MAIS 202 - Data Selection Proposal

Zahur Ashrafuzzaman, Jules Barbe, Benjamin Segall

1 Introduction

The goal of our project is to create an algorithm that, when given an article, will be able to 1. tell if the article is fake and 2. redirect the user to a similar article from a trustworthy newsource.

2 Dataset

We chose this dataset from Kaggle (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset). It contains about 18 000 fake articles scraped from various sources and 21 000 true articles scraped from Reuters.com. The fake news articles were collected from unreliable websites that were flagged by Politifact (a fact-checking organization in the USA) and Wikipedia. The dataset contains different types of articles on different topics, however, the majority of articles focus on political and World news topics. The dataset is already in CSV form, with the title, text, subject and date all in different columns. We will mainly be using the title and text for our algorithm, where we will need to do typical text pre-processing, like setting all letters to lowercase, extracting words, removing useless function words, etc.

3 Methodology

We first want to determine what makes an article true or false purely from the way it is written, mainly from the vocabulary used and the tone of the article. We are thinking of using some type of classification model, where we will able to make a link between certain words and combinations of words and the truth-value of the article. However we also want to have a continous metric for both parts of our algorithm, where we could say the article is x% trustworhy, and where our recommended alternate article would have y% accuracy compared to the original article content-wise. We aren't sure how to implement this at the moment, but we're guessing it has something to do with error probability or a loss metric within our algorithm. Since we have a classification problem and based on what we've seen so far in the bootcamp, we are thinking of using a confusion matrix as our evaluation metric.

4 Application

As for the application, we could do a web app where the user simply gives out a link and the algorithm parses the page and if its format is article-like, decides if it is fake or not. We could also make a chrome extension that would automatically do this process, such that if the user enters what seems like a news site, the algorithm would automatically determine the current page's truthfulness and recommend alternate news sources in the case it goes under a certain threshold.