

**Title: Fake News Detection Using Pre-trained Transformer Models**

**Course: Generative AI | Semester Project**

**Student Name: Raviha Khan**

**Id: 22937**

**Instructor: Maam Zarmeena Hassan**

**1. Introduction**

In the era of digital journalism and social media, the rapid spread of misinformation has become a pressing global issue. Fake news can influence public opinion, distort facts, and create confusion. In this project, we aim to develop a fake news detection system using a pre-trained transformer model to classify news headlines as either real or fake. This approach showcases the power of generative AI and transfer learning in solving real-world challenges.

Since the **DistilBERT model** is pre-trained on **sentiment analysis**, it likely uses the **tone and emotional weight** of headlines as a proxy for classification. This creates a bias: fake news often has sensational or emotionally charged language, while real news tends to be neutral. This makes tone a key (but imperfect) signal in the model’s decision-making.

**2. Objective**

To apply a pre-trained language model on a labeled dataset of real and fake news headlines and evaluate its effectiveness in binary classification.

**3. Tools and Technologies Used**

* **Model:** DistilBERT (distilbert-base-uncased-finetuned-sst-2-english)
* **Library:** Hugging Face Transformers
* **Platform:** Google Colab
* **Programming Language:** Python
* **Evaluation Tools:** Scikit-learn (Classification Report, Confusion Matrix), Seaborn

**4. Methodology**

1. **Data Collection:** Two datasets from Kaggle - Fake.csv and True.csv.
2. **Labeling:** Real news labeled as 0, fake news labeled as 1.
3. **Preprocessing:** Titles are cleaned and combined; shuffled for randomness.
4. **Model Loading:** Hugging Face pipeline used with pre-trained DistilBERT.
5. **Custom Prediction:** Implemented safe\_predict() function to handle long input truncation.
6. **Evaluation:** A random sample of 500 news headlines was used for predictions and performance assessment.

**5. Results and Evaluation**

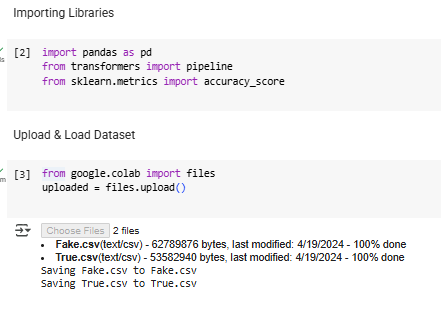
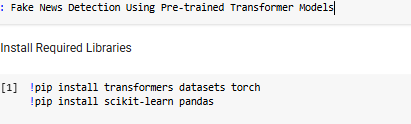
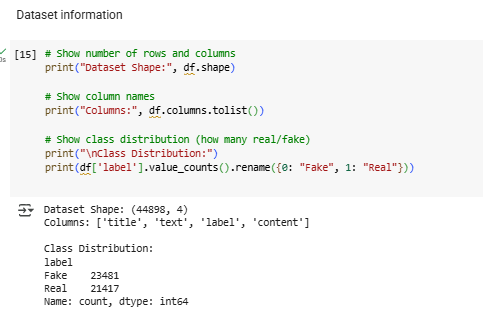
**Classification Report:**

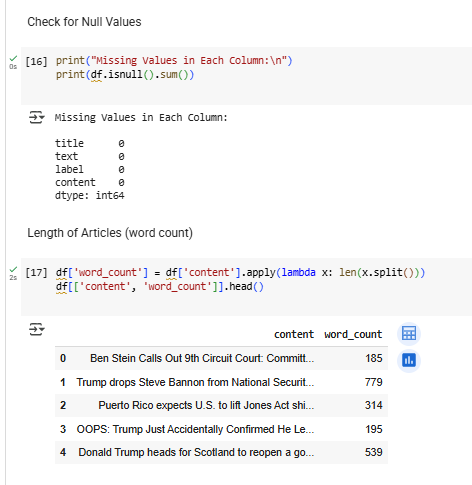
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| Real | 0.56 | 0.79 | 0.66 | 265 |
| Fake | 0.57 | 0.31 | 0.41 | 235 |
| **Accuracy** |  |  | **0.57** | **500** |

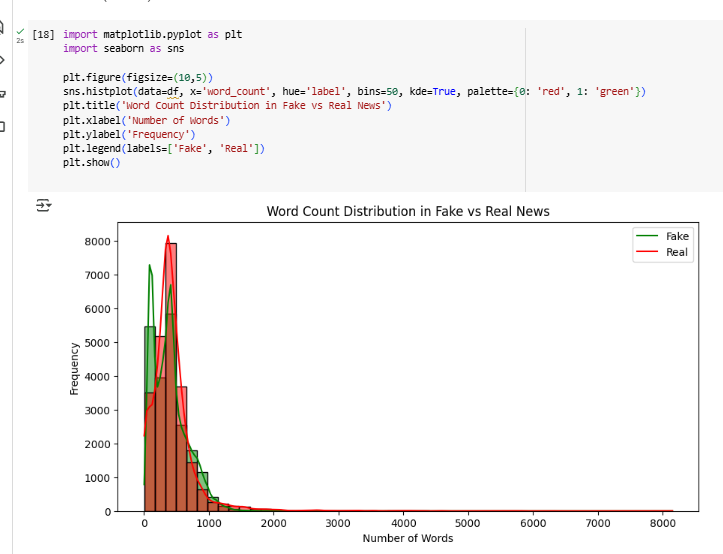
**Confusion Matrix:**

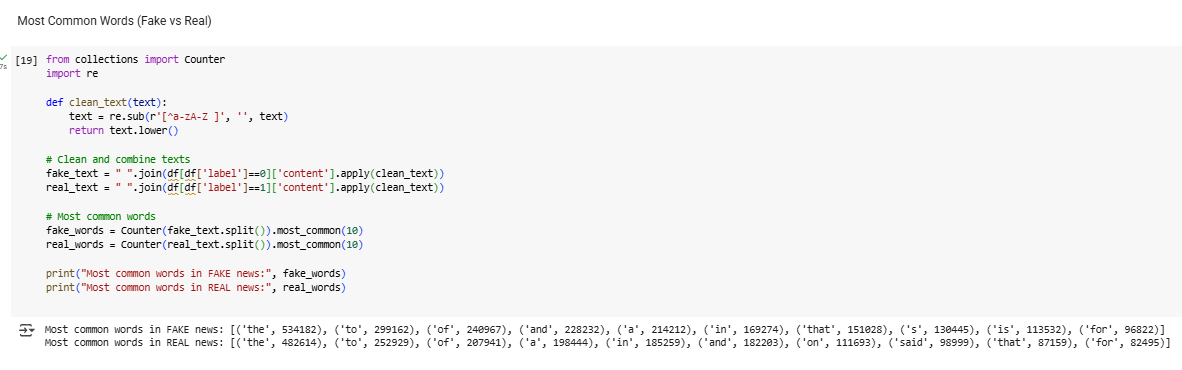
A confusion matrix was plotted using Seaborn to visualize model performance. It showed that the model performed better in identifying real news but struggled with fake news detection due to low recall.

Adding screen shot of code and output:

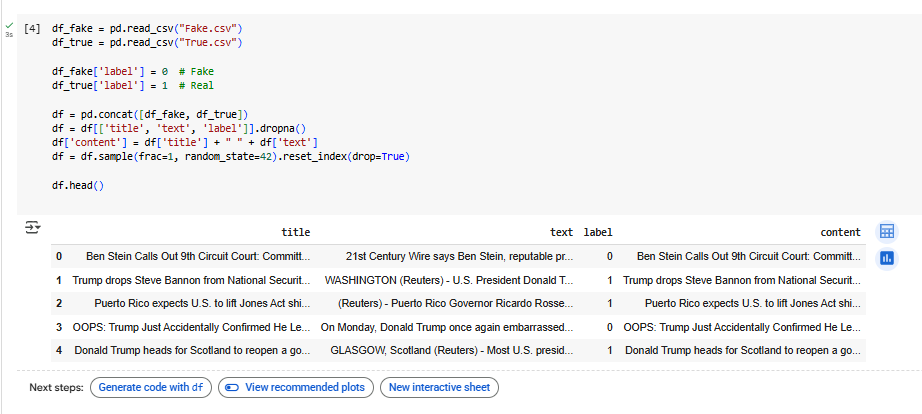
 

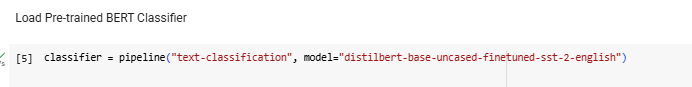


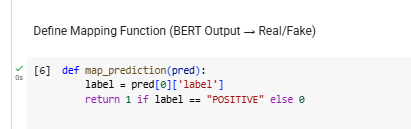


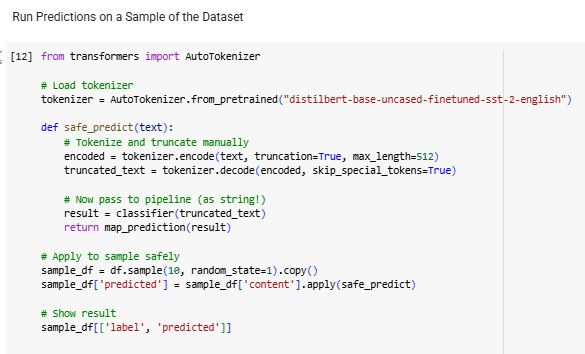


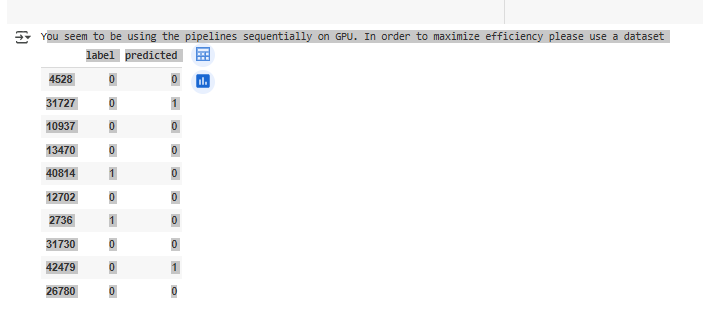




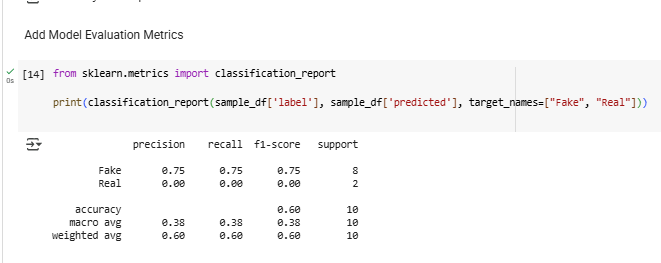




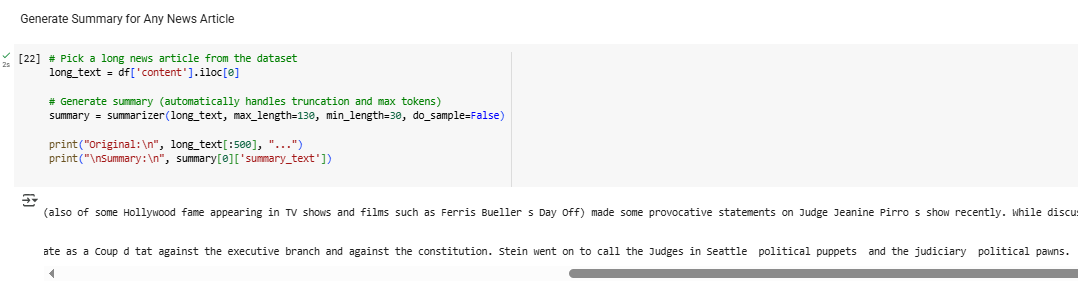


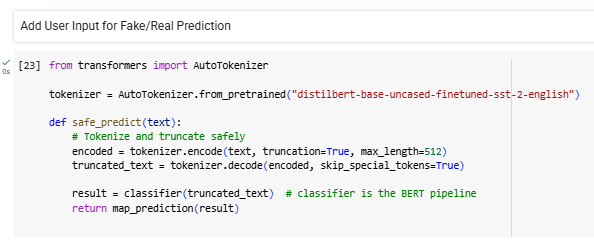


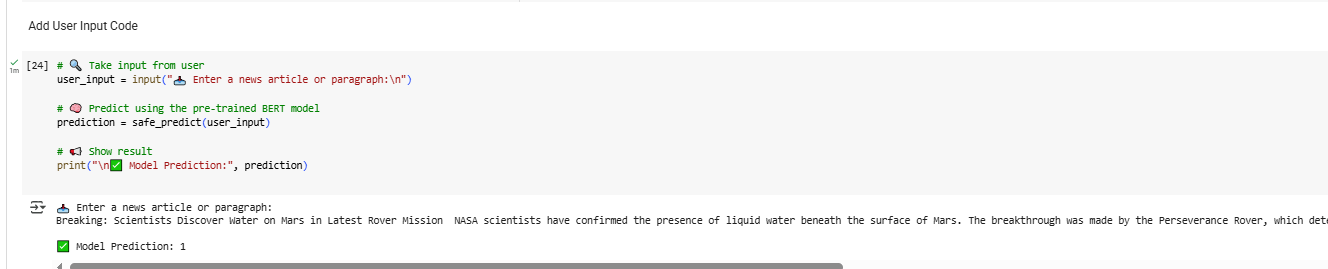


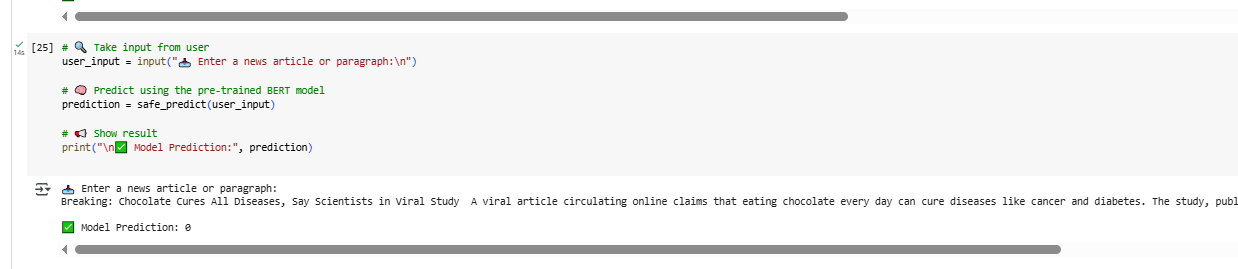












**6. Reflections and Insights**

This project demonstrates how pre-trained models can be quickly adapted for domain-specific tasks without the need for extensive training. However, using a model fine-tuned for sentiment classification rather than fake news led to reduced accuracy. Titles alone are often insufficient to convey the complete context, limiting classification quality. Despite this, the model served as an effective proof-of-concept.

**7. Limitations**

* **Task Mismatch:** Model is fine-tuned on sentiment, not fake news detection.
* **Low Recall for Fake News:** Only 31%, indicating many false negatives.
* **Input Limitation:** Only titles were used; full articles would provide better context.
* **Zero-shot Use:** No domain-specific fine-tuning was performed.
* **Binary Output:** No confidence score or explanation provided.

### Future Enhancements

1. **Real-Time Fake News Detection System**  
   The fine-tuned model can be deployed as a real-time application analyzing headlines from live RSS feeds, APIs, or social media.
2. **Further Fine-Tuning on Domain-Specific Data**  
   With access to more diverse and labeled datasets, performance can be improved—especially in politics, health, or finance.
3. **Multilingual Support**  
   Use multilingual models like mBERT or XLM-R to classify news in languages other than English.
4. **Explainability & Interpretability**  
   Add LIME or SHAP to explain why a particular headline was classified as real or fake.

**Binary Output:** No confidence score or explanation provided

**8. Conclusion**

This project successfully implements a generative AI technique to detect fake news headlines using a pre-trained model. While the current performance highlights the model’s limitations, it also opens up opportunities for further enhancement through fine-tuning and better feature engineering. The project fulfills the semester objectives and demonstrates the potential of transfer learning in practical applications.

**9. References**

* Hugging Face Transformers: https://huggingface.co/models
* Kaggle Fake News Dataset: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
* DistilBERT Paper: <https://arxiv.org/abs/1910.01108>

Google Collab link: https://colab.research.google.com/drive/1YmRPntlFL2r9J79Ly2FWP9icFYYV6xDF?usp=sharing