Final Project: Classifying Fake News

A Kaggle-Inspired Project

By
Eddie Dew, Ryan Malmfelt, Stephen Butters, Kyle Jeffrey, Trey Hall, Stephen Rivas

December 11th, 2024

Introduction

The internet has paved the way for spreading information like wildfire, giving rise to online news sources and social media platforms that make communicating information to the public easier. A 2024 Pew Research study found that more than 54% of US adults get the news from social media platforms and about 27% from podcasts (Aubin, Christopher St., and Jacob Liedke 2024). Half of America interprets the news from quick and easy-to-follow social media platforms meant to be addictive. This has caused widespread misinformation as news platforms have found an algorithm that generates clicks: catchy, misleading titles. Unfortunately, people engage (knowingly or unknowingly) with these misleading news posts daily. Orbanek, Stephen (2021) found that about 21%, or 1 in 5 adults, obtain their news information from influencers. This has become a problem in the present and has caused higher division amongst the US and more people accepting conspiracy theories. What if there was a way to inform people that they are clicking on a news platform spreading information? This paper attempts to make predictions on classifying whether a news article is reliable or not. This paper focuses on the scope of news content, trying to identify a real or fake news article from the get-go to help users question whether it's worth viewing the report.

Dataset Description

The dataset used for this model is sourced from Kaggle and titled "fake-and-real-news-dataset." It is presented as two files, one containing all the fake news and the other containing the *true* news. Each dataset contains the news article's title, the respective document's text, the subject, and the posting date. The posting dates range from 2016 to 2017. There are no missing values from either dataset. To keep track of the *true* values, an additional column was made that denoted *true* news as a one and fake news as a 0. Combining this dataset forms a total of 44,898 news articles, with 48% of the data being true and 52% being fake, making it relatively balanced. The most common news subject was politics.

Exploratory Data Analysis

A column inspection confirmed the data set consisted of 5 features, ‘title’, ‘text’, ‘subject’, ‘date’ and the newly added column ‘True/Fake’. Checking for any null values and getting the data type was the first step, what was initially explored was the ‘title’ column. The inquiry about investigating the title was what words were typically associated with real and fake news. This was not as revealing as anticipated so to get more in depth the ‘title’, ‘text’ and ‘subject’ were combined into one feature called ‘content’ for the sake of further analysis. This allows the TF-IDF vector to perform a more in-depth analysis. The text being combined led to the problem that multiple stop words such as ‘the’ were appealing as the most influential , to combat this common English terms like ‘the’ were removed to avoid skewing the results. Figure 1 shows what words appeared most frequently in real and fake news titles.

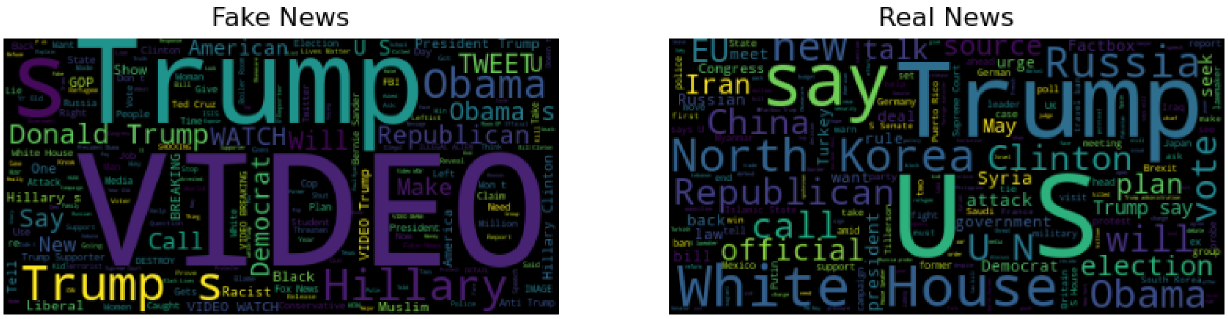


Figure 1: Word Cloud for Fake and Real News Titles

What’s interesting is that Trump (indicating Donald Trump) is involved a lot in this dataset as he’s referred to a lot in the titles for real and fake news. It also appears titles with the word video tend to be fake. Maybe indicating video evidence in the title of an article is not recommended?

It’s hard to conclude, but it is still an interesting observation. The word cloud consisted of mainly political based news, given the dataset is 2016 - 2017 it's safe to assume it's being pulled right after the 2016 election. With this information it could potentially explain why you see words like “Trump” or ‘president’ appearing on both sides, just from the sheer amount they are mentioned in the media. A potential drawback of the analysis is the lack of transparency regarding the sources. An example could be, fake news articles mentioning "president" could artificially inflate the perceived importance of certain terms, ultimately skewing results.

Tf idf Vectorizer

The purpose of TF-IDF is to highlight important words within a specific document that are less common across all documents in a corpus. The process begins by calculating frequency (TF), which measures how frequently a word appears in a single document. It then adjusts this value using inverse document frequency (IDF), which decreases the weight of words that occur frequently across many documents. In the context of this dataset, TF-IDF will examine the words in the ‘content’ column, identifying commonly used terms such as ‘Trump’, ‘North Korea' , and ‘Video’ (as shown in Figure 1). However, words that are less frequently used across the dataset, such as ‘attack’, will be assigned a higher TF-IDF score, indicating their relative importance within individual comments. When this is fit to the ‘content’ column it returns a sparse matrix where each word from each row is scored and a total of 5,000 words are used. By looping through the iterations and summing up the scores, we can find the terms most frequently used in this vectorizer and their score. Figure 2 shows the top 5 words with their summed score.

Word	TF-IDF Score
Trump	3020.421
Said	1979.741
President	1146.776
Clinton	958.106
Obama	928.913

Figure 2: Top 5 words based on summed TF-IDF Score

Depending on how their scores were for each row will be how they are valued. For example, if the first row is a real news article and 'Trump' is valued the highest, the model will learn to associate 'Trump' with real news.

Model

Two machine learning models, Logistic Regression and Naive Bayes, were used to predict real and fake news. Both models are well-suited for balanced datasets, which implies they should yield reliable predictions. Logistic Regression, in particular, works by fitting an S-curve to the data while applying L2 regularization. This regularization helps mitigate bias by shrinking variable weights, thus improving the generalizability of the model. Naive Bayes works by finding the probability of each class occurring and picks the highest probability. Both models worked exceptionally well with Logistic Regression having a 99% accuracy and Naive Bayes having a 94% accuracy. An ROC curve can be used to convey how well these predictions are leaning towards a correct prediction. The curve does this by plotting the True-Positive Rate against the False-Positive Rate, and in this case the AUC value, which is at 1 for Logistic Regression and 0.98 for Naive Bayes shown in Figure 3. With the AUC being at such a high value, it is able to tell us that the model is able to distinguish real and fake news with a high success rate with minimal errors. And the curve is closer to the top left corner indicates that the model has a strong performance even while having a few false positives, but overall successful.

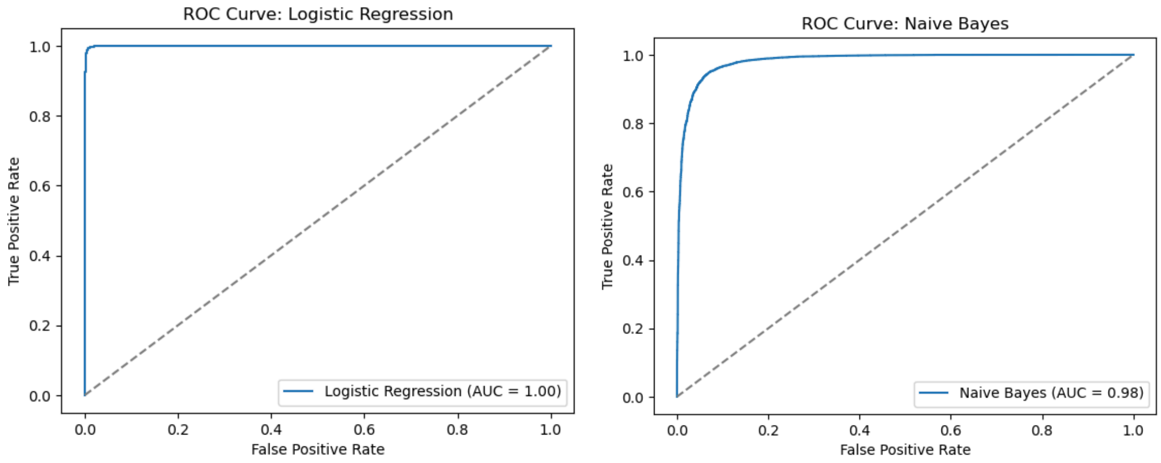
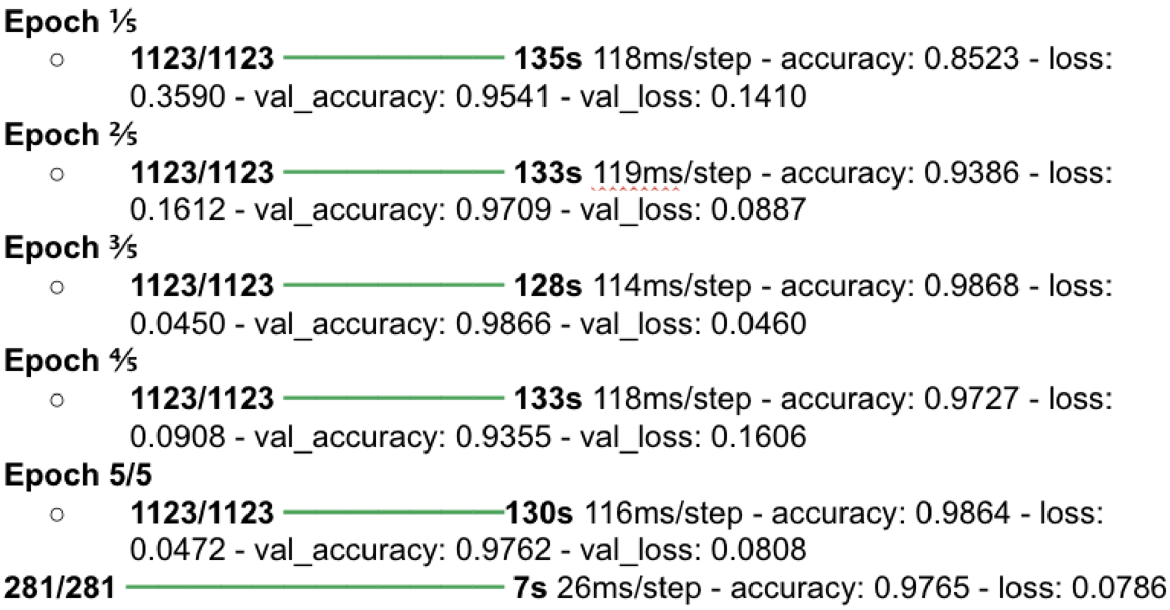


Figure 3: ROC Curve for Logistic and Naive Bayes Models

The next model tested was a deep learning model called Long-Short Term Memory (LSTM), designed to work with sequential data such as text. How it handles the sequential data is by storing past points in memory as it loops through the input, which can increase computational cost. The model was trained and evaluated across 5 epochs, using accuracy, log loss, validation accuracy, and validation loss as the metrics evaluated. Training accuracy measures how well the model is performing based on the datasets it learned from. Validation accuracy indicates how well the model operates when using unseen data. Training loss represents how well the model is optimizing its weights in order to minimize any failures or errors that could occur. Lastly, validation loss is the errors between the model's prediction and true labels on the validation dataset. Each epoch is one complete pass through the entire training dataset, allotting the model to learn and improve its predictions after each one. 5 epochs or “rotations” were used because too little can lead to underfitting while too many can lead to overfitting.

Figure 4 shows the training loop results for the LSTM model.



```
Epoch ⅕
    ○      1123/1123 ━━━━━━━━━━━━━━━━━━━━ 135s 118ms/step - accuracy: 0.8523 - loss:
           0.3590 - val_accuracy: 0.9541 - val_loss: 0.1410
Epoch ⅖
    ○      1123/1123 ━━━━━━━━━━━━━━━━━━━━ 133s 119ms/step - accuracy: 0.9386 - loss:
           0.1612 - val_accuracy: 0.9709 - val_loss: 0.0887
Epoch ⅗
    ○      1123/1123 ━━━━━━━━━━━━━━━━━━━━ 128s 114ms/step - accuracy: 0.9868 - loss:
           0.0450 - val_accuracy: 0.9866 - val_loss: 0.0460
Epoch ⅘
    ○      1123/1123 ━━━━━━━━━━━━━━━━━━━━ 133s 118ms/step - accuracy: 0.9727 - loss:
           0.0908 - val_accuracy: 0.9355 - val_loss: 0.1606
Epoch 5/5
    ○      1123/1123 ━━━━━━━━━━━━━━━━━━━━130s 116ms/step - accuracy: 0.9864 - loss:
           0.0472 - val_accuracy: 0.9762 - val_loss: 0.0808
281/281 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7s 26ms/step - accuracy: 0.9765 - loss: 0.0786
```

Figure 4: LSTM Results across Epochs

During the first epoch, the training accuracy starts at 0.8523, not bad. The second epoch saw an increase to 0.9386, showing the running through the dataset again the model is slowly learning and improving overall. The model kept improving and by the 5th epoch, both the accuracy and the valuation accuracy increased to 0.9864 and 0.9355 respectively, with the accuracy being only 0.0004 difference from epoch 3. The final results accuracy was 0.9765 with a loss of 0.0786. Overall this model is working as intended and is very successful in determining which are real and which are fake depending on many factors. The model also had a 0.98 AUC shown in Figure 5.

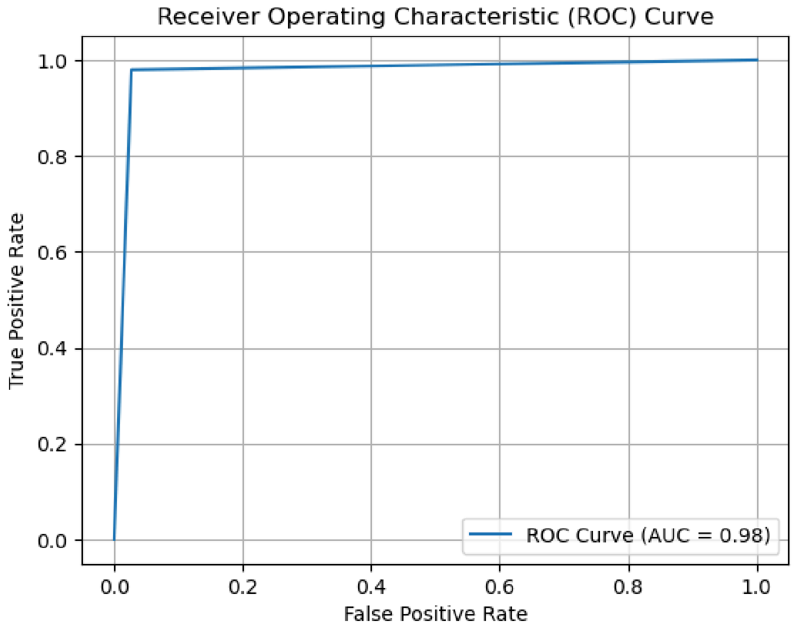


Figure 5: ROC Curve for LSTM Model

Overall, all of the models performed well with Logistic Regression being the best performing model. Its simplicity with a clean dataset proved trivial at making predictions compared to Naive Bayes and LSTM. What could've impeded the LSTM model was the tf-idf tokenizer creating unwanted noise for a sequential model so to prevent this in the future, it would be recommended to create word embeddings to give context to that word by looking at the words to the left and right of it.

Storytelling/Impact

Through this project, we gained several valuable insights into the power and limitations of machine learning models for classifying real versus fake news. One of the key takeaways is the importance of preprocessing steps, such as handling missing data and applying TF-IDF vectorization, to prepare the text data for machine learning models. By using these techniques, we were able to extract meaningful features that allowed the models to effectively differentiate between true and fake articles. Another important insight was the significant difference in performance between the Logistic Regression and Naive Bayes classifiers. While both models performed well, Logistic Regression was slightly more accurate, which underscores the importance of selecting the right model for the task at hand.

In terms of answering our initial problem, we successfully developed a machine learning model capable of distinguishing between real and fake news articles with high accuracy. This model achieved a very impressive 99.26% accuracy using Logistic Regression, which shows that it can be a valuable tool in identifying misinformation. However, we also realized that while accuracy is important, other evaluation metrics like precision, recall, and F1-score should be taken into consideration to ensure the model's overall effectiveness, especially in real-world applications where false positives and negatives can have serious consequences. What is limited by this model is it doesn't have news from more popular social media platforms e.g. Instagram, Tik Tok and some big news platforms such as the New York Times are not part of this dataset.

The impact of this project is significant both socially and ethically, as it addresses the growing problem of misinformation in society. Fake news has the potential to influence public opinion, shape elections, and even cause harm by spreading false or misleading information. By creating a model capable of distinguishing between real and fake news articles, this project can help mitigate the spread of false information, ensuring that individuals have access to more reliable, trustworthy sources. This could enhance public awareness and promote a more informed citizenry, which is crucial for a functioning democracy.

However, there are ethical considerations to be aware of in the development and deployment of such a model. One potential risk is the reliance on machine learning models to make decisions about what constitutes "truth" or "fake." These models can only be as good as the data they are trained on, and if the data is biased or incomplete, the model could reinforce existing biases or overlook important context. For example, certain political or social perspectives could be underrepresented, leading the model to label articles from these viewpoints as "fake" unfairly. Additionally, using such models to flag news articles as fake could raise concerns about censorship or the suppression of free speech if the model is not transparent or accountable in its decision-making process. Another potential negative impact involves the misuse of the model. If placed in the hands of those who wish to manipulate information, this tool could be used to discredit legitimate sources and promote biased narratives. It’s important to consider safeguards to ensure that the technology is used responsibly and ethically, and that the outcomes are regularly audited for fairness. Ultimately, while the project has the potential to make a positive societal impact by combating misinformation, it is crucial to balance these benefits with the potential risks.

References

Aubin, Christopher St., and Jacob Liedke. “News Platform Fact Sheet.” *Pew Research Center*, Pew Research Center, 17 Sept. 2024, www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/.

Clmentbisaillon. “Fake-and-Real-News-Dataset.” *Kaggle*, 19 Apr. 2024, www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/data.

Orbanek, Stephen. “Study Shows Verified Users Are among Biggest Culprits When It Comes to Sharing Fake News.” *Temple Now*, Temple University, 10 Nov. 2021, news.temple.edu/news/2021-11-09/study-shows-verified-users-are-among-biggest-culprits-when-it-comes-sharing-fake.