Fake News Classification

Roshni Koduri
roshnikoduri@gmail.com

ABSTRACT

As fake news becomes more of a problem across the US, its consequences are becoming more and more damaging. Many viewers who interact with information online are unable to distinguish the difference between fake information and real information, especially with social media worsening the spread of fake articles. This project aims to combat this disinformation through an artificial intelligence algorithm that can classify real articles versus fake ones. The model utilizes a BERT model and is tested on two different datasets, the Kaggle dataset and the LIAR dataset. The results are also presented through an app where readers can input any article and immediately find out whether it contains false information or not.

1. INTRODUCTION

Over the past decade, disinformation has become one of America's most prevalent and damaging issues. With the amount of fake articles rapidly increasing each day and becoming harder and harder to distinguish from real articles, it is not surprising that false news spreads faster than true news on many social media platforms. This disinformation has extremely dangerous impacts as it causes readers to have a false view of the world around them. It also increases polarization, as readers only interact with news that agrees with their views — even though that information may not be real — and become unwilling to listen to opposing viewpoints. Thus, fake news has exacerbated division and radicalization, and has worsened the nation's ability to unify and solve pressing issues.

This project aims to combat disinformation through a machine learning model that is trained to classify real news vs fake news. Since the first step to stopping the spread of fake news is recognizing it in the first place, this algorithm would be very useful in helping readers choose what kinds of information is trustworthy to read. Classifying an article as accurate/not accurate would help readers understand the kind of information they're interacting with and be less persuaded by fake news.

2. LITERATURE REVIEW

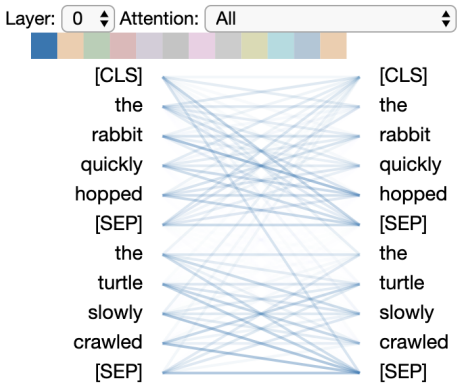
There are several models already built that classify fake news. These models all use different methods, such as Random Forest Classifiers or BERT models to process and classify text. Unfortunately, most of the models only have an accuracy around 75%, with the highest accuracy model (that I've been able to find on the Internet) being 88%. A big limitation of many of these research studies is the problem of a limited dataset; for example, one of the studies used a dataset with only 260 articles. Unfortunately, the New York Times explains that although fake news detection algorithms can get more and more accurate, it is unlikely that they will have the capability to fully replace humans. Fake articles are becoming more resemblant to true news as authors get better at hiding and spinning fake information. Since the computer just predicts fake news based on similarities it has seen in previous fake articles, it has no understanding of what the "truth" actually is, making it hard for it to adapt to changes it has never seen before or compare fake information with real information.

3. METHODS

3.1 BERT

In order to build the classifier, we used a pre-trained BERT model, which we then tweaked to fit our project. BERT, which stands for Bidirectional Encoder Representations from Transformers, was developed by Google and is a model for computers to understand human language (known as Natural Language Processing). BERT has two major components that it uses to train the computer to decipher language: Mask Language Model and Next Sentence Prediction.

Mask Language Model predicts a missing word in a sentence based on the other words in the sentence. This allows the computer to learn the meaning of various words and how they are used in context. The computer also understands words that are similar and different to each other depending on how well they can replace each other in a sentence (for example, a word like "big" can easily replace "large" in a sentence, whereas the word "small" cannot). Next Sentence Prediction then has the computer predict the next logical sentence given an input of one or more sentences. This trains the computer to understand the relationship between different sentences and ideas.



The visual above highlights how BERT learns how to process language by understanding the relationship between different words (for example, the word "rabbits" associated with "hopped" and "quickly").

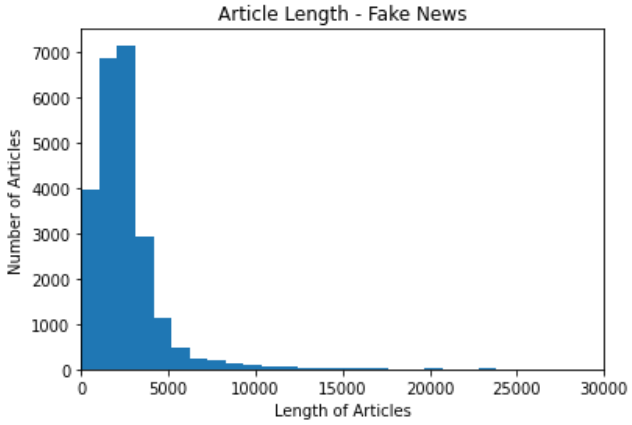
This project utilized transfer learning, which means we drew from pretrained models and tweaked them instead of building an NLP

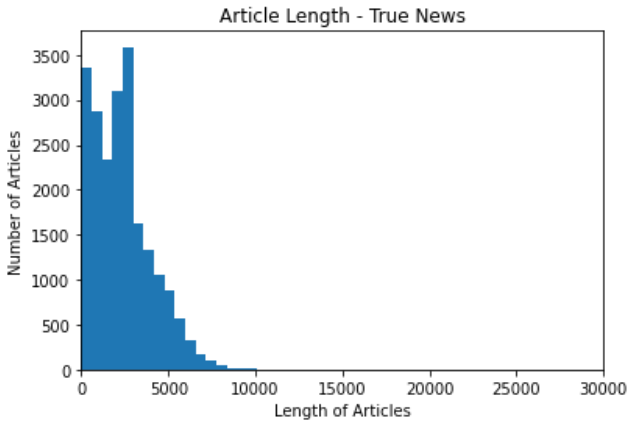
model from scratch. Specifically, we tailored BERT to be able to classify fake news. Our tokenization model, which separates chunks of texts into smaller pieces called tokens, was the distilibert-uncased model.

3.2 DATASET

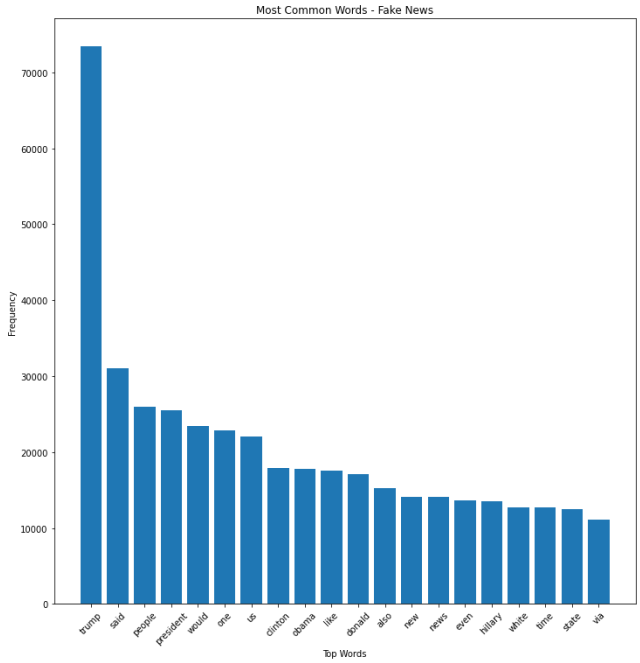
The dataset I used for the project was from Kaggle. It consists of two csv files, one with fake articles and the other with true articles, and has a total of ~17,600 pieces of data. With each article, the dataset also includes metadata such as the title, author, subject, and date (which were ultimately not used for the classification parameters). Since the dataset didn't have a way to numerically classify the fake and true news, I added numerical labels to classify each type of news, with fake news being a 0 and true news a 1.

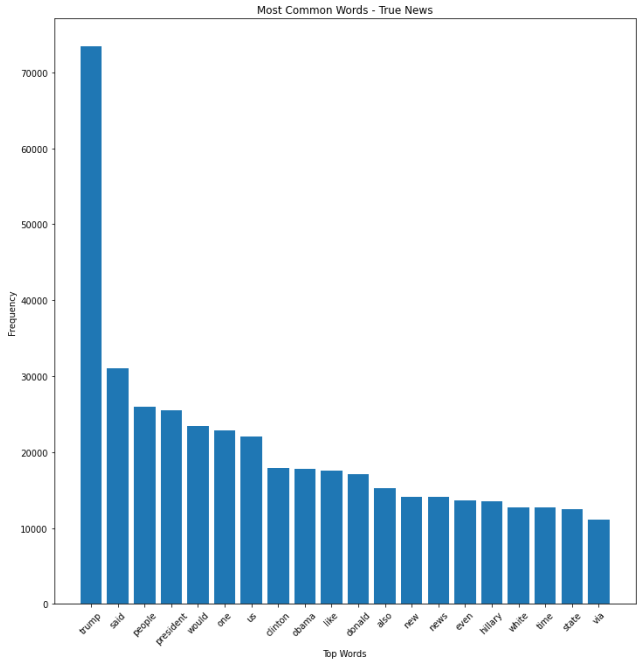
Comparing the lengths of the articles in each dataset was surprising. When we plot the lengths on a histogram, as shown below, we can see that the fake articles are actually longer than the true articles, which is unexpected.





Another visualization I did on the data is plotting the most common words. I omitted all "stop words", which are very common words such as "the" or "and". While one would think that both articles would have different content, the most used words in each article were surprisingly the same. This highlights the fact that fake news and true news are presented increasingly similarly, with fake articles being harder to distinguish from real articles.





As shown above, both the fake articles and true articles contain a lot of misinformation about Donald Trump and the government in general. Another interesting word that is present in both articles is "us", which suggests that many information sites want to make their viewers feel part of a group. This unifies readers who agree

with the news sites' views and alienates those who don't, ultimately catering to what their viewers want to hear instead of giving them accurate information.

4. RESULTS

4.1 Kaggle Dataset

The accuracy for the Kaggle Dataset is 99.98% on the validation set, which means the algorithm correctly classified almost all the articles it tested with. The precision, which determines the number of true positives versus false positives, was 1.0. This means that the algorithm had no false positives — everything it labeled as fake news was, in fact, fake. The recall score tests whether the algorithm caught every piece of fake news or if some of it was mislabelled as true. In this case, the recall score was 0.9997, meaning the algorithm caught virtually every piece of fake news. Lastly, the F1 score, which is essentially a combination of the precision and recall, was 0.9998.

4.2 LIAR Dataset

When testing on the LIAR dataset, the accuracy was only ~61%. Although this may seem like a much lower score compared to the first dataset tested, there are some differences with the LIAR set compared to the Kaggle set. Most importantly, each piece of data from LIAR is only a sentence instead of a full article. The computer is trained to determine fake news based on an entire article, which explains its lower accuracy for classifying only a couple words. Additionally, the LIAR dataset has a much more complicated system for labeling fake and true news; it has five different categories, including "half true", "mostly true", and "completely true". These categories make it harder to label a piece of information (for example, what's the difference between "half true" and "mostly true"?). I cut down the number of labels to 2 ("true" and "false"), meaning the algorithm likely had some confusion on how to label everything. These issues show how the LIAR dataset may not be a good benchmark dataset to train/evaluate a fake news detection algorithm.

5. LIMITATIONS OF THE STUDY

This study didn't compare different methods of classification besides BERT — there could be a better way to identify the articles. An example could be using metadata (news source, date, topic) instead of the actual content of the article.

6. REFERENCES

[1] Raza, Shaina, and Chen Ding. "Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach - International Journal of Data Science and Analytics." *SpringerLink*, Springer International Publishing, 30 Jan. 2022, https://link.springer.com/article/10.1007/s41060-021-00302-z.

[2] Paialunga, Piero. "Fake News Detection with Machine Learning, Using Python." *Medium*, Towards Data Science, 14 Feb. 2022, https://towardsdatascience.com/fake-news-detection-with-machine-learning-using-python-3347d9899ad1.

[3] Glen, Stephanie, and Stephanie Glen. "Machine Learning and the Challenge of Predicting Fake News." *Data Science Central*, 18 Dec. 2021, https://www.datasciencecentral.com/fake-news-prediction-with-ml/.

[4] Mishra, Aditya. "Metrics to Evaluate Your Machine Learning Algorithm." *Medium*, Towards Data Science, 28 May 2020, https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

[5] Alammar, Jay. "A Visual Guide to Using Bert for the First Time." *A Visual Guide to Using BERT for the First Time – Jay Alammar – Visualizing Machine Learning One Concept at a Time.*, https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/.