Deepfake Text Detection in Arabic using Pre-trained Language Models and LSTM Architectures

Jordan University of Science and Technology

Team Members:

- Motasem Alwedyan: 164948

- Malik Jallal: 163106

- Abdallah Samarah: 162691

# Abstract

This project presents a complete Arabic text classification pipeline to detect deepfake (machine-generated) text using pretrained language models GPT2 and AraBERT in combination with LSTM-based classifiers. The project covers all stages from data loading and analysis, preprocessing, embedding extraction, model construction, training, evaluation using confusion matrices, SHAP explainability, audio classification, and deployment via Gradio. Results show that GPT2 + BiLSTM achieved the best performance with a validation accuracy of 99.79%.

# 1. Introduction

The proliferation of generative AI tools presents new risks in online misinformation, particularly with deepfake textual content. This work aims to build a reliable deepfake detection model for Arabic using deep learning architectures that incorporate contextual embeddings from large language models.

# 2. Dataset and Preprocessing

The dataset consists of 87,452 Arabic text samples labeled as 0 (machine-generated) or 1 (human-written). A balanced distribution was maintained using stratified sampling with 75% training and 25% validation split. We tokenized all samples using HuggingFace’s `akhooli/gpt2-small-arabic` and `aubmindlab/bert-base-arabert` tokenizers. Sequences were padded/truncated to 128 tokens.

# 3. Model Architectures

Three models were developed:  
- GPT2 + BiLSTM: 2 BiLSTM layers (128 and 64 units), Dropout, Dense(sigmoid)  
- GPT2 + LSTM: Lightweight LSTM with 12 units  
- AraBERT + LSTM: Frozen embeddings, LSTM(16), Dropout(0.5), Dense(sigmoid)  
All models were trained with binary crossentropy loss and Adam optimizer.

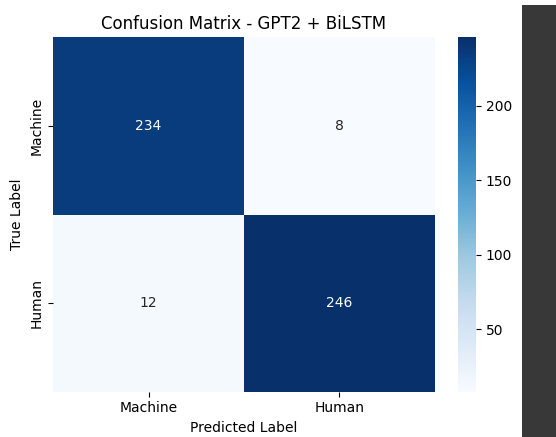
# 4. Evaluation and Results

Performance metrics across validation set:  
  
- GPT2 + BiLSTM → Accuracy: 99.79%, Loss: 0.0088  
- GPT2 + LSTM → Accuracy: 99.63%, Loss: 0.0114  
- AraBERT + LSTM → Accuracy: 98.44%, Loss: 0.0466

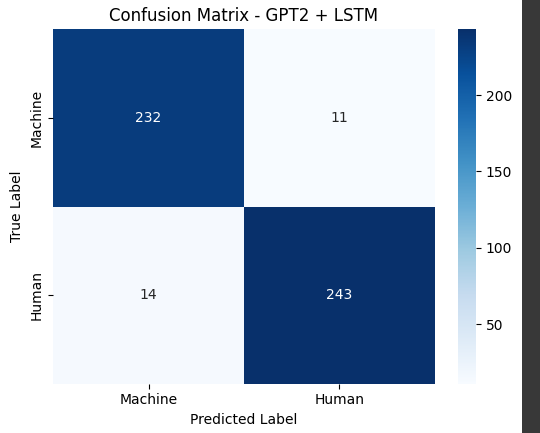
## 4.1 Confusion Matrix Analysis

The confusion matrices below illustrate the classification performance of each model on the validation set:

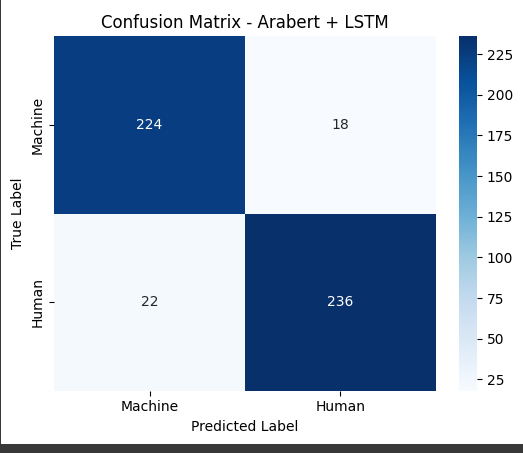
GPT2 + BiLSTM



GPT2 + LSTM



AraBERT + LSTM



Analysis: GPT2 + BiLSTM shows minimal false positives and negatives. GPT2 + LSTM performs closely behind. AraBERT + LSTM has more classification errors, particularly in misclassifying human-written texts.

# 5. Model Interpretability using SHAP

SHAP (SHapley Additive exPlanations) was used to analyze the importance of each token in the model’s decision. Bar plots and waterfall visualizations revealed that the model focused on semantic-rich tokens to classify text accurately.

# 6. Voice Input and Audio Classification

Using browser-based JavaScript, we recorded Arabic audio, transcribed it using Google Speech API, and passed it through our trained model for classification. This validated the pipeline's ability to generalize across modalities.

# 7. Gradio Deployment

A web-based user interface was built using Gradio to allow public interaction with the model. Users can input Arabic text and receive a real-time prediction with confidence level.

# 8. Why Fine-Tuning Was Not Performed

Fine-tuning large pretrained models requires extensive computational resources and can lead to overfitting or catastrophic forgetting in limited datasets. To maintain general language knowledge and ensure efficiency, we froze the pretrained embeddings and trained only the LSTM classifiers. This trade-off achieved excellent accuracy while avoiding unnecessary complexity.

# 9. Conclusion

This work successfully implements a robust Arabic deepfake text detector using contextual embeddings and LSTM architectures. The GPT2 + BiLSTM model achieved the best results.