IEOR142 Machine Learning & Data Analytics Project Proposal: Getting Real about Fake News

(A team consisting of Aditya Tyagi, Raleigh Lukas, Han Zeng)

I. Motivation

An interesting recent development is the proliferation of deliberate misinformation or hoaxes via traditional print and broadcast news media or online social media – informally termed 'fake news'.[1] Accusations of 'fake news' were prolifically used in the recent U.S. Presidential Elections and have become a frequently heard term in the broader national conversation.[2] And so, identifying when a news story may be authentic or potentially 'fake' is of great importance to the mindful reader.

Naturally, the machine learning community has sought for ways to contribute a solution to this problem. For example, the recently concluded 'Fake News Challenge' challenged researchers to identify the 'stance' of a particular news article ("stance detection").[3] In particular, a popular Chrome plugin ("BS Detector") helps users identify whether a particular webpage contains information that is 'fake', 'biased', 'a conspiracy', etc.[4] To accomplish this, it uses a manually compiled list of untrustworthy websites and then mechanically checks the target webpage for links/references to such dubious websites. However, given the sheer size of the World Wide Web, and the expected increase in such websites, such a manual/hand-labelling approach may be infeasible.

In a nutshell, our project aims to solve this problem by using the automated ML approaches we have learned in class. Thematically, this is a multi-class classification problem where the labels are "fake", "conspiracy", "biased", etc. and the features consist of metadata about the website, as well as the text of the headline.

II. Data

Broadly speaking, our dataset incorporates features of three main types:

1. Metadata about the website: date of creation, spam score, language/country of origin, etc.
2. Social Media User Engagement: no. of likes/comments/shares on Facebook.
3. Text of the website headline ("BOOM! Math shows Trump would have beaten Obama in Romney-Obama Election!")

The data was sourced from Kaggle. Originally, the data was collected by webscraping known 'fake news' websites to collect website metadata, social media user engagement, and headline text. A few features from external sources are also present such as the web domain rank, etc. Instances were labelled using Chrome's BS Detector plugin.

Currently, the classification labels only include various types of fake news websites ("conspiracy theory", "bias", "satire", etc.). However, to really evaluate our model's discriminative power, we also would like to add some instances of 'non-fake news' websites into the dataset.

Feature Engineering

We also plan to add some more features to improve the predictive capability of the model:

- Time of day of article writing
- Alexa statistics about website: traffic rank, top country visitors, daily average time on site per user, etc.
- Sentiment of headline (positive/negative, etc.)

III. Analytics Models

At present, we have sketched out an ensemble learning model building on two base models:

- A NLP-based model for the text aspect of a website
- A Supervised Learning technique (Random Forest, LDA, Multinomial Logistic, k-NN, etc.) using the other features.

We still need to figure out the exact mechanics of prediction. Some possibilities are:

- Probability Prediction for each class, and we pick the class with maximum probability
- Majority vote/combination of the outputs of the individual models forming the ensemble

An immediate operational problem is the existence of numerous NULL values in the dataset. We still need to think about a suitable means to impute values into them (or perhaps delete those instances altogether?).

Prior to using NLP techniques, we also intend to use standard text preprocessing (stemming, conversion to lower case, stop word removal, Zipf's Law word reduction, etc.).

IV. Impact

In the era of 'fake news', our automated approach to fake news detection has enormous value. Some key people we expect to benefit from our model are:

1) News Desks at various publishers/news corporations: Helps verify whether an incoming news story is genuine or not
2) Netizens who read get their news information online: Identify misinformation (from blogs, smaller websites, etc.)
3) Social Media Administrators at Facebook, Twitter, etc.: Locate disingenuous stories on their network and monitor their spread.
4) Campaign Consulting Firms: Assist in Web monitoring for potentially libelous information about their candidate

Some immediate real-world impact we hope will result from our project are:

- Creation of a Chrome plug-in based on our model that helps users identify whether a story they are viewing is trustworthy or not.
- Possible sharing of model with the makers of BS Detector, to improve their work.

V. References

[1] https://en.wikipedia.org/wiki/Fake_news

[2] http://www.cnn.com/2017/10/08/politics/trump-huckabee-fake/index.html

[3] http://www.fakenewschallenge.org/

[4] http://bsdetector.tech/