A Sentence-level AI-Generated Text Detection Approach with Fine-Tuned BERT*

Notebook for the PAN at CLEF 2025

Fuchuan Ye, Zhongyuan Han* and Haojie Cao

Foshan university,Foshan,China

Abstract

The rapid advancement of large language models (LLMs) has raised growing concerns about distinguishing Al-generated text from human-written content. In this paper, we tackle the fine-grained AI text detection task by transforming document-level inputs into sentence-level classification. We propose a BERT-based model fine-tuned on segmented sentences from the PAN 2025 Subtask 2 dataset, which involves six specific categories of human-AI authorship combinations. Our method employs NLTK for sentence tokenization, filters out trivial inputs, and leverages Huggingface's Trainer API for efficient training. Experimental results show that our approach achieves strong performance, with an accuracy of 77.01% and a weighted F1-score of 77.13%, demonstrating the effectiveness of sentence-level modeling for nuanced authorship detection.

Keywords

AI-generated text detection, human-AI collaboration, sentence-level classification, BERT, authorship verification, PAN 2025

1. Introduction

With the rapid proliferation of large language models (LLMs), the boundary between human-authored and Al-generated text is becoming increasingly blurred. As LLMs are frequently integrated into content creation workflows, distinguishing the degree of AI involvement in text production has become a critical challenge for authorship verification, academic integrity, and digital forensics. The 2025 PAN shared task on Generative AI Authorship Verification addresses this issue through two subtasks[2] .In this paper, we focus on Subtask 2: Human-AI Collaborative Text Classification, which aims to classify documents based on the nature and extent of collaboration between human authors and AI model[3].

Specifically, Subtask 2 defines six nuanced categories that describe how a given text may have been co-authored: fully human-written, human-initiated then machine-continued, human-written then machine-polished, machine-written then human-edited, machine-written then machine-humanized, and deeply mixed texts. This classification problem is substantially more complex than traditional binary authorship detection, as it requires systems to recognize subtle textual cues indicating machine involvement at different stages of composition.

Previous approaches to this task have mostly relied on document-level analysis or zero-shot LLM- based detectors such as DetectGPT.While these models can achieve high-level accuracy, they often overlook fine-grained authorship signals embedded within sentences, especially in mixed or obfuscated scenarios. To address this gap, we propose a sentence-level classification approach using a fine-tuned BERT model. By segmenting texts into individual sentences and training on labeled

*CLEF 2025 Working Notes, 9 -- 12 September 2025, Madrid, Spain

1* Corresponding author.

† These authors contributed equally.

yfc2097639788@gmail.com (F. Ye); hanzhongyuan@gmail.com (Z. Han); caohaojie0322@163.com (H. Cao)

0009-0002-1467-0358 (F. Ye); 0000-0001-8960-9872 (Z. Han); 0000-0002-8365-168X (H. Cao)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

sentence-level data, our method is able to capture local linguistic patterns indicative of specifiy human-AI collaboration modes.

We reformulate the document-level six-class detection task as a sentence-level classification problem, enabling finer granularity and improved interpretability. By fine-tuning a pre-trained BERT-base-uncased model on over 250,000 labeled sentences derived from the official PAN 2025 dataset, we demonstrate that sentence-level modeling achieves competitive performance compared to document-level baselines, with promising results on validation data. This work provides a foundation for future research on explainable and fine-grained AI authorship detection, contributing both a practical detection method and a framework for deeper analysis of human-AI collaboration.

Prior work on AI-generated text detection has mainly focused on binary classification, distinguishing fully human- from fully machine-written content. Zero-shot approaches such as DetectGPT[1] and Binoculars[4] leverage language model behavior to identify machine outputs, while traditional baselines include TF-IDF+ SVM and compression-based methods. However, these methods typically overlook the complex nature of human-AI collaboration, where authorship may be blended or sequential. Recent efforts, such as the PAN shared tasks, have emphasized multi-class detection. Still, most models operate at the document level, missing finer linguistic cues. Our work addresses this gap by applying sentence-level classification with a fine-tuned BERT model, enabling more granular and interpretable detection across six collaboration types.

Recent shared tasks such as those organized at PAN have emphasized multi-class detection of human-AI collaborative texts. Our work builds upon these efforts by introducing sentence-level granularity.

This work is part of our participation in the PAN 2025 Voight-Kampff Generative AI Detection task, organized as part of the CLEF 2025 evaluation campaign[2][3]

2. Experimental Setup

To enable sentence-level classification, we first remove line breaks and redundant whitespace, then split each document into individual sentences using the Punkt tokenizer from the NLTK toolkit. Sentences shorter than five characters are discarded to reduce noise, and each remaining sentence is assigned the original document's label.

After sentence segmentation and filtering, we obtain a dataset of labeled sentences. The corpus is shuffled using a fixed random seed and split into training (90%) and validation (10%) subsets.

3. Method

This section presents our sentence-level classification framework for detecting human–AI collaborative authorship, including the model architecture, training configuration, and evaluation strategy.

We adopt the bert-base-uncased model from the Huggingface Transformers library as our core encoder[5]. This transformer model has been pre-trained on large-scale English corpora and is well-suited for downstream classification tasks. To adapt BERT for multi-class classification, we append a linear classification head to the [CLS] token output of the final hidden layer. The head outputs a 6-dimensional vector, representing the six predefined categories: (1) Fully human-written, which refers to text authored entirely by humans without any machine assistance; (2) Human-written, then machine-polished, where human-authored content is refined by machine tools such as grammar or style correctors; (3) Machine-written, then machine-humanized, where machine-generated text is passed through another machine process to make it appear more human-like; (4) Human-initiated, then machine-continued, where a human-written fragment is extended by a ma-chine; (5) Deeply-mixed, where human and machine contributions are heavily interwoven; and (6) Machine-written, then human-edited, where a machine generates the initial text which is subse-quently edited by a human for coherence or fluency. Sentence-level classification is performed independently, enabling the model to focus on localized stylistic patterns rather than relying on document-level context. The model outputs a probability distribution over the six classes through a softmax layer.

We fine-tune the model using the Huggingface Trainer API, which provides scalable and efficient training utilities. The key training settings include a cross-entropy loss function, AdamW optimizer, a learning rate of 2e-5, a batch size of 24 per device, and three training epochs. Mixed-precision training (fp16) is enabled if supported by the hardware. Evaluation is performed every 5,000 steps using weighted F1-score as the early stopping criterion, and model checkpoints are saved every 10,000 steps with the best checkpoint selected based on validation F1. Each sentence is tokenized with padding and truncation enabled, and the maximum sequence length is set to 128 tokens.

We explored commonly used hyperparameter configurations by conducting limited manual tuning over the learning rate (1e-5, 2e-5, 3e-5), batch size (16, 24, 32), and number of epochs (3, 4, 5). Based on validation performance measured by weighted F1-score, we selected a learning rate of 2e-5, batch size of 24, and 3 training epochs as the best-performing setting. Due to resource con-straints, we did not conduct exhaustive grid search and instead relied on prior empirical evidence and small-scale experiments. We also refrained from using additional handcrafted features beyond sentence segmentation and basic text cleaning, as our focus was to evaluate the sentence-level representations captured by the pre-trained BERT model. Future work may consider incorporating syntactic, stylistic, or edit-based features to further enhance performance.

Model performance on the validation set is assessed using accuracy, weighted precision, recall, and F1-score to handle class imbalance, as well as the confusion matrix to analyze class-wise misclassification trends. This evaluation setup ensures a comprehensive and robust assessment of the model's ability to detect nuanced patterns in human–AI collaborative writing.

4. Results and Analysis

We evaluate our sentence-level BERT classifier on the training set of PAN 2025 Subtask 2. The model achieves an accuracy of 77.01% and a weighted F1-score of 77.13%, indicating that fine-grained sentence-level modeling is effective in capturing human–AI collaboration signals.

Overall performance: The model demonstrates a well-balanced performance on the training data. It achieved an accuracy of 77.01%, meaning that 77.01% of all predictions matched the ground truth labels. The precision is 77.05%, indicating that among all predicted instances of each class, 77.05% were correctly identified. The recall is also 77.01%, suggesting that the model successfully retrieved 77.01% of the actual relevant instances. The F1-score, which harmonizes precision and recall, reaches 77.13%, reflecting consistent performance across all classes.

Class-wise Performance: On the training set, the model demonstrates relatively consistent performance across all six class labels. It achieves the highest F1-score of 68.9% on the "Fully human-written" class, suggesting strong capability in identifying purely human-authored texts. The "Machine-written, then human-edited" class follows closely with an F1-score of 68.0%, indicating the model's effectiveness in recognizing machine-generated content that has been revised by humans.

In terms of hybrid texts, the model performs moderately well, reaching 66.8% on "Deeply-mixed text", 66.2% on "Human-written, then machine-polished", and 65.3% on "Human-initiated, then machine-continued". These results reflect the model's ability to handle varying degrees of human-AI collaboration. The lowest F1-score, 64.5%, is observed on the "Machine-written, then machine-humanized" class, indicating that this category poses the greatest challenge for accurate classification.

Overall, these training set results show that the model maintains a fairly balanced performance across diverse forms of human and machine authorship.

Error analysis of misclassified samples reveals several key patterns. Lexical ambiguity arises when human-edited machine texts incorporate complex syntactic structures typically associated with human writing. Style convergence in deeply mixed texts, where strong stylistic continuity is lacking, can lead to mislabeling due to the limited context at the sentence level. Additionally, boundary dilution occurs when polished or humanized AI-generated texts closely resemble native human prose, making it difficult for the classifier to distinguish between them.

5. Conclusion and Future Work

Here is a revised and unified paragraph that integrates your contributions, limitations, and future directions without bullet points:

Our contributions include leveraging robust preprocessing techniques and carefully configured training setups to establish a strong baseline, along with providing a detailed analysis of class-wise performance and prevalent error patterns. However, our approach has certain limitations: sentence-level classification overlooks inter-sentence dependencies that may be critical for detecting shifts in authorship, and some class boundaries—such as those between human-edited and machine-humanized texts—are inherently ambiguous. To address these challenges, future work will explore hierarchical models that integrate sentence-level predictions with document-level context, incorporate prompt metadata or editing history when available to better capture collaboration dynamics, extend the methodology to multilingual settings for broader applicability, and continue refining sentence-level modeling as a means to enhance granularity and resolution in document-level classification tasks..

Acknowledgements

This work is supported by the Social Science Foundation of Guangdong Province, China (No.GD24CZY02).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Clark, E., August, T., & Smit, N. A. (2023). DetectGPT: Zero-Shot Detection of Machine-Generated Text. ACL.

[2] Bevendorff, J., Dementieva, D., Fröbe, M., Gipp, B., Greiner-Petter, A., Karlgren, J., Mayerl, M., Nakov, P., Panchenko, A., Potthast, M., Shelmanov, A., Stamatatos, E., Stein, B., Wang, Y., Wiegmann, M., & Zangerle, E. (2025). Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*. Springer, Madrid, Spain.

[3] Bevendorff, J., Wang, Y., Karlgren, J., Wiegmann, M., Tsivgun, A., Su, J., Xie, Z., Abassy, M., Mansurov, J., Xing, R., Ta, M. N., Elozeiri, K. A., Gu, T., Tomar, R. V., Geng, J., Artemova, E., Shelmanov, A., Habash, N., Stamatatos, E., Gurevych, I., Nakov, P., Potthast, M., & Stein, B. (2025). Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025. In Working Notes of CLEF 2025 -- Conference and Labs of the Evaluation Forum. CEUR-WS.org, Madrid, Spain.

[4] Goldfarb-Taran, S., et al. (2023). Binocular: Fast and Accurate Detection of Machine-Generated Text. EMNLP.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[6] PAN@CLEF 2025. Generative AI Authorship Verification Task. [Online Resource].