Detecting LLM-Generated Text Using Fine-Tuned Transformers: A Comparative Study

Aditya Ratan Jannali
Khoury College of Computer Sciences
Northeastern University
jannali.a@northeastern.edu

Abstract

The ability to detect text generated by large language models (LLMs) is an important problem with implications for academic integrity, content moderation, intellectual property rights, and more. As LLMs become more advanced at generating human-like text, efficient detection methods are needed. In this work, we finetune and evaluate two compact but powerful transformer models, ALBERT and DeBERTa-XS, on the DAIGT Proper Train dataset to detect LLM-generated essays. Our experiments show that both models perform quite well, with DeBERTa-XS achieving 99.43% accuracy and ALBERT achieving 96.57% accuracy. We analyze the results and discuss potential directions for future work. Overall, these findings demonstrate the feasibility of using finetuned transformer language models as a potential solution for LLM text detection.

1 Introduction

LLMs have recently gained immense popularity for their ability to produce relevant responses to questions. Their rise in popularity has raised concerns about the potential for misuse, such as generating deceptive or misleading content, academic dishonesty, intellectual property violations, etc. However, existing tools often fail to identify LLM-generated text accurately, even when minimal changes are applied to the generated content.

Several prior works have explored techniques for the task of detecting LLM generated text. One approach is GLTR [1], which uses statistical analysis and calculates features like burstiness and rank statistics over texts to identify machine-generated outputs. A different line of work uses natural language inference (NLI) for detection - the ANLI benchmark [2] was developed to evaluate NLI models' sensitivity to distribution shifts that could indicate machine generation. Developing LLM discriminators in an adversarial training[3] loop to improve robustness over time is another method to detect llm generated text. Developing reliable, efficient, and scalable methods to accurately detect whether a given text is human-written or generated by an LLM across diverse domains is an important problem that must be tackled.

In this paper, we take a different approach by finetuning and evaluating two compact yet powerful transformer language models, ALBERT [4] and DeBERTa-XS [5], on a benchmark dataset of human and LLM-generated essays. Our key findings are that after finetuning, both models can achieve a very high accuracy of 99% for DeBERTa-XS and 96% for ALBERT on the test set. We analyze the results, compare the two models, and discuss potential next steps.

2 Methodology

2.1 Dataset

For our experiments, we use the DAIGT V2 Train Dataset dataset [6] which contains 44868 essays. Out of these, 27371 are essays written by humans which was collected to create the PERSUADE 2.0 corpus. 17497 of the essays have been generated by popular large language models like ChatGPT, LLAMA2, and Mistral. Every entry in the dataset consists of the essay text, the essay prompt, and a label indicating whether the essay was AI-generated or not. Additionally, the dataset includes metadata such as essay topics and lengths, which can provide insights into the characteristics of LLM-generated text.

The dataset was split into training (80%) and testing (20%) sets to train and evaluate the models' performance.

2.2 Models

Two transformer-based models, ALBERT and DeBERTa, were chosen to perform the classification of documents as human or AI-generated. Both transformer based language models are based on BERT (Bidirectional Encoder Representations from Transformers) which has been pre-trained on a large corpus of text data, allowing it to capture complex linguistic data and patterns from various domains. This pre-trained knowledge can be leveraged during fine-tuning to detect subtle differences between LLM-generated text and human-written text.

Details of the two transformers are provided below

1. ALBERT (A Lite BERT) is a transformer-based language model developed by Google AI. It is designed to address the computational inefficiency of BERT (Bidirectional Encoder Representations from Transformers) by reducing model size and training time while maintaining performance. ALBERT achieves this by implementing parameter sharing across layers and factorized embedding parameterization. Despite its reduced size of only 11M parameters, ALBERT's performance is suitable for deployment in resource-constrained environments.
2. DeBERTa (Decoding-Enhanced BERT with Disentangled Attention) is a transformer-based language model introduced by Microsoft Research. It improves upon BERT by incorporating two key innovations: disentangled attention and enhanced decoding. Disentangled attention allows the model to attend to different aspects of the input independently, enhancing its ability to capture complex relationships within the data. Enhanced decoding enables more effective generation of output sequences by leveraging both left-to-right and right-to-left decoding strategies. These innovations make even the 22M parameter DeBERTa-XSMALL very effective at natural language understanding and text generation.

2.3 Preprocessing and Tokenization

Before finetuning, the dataset of human and LLM-generated essays first went through a preprocessing and tokenization step.

For ALBERT, the ALBERT tokenizer which is a SentencePiece tokenizer trained on a large text corpus, was employed. It tokenizes text into WordPiece tokens. For DeBERTa-XS, the DeBERTa tokenizer was used, which is also a WordPiece tokenizer.

The tokenized sequences were then padded and truncated to a fixed maximum length determined by the model's input size limits.

2.4 Training

The pretrained ALBERT and DeBERTa-XS models from Hugging Face Transformers are instantiated with the pretrained weights acting as initialized parameters. The models are then finetuned on the DAIGT Proper Train dataset for text classification.

For this classification task, Cross-entropy was used to calculate the loss and the Adam optimizer was used to optimize the model weights during the finetuning process. A learning rate of 2e-5, a batch size of 16 was used to fine-tune both transformers for 2 epochs.

3 Results

The finetuned DeBERTa-XS model achieves 99.43% accuracy on the test set while ALBERT achieves 96.57% accuracy, for the binary classification task of distinguishing between human-written and LLM-generated text.

Figure 1 shows the receiver operating characteristic (ROC) curves for the two models on the test set. DeBERTa-XS achieves an impressive AUC of 0.99, marginally higher than ALBERT's AUC of 0.97, both of which indicate an excellent ability to distinguish between the two classes.

The confusion matrices in Figure 2 provide further insight into the models' predictions. For DeBERTa-XS, we see very few misclassifications. ALBERT has more errors but still performs well overall.

An interesting result found was that ALBERT misclassified human-written text with foul language. ALBERT also tends to misclassify shorter human-written essays. The model predicted both such cases to be LLM-generated. DeBERTa-XS, however, correctly classified both these cases as human-generated.

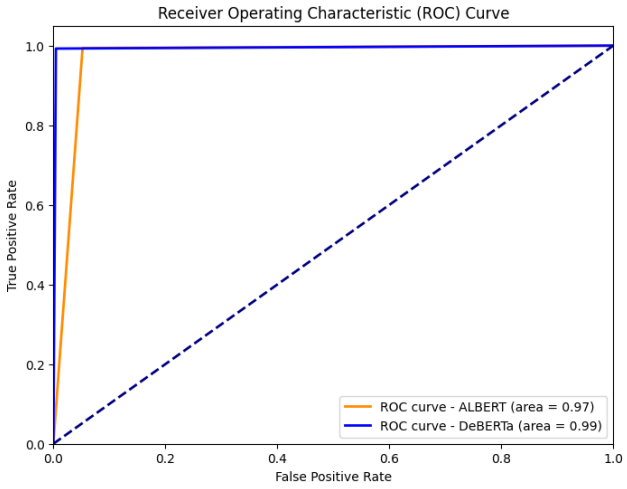
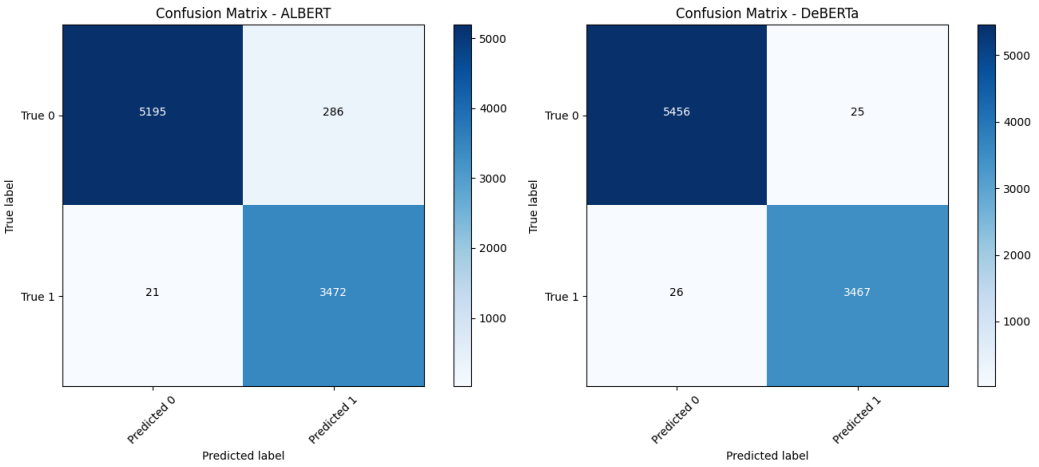


Figure 1: ROC curves for ALBERT and DeBERTa



(a) ALBERT

(b) DeBERTa

Figure 2: Confusion matrix

4 Limitations and Future work

While the results are promising, it must be noted that the models were only trained on essays of a narrow domain. For a more holistic understanding of the feasibility of transformer-based language models, essays from broader domains should be tested against bigger transformer models. It must also be noted that as LLMs continue to evolve and improve, these transformer-based models may not be sufficient on their own.

Future directions for this experiment can include expanding the evaluations to include more domains. Models should be analyzed more deeply to provide a deeper understanding of the limitations of these methods. Bigger language models can also be trained and analyzed.

5 Conclusion

In this work, we demonstrated that compact finetuned transformer models can achieve a high accuracy in detecting LLM-generated text. The DeBERTa-XS model achieved 99.43% accuracy, outperforming ALBERT at 96.57% on this task. We discussed the limitations of our approach as well as promising future directions. Overall, our findings illustrates the immense potential of using fine-tuned language models as an effective solution for LLM text detection.

References

[1] Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043.

[2] Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599.

[3] Chen, H., & Ji, Y. (2022, June). Adversarial training for improving model robustness? Look at both prediction and interpretation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 10463-10472).

[4] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[5] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

[6] https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset