Integrating Dual BERT Models and Causal Language Models for Enhanced Detection of Machine-Generated Texts

Notebook for the PAN Lab at CLEF 2024

Jitong Chen, Leilei Kong*

Foshan Univisity, Foshan,China

Abstract

In the task of detecting machine-generated texts, accurately distinguishing between those created by artificial intelligence and those authored by humans is crucial. In this evaluation, we use a method that integrates two BERT models with a causal language model specifically trained on in-distribution (ID) samples. This integration enhances the performance of individual models in distinguishing between machine-generated and human-authored texts. Experimental results indicate that our method achieves a certain level of effectiveness in distinguishing between machine-generated and human-authored texts.

Keywords

Detecting Machine-generated Texts, BERT Model, Causal Language Model

1. Introduction

In recent years, the rapid development of large-scale language models such as ChatGPT and Claude has increased the awareness and usage of artificial intelligence technologies among a wider audience. While these advancements have improved efficiency in daily tasks, they have also brought about some negative consequences, such as the unethical use of AI to generate academic papers or complete assignments[1]. Therefore, detecting machine-generated text has become crucial.

The goal of machine text generation detection is to discern whether a text is generated by a machine or by a human[2]. Defined by Voight-Kampff Generative AI Authorship Verification 2024, this task involves evaluating a pair of texts to determine which one is human-generated and which one is machine-generated. Each pair is assigned a score; a score below 0.5 indicates that the first text is human-generated, while a score above 0.5 indicates that the second text is human-generated[3, 4, 5]. In major AI competitions such as Kaggle, participants have showcased inspiring methods for machine text generation detection. These methods mainly involve fine-tuning models in a supervised manner, enriching datasets to maintain balanced data distribution and ensure effective feature learning, and combining multiple models to distinguish between machine-generated and human-authored texts[6].

Building upon existing work and considering the excellent performance of BERT models in capturing contextual information and the ability of causal language models to capture the causal relationships in text generation better, we hypothesize that integrating BERT models and causal language models can complement their weaknesses and reduce biases and errors that might occur with a single model. Therefore, we use a method that integrates two BERT models with a causal language model specifically trained on in-distribution (ID) samples.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ 769855679@qq.com (J. Chen); kongleilei@fosu.edu.cn (L. Kong)
iD 0009-0004-0030-575X (J. Chen); 0000-0002-4636-3507 (L. Kong)

2. Method

Our methodology, as depicted in Figure1, involves training two distinct BERT models, namely BERT Model A and BERT Model B, utilizing the same dataset but employing different processing techniques. Subsequently, we exclusively train a causal language model using machine-generated texts, which can compute the perplexity of a text segment. Perplexity is a metric used to evaluate the performance of a language model. It represents the model's ability to predict the next word[1]. Due to slight differences in causal logic between machine-generated and human-authored texts at the sentence or word level, we classify a segment as machine-generated if the perplexity score computed by the causal language model is low; otherwise, it is classified as human-authored[7, 8]. We then integrate these three models to identify machine-generated text[9].

Here's the detailed workflow: Initially, our method takes two text segments (referred to as text1 and text2) as input. These segments are then separated by [sep] and fed simultaneously into BERT Model A and BERT Model B for evaluation, yielding label A and label B. Meanwhile, text1 and text2 are independently input into a causal language model. This model evaluates each input text segment separately, computing their respective perplexity scores (Perplexity1 and Perplexity2). Subsequently, using a function to generate the output result label C of the causal language model based on perplexity1 and perplexity2, if Perplexity1 is less than Perplexity2, label C is set to 1; otherwise, it is set to 0.

Each model generates a label based on its analysis results, indicating which segment of text it believes is more likely to be human-authored. Finally, through a voting mechanism, the evaluations from all models are aggregated to determine the final output, identifying which segment of text is authored by a human[10].

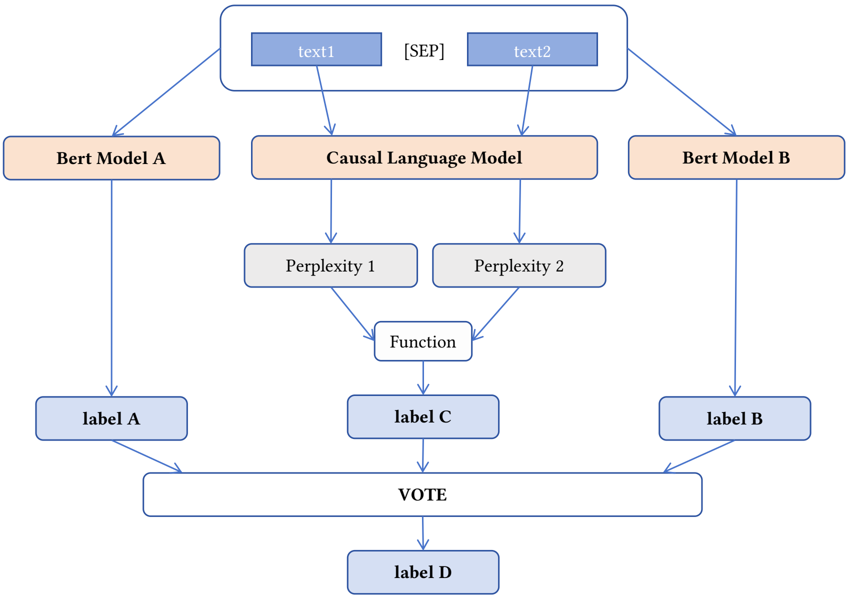

Figure 1: Detailed Methodology Flowchart

[1] https://huggingface.co/learn/nlp-course/chapter7/3

3. Experiments

3.1. Datasets

The datasets are sourced from the PAN24 generation-author-news and Kaggle's DAIGHT-v2-train-dataset[2]. The PAN24 generation-author-news dataset is provided by the evaluation authority, while the DAIGHT-v2-train-dataset is contributed by a participant in Kaggle's machine-generated text detection competition. The PAN24 dataset includes texts generated by 13 large language models and one text authored by a human. Notably, these 13 machine-generated texts and the single human-written text cover 1,087 descriptions of the same topics. The daigt-v2-train-dataset contains segments of machine-generated and human-authored texts on 15 same topics.

Considering the token length limitations of large models, we have conducted a token count on these data.

Table 1
Token distribution by text origin and length

Origin	Token Range	Number of Texts
Machine	<512	24458
Machine	512 - 1024	3854
Machine	>1024	4
Human	<512	9968
Human	512 - 1024	5115
Human	>1024	189

Table 1 presents an overview of these two datasets. In machine-generated texts, the proportion of texts exceeding 512 tokens is 13.6%, while in human-authored texts, the proportion of texts exceeding 512 tokens is 34.7%. However, since BERT and GPT-2 can only accept up to 512 and 1024 tokens, respectively, we need to preprocess this data accordingly. The specific preprocessing steps will be discussed in the Data Processing section.

3.2. Causal Language Model Data Processing

Our analysis of the dataset revealed that some texts exceed 512 tokens. To address this issue, we propose a solution that involves segmenting the texts based on their length and then averaging the perplexity scores of these segments to determine the final result. This method not only overcomes the token length limitation but also has the potential to improve model accuracy. We established text segmentation rules where each segment does not exceed 500 characters and is divided by punctuation marks.

Through experiments, we observed that shorter texts significantly impact model accuracy, so we discarded text segments shorter than 50 characters after segmentation. We randomly selected 1,087 unique topic text segments from 13 machine-generated text datasets and combined them with a human-authored text dataset to form a new dataset. This new dataset was then segmented according to the text segmentation rules to create a validation set. The remaining machine-generated texts were also segmented according to the same rules and used as the training set.

3.3. Bert model Data Processing

In the PAN24-Generation-Author-News dataset, we paired the human-generated dataset with 13 machine-generated datasets by matching texts on the same topic. The human-generated texts were designated as text1, and the machine-generated texts as text2, with a label assigned as 0. We applied the same pairing method in the DAIGHT-v2-train dataset, combining human-generated texts with machine-generated texts based on similar topics. After completing all pairings, we swapped the content

[2] https://www.kaggle.com/competitions/llm-detect-ai-generated-text/discussion/455517

of text1 and text2 for half of the data and changed the label to 1, thus forming Dataset A. Considering the maximum number of tokens the model can accept and the relationship between the two text segments, we swapped the content of text1 and text2 in Dataset A again and inverted the label values, creating Dataset B. Both Dataset A and Dataset B were then split into training and validation sets in a 7:3 ratio. Texts were always truncated at the end after concatenation. If the first text alone exceeded the BERT token limit, it was treated as a single-text classification problem. The aforementioned text swapping operations mitigated the potential adverse effects of this issue to some extent.

3.4. Causal Language Model Experience Setting

We employ the GPT-2 model as a causal language model to perform the task of predicting text perplexity. During the training process, we opted for the AdamW optimizer, which utilizes an adaptive learning rate suitable for weight decay regularization. In setting the optimizer parameters, we distinguished between parameters that required decay and those that did not: 'bias' and 'LayerNorm.weight' parameters within the model were exempt from weight decay, whereas all other parameters were subjected to it. The initial learning rate was set at 5e-5, and we implemented a 'constant_with_warmup' learning rate schedule. This strategy starts with a warmup phase and subsequently maintains a constant learning rate. We calculated the number of update steps per training epoch based on the batch size and the gradient accumulation steps. Training cycles were adjusted inversely based on the maximum training steps to ensure the completeness and consistency of the training. Each batch of data is first processed on the device, followed by bidirectional forward propagation, meaning each batch undergoes two separate forward propagations, yielding two sets of outputs. Additionally, to prevent gradient explosion, gradient clipping techniques were applied. All experiments are conducted on NVIDIA A800 GPU with 80GB memory with a batch size of 8.

3.5. Bert Model Experience Setting

We trained the BERT model on Dataset A and Dataset B separately, resulting in Model A and Model B, respectively. The BERT models were optimized using the AdamW optimizer, with the learning rate set at 3e-5. The batch size was set to 8, and the models were trained for 3 epochs. Notably, our experiments revealed that although this appears to be a binary classification task, setting the model's output categories to three significantly enhances its performance.

4. Results

In this evaluation task, our method surpasses the baseline in both the minimum and median scores, and exceeds most baselines in the 25th quantile, 75th quantile, and maximum scores, as shown in Table 2.

Table 2
The results on official test set

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
Baseline Binoculars	0.342	0.818	0.844	0.965	0.996
Baseline Fast-DetectGPT (Mistral)	0.095	0.793	0.842	0.931	0.958
Baseline PPMd	0.270	0.546	0.750	0.770	0.863
Baseline Unmasking	0.250	0.662	0.696	0.697	0.762
Baseline Fast-DetectGPT	0.159	0.579	0.704	0.719	0.982
Casual Language Model	0.040	0.729	0.782	0.820	0.970
Bert Model	0.556	0.696	0.795	0.883	0.893
Our Method	0.575	0.775	0.891	0.903	0.923

5. Conclusion

This paper discusses a method used in the authorship verification task of generative AI in the PAN@CLEF 2024 competition, which integrates two trained BERT models with a causal language model. The results indicate that our method is effective in distinguishing between machine-generated and human-authored texts, leveraging the strengths of both types of models. However, our study has certain limitations, and there is room for further improvement in the performance of individual models. This could involve exploring alternative base models or adopting different training strategies to optimize the results.

Acknowledgments

This work is supported by the National Social Science Foundation of China (No.22BTQ101) .

References

[1] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, M. Maniatakos, Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis, arXiv preprint arXiv:2305.18226 (2023).

[2] M. Chakraborty, S. Tonmoy, S. Zaman, K. Sharma, N. R. Barman, C. Gupta, S. Gautam, T. Kumar, V. Jain, A. Chadha, et al., Counter turing test ctˆ 2: Ai-generated text detection is not as easy as you may think–introducing ai detectability index, arXiv preprint arXiv:2310.05030 (2023).

[3] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[4] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[5] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[6] Y. Fang, Automatic detection of machine-generated text using pre-trained language models, in: Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association, 2023, pp. 159–163.

[7] Q. Wu, H. Jiang, H. Yin, B. F. Karlsson, C.-Y. Lin, Multi-level knowledge distillation for out-of-distribution detection in text, arXiv preprint arXiv:2211.11300 (2022).

[8] Z. Junhui, W. Mengyan, Y. Erhong, N. Jingran, W. Yujie, Y. Yan, Y. Liner, —— chatgpt (a comparative study of language between artificial intelligence and human: A case study of chatgpt), in: Proceedings of the 22nd Chinese National Conference on Computational Linguistics, 2023, pp. 523–534.

[9] F. Harrag, M. Debbah, K. Darwish, A. Abdelali, Bert transformer model for detecting arabic gpt2 auto-generated tweets, arXiv preprint arXiv:2101.09345 (2021).

[10] G. Dhaou, G. Lejeune, Comparison between voting classifier and deep learning methods for arabic dialect identification, in: Proceedings of the Fifth Arabic Natural Language Processing Workshop, 2020, pp. 243–249.