P.4: Final Report

Aaron Tsui

MSDS 453

March 16, 2025

Introduction

The rapid dissemination of information in the digital age has led to the widespread proliferation of fake news, which poses significant challenges to society, including the distortion of public opinion, manipulation of political processes, and erosion of trust in media. Automating the detection of fake news offers a promising solution to mitigate its impact. This project proposes the development of a model to classify news articles as either “real” or “fake” using the Kaggle dataset of similar name.

Data

My data source is from Kaggle, titled “Fake and Real News Dataset,” which is separated into 2 CSV files, each containing solely fake news and solely real news, respectively. There are 23,502 fake news articles and 21,417 real news articles in the dataset. Each file in the dataset has 4 features, from the title to the text to the subject and publication date of the article.

Methods

I preprocessed the data by removing unnecessary characters, then converting the text to entirely lowercase, and removing stopwords. Then I tokenized the text and applied stemming to reduce words in the dataset to their base forms. Then I will convert the cleaned data into numerical features using TF-IDF and Word2Vec.

To accurately distinguish fake news from real news using the previously mentioned Kaggle dataset, I used a combination of machine learning techniques like Logistic Regression, Naive Bayes, and XG Boost and hybrid models. Initially, both datasets (Fake News dataset and Real News dataset) were concatenated to formulate one large dataset of news text documents,

which was preprocessed as previously mentioned. Then I built a logistic regression model, with the combined dataset split between training and test sub-datasets. To evaluate model performance, key metrics like precision, recall, accuracy, weighted average accuracy, and F1-score were used. The logistic regression model performed with a 98.88% accuracy and a 99% precision, recall, and F1-score across the board. Next, Naive Bayes model was performed on the training data, resulting in a slightly worse but still quite accurate accuracy percentage of 93.43%, with a weighted average accuracy of 93%. Next, XG Boost was the utmost successful model, performing at a 99.67% accuracy, with near-perfect precision and recall.

As for the deep learning methods, I used the Long Short-Term Memory model through Pytorch to classify real vs. fake news, which performed with an accuracy of 99%. The model ran through 5 epochs, and I created a subroutine to process the dataset vocabulary size as a parameter in the model. The first epoch's loss value of .69 suggests that the model is a moderate improvement over random guessing, with following epochs having gradually decreasing loss values, suggesting that the model is performing more efficiently and accurately with each additional epoch.

This proposal is novel because it integrates production-grade machine learning techniques, especially XG Boost, with deep learning and transformer-based models, enhancing fake news detection through both statistical and semantic text representations. By enhancing gradient boosting through optimized computational speed and predictive accuracy, XG Boost becomes a powerful and widely adopted machine learning algorithm. Additionally, Long Short-Term Memory (or LSTM) models are highly effective and novel because they are designed to capture long-range dependencies in sequential text data, which poses a challenge for traditional neural networks. LSTMs are equipped with special memory cells and gating

mechanisms that allow for better retention of important information, contextualized in the dataset. Lastly, the use of advanced evaluation metrics ensures robust assessments of each model's performance.

Challenges

In the process of working on this project, I encountered several technical challenges that required careful troubleshooting. The first issue arose during the installation of XGBoost, where a persistent "ModuleNotFound" error occurred. After extensive research, I resolved this by importing the sys module and configuring it as an executable in Jupyter Notebook to ensure proper library recognition.

Subsequently, I faced an Internal Tensorflow AttributeError, which stemmed from a version incompatibility between Tensorflow and Numpy. To address this, I executed the command line as an administrator and aligned the versions of both libraries, ensuring compatibility. This, however, did not fix the issue. I ended up running the LSTM model through torch as I was not able to find an implementation solution that fixed my previous issue. Despite this, these challenges highlight the importance of version management and system configuration in overcoming technical obstacles in workflows.

Conclusion

This project successfully developed and evaluated multiple machine learning and deep learning models to classify news articles as real or fake using the designated Kaggle dataset. This logistic regression and XG Boost models demonstrated exceptional performance, with accuracies of 98.88% and 99.67% respectively, while the LSTM model with an accuracy of 99% provided additional insights into capturing long-range dependencies in text data. Despite encountering technical challenges, such as library compatibility issues, the project highlights the effectiveness of combining machine learning techniques with advanced deep learning approaches for fake news detection. This work underscores the potential of automated systems to combat misinformation and contributes to the ongoing efforts to enhance trust in digital media. Future work could expand the dataset to further improve classification accuracy and robustness.

Appendix

Logistic Regression

```
Accuracy: 98.8864%

Classification Report:
              precision    recall  f1-score   support

   Real News       0.99      0.99      0.99      4733
   Fake News       0.99      0.99      0.99      4247

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980


Confusion Matrix:
[[4681   52]
 [  48 4199]]
```

Naive Bayes

```
Naive Bayes Accuracy: 93.4298%
Naive Bayes Classification Report:
              precision    recall  f1-score   support

   Real News       0.94      0.94      0.94      4733
   Fake News       0.93      0.93      0.93      4247

    accuracy                           0.93      8980
   macro avg       0.93      0.93      0.93      8980
weighted avg       0.93      0.93      0.93      8980
```

XG Boost

```
XGBoost Accuracy: 99.6659%
XGBoost Classification Report:
              precision    recall  f1-score   support

   Real News       1.00      1.00      1.00      4733
   Fake News       1.00      1.00      1.00      4247

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

LSTM Epochs

```
Epoch 1/5, Loss: 0.6862496733665466
```