NEW YORK UNIVERSITY

FDS

Fake News Detector

Karanpreet Singh Wadhwa (ksw352)

Arjan Singh Narula (asn419)

Oct 2019

Version 1.5

Fake News Detector

Background: The widespread propagation of false information online is not a recent phenomenon, but its perceived impact in the 2016 U.S. presidential election has thrust the issue into the spotlight. Apart from this, fake news has been the cause of many agitated situations and even fatalities in many countries.

In General, The four observed flavors of "fake news":
1) *Clickbait* — Shocking headlines meant to generate clicks to increase ad revenue. Frequently these stories are highly exaggerated or totally false.
2) *Propaganda* — Intentionally misleading or deceptive articles intended to promote the author's agenda. Often the rhetoric is hateful and incendiary.
3) *Commentary/Opinion* — Biased reactions to current events. These articles frequently tell the reader how to perceive recent events.
4) *Humor/Satire* — Articles written for entertainment. These stories are not meant to be taken seriously.

Potential Question(s):

- How to recognize whether a given new is real or fake?
- Are there any words or groups of words (Probably Bait) that are plentiful in fake news?

Predictor and Target Variable(s):

- Predictor variable: News Title, Author Name, Site published on
- Target Variable: whether the news is fraudulent (1) or not fraudulent (0).

Data sourced from:

- Reddit has subreddit called "The Onion" (Fake news) and "Not The Onion" (Real News).
  - http://www.theonion.com/
  - https://www.reddit.com/r/nottheonion/
- Different dataset set for fake news provided by the University of Washington.
  - https://homes.cs.washington.edu/~hrashkin/factcheck.html

Data preparation:

Since I am working with the text, I need to use NLP methodology to scrutinize the text. Two of them would be.

- Count Vectorizer
- Term Frequency- Inverse Document Frequency

EDA:

Before leaping into making a model, I would take some time to do exploratory data analysis. To get more familiar with the data and possibly make some initial hypotheses.

Modeling:

This project Falls under the supervised learning Classification problem, as the dataset contains the target labels and labels are discrete 0 or 1.

I would try out different classification models. To name a few:

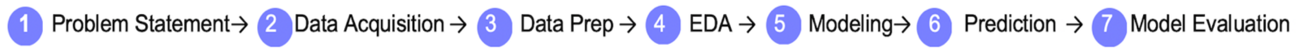
- Logistic regression
- SVM

Model Evaluation:

As I am working with a binary classification problem, I would use these Model Evaluation metrics:

- Accuracy
- F-Measure
- AUROC

Workflow:



1 Problem Statement→ 2 Data Acquisition → 3 Data Prep → 4 EDA → 5 Modeling→ 6 Prediction → 7 Model Evaluation

Quick Summary:

Title: Fake News Detector		
1) Problem Statement: • How to recognize whether a given new is real or fake? • Are there any words or groups of words (Probably Bait) that are plentiful in fake news?	**2) Predictor and Target Variable(s):** • Predictor variable: News Title, Author Name, Site published on • Target Variable: whether the news is fraudulent (1 or not fraudulent (0).	**3) Data source:** • Reddit • University of Washington site
4) Data Preparation and EDA: • Since I am working with the text, I need to use NLP methodology to scrutinize the text. Two of them would be. ○ Count Vectorizer ○ Term Frequency- Inverse Document Frequency • Get more familiar with the data and possibly make some initial hypotheses.	**5) Modeling:** This project Falls under the supervised learning Classification problem, as the dataset contains the target labels and labels are discrete 0 or 1. To name some: • Logistic regression • SVM	**6) Model Evaluation:** As I am working with a binary classification problem, I would use these Model Evaluation metrics: • Accuracy • F-Measure • AUROC