Task: LLM – Detect AI Generated Text Team ID: 81 Team Member: 沈韋翔 林士翔 張瑋倫 莊翔凱

Detecting AI-Generated Text Using LLMs

1. Introduction

With the rapid advancement of AI, modern large language models (LLMs) can already generate articles that are very similar to those written by humans. However, such technological advances also raise concerns in the education and academic fields. For schools, students may rely on AI-generated articles instead of developing their own opinions, affecting their learning foundation. In the academic field, this may lead to an increase in plagiarism problems. The purpose of this project is to design a model that can accurately distinguish whether an article is generated by LLM or written by a middle school student, and is committed to improving the accuracy of classification.

2. Methodology

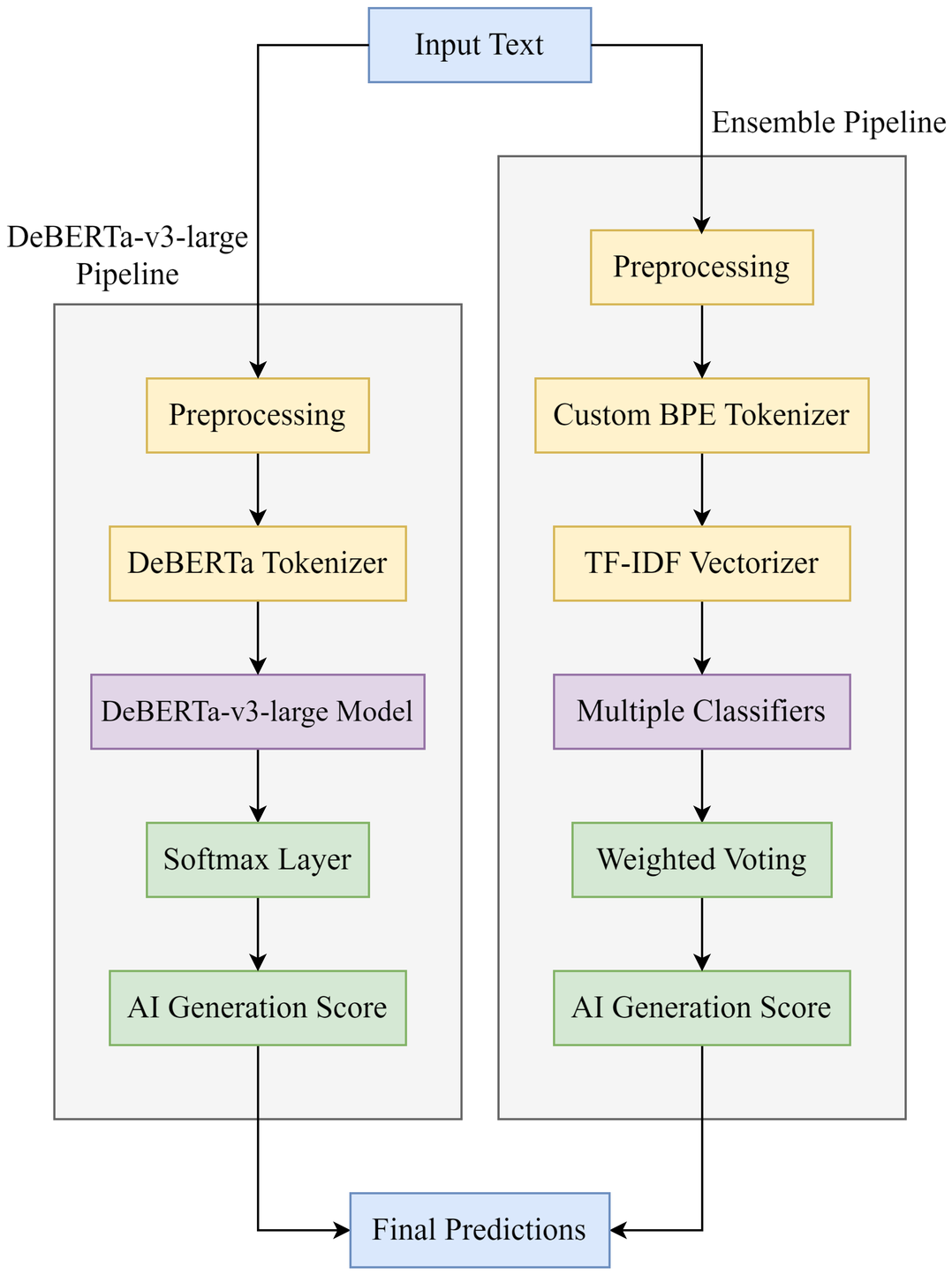

Figure 1. AI-Generated Text Detection System Architecture

DeBERTa-v3-large Pipeline: In this pipeline, inputs are passed to a DeBERTa-v3-large model, which is fine-tuned on a dataset combining generated texts from SlimPajama and high-quality human essays from Persuade 2.0. The model was trained as a binary classifier to distinguish between human and AI-generated text. The final prediction probability is obtained through a softmax layer.

Ensemble Pipeline: The pipeline combines multiple machine learning classifiers. Text inputs are preprocessed, tokenized using a custom BPE tokenizer, and then converted to TF-IDF features using 3-5 character n-grams. These features are analyzed by four classifiers: MultinomialNB, SGDClassifier, LGBMClassifier, and CatBoostClassifier. The outputs are combined through weighted voting for prediction.

3. Experiments and Results

Evaluation Metrics: Private score uses AUC-ROC as the metric to assess detection performance. The score explains the capacity of the model in differentiating human texts from those generated by AI tools looking as if it was human written, where a score approaching 1.0 implies better integration.

Table 1: Performance comparison of different detection methods.

Model / Method	Private Score (AUC)
Fine-tuned DeBERTa	0.965134
Ensemble Only	0.872208
Hybrid (Threshold-Based: 30-70)	0.967799

4. Discussion

The DeBERTa model provides a strong baseline (score 0.965), but the hybrid approach, which combines DeBERTa with an ensemble, demonstrates the power of uncertainty-driven modeling (score 0.967). By using DeBERTa's confidence to guide ensemble re-evaluation, we address the limitations of relying solely on a single model or a TF-IDF based model approach. This approach targeted a subset of data where a more nuanced approach was needed to classify the text.

However, the TF-IDF approach lacks the ability to capture contextual meanings and some language patterns critical for distinguishing ai-generated text from human writing. We could explore more advanced tokenization and contextualized word embeddings (e.g. BERT, RoBERTa) for richer semantic and syntactic representations.

5. Conclusion and Future Work

This project focuses on detecting the AI-generated text, which combines benefits of DeBERTa model with ensemble using uncertainty quantification techniques. The hybrid approach suggests a more robust way of detection.

Future works will focus on optimizing the confidence threshold in the hybrid model, and exploring alternative ensembling methods. Additional directions may include:

- Leveraging perplexity scores from multiple LLMs as complementary features
- Combine BPE Tokenization with Contextual Embeddings

6. References and Acknowledgments

Kaggle Competition: “**LLM - Detect AI Generated Text**” https://www.kaggle.com/competitions/llm-detect-ai-generated-text/overview
YEVHENII MASLOV: “[LLM-DAIG] 3rd place solution” https://www.kaggle.com/code/evgeniimaslov2/llm-daig-3rd-place-solution
HAO MEI: “DAIGT-DeBERTa” https://www.kaggle.com/code/tailen/daigt-deberta