LLM Detection

Machine Learning Approach

Department of Mathematics and Compute Science

Team members:

Mobin Nesari

Parmida Jabari

Supervisors:

Hadi Farahani

Shide Sharif

Niloofar Shabani

January 30, 2024

Abstract

In this project, we present an implementation of LLM detection using the DeBERTa model. The objective is to develop a model capable of detecting AI-generated text, which has a role in the field of natural language processing. This report starts by describing the dataset used and then continues by describing the DeBERTa architecture. Then, we introduce our custom model implementation for the specific requirements of LLM detection.

Our model shows promising outcomes in the task of identifying AI-generated text instances. We used the evaluation metrics and showcase the model's efficiency in distinguishing between real and AI-generated content. The findings highlight the effectiveness of the DeBERTa-based model in addressing the challenges in AI-generated text. Overall, this project uses the advancement of LLM detection techniques, showing the capabilities of DeBERTa in handling this task in the domain of natural language processing.

Acknowledgements

We would like to express our sincere gratitude to our supervisors, Professor Hadi Farahani and Teaching Assistants Shide Sharif and Niloofar Shabani, for their invaluable guidance and support throughout the development of this movie recommendation system. Their expertise and insights were essential in ensuring the quality and accuracy of the system.

We also extend our thanks to Professor Seyed Ali Katanforush for teaching us data structures and algorithms, which provided the necessary foundation for our understanding of the fundamental concepts that underlie this project.

Finally, we would like to thank Dr. Saeedreza Kherad Pishe for teaching us programming and deep learning, which were critical skills for the development of this recommendation system. Thank you all for your contributions to our education and success.

Contents

Chapter 1

Dataset

1.1 Introduction

With the growing use of AI, particularly with the recent introduction of ChatGPT, a new challenge has been emerged, distinguishing between real and AI-generated content. This challenge holds significance, especially in academic contexts to ensure data reliability and fair assessment. In response to this challenge, our project implements a DeBERTa LLM for the text detection task, using its advanced language understanding.

1.2 Dataset Description

The main selected dataset for LLM detection task consists of 1,378 essays, each labeled as either "real" or "fake". "fake" label indicates that the essay is AI-generated. The dataset has 6 columns each representing specific attributes. The 'id' column represents a unique identifier assigned to each essay Uniquely, while 'prompt_id' shows the prompt related to the essay. The 'text' column contains the actual content of the essay, providing the input text for the detection task. The 'generated' column shows whether the content is machine-generated ('1' for AI-generated and '0' for real). The 'label' column is the target label for the detection task, with '1' indicating AI-generated ('fake') and '0' indicating real content. The 'name' column is a categorical representation of the label, mapping '0' to 'real'.

It is important to mention the class imbalance in this dataset, where all of the data instances belong to the "real" class, creating a challenge for the model's generalization. To overcome this issue, we presented two additional datasets. The first, named the Proper Train Dataset, [1], includes a set of AI-generated content. We in-

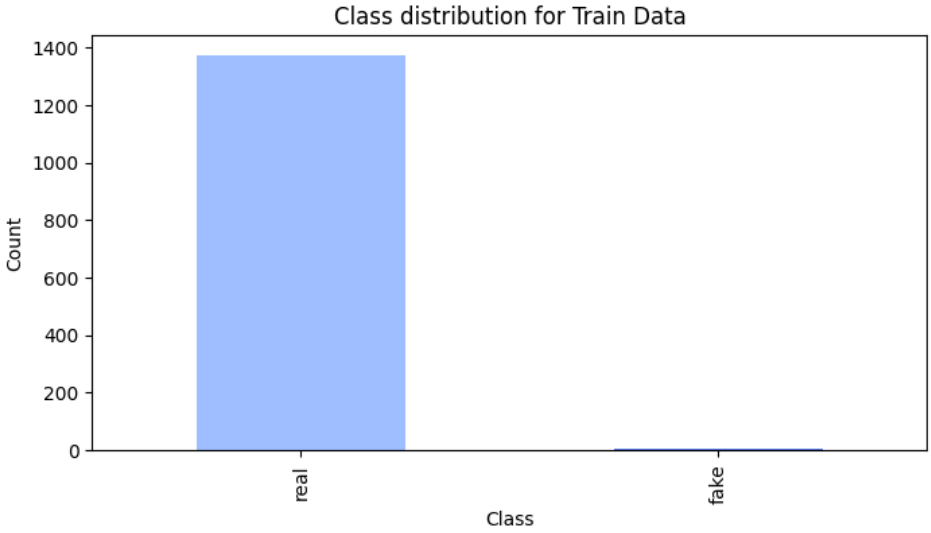

Figure 1.1: Main Dataset Distribution

cluded a subset of 10,000 samples from the 'persuade_corpus' source and all samples from other sources to ensure a equal representation. The second dataset, ArguGPT, [2], enriches the training set with different text samples. With a combined total of 28,210 samples from these external datasets, our approach solves class imbalance problem and improve the performance of the LLM model in detecting AI-generated text.1.2

Figure 1.2: Dataset Distribution After

Chapter 2

Language Models

2.1 Large Language Model (LLM)

The Large Language Model (LLM) is a form of language model that helps in natural language processing (NLP). This models, often built on a transformer, has the capability to understand and produce text based data on human language. These LLMs models such as OpenAI's GPT series and BERT, are pre-trained on large amounts of text data, allowing them to learn underlying patterns and linguistic structure of language. Following the initial steps, certain LLMs undergo training and refinement through a self-supervised learning approach. This involves partial data labeling, enhancing the model's ability to recognize different concepts more precisely. This models tend to perform well in various language tasks, such as translation, summarization, sentiment analysis, and more.

2.2 DeBERTa

DeBERTa, short for Decoding-enhanced BERT with Disentangled Attention, is an advanced language model based on BERT (Bidirectional Encoder Representations from Transformers), developed by Microsoft Research Asia. DeBERTa improves how the model pays attention to words in a sentence. It cleverly separates different aspects of the text, making it more effective in understanding context.

One special thing about DeBERTa is its disentangled attention. This means it can focus better on important details by keeping track of the order of words and their actual meaning separately. This helps the model better understand the relationships between words. DeBERTa uses a decoding strategy during training. This encourages the model to predict missing parts of the input, helping it grasp

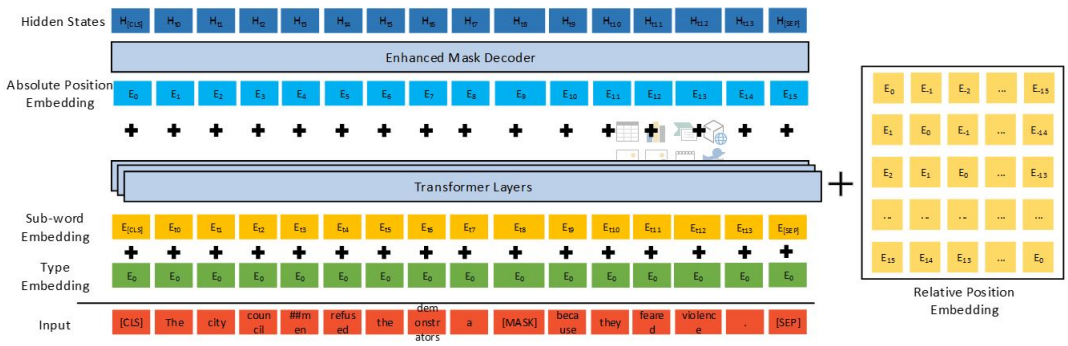

Figure 2.1: DeBERTa Architecture

context from both directions. These improvements make DeBERTa really good at understanding connections between words in longer texts, making it a strong choice for various language-related tasks.

2.3 Our Model

Our chosen model for the AI-generated text detection task is based on DeBERTaV3, an advanced variant of BERT (Bidirectional Encoder Representations from Transformers). DeBERTaV3, with its disentangled attention mechanism, performs well in capturing relationships between words in a sentence, making it suitable for understanding the concepts of AI-generated text.

The model architecture consists of a pre-trained DeBERTaV3 base model, obtained from the Hugging Face model hub. Additional layers has been added to the base model for fine-tuning on our specific task of detecting AI-generated text. The last layer uses a sigmoid activation function, allowing the model to output probabilities indicating the likelihood of text being AI-generated. The raw text data goes thorough a preprocessing step using the DeBERTaV3 preprocessor. This involves tokenization, converting input strings into sequences of token IDs. The sequences are then padded to a fixed length, ensuring efficiency during the training process. After the preprocessing step, The model is trained using a binary cross-entropy loss function with label smoothing (0.02) and the AdamW optimizer. During training, we use a learning rate scheduler that starts with a warm-up phase, followed by a cosine annealing schedule to refine the model's performance over epochs.

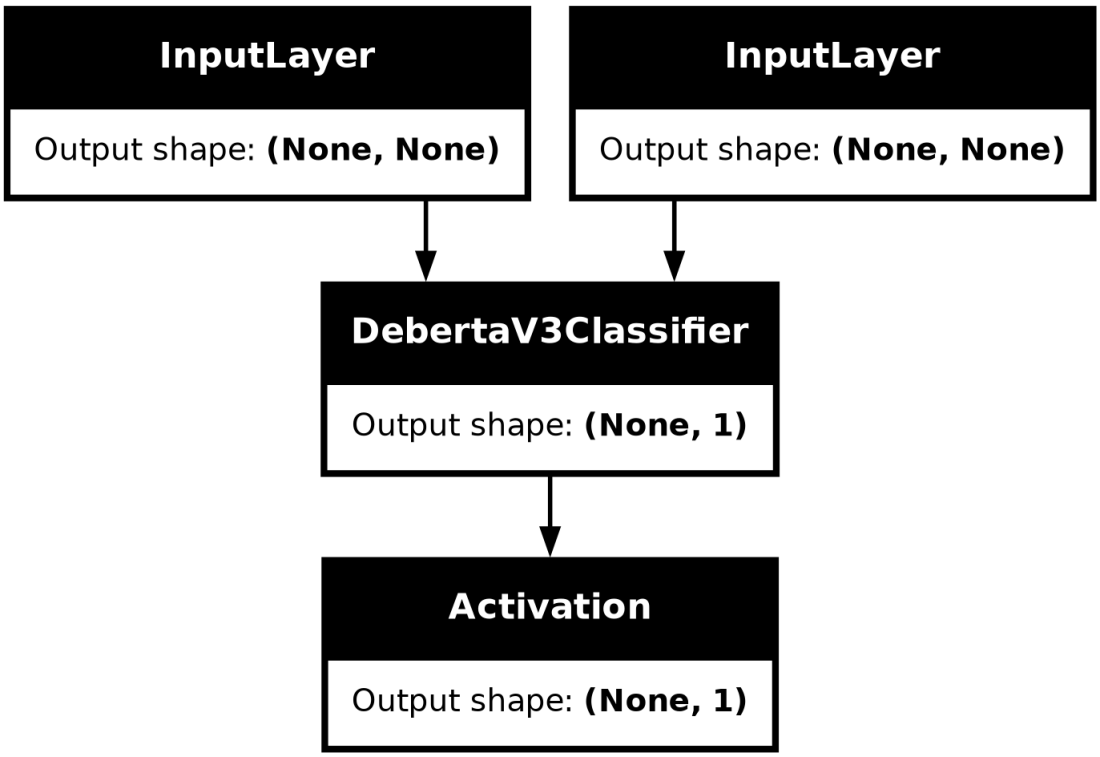

Figure 2.2: The Model Architecture

Chapter 3

Conclusion

3.1 Result

To evaluate the model's performance, we use the area under the receiver operating characteristic curve (AUC) as the primary evaluation metric. This metric provides a measure of the model's ability to distinguish between real and AI-generated text. Also because of the initial class imbalance this metric gives better understanding of the models performance.

The training is done across different folds to evaluate the model's consistency and generalization. Each fold represents a unique split of the data into training and validation sets. To ensure the efficiency of our model and also handle imbalance classes, we evaluate its performance on external datasets, including the Proper Train Dataset and ArguGPT. These datasets consists of various AI-generated content, helping the model's understanding of different patterns.

Beyond the evaluation metrics, the real-world applicability of our model is evaluated by analyzing its performance on unseen data, simulating scenarios where the model faces new examples of AI-generated text. This helps us to gain insights into the model's potential for deployment in different contexts, such as content moderation and fake news detection.

To provide real-time insights into the progress of our LLM detection task, we have set up a live tracking dashboard on the Weights and Biases platform. This dynamic dashboard is accessible through the following link: LLM Detection Live Dashboard.This dashboard offers a detailed view of metrics, training curves, and other relevant visualizations. By checking this dashboard, you can stay informed about the ongoing training process, model performance, and any developments.

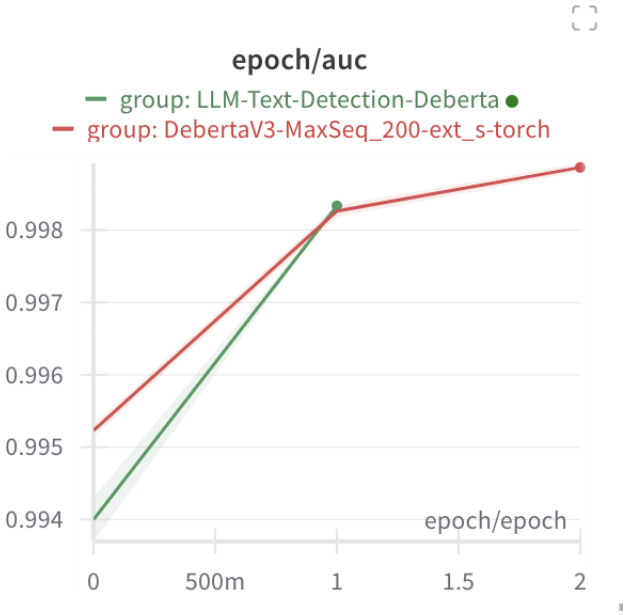

Figure 3.1: Sample Plot of Dashboard

3.2 Future Work

Our model's performance lays the groundwork for further research and development in AI-generated text detection. Future works include exploring ensemble methods, fine-tuning on domain-specific data, and continuous improvement to address challenges in the AI-generated content.

The evaluation and results presented here shows the potential of our DeBERTaV3 model to perform well in the field of AI-generated text detection.

3.3 Accessing the Code

To access the code for this project, please visit the following GitHub repository: https://github.com/MobinNesari81/LLM-Detection.

Bibliography

[1] @thedrcat. (n.d.). *Proper Train Dataset*. (https://www.kaggle.com/datasets/thedrcat/daigt-proper-train-dataset/

[2] @alejopaullier. (n.d.). *ArguGPT*. (https://www.kaggle.com/datasets/alejopaullier/argugpt