Team Deloitte at PAN: A Novel Approach for Generative AI Text Detection

Harika Abburi[1], Nirmala Pudota[1], Balaji Veeramani[2], Edward Bowen[2] and Sanmitra Bhattacharya[2]

[1]*Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited, India*
[2]*Deloitte & Touche LLP, USA*

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating text that closely resembles human writing across wide range of styles and genres. However, such capabilities are prone to potential misuse, such as fake news generation, spam email creation, and misuse in academic assignments. Hence, it is essential to build automated approaches capable of distinguishing between artificially generated text and human-authored text. In this paper, we propose an architecture which includes three components: transformer model, token-level features, and state-of-the-art embeddings. This approach achieves a mean score of 0.973 on PAN dataset, demonstrating its effectiveness in identifying AI-generated text.

Keywords

AI-generated text, Large language models, Text classification

1. Introduction

The domain of Natural Language Generation (NLG) is witnessing a remarkable transformation with the emergence of Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT-4) [1], Large Language Model Meta AI (LLaMA-3), and Mistral LLMs. LLMs, characterized by their large parameter size, have shown state-of-the-art (SOTA) capabilities in generating text that closely mirrors the verbosity and style of human language. They have been shown to outperform traditional Natural Language Processing (NLP) approaches across applications ranging from question answering to code completions [2, 3]. While LLMs' ability to generate human-like text is impressive, it concurrently poses a growing risk in various sectors, including the proliferation of misinformation, phishing email generation, and the preservation of academic integrity [4, 5, 6]. To address these challenges, it has become increasingly crucial for both humans and automated systems to detect and distinguish AI-generated text. This calls for ongoing research and the development of reliable detection methods to promote the responsible and ethical use of LLMs [7, 8].

Diverse modeling strategies, ranging from simple statistical techniques to cutting-edge Transformer-based architectures [9, 10], have been investigated to help develop solutions capable of distinguishing AI-generated text from those written by humans. For instance, Gehrmann et al. [11] proposed straightforward statistical methods which capitalize on the assumption that AI systems tend to rely on a limited set of language patterns with high confidence scores. Liu et al. [12] proposed a model which extracts Robustly optimized Bidirectional Encoder Representations from Transformers (BERT) approach (RoBERTa) embeddings and combines them with sentence-level graph representations. In contrast to individual detection models, we recently proposed ensemble modeling approaches for detecting AI-generated text where the probabilities from various constituent pre-trained language models are concatenated and passed as a feature vector to machine learning classifiers [10, 13]. This approach resulted in improved predictions compared to individual classifiers, highlighting the benefits of combining multiple models.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
abharika@deloitte.com (H. Abburi); npudota@deloitte.com (N. Pudota); bveeramani@deloitte.com (B. Veeramani); edbowen@deloitte.com (E. Bowen); sanmbhattacharya@deloitte.com (S. Bhattacharya)

Recently, there has been a notable increase in research focused on zero-shot detection techniques for AI-generated text. These methods predominantly involve the analysis of outputs from LLMs, utilizing features such as entropy, log-probability scores, and perplexity [14, 15, 16, 17] to help distinguish between human-authored and AI-generated content. However, the zero-shot detection methods can be more effective when there is direct access to the internal specifics of the LLM that generated the text. This limits the robustness of zero-shot detection methods across different scenarios [18, 19].

To boost this area of research further, PAN 2024 workshop introduced 'generative AI authorship verification' shared task, which focuses on determining whether a given text is human-authored or AI-generated. In response to this challenge, we proposed an architecture which leverages a pre-trained RoBERTa-base AI-text detector [20], a Bidirectional Long Short-Term Memory (BiLSTM) attention layer for processing token-level perplexity and word-frequency features, and a state-of-the-art EmbEddings from bidirEctional Encoder rEpresentations (E5) model [21]. Our experiments show that our proposed approach outperforms several state-of-the-art approaches based on established metrics.

2. PAN Dataset

The PAN dataset, released by the PAN shared task organizers, contains both human-authored and AI-generated text. It includes a total of 15,190 samples, consisting of 1,087 human-authored texts and 14,103 AI-generated texts produced using thirteen different LLMs, namely: (i) alpaca-7b, (ii) bigscience-bloomz-7b1, (iii) chavinlo-alpaca-13b, (iv) gemini-pro, (v) meta-llama-llama-2-70b-chat-hf, (vi) meta-llama-llama-2-7b-chat-hf, (vii) mistralai-mistral-7b-instruct-v0.2, (viii) mistralai-mistral-8X7b-instruct-v0.1, (ix) qwen-qwen1.5-72b-chat-8bit, (x) text-bison-002, (xi) vicgalle-gpt2-open-instruct-v1, (xii) gpt-3.5-turbo-0125, and (xiii) gpt-4-turbo-preview. More details about the dataset can be found in the PAN overview paper [22].

3. Proposed Approach

Our framework consists of three major components for feature representations of input text:

(i) RoBERTa base Open AI detector [20] produces document-level representations that capture the overall content's meaning;

(ii). Token-level features[23] are extracted from various GPT2 variants (DistilGPT2, GPT-2, GPT-2 Medium, and GPT-2 Large) to analyze both the predictability of the word sequence and word frequency. The token level features include: log-probability of the observed token, log-probability of the most likely token, entropy of the token probability distribution at a given position, and word frequency. A BiLSTM with attention layer processes these token-level features to create combined document-level representations;

(iii). Document-level feature representation are also extracted using the E5 model.

Finally, the three document-level representations are concatenated into a single representation. The combined representation is then fed into a fully connected layer to generate the final probabilities.

Table 1
Results on PAN test set.

Model	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
Our approach	**0.987**	**0.967**	**0.97**	**0.97**	**0.97**	**0.973**
Baseline Binoculars [17]	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 Fast-DetectGPT [16]	0.668	0.776	0.695	0.69	0.691	0.704

3.1. Results

In this section, we present an evaluation of our AI-generated text detection experiments. In our experiments, 20% of the training data was used for validation. For test run submissions, the validation set was merged back with the training set. We report results using well established metrics [22], and compare our model's performance with state-of-the-art models, as shown in Table 1. After predicting the label for each input, we produced the final scores of each text pair as recommended by the organizers [22]. The results indicate that our method surpasses the state-of-the-art models, achieving a modest improvement of around 1% over the mean score of Binoculars model.

4. Conclusion

In this paper, we described our submission to the PAN shared task for detecting the generative AI content. Our experiments demonstrated that our generative AI text detection approach performs well compared to other state-of-the-art approaches in this domain. For future work, we aim to enhance the generalizability of our model by testing it on diverse datasets to evaluate its robustness in real-world applications.

References

[1] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[2] Z. Tang, J. Ge, S. Liu, T. Zhu, T. Xu, L. Huang, B. Luo, Domain adaptive code completion via language models and decoupled domain databases, arXiv preprint arXiv:2308.09313 (2023).

[3] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, PLoS digital health 2 (2023) e0000198.

[4] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, Turingbench: A benchmark environment for turing test in the age of neural text generation, arXiv preprint arXiv:2109.13296 (2021).

[5] M. Weiss, Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions, Technology Science 2019121801 (2019).

[6] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359 (2021).

[7] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, A survey on llm-gernerated text detection: Necessity, methods, and future directions, arXiv preprint arXiv:2310.14724 (2023).

[8] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, et al., M4gt-bench: Evaluation benchmark for black-box machine-generated text detection, arXiv preprint arXiv:2402.11175 (2024).

[9] V. Verma, E. Fleisig, N. Tomlin, D. Klein, Ghostbuster: Detecting text ghostwritten by large language models, arXiv preprint arXiv:2305.15047 (2023).

[10] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, S. Bhattacharya, Generative ai text classification using ensemble llm approaches, in: IberLEF@ SEPLN, 2023.

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

[12] X. Liu, Z. Zhang, Y. Wang, Y. Lan, C. Shen, Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning, arXiv preprint arXiv:2212.10341 (2022).

[13] H. Abburi, K. Roy, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, S. Bhattacharya, A simple yet efficient ensemble approach for AI-generated text detection, in: Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Association for Computational Linguistics, Singapore, 2023, pp. 413–421. URL: https://aclanthology.org/2023.gem-1.32.

[14] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, arXiv preprint arXiv:2306.05540 (2023).
[15] 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).
[16] G. Bao, Y. Zhao, Z. Teng, L. Yang, Y. Zhang, Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, arXiv preprint arXiv:2310.05130 (2023).
[17] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, arXiv preprint arXiv:2401.12070 (2024).
[18] Y.-F. Zhang, Z. Zhang, L. Wang, R. Jin, Assaying on the robustness of zero-shot machine-generated text detectors, arXiv preprint arXiv:2312.12918 (2023).
[19] X. Yang, L. Pan, X. Zhao, H. Chen, L. Petzold, W. Y. Wang, W. Cheng, A survey on detection of llms-generated content, arXiv preprint arXiv:2310.15654 (2023).
[20] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, arXiv preprint arXiv:1908.09203 (2019).
[21] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).
[22] 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.
[23] P. Przybyła, N. Duran-Silva, S. Egea-Gómez, I've seen things you machines wouldn't believe: Measuring content predictability to identify automatically-generated text, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain, 2023.