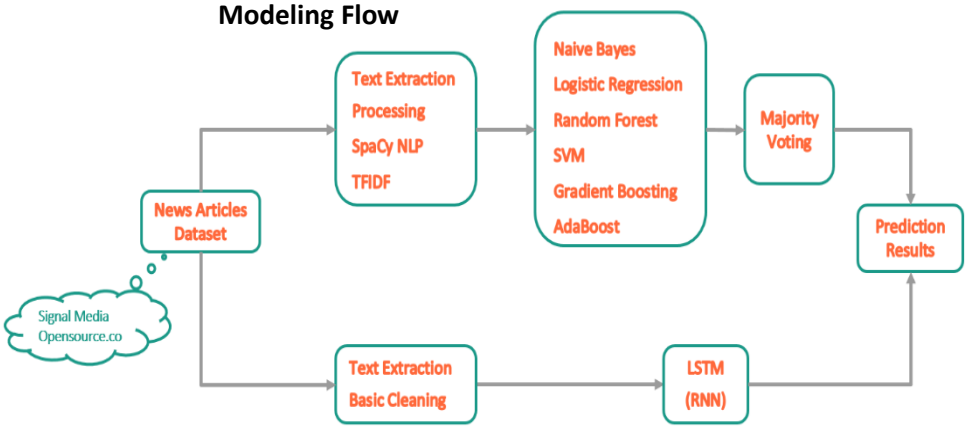
Summary

Fake news misleads readers and manipulates them into believing something that isn't real. The prevalence of fake news on social media lately became a very big problem. In this project, I worked on detecting fake news using data science and machine learning algorithms.

Data

Dataset published by Signal Media was used to obtain news articles from a variety of news sources. 'Fake' news sources were identified using OpenSource.co and 'Real' news sources were identified using newsapi.org. Text from news content was cleaned based on content exploration and the SpaCy library was used for text cleaning and Named Entity Recognition.

Modeling Flow



Final dataset was built as a corpus of labeled real and fake news articles and included about 6500 real news and 3200 fake news. Next, TFIDF vectorizer was used along with bi-gram to generate relevant feature matrix.

Models

Various classification models (Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting and AdaBoost) were implemented and their performance was compared. In addition, results from an ensemble model (majority voting) were also obtained to try and achieve better performance.

Based on these results, many of these models seem to work decently well for the small dataset used, with Gradient Boosting model performing the best. It would be interesting to see the performance of these classification algorithms on a larger and much better

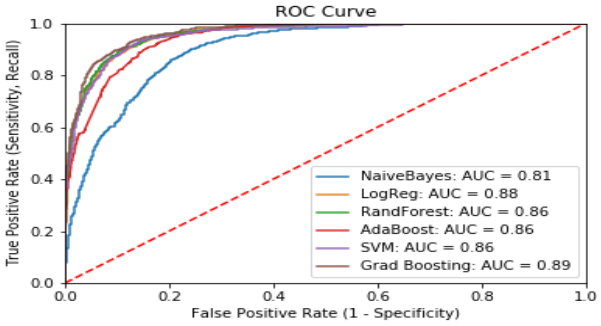
Future Work

- Gathering more data with better definition of fake news (currently relying on Opensource tagging)
- Use of others features (e.g., source, URL, publishing medium etc.) to improve classification
- Use pre-trained embedding's such as Word2Vec and GloVe in RNN
- Learn more on tuning hyper-parameters of deep-learning models

Deep Learning

Since the order of the words can carry useful information in a text classification, I also explored a recurrent neural network using LSTM which is considered to work well with serialized data such as text. I transferred original text to a fixed length integer vector and only considered 500 words and padded short length articles with 0. I also used word embedding to transfer each word id to 32-dimension vector.

MODEL	ACCURACY	F1 SCORE	PRECISION	RECALL	AUC
Naive Bayes	81.97	74.00	70.67	77.65	0.81
Logistic Regression	89.32	83.69	84.44	82.95	0.88
Random Forest	89.15	82.41	88.73	76.92	0.86
Support Vector Machine	88.91	82.47	86.27	79.00	0.86
Gradient Boosting	91.00	85.93	88.89	83.16	0.89
AdaBoost	87.05	80.74	79.40	82.12	0.86
Ensemble Model	91.42	86.57	89.56	83.78	0.89
RNN with LSTM	89.78	84.50	85.26	83.66	0.88



Tech Used

Python | Pandas | Scikit-learn | SpaCy | NLP | Keras | Matplotlib

https://github.com/soniampub/Fake_News_Detection