Munchkin

Ajitesh Bansal
Weizhi Zhang
Ishwa Tushar Shah

April 7, 2023

Problem Introduction

Fake news is a growing concern for individuals and organizations alike. With the advent of social media and easy access to news on the internet, it is becoming increasingly difficult to distinguish between real and fake news.

In this project, we will classify news articles into real or fake news. The source of our news articles is the News API and news dataset from Kaggle. We plan to use two different models toclassify the articles: Passive Aggressive Classifier and Multinomial Naive Bayes. We will then compare the results of both models to determine which models performs better.

In this report, we will explain the development process of a fake news detector. The project involves:

- Importing various modules
- Creating a method to get news data from the API
- Getting news using multiple sources,
- Loading and concatenating the DataFrame
- Creating a model for detecting the fake news

Progress

The following steps have been taken to develop the fake news detector so far:

- ***Importing the Modules***

  We have imported the necessary modules, including NewsApiClient from NewsApi, CountVectorizer from sklearn.feature_extraction.text, PassiveAggressiveClassifier, MultinomialNB, amongst others.

- ***Method to Get News***

  We created a method to get the news data from the News API. To interact with the News API, an API key is required. We used the get_everything() method from the

NewsApiClient to get data and passed various parameters to the method, including sources, domains, etc.

- ***Get the News Sources***

  We got all the sources from the News API and added the ID of each source to a list. We then truncated the list to a size of 10 and got news from those sources.

- ***Get News Using Multiple Sources***

  We have used multiple sources to gather the news articles. After retrieving the news sources, we then used the getNews() method defined earlier to get the news from the API. Next, we created a new data frame using the news list and further added column headings to the data frame.

- ***Load and Concat the DataFrame***

  We loaded and concatenated the data frame using a CSV file containing both the fake news and real news to train the model.

- ***Creating the Model***

  We created a model for detecting fake news using PassiveAggressiveClassifier and MultinomialNB from the sklearn module. We then trained and tested the model using the train_test_split method and calculated the model's accuracy.

Issues and Difficulties

One of the main issues we encountered while working on the project, was with the News API. We had to use an API key to interact with the API, which made it difficult to share the code.

Additionally, we had to get all the news sources from the API and truncate the list to a size of 10 to get news from those sources. This meant that some important news sources were missing.

Another major task at hand was to maintain the correctness and quality of the self-created dataset. Data from different sources can vary in terms of quality.

Some sources may have missing or incorrect data, while others may have inconsistent or biased data. To tackle this issue, we carefully evaluated each source and cleaned the data as needed.

Another difficulty was in creating the model for detecting fake news. We had to use various modules from sklearn and had to train and test the model using the train_test_split method. Additionally, we calculated the model's accuracy using the accuracy_score and confusion_matrix methods.

Results so far

After performing the experiment, we found that the Passive Aggressive Classifier performed better than the Multinomial Naive Bayes algorithm. The accuracy of the Passive Aggressive Classifier was around 97%, while the accuracy of the Multinomial Naive Bayes algorithm was around 91%. The confusion matrix for the Passive Aggressive Classifier showed that it had a high true positive rate and a low false positive rate, indicating that it was better at correctly identifying real news articles than fake news articles. On the other hand, the confusion matrix for the Multinomial Naive Bayes algorithm showed that it had a high false positive rate, indicating that it was less accurate in identifying fake news articles. In conclusion, our experiment shows that the Passive Aggressive Classifier is a better model for classifying news articles into real or fake.

Conclusion

In conclusion, we have successfully developed a fake news detector. We imported various modules, created a method to get news data from the API, got news sources, got newsusing multiple sources, loaded and concatenated the DataFrame, and created a model for detecting fake news. While we faced some issues and difficulties during the development process, we were able to overcome them and create an effective solution for detecting fake news.