1.Introduction

In this analysis report, we built a fake news detection data product using BERT, GPT-2, and LSTM models. Our goal was to demonstrate the effectiveness of these models by employing a Global Surrogate approach: training a logistic regression model using the predicted labels from our models on the training dataset. We further enhanced the interpretability of these black-box models using LIME and SHAP techniques. We tested our models using a news repository (https://www.kaggle.com/datasets/sonalgarg174/ifnd-dataset) containing news categories and timestamps. We subsequently compared the performance of the predicted labels across different time periods and news categories.

2.Methodology

We used the dataset with pred_labels generated by BERT, GPT-2, and LSTM models to train and evaluate a logistic regression model for fake news detection. We applied LIME and SHAP techniques to interpret the model. Here's a brief explanation of each step:

Logistic Regression Model:

Load the dataset and preprocess it to create binary labels.
Split the dataset into train and test sets, and vectorize the text using TfidfVectorizer.
Train a logistic regression model and evaluate its performance using the classification report.
Use the logistic regression model as a surrogate model for the black-box models (BERT, GPT-2, and LSTM) to improve interpretability.
Interpretability Techniques (LIME and SHAP):

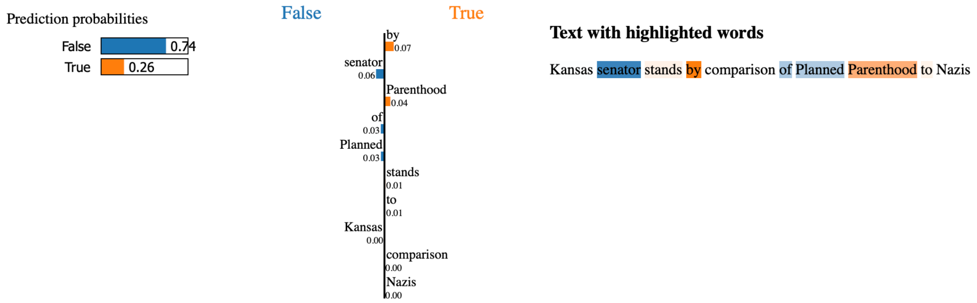
LIME:
Define a function to predict probabilities for the LIME explainer.
Create a LimeTextExplainer object and choose an instance from the test set to explain.
Use the LIME explainer to generate an explanation for the chosen instance.
Visualize the LIME explanation to understand the impact of each feature (word) on the prediction.
SHAP:
Create a SHAP explainer using the logistic regression model and calculate SHAP values for the test set.
Visualize the SHAP values using a beeswarm plot to understand the contribution of each feature (word) to the predictions.
Testing the models using the news repository:

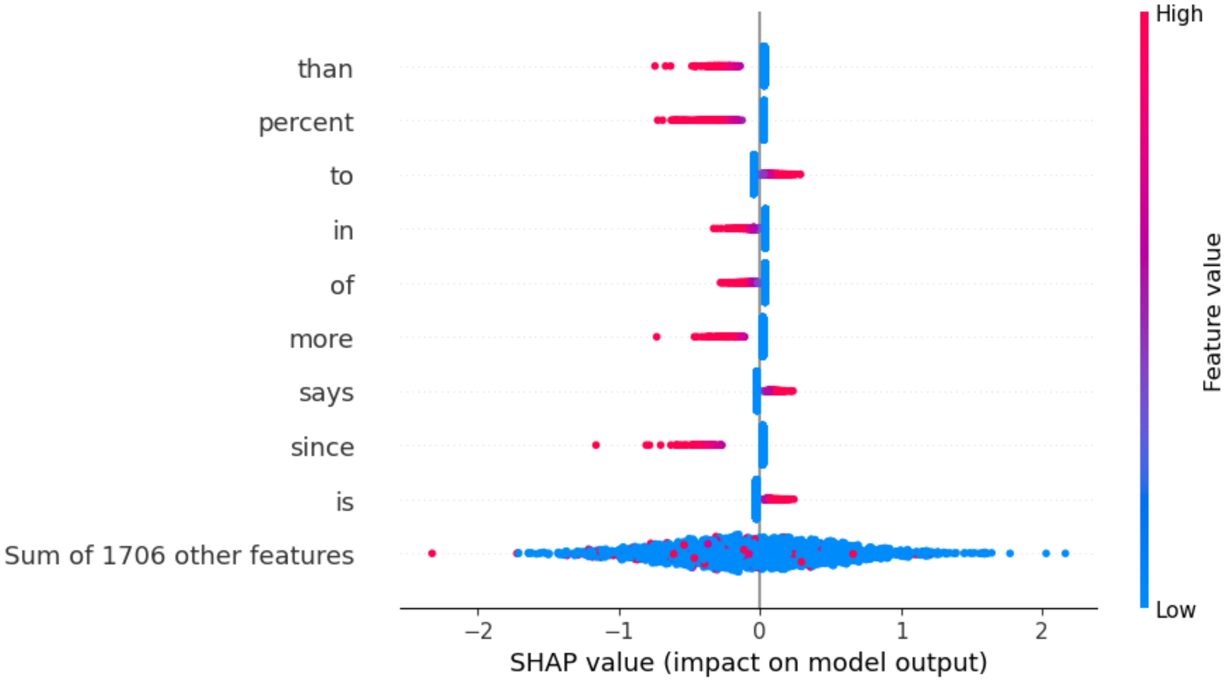
Extract unique categories from the dataset and filter the data based on each category.
Create a contingency table between true and predicted labels.
Perform a chi-squared test on this table to assess the model's performance across categories, and print the test results, including the chi-squared statistic, p-value, degrees of freedom, and expected frequencies.
Comparing performance across different time periods and news categories:

Convert the 'Date' column to a datetime object and filter the data to include only dates after 2018.
Remove rows where both 'Label' and 'Pred_Label' are equal to 0.
Group the data by date, label, and pred_label, counting the number of articles.
Create a pivot table with 'Date' as the index and 'Label' and 'Pred_Label' as columns.
Plot a bar chart comparing 'Label' and 'Pred_Label' for each date to visualize the model's performance across different time periods and categories.

3.Results

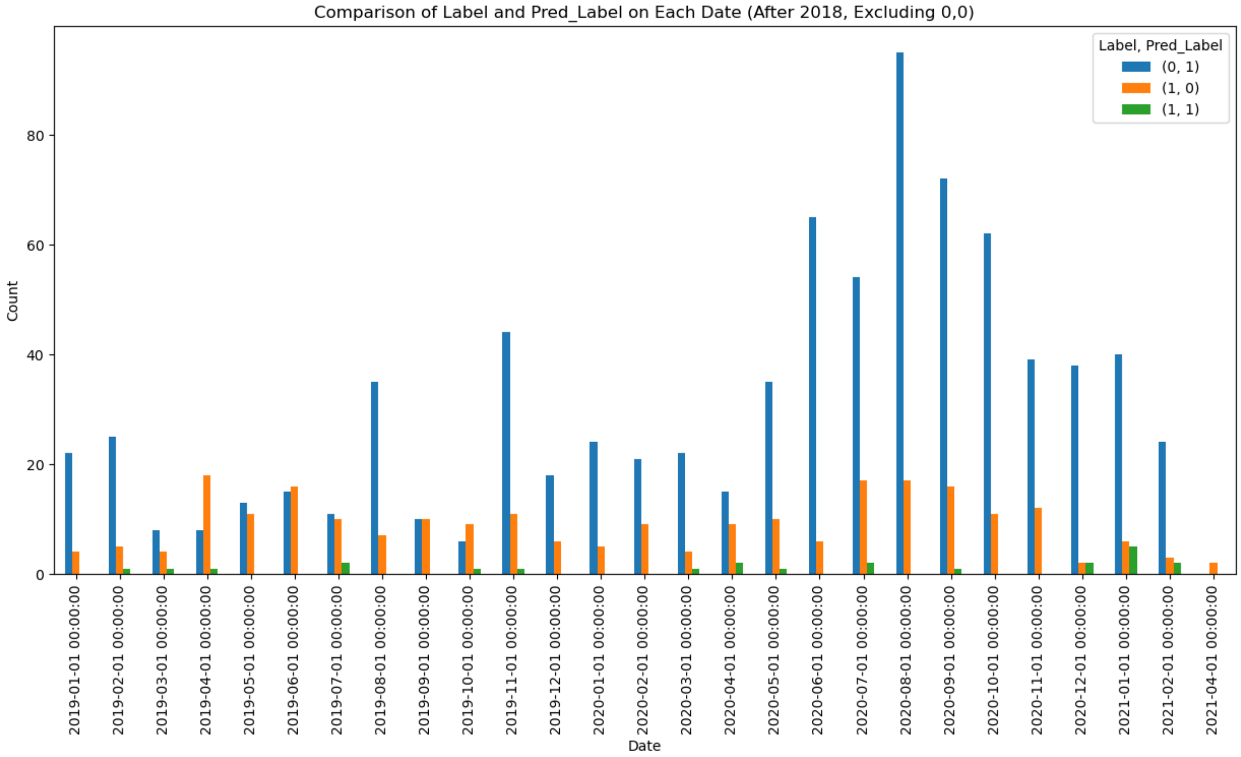
Interpretability Results: LIME and SHAP results provide insights into the most important features contributing to the predictions. LIME produces instance-by-instance explanations, which may seem random and harder to interpret. In contrast, SHAP offers a comprehensive evaluation, revealing features like "video" positively affect predictions, possibly suggesting that visual evidence makes news more reliable. Features like "http" and "says" negatively affect predictions, indicating online rumors and hearsay might be prevalent in these cases. Analyzing these explanations helps us better understand the factors considered by the models when detecting fake news.





Prediction Performance Comparison:
Comparing prediction performance across different news categories:
Results show significant associations for COVID-19, VIOLENCE, ELECTION, GOVERNMENT, POLITICS, and TRAD categories, suggesting that the model's performance varies across these

categories. For categories with no significant association (TERROR, MISLEADING, and MISLEADIND), it's important to note that they seem to have insufficient data, as there is only one label present in each of these categories. This lack of data might be affecting the reliability of the test in these cases.



Comparing prediction performance across time periods:
The analysis focuses on data from 2018 onwards, excluding instances where both the true label and predicted label are 0 (i.e., true positives). The observed patterns show that the number of (1,0) cases, where the news is fake but predicted as true, is much larger than the number of (1,1) cases, where the news is correctly identified as fake. This suggests that the model's generalization performance for predicting fake news is limited and could be improved.

4.Conclusion

In this analysis, we demonstrated the effectiveness of BERT, GPT-2, and LSTM models for fake news detection. We employed a logistic regression model as a global surrogate to improve the interpretability of these black-box models. Additionally, we utilized LIME and SHAP techniques to enhance our understanding of the factors that influence the predictions.

The results show that the model's performance varies across different news categories, with some categories displaying a significant association. In contrast, categories with insufficient data did not exhibit a clear relationship. When comparing the prediction performance across time periods, it was observed that the model's generalization ability in predicting fake news could be improved.

This analysis has provided valuable insights into the performance of our fake news detection data product. However, there are limitations to the current approach, including the reliance on a surrogate model and the varying performance across news categories and time periods. Future work could focus on improving the generalization capabilities of the models, exploring alternative interpretability techniques, and investigating the impact of data quantity and quality on the performance of these models. Additionally, it would be beneficial to explore other features or metadata that could further enhance the fake news detection process.