Predicting Real versus Fake News

Keshav Mantha

Problem Statement: Can we predict whether a news article is “fake” or not given existing data on real and fake news?

Context: Media bias is rampant today more than ever. Figuring out whether information published online is real or not can be a huge problem, and this issue has arguably been the crux of social and political divisiveness, both in the USA and around the world. Fake news can reduce the impact of real news by competing with it; a Buzzfeed analysis found that the top fake news stories covering the 2016 U.S. presidential election received more engagement on Facebook than top stories from certified major media outlets. Fake news also has the potential to undermine trust in serious media coverage.The term has at times been used to cast doubt upon legitimate news, and recently has been used even by the political and business elite to describe negative press coverage of themselves. These impacts can be disastrous to the moral and social fabric of any democratic nation.

Criteria for Success: If we can achieve a strong fit between our train and test data by splitting up the datasets, then we can predict the viability of a news article as “real” or fake”. Our goal is to train a highly accurate model on the fake and real news data to fit the test data well under cross-validated metrics.

Scope of Solution: This model could be implemented to check other news sources and sites for an initial level of validity.

Constraints: Our dataset comes has been scraped for articles across the internet, and has conveniently been split into 2 CSV files for easier readability.

Stakeholders: Any individuals that consume news and media to get their information and current events knowledge. This amounts to entire national populations.

Notable Features: The most important (and consistent) features across these data are “Title”, “Text”, and “Author”. We also have indications of the subject of the article as well as the publication date. Text classification will likely be the most notable approach to modeling. Notable algorithms and techniques would include LSTM, Bag Of Words, and TF-IDF alongside regression methods like Random Forest and Logistic Regression.

Data Sources: Split datasets with news classified as either fake or real from kaggle: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Evaluation Metrics: The most notable prediction metric will be a strong R-squared value, which will indicate model performance accuracy. We want the data-fitting to be quite close to ensure that when test data is fed into the model, the real-vs-fake classifier works as intended.