Fake News: A Study of the Differences between Real and Fake News

Data Mining – CSPB4502

Group 3: Claudia Hidrogo, Madeline Odom, Jordan Sims, and Xiomara Winkler

1. Introduction and Project Background

According to the New York Times, “We are in an era of endemic misinformation -- and outright disinformation”, (Fisher, 2021). One very important facet in the battle against misinformation is the vital step of identification. As time goes on, it appears that misinformation is becoming harder and harder to spot to the naked eye. The goal of this project is to study AI-identified fake news, as well as real news, to determine key differences that could help the average newsreader be able to identify the differences between fake and real news.

Fake news and misinformation have been topics of growing conversation and interest over the past few years, and the field of study on this topic has been growing as well. In their paper “Fake News Detection Using Machine Learning Approaches”, authors Khanam, Alwasel, Safari, and Rashid discuss advanced methods of building AI machine learning models that can effectively differentiate fake news from real news. Shu, Silva, Wang, Tang, and Liu’s “Fake News Detection On Social Media: A Data Mining Perspective” takes a similar view to the fake news detection. Both of these works focus on the problem from the perspective of computer science, which is decidedly different from the objective of this project.

We will be using similar methodologies to classify our data sets, but from there, we will be analyzing the real and fake articles for key similarities and differences on the body text of the articles alone. We are not primarily focusing on the AI driven classification of real vs fake news, but analyzing the articles once the classification has been made for differences that could potentially be caught by the average news reader.

2. Data Preprocessing

2.1 Pre-processing

To start our project, our group gathered multiple data sets from across a variety of sources (See Section 2. Datasets). The first thing that we need to do is to combine each of these data sets into one Master set. This involves getting each of these data sets ready to be integrated with one another. For example, the dataset from Common Crawl titled “News dataset containing news from news sites all around the world.” was a dataset containing around 700,000 news articles in English from various news sites hosted around the world. The problem was that each of these data points was provided as its own, separate CSV file. So, to get all of this data into one place, a custom python program was written by the team to gather all of the relevant data from each of these individual files and writing them to one large CSV file. A similar process is needed for each of the data sets to prepare them to be all converted into one large data set to be evaluated.

2.2 Data Cleaning

The next step in preparing the individual data sets is cleaning them. This includes multiple steps, such as removing duplicates and unnecessary data . We will be identifying and removing the words from every news article that do not have a significant meaning. The list of words can be found by using “stop words” from the Natural Language Tool Kit (nltk) library (See Section 5.8 : Natural Language Tool Kit). Stop words are words commonly used that many programs and projects (such as ours) are programmed to ignore as they do not add any value or context to the text (*Removing stop words with NLTK in python* 2021). Some examples of stop words are “the”, “and”, “an”, “because”, and “of”.

2.3 Data Classification

Many articles from the datasets we are evaluating are pre-classified as fake news and real news, but not all of them. To prepare the data for integration, we will be using a Python AI-Model to classify all articles that have not already been classified as real or fake.

2.4 Data Integration

In this step, we will combine all the datasets labeled as fake and true news articles to obtain one large dataset. We will host the dataset in a MySQL database which will be hosted on AWS. We decided to use a cloud option so that the dataset would not need to be individually downloaded on each member’s system. Having a single repository would enable us to make any changes or updates to the data once without the need to individually update all members’ systems. Hosting on AWS also allows us to back up our database as well.

2.5 Data Transformation

After combining all the datasets and obtaining the lists of words of each news article, the strings will be converted to tokens (numbers) for easy data manipulation. This is known as tokenization.

2.6 Data Reduction

Lastly, the data will be reduced by eliminating repeated words and keeping only the unique words contained in each news article.

3. Datasets

Title	**Source**
A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles	Harvard Dataverse
News dataset containing news from news sites all around the world.	Common Crawl
Fake news detection dataset consisting of train and test data for fake news model development.	Kaggle: Samrat Sinha
Dataset consisting of fake and real news articles.	Kaggle: Clément Bisaillon
Generic word embedding dataset for fake news detection consisting of real and fake news from popular news sources.	Zenodo: Radu Prodan, Prateek Agrawal, Pawan Verma

4. List of Tools

The programming language Python will be the primary tool used to perform the data preprocessing of the news articles. The following Python libraries will be imported to create, train, and analyze the model.

4.1 Pandas

“Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.” (*Pandas*) We will be using it for its data frame manipulation and analysis tools.

4.2 Numpy

A power tool built for Python with a wide variety of mathematical applications. (*NumPy*) We will be using it for numeric analysis.

4.3 Matplotlib

“Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.” (*Visualization with python*). We will be using it for data visualization in our project.

4.4 Plotly

“Plotly's Python graphing library makes interactive, publication-quality graphs.”(*Plotly python graphing library*) We will be using it in our project to create interactive plots.

4.5 Keras

Keras “offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides.“ (*Keras*) We will be using it for deep learning in our AI model to classify fake vs. real news.

4.6 Natural Language Tool Lit for Python

This is a library made for Python that is intended “to work with human language data” (*NLTK*).

5. Evaluation

Confusion Matrices will be used to view the performance of our classification model. We will use our set of test data, for which we know the true values, to correct against our model results. Assembling the Confusion Matrix will allow us to quantifiably evaluate the results of our model. We will use the metrics of Accuracy and Precision, to measure exactness.

We will use a Recurrent Neural Network (RNN) to train our model of tokenized word vectors by utilizing the added temporal or sequential aspect that RNN allows us. We will also augment our RNN with Long-Short Term Memory (LSTM) to be able to establish long term dependencies between entailments and presuppositions in utterances. Using these modified Neural Networks will allow us to better process and correlate related semantic and syntactic features.

6. Milestones

Task	**Completion**
Data Preprocessing	November 1, 2021
Data Integration	November 8, 2021
Progress Report	November 20, 2021

Project Submission	December 10, 2021

7. Sources

E. Bender and A. Lascarides. *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Morgan & Claypool, Toronto, 2019.

Fisher, M. (2021, May 7). *'belonging is stronger than facts': The age of misinformation*. The New York Times. Retrieved October 24, 2021, from https://www.nytimes.com/2021/05/07/world/asia/misinformation-disinformation-fake-news.html.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012040. https://doi.org/10.1088/1757-899x/1099/1/012040

NLTK. (n.d.). Retrieved October 24, 2021, from https://www.nltk.org/.

NumPy. (n.d.). Retrieved October 24, 2021, from https://numpy.org/.

Pandas. pandas. (n.d.). Retrieved October 24, 2021, from https://pandas.pydata.org/.

Plotly python graphing library. Plotly. (n.d.). Retrieved October 24, 2021, from https://plotly.com/python/.

Removing stop words with NLTK in python. GeeksforGeeks. (2021, May 31). Retrieved October 24, 2021, from https://www.geeksforgeeks.org/removing-stop-words-nltk-python/.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Keras Team. (n.d.). Keras. Retrieved October 24, 2021, from https://keras.io/.

Visualization with python. Matplotlib. (n.d.). Retrieved October 24, 2021, from https://matplotlib.org/.