Automatic Detection of Machine-Generated Text Using Pre-Trained Language Models

Yunhao Fang
School of Computer Science and Information Systems
The University of Melbourne
yunhfang@student.unimelb.edu.au

Abstract

In this paper, I provide a detailed description of my approach to tackling the ALTA 2023 shared task whose objective is to build an automatic detection system to distinguish between human-authored text and text generated from Large Language Models. By leveraging several pre-trained language models through model fine-tuning as well as the multi-model ensemble, the system managed to achieve second place on the test set leaderboard in the competition.

1 Introduction

Large Language Models (LLMs) have experienced a drastic advancement over the past few years and brought a revolution to the domain of Natural Language Processing (Gordijn and Have, 2023). Through the expansion of model parameters and the intensive pre-training on a large corpus, recent LLMs such as GPT-4 (OpenAI, 2023) and Llama2 (Touvron et al., 2023) have shown their capability to understand the human language and generate high-quality text.

However, the growing attention to LLMs and their increasing availability to the public nowadays has inevitably led to some concerns as these models can be used in an inappropriate manner to cause harm to society. This includes fake news generation (Zellers et al., 2019), fake product reviews generation (Adelani et al., 2020) and plagiarism (Dehouche, 2021). Therefore, this calls for the construction of a reliable machine-generated text detection system to regulate the use of LLMs so that we can make the most of them. To explore the effective ways that can achieve this objective, ALTA 2023 (Molla et al., 2023) organised a shared task with the goal of constructing an automatic detection system to distinguish between the human-authored text and text generated by the LLMs. The task is formed as a binary classification problem.

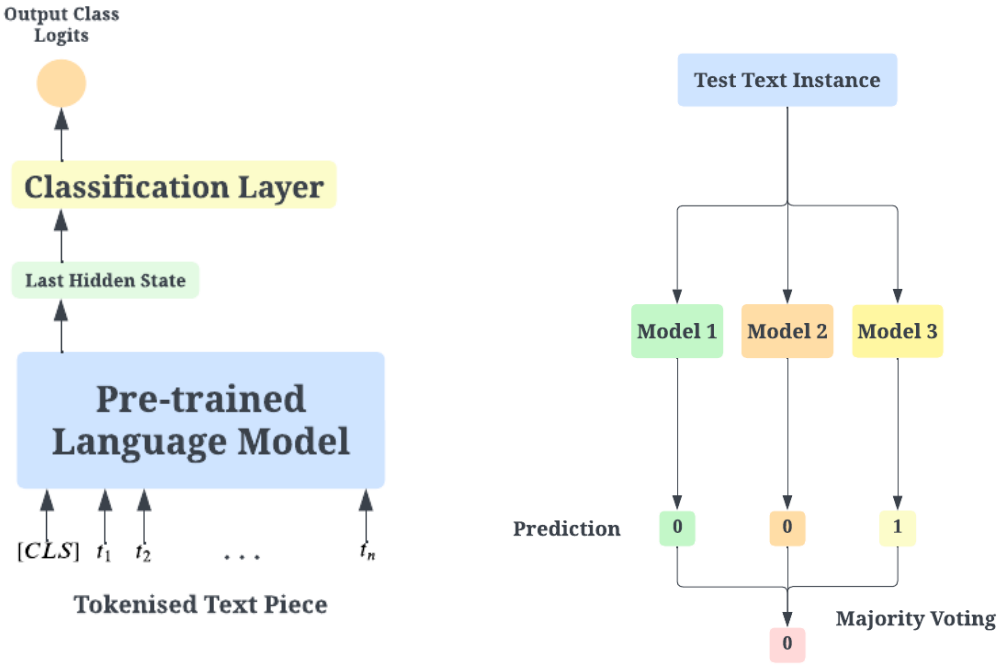
My team handled this task through the utilisation of some representative pre-trained models to tackle the classification problem for machine-generated vs human-authored text given the fact that they have already exhibited their strength in various Natural Language Processing tasks. The models I experimented with include the vanilla BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTaV3 (He et al., 2022a) which represent the chain of improvement for the BERT-based models. I also implemented an ensemble model via majority voting over the best models to further enhance the performance. The rest of the paper will provide a detailed explanation of the design of my system as well as the performance with respect to the task.

2 Related Work

2.1 Machine-Generated Text Detection

Recent studies related to the construction of automatic machine-generated text detection systems focus on the utilisation of the source generator to assist the detection. One area of research intended to rely on internal information from the generative models, such as the probability distribution of tokens or text sequences assigned by the generator, to construct the detector (Mitchell et al., 2023). The other group of researchers proposed the incorporation of the watermarking technique into the generative models by introducing some signals inside the text that cannot be perceived by humans but are detectable by machines. (Kirchenbauer et al., 2023; He et al., 2022b). However, these approaches suffer from their practicality since there exist numerous proprietary LLMs in the industry where the developers are reluctant to expose the internal details of their models, and it is also difficult to guarantee that every LLM developer agrees on the incorporation of watermarking into their models. Therefore, my detection system aims to obtain a good performance under the "black-box" scenario where only the generated text from the generative models is accessible.

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 159–163
November 29 - December 1, 2023



(a) The architecture of the detection system based on Pre-trained Language Models

(b) Multi-model ensemble through majority voting

Figure 1: Illstration of the automatic detection system

2.2 Pre-Trained Language Models

My detection system took advantage of several pre-trained language models by constructing the classifiers upon these models to differentiate machine-generated and human-authored text. This section will provide a description of the models that have been applied during the model development phase.

2.2.1 BERT

BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representation from Transformer, aims to learn the deep bidirectional contextual representation of the language through pre-training on a large text corpus. It attains this objective through the conduction of unsupervised tasks during pre-training to learn the language patterns from the text, which includes the Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM intends to predict the tokens that are masked randomly in the text to capture the bidirectional information of the token, while NSP attempts to understand the relationship between two sentences by predicting whether one sentence follows the other.

2.2.2 RoBERTa

RoBERTa (Liu et al., 2019) is an extension of the vanilla BERT with the goal of optimising the design choices and training strategies of BERT to boost the performance on the downstream tasks. It replaced the static masking in BERT with dynamic masking to avoid duplicated masks and removed the NSP objective from BERT. In addition to this, RoBERTa is also pre-trained on a higher volume of data for a longer time and over a larger batch size compared to BERT.

2.2.3 DeBERTaV3

The original DeBERTa model (He et al., 2020) managed to make a further enhancement on both BERT and RoBERTa through the introduction of two novel techniques: disentangled attention and enhanced mask decoder. A recently upgraded version of DeBERTa called DeBERTaV3 (He et al., 2022a) was proposed by the authors to replace the MLM objective from BERT with Replaced Token Detection (RTD), where a generator is employed to generate corrupted tokens inside the text and the model is trained as a discriminator to determine whether the token is the original one or has

been corrupted. It also proposed a method called gradient-disentangled embedding sharing (GDES) to handle the embeddings from the generator and the discriminator in an effective way.

3 Dataset

The dataset provided by the ALTA 2023 shared task (Molla et al., 2023) consists of text pieces of human-authored and machine-generated text across a wide range of domains. The machine-generated text inside the dataset originates from different types of LLMs. The statistics of the dataset are presented in Table 1. The labels are only contained in the training set where the label assigned to each text piece is either 1 or 0, with 0 indicating that the text is generated by the machine and 1 indicating that the text is written by the human. The distribution of the labels inside the training dataset is 50% for machine-generated and 50% for human-authored which is well-balanced.

Category	Size
Training	18,000
Development	2,000
Test	2,000
Total	22,000

Table 1: Statistics of the dataset for ALTA 2023 shared task

The training set and the development set are released at the same time for model development and the test set is used for the final evaluation of the models and the determination of the rank in the competition.

4 Methodology

Following the process explained in BERT (Devlin et al., 2019), the pre-trained language models discussed in Section 2.2 are adopted to build binary classifiers by adding a single classification layer on top of the last hidden state of the first token (the special '[CLS]' token added by these pre-trained language models) for each of them, which is the contextual representation of the full text. The model architecture is shown in Figure 1a. The original text pieces are tokenised using the corresponding tokeniser for each model and the tokens are input into the classifier. The classifiers are then fine-tuned on the provided training set so that they can learn the language patterns inside the data. The resulting models will be applied to make predictions about the development and test set to gain insight into their performance.

Besides the employment of each single pre-trained language model to perform classification and obtain the results, I've further performed the multi-model ensemble through majority voting over the prediction results from the 3 models that express the best performance. The process is demonstrated in Figure 1b. The voting is conducted as a hard voting where for each instance of the text pieces inside the test set, the label that is assigned to the text by most of the classifiers will be selected as the final label. The logic behind this is to improve the robustness of the detection system by combining the results from multiple models.

Hyperparameter	Value
Learning rate	2e-5
Batch size	64
Training epochs	5
Max length	100

Table 2: Hyperparameter Setting in the experiment

5 Experiments

5.1 Experimental Settings

During the experimental stage, I utilised the pre-trained language models from huggingface to build the classifiers and perform fine-tuning, which includes the models discussed in Section 2.2 with varied size: 1) *bert-base-cased* [1], 2) *bert-large-cased* [2], 3) *roberta-base* [3], 4) *roberta-large* [4], 5) *microsoft/deberta-v3-base* [5], 6) *microsoft/deberta-v3-large* [6]. I used BCEWithLogitsLoss [7] as the loss function and AdamW (Loshchilov and Hutter, 2019) as the optimizer during the model training phase. The setting of the hyperparameters used for the experiment is indicated in Table 2. All the implemented models applied the same experimental settings to compare the performance between each other.

[1] https://huggingface.co/bert-base-cased
[2] https://huggingface.co/bert-large-cased
[3] https://huggingface.co/roberta-base
[4] https://huggingface.co/roberta-large
[5] https://huggingface.co/microsoft/deberta-v3-base
[6] https://huggingface.co/microsoft/deberta-v3-large
[7] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

Model	Version	Development set	Test set
BERT	*bert-base-cased*	0.986	0.976
	bert-large-cased	-	0.980
RoBERTa	*roberta-base*	0.985	0.981
	roberta-large	0.991	0.985
DeBERTaV3	*microsoft/deberta-v3-base*	0.984	0.978
	microsoft/deberta-v3-large	**0.992**	0.982
Ensemble	-	-	**0.990**

Table 3: Classification accuracy of different models on development and test set

The performance of the resulting models is evaluated using the accuracy_score [8] from scikit-learn as specified by the ALTA 2023 shared task.

5.2 Results

Table 3 shows the classification accuracy of all the fine-tuned pre-trained language models as well as the ensemble model involved in the experiment over the development set and test set. As indicated in the table, for all types of pre-trained language models, the large version of the models obtain a better performance compared to the base ones on both the development and the test set. This illustrates the fact that larger models with more parameters have the ability to learn more language patterns from the text to distinguish between human-authored and machine-generated text. Additionally, all versions of BERT underperform RoBERTa and DeBERTaV3 on the test set, while RoBERTa and DeBERTaV3 express a comparable performance between each other. This suggests that the evolution of the BERT model makes contributions to the classification of machine-generated and human-authored text similar to most of the NLP tasks. The results from the table also demonstrate the effectiveness of the multi-model ensemble as the ensemble model using majority voting outperforms all the single models by a certain amount on the test set.

6 Conclusion

In this paper, I've presented my automatic detection system for the ALTA 2023 shared task that classifies machine-generated and human-authored text. The capability of pre-trained language models in handling the task is demonstrated by fine-tuning them on the dataset and constructing the classifiers. The benefits that the multi-model ensemble brings to the performance of the detector are also indicated by the experiment results. As a result, the best system achieves second place in the ALTA 2023 shared task.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative for providing computational resources and I also thank the ALTA 2023 shared task organisers for providing the dataset and organising the task.

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020)*, pages 1341–1354. Springer.

Nassim Dehouche. 2021. Plagiarism in the age of massive generative pre-trained transformers (gpt-3). *Ethics in Science and Environmental Politics*, 21:17–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bert Gordijn and Henk ten Have. 2023. Chatgpt: evolution or revolution? *Medicine, Health Care and Philosophy*, 26(1):1–2.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022a. Debertav3: Improving deberta using electra-style pre-

training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. 2022b. Protecting intellectual property of language generation apis with lexical watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10758–10766.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 alta shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association (ALTA 2023)*.

OpenAI. 2023. Gpt-4 technical report.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.