Nat Hawkins
A49510775

CSE802 Project Proposal

It is important for individuals in the modern age to stay well informed with current events. As the amount of information available to consumers continues to exponentially grow, it is becoming increasingly important for an individual to be able to distinguish legitimate information from illegitimate “fake news.” One area of news, specifically, where this gets a lot of attention is in political news where the term “fake news” is frequently used to discredit articles or sources altogether. Despite the social and political undertones associated with the term “fake news,” the question at hand is an important one: how does one determine whether the information that they are reading is true or not? In this project, we will analyze a large number of political news articles and develop a method that accurately and efficiently identifies whether a piece of information is true or not by examining the article titles alone. Once developed, this binary classifier can be applied to any number of articles as an initial filter for flagging false information in media sources and provide consumers with a tool for staying correctly informed on current events.

For this project, we will utilize a publicly available dataset from kaggle. This dataset contains the title and full article text for approximately 20,000 legitimate and fake articles collected over a span of approximately 3 years. These articles are primarily focused on political news. As such, the classifier that we train will be predominantly applicable to vetting political news sources once complete.

To carry out this task, we will convert the titles of these articles into a one-hot vector. The length of this vector will be equal to the size of the vocabulary of these articles (all unique words across all articles). A 1 in an index position denotes that word appears in the title somewhere, where a 0 denotes the word is not present. This strategy utilizes a bag of words (BOW) assumption where the presence of a word or words, not necessarily the context or frequency, is the determining characteristic used for classification. Our final feature matrix will be of size N x D, where N is the number of samples (articles from both classes), and D is the size of the vocabulary. From there, we will then use a non parametric Bayesian estimation algorithm to train a classifier to predict whether an article is true or not using a BOW representation of its title alone.

Using only the title as opposed to the title and abstract has a number of benefits. First, including the full article text increases D dramatically. Increasing the number of features for a fixed training invokes the curse of dimensionality. Second, using a smaller feature spaces lends itself to improved computational efficiency for training and testing as well as implementing this on novel articles on the user side. Third, by examining only the titles, we aim to answer the question of whether titles are indicative enough to distinguish fake news from true news, or if the body of the text is inherently necessary. The conclusion for point three will be a qualitative answer determined based on interpreting the performance metrics for our experiment.

To evaluate our method, we will use 5 times 10-fold cross validation ensuring approximately equal class distributions across folds. We will report an array of performance metrics including: accuracy score, area under the precision-recall curve (auPRC), area under the receiver operating characteristic curve (auROC), F1-score, precision, and recall. Each metric will tell us something different about the classifier we train, and as such, we want to report as many metrics as possible in order to provide a full picture of this model’s practical usefulness.