Team lm-detector at PAN: Can NLI be an Appropriate Approach to Machine-Generated Text Detection

Notebook for the PAN Lab at CLEF 2024

Guojun Wu[1,†], Qinghao Guan[1,†]

[1]University of Zurich, Zurich, 8050, Switzerland

Abstract

The ability to accurately detect machine-generated text is becoming increasingly important in various fields, including academia, journalism, and online security. In this study, we propose a novel method for detecting machine-generated text, predicated on the hypothesis that the probability of reasoning from human-generated text to machine-generated text is inherently higher. Our approach is inspired by the principles of Natural Language Inference (NLI), leveraging the differences in logical consistency and contextual coherence between human and machine-generated texts. However, our experimental results indicate that this method may not be as effective as anticipated. Despite the theoretical foundation, the practical application of our method revealed significant limitations, suggesting that it might not be a reliable solution for detecting machine-generated text. Further research and refinement are necessary to enhance the efficacy of detection techniques.

Keywords

Machine-Generated Text Detection, Natural Language Inference, Probability

1. Introduction

The rapid advancement of artificial intelligence has led to the widespread use of machine-generated text in various domains [1]. Recent development of Large Language Models, such as ChatGPT [2], LLaMA2 [3], can generate human-like texts for various downstream tasks. The performance has been proven to be better than humans in some specific tasks. From automated news articles to customer service chatbots, these texts are becoming indistinguishable from those written by humans [4]. While this technological progress brings many benefits, it also poses significant challenges, particularly in the realm of text authenticity and content verification.

Detecting machine-generated text is crucial for maintaining the integrity of information. In academia, it helps prevent plagiarism and ensures the originality of scholarly work. In journalism, it safeguards against the dissemination of fake news and misinformation. In online platforms, it enhances security by identifying automated accounts and reducing the spread of malicious content. Despite the growing need for effective detection methods, current techniques often fall short. Traditional approaches typically focus on stylistic and linguistic features, which can be easily manipulated by advanced language models. As a result, there is a pressing need for more robust and reliable detection methods.

In this study, we propose a novel approach inspired by Natural Language Inference (NLI). Our method is based on the hypothesis that we are able to judge which text is generated by human by comparing the probability of reasoning (See Section 4). By leveraging the logical consistency and contextual coherence differences between human and machine-generated texts, we aim to develop a more accurate detection model.

However, our experimental results suggest that this method may not be as effective as initially anticipated. Despite its theoretical promise, practical application revealed significant limitations, highlighting the complexity of the detection problem. This paper presents our findings and discusses the implications for future research in this area.

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

† Authors contributed equally

guojun.wu@uzh.ch (G. Wu); qinghao.guan@uzh.ch (Q. Guan)

0000-0003-0062-4502 (Q. Guan)

2. Background

This work was developed for the PAN task——Generative AI Authorship Verification [5] [6] [7] [8] ——where we are given two texts, one authored by a human, and another by a machine, and our target is to pick out the one generated by a human. The dataset was generated by the PAN organizers which is another PAN task, where the participants were asked to build models that can create texts as similar as human-written. The bootstrap dataset consists of multiple text genres, including news articles, Wikipedia texts, and fiction.

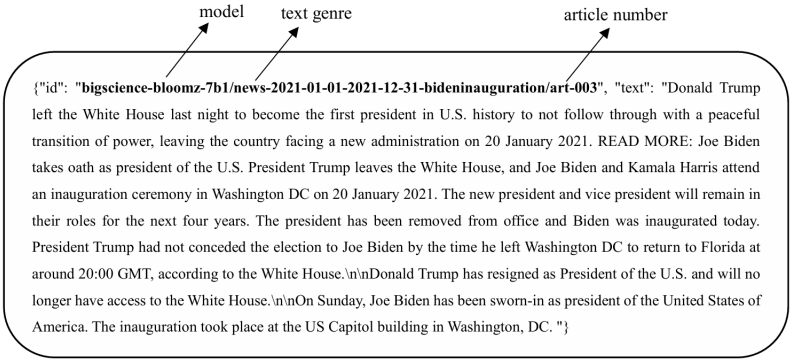

Figure 1: An example of the bootstrap dataset

3. Previous Work

Several studies have approached the detection of AI-generated text as a binary classification problem using neural network-based detectors [9]. For instance, OpenAI has fine-tuned RoBERTa-based GPT-2 detector models to differentiate between human-generated and GPT-2-generated texts [10]. Also, some researchers explored the zero-shot detection method for AI-generated text, such as [11], who noted that AI-generated passages typically exhibit negative curvature in the log probability of texts and proposed DetectGPT, a zero-shot detection method that capitalizes on this observation.

However, relying on neural networks for detection can expose these methods to adversarial and poisoning attacks [12] [13]. To address this, some researchers have explored watermarking AI-generated texts to facilitate detection [14] [15]. Watermarking involves embedding specific patterns in the text, making detection easier. This method provides consistent detection across various contexts and model updates, maintaining its effectiveness without the need for frequent re-training. Watermarking is computationally efficient, requiring minimal additional resources during text generation and enabling quick verification processes. Furthermore, it enhances security by complicating adversarial attempts to alter the text undetected and supports traceability by linking the generated content back to the specific model instance, aiding in accountability and auditing efforts. Overall, watermarking presents a low-overhead, resilient, and scalable approach to managing the challenges of AI-generated text detection.

4. System Overview

NLI is an NLP task involving determining the relationship between two sentences: whether one sentence (the hypothesis) can be inferred from another sentence (the premise). It has been proven that NLI can be used for inconsistency detection in summarization where the source document acts as the premise, and the generated summary acts as the hypothesis [16]. The NLI model evaluates whether the information in the summary can logically be inferred from the source document.

Inspired by the usage of NLI in the summarization task, we detect the machine-generated text by ways of detecting the logical relationship between the premise and hypothesis.

The model checks for three possible relationships between the premise and hypothesis:

Entailment: The hypothesis (summary) logically follows from the premise (source text).

Contradiction: The hypothesis contradicts the premise.

Neutral: There is no clear logical relationship, meaning the hypothesis might add information not present in the premise or omit critical information.

Given two texts, text_a and text_b, one authored by a human and the other generated by a machine, we calculated the probability of reasoning for each text pair independently. Assume that the probability of reasoning from text_a to text_b is $P(T_{\text{text\_a}} \rightarrow T_{\text{text\_b}})$ while the probability of reasoning from text_b to text_a is $P(T_{\text{text\_b}} \rightarrow T_{\text{text\_a}})$. If $P(T_{\text{text\_a}} \rightarrow T_{\text{text\_b}})$ is larger than $P(T_{\text{text\_b}} \rightarrow T_{\text{text\_a}})$, we could assume that text_a was written by human. It is worth noting that we did not conduct any pre-processing (i.e. segmentation) in order to provide sufficient contexts for ratiocination by our model. Our hypothesis is as follows.

As known, the premise provides the basis or groundwork for a conclusion while the hypothesis, in a logical structure, is a statement whose validity is supported by the premise. On the one hand, the machine-generated text in our task was generated based on the human-written text, which means that the human-generated text provides the foundation thus the human-written text should be the premise. On the other hand, the text generated by AI may not match human authors in terms of semantic coherence and logical depth [17]. Accordingly, it is impossible to derive the human-generated text on the basis of the machine-written one.

Beside, we define that if the difference between $P(T_{\text{text\_a}} \rightarrow T_{\text{text\_b}})$ and $P(T_{\text{text\_a}} \rightarrow T_{\text{text\_b}})$ is lower than 0.05, their relation is neutral, meaning there is no clear logical relationship between these two texts.

The language model for NLI was *DeBERTa-v3-large-mnli-fever-anli-ling-wanli* which is a fine-tuned model specifically for NLI tasks [18] for the reason that this model was fine-tuned on distinct datasets including FEVER (Fact Extraction and VERification), ANLI (Adversarial NLI), and WANLI (Weakly-supervised ANLI).

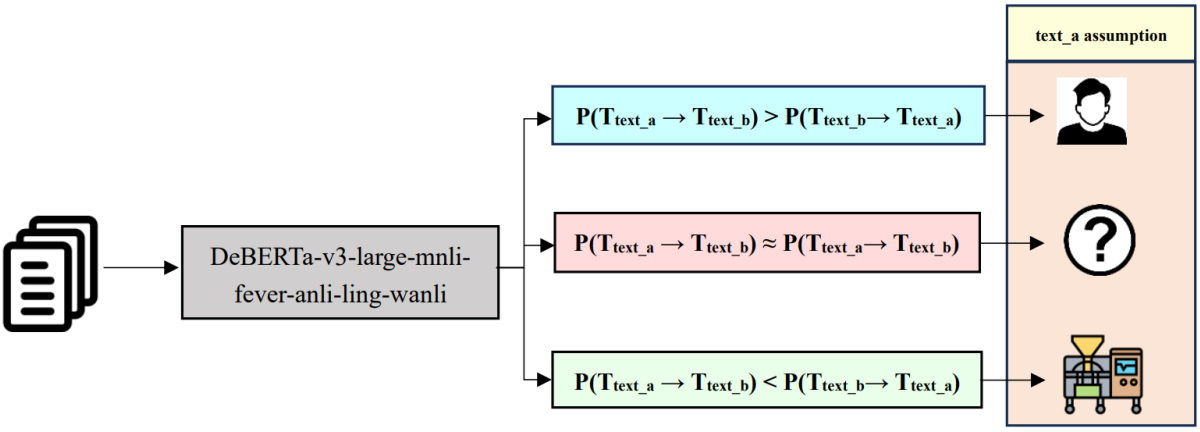

Figure 2: Figure 2: Pipeline of our detector model

5. Results

We compared our model, detector, with the baseline models. The performance metrics indicates that the detector model significantly underperforms compared to all baseline approaches. Specifically, the *detector* (our model) achieved a ROC-AUC of 0.548, which is the lowest among all models, indicating poor discriminative ability. Its Brier score is 0.622, suggesting less accurate probabilistic predictions, while its C@1 score of 0.489 is the lowest, reflecting suboptimal performance. The detector's F1 score of 0.442 and F0.5u score of 0.461 are also the lowest, indicating poor balance and precision-focused performance, respectively. In contrast, Baseline Binoculars exhibits the highest performance across

most metrics, with a ROC-AUC of 0.972, a Brier score of 0.957, and C@1, F1, and F0.5u scores all around 0.965. The overall mean score of Baseline Binoculars is 0.965, compared to the detector's mean of 0.512. The Fast-DetectGPT (Mistral) baseline also performed well, with a ROC-AUC of 0.876 and a mean score of 0.866. Quantile-based evaluations show the 95-th quantile achieving the highest scores, with a ROC-AUC of 0.994 and a mean score of 0.990, underscoring the best performance of the top 5 percentages of models.

Table 1
Overview of the accuracy in detecting if a text is written by an human in task 4 on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). We report ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
detector	0.548	0.622	0.489	0.442	0.461	0.512
Baseline Binoculars	0.972	0.957	0.966	0.964	0.965	0.965
Baseline Fast-DetectGPT (Mistral)	0.876	0.8	0.886	0.883	0.883	0.866
Baseline PPMd	0.795	0.798	0.754	0.753	0.749	0.77
Baseline Unmasking	0.697	0.774	0.691	0.658	0.666	0.697
Baseline Fast-DetectGPT	0.668	0.776	0.695	0.69	0.691	0.704
95-th quantile	0.994	0.987	0.989	0.989	0.989	0.990
75-th quantile	0.969	0.925	0.950	0.933	0.939	0.941
Median	0.909	0.890	0.887	0.871	0.867	0.889
25-th quantile	0.701	0.768	0.683	0.657	0.670	0.689
Min	0.131	0.265	0.005	0.006	0.007	0.224

Table 2 also shows the results, initially pre-filled with the official baselines provided by the PAN organizers and summary statistics of all submissions to the task (i.e., the maximum, median, minimum, and 95-th, 75-th, and 25-th percentiles over all submissions to the task).

We analyzed the reason why our model has such bad performance.

Firstly, our method relies on a single feature—logical inference—which might be insufficient for a comprehensive detection mechanism. Successful detection methods typically incorporate multiple features, including linguistic, syntactic, and semantic analysis, to capture the multifaceted nature of human versus machine-generated text. It suggests that we should establish more comprehensive classification features.

Besides, modern AI models like GPT-3 and GPT-4 are designed to generate text that closely mimics human writing, including coherence and detail. Consequently, the distinction between detailed AI-generated text and detailed human text becomes blurred. Human writers can also produce highly detailed and coherent text, especially in structured or formal contexts. This overlap reduces the effectiveness of using coherence and detail as discriminative features.

Human-generated text can also exhibit inferential relationships, especially in informative or explanatory writing. For instance, when humans explain concepts or provide detailed descriptions, their sentences can logically infer one another. As mentioned, the dataset involves multiple genres. Our method might frequently misclassify detailed and coherent human text (such as news articles) as AI-generated, leading to a high rate of false positives.

From the NLI model's perspective, the method we use is zero-shot which means that our model has not been specifically trained or fine-tuned on a dataset of human vs. AI-generated texts. Also, DeBERTa's strength in recognizing logical relationships might lead it to frequently detect coherent inferences in both human and AI texts, making it difficult to distinguish between them based solely on coherence. This means it may not be optimized to distinguish the subtle differences between the two types of text.

Table 2
Overview of the mean accuracy over 9 variants of the test set. We report the minumum, median, the maximum, the 25-th, and the 75-th quantile, of the mean per the 9 datasets.

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
detector	0.405	0.505	0.521	0.571	0.622
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
95-th quantile	0.863	0.971	0.978	0.990	1.000
75-th quantile	0.758	0.865	0.933	0.959	0.991
Median	0.605	0.645	0.875	0.889	0.936
25-th quantile	0.353	0.496	0.658	0.675	0.711
Min	0.015	0.038	0.231	0.244	0.252

6. Further Direction

To enhance the performance of AI-generated text detection method, it is crucial to fine-tune the DeBERTa model specifically on a dataset tailored for distinguishing human and AI-generated text. This specialized training will help the model learn the unique patterns and nuances of the task. Additionally, incorporating a broader feature set, including stylistic markers, syntactic complexity, and lexical diversity, can provide a more robust classification framework. Employing ensemble methods that combine zero-shot NLI models with supervised models trained on the detection task can further improve performance by leveraging the strengths of different approaches. Regular evaluation and refinement using diverse and updated datasets will ensure the model adapts to new patterns in text generation. Lastly, utilizing contextual embedding techniques can capture richer text representations, enabling deeper contextual analysis beyond simple logical inference.

7. Conclusion

In this study, we explored the potential of using Natural Language Inference (NLI) to detect machine-generated text by examining the logical relationship between premises and hypotheses. Our hypothesis was that machine-generated text, being more detailed and coherent due to probabilistic generation, would differ significantly from human text in inferential relationships. However, our experimental results revealed significant limitations in this approach. Specifically, our zero-shot method using the "DeBERTa-v3-large-mnli-fever-anli-ling-wanli" model underperformed across various metrics, including ROC-AUC, Brier score, C@1, F1, and F0.5u scores, when compared to baseline models. The primary reasons for this underperformance include the overlap in coherence and detail between human and AI-generated texts, the limitations of a single-feature approach based solely on logical inference, and the model's lack of fine-tuning on a task-specific dataset. Our findings suggest that successful detection of AI-generated text requires a multifaceted approach, incorporating diverse linguistic features and specialized training. Future work should focus on fine-tuning models on relevant datasets and integrating additional classification features to improve the robustness and accuracy of detection methods.

Acknowledgments

We appreciate the help from Simon Clematide and Andrianos Michail who provided their suggestions to improve our work. We would also extend our sincere gratitude to the anonymous reviewer whose insightful comments and suggestions significantly contributed to the improvement of this manuscript.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models (2022) 10684–10695.

[2] OpenAI, Chatgpt: Optimizing language models for dialogue (2022).

[3] H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. V. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, arxiv preprint arXiv:2307.09288. (2023).

[4] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, C. Callison-Burch, Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text, AAAI (2022). arXiv:2212.12672.

[5] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, 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.

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

[7] 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.

[8] 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.

[9] G. Jawahar, M. Abdul-Mageed, L. V. Lakshmanan, Automatic detection of machine generated text: A critical survey, arXiv preprint arXiv:2011.01314 (2020).

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[11] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, arXiv preprint arXiv:2301.11305 (2023).

[12] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv

preprint arXiv:1412.6572 (2014).
[13] V. S. Sadasivan, M. Soltanolkotabi, S. Feizi, Cuda: Convolution-based unlearnable datasets, arXiv preprint arXiv:2303.04278 (2023).
[14] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, T. Goldstein, On the reliability of watermarks for large language models, arXiv preprint arXiv:2303.04278 (2023).
[15] X. Zhao, Y.-X. Wang, L. Li, Protecting language generation models via invisible watermarking, arXiv preprint arXiv:2302.03162 (2023).
[16] P. Laban, T. Schnabel, P. N. Bennett, M. A. Hearst, Summac: Re-visiting nli-based models for inconsistency detection in summarization, Transactions of the Association for Computational Linguistics 10 (2022) 163–177.
[17] O. Marchenko, O. Radyvonenko, T. Ignatova, P. Titarchuk, D. Zhelezniakov, Improving text generation through introducing coherence metrics, Cybernetics and Systems Analysis 56 (2020) 13–21.
[18] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).