Bert_T for Human-AI Collaborative Text Classification

Notebook for PAN at CLEF 2025

Weidong Wu[1], Wenyin Yang[1,*], Zhen Shen[2], Meifang Xie[1], Zhiliang Zhang[1], Miaoji Zheng[1], Tufeng Xian[1], Qiyuan Sun[1]

[1]*Foshan University, Foshan, China*
[2]*Cargosmart (Zhuhai) Co., LTD, Zhuhai, China*

Abstract

As generative language models become increasingly integrated into writing processes, distinguishing between human-written, AI-generated, and collaboratively authored texts has emerged as a critical challenge. This paper explores neural approaches to text classification in Human-AI co-authorship contexts, aiming to identify the varying degrees and patterns of collaboration within texts. We propose a hybrid neural architecture that combines the contextual strength of BERT with the sequence modeling capabilities of Transformer layers, tailored to capture subtle signals of authorship. The model is evaluated on a specially constructed dataset reflecting diverse collaborative scenarios, including pure human writing, fully AI-generated content, and human-AI co-authored texts. Experimental results demonstrate that this approach consistently outperforms standard baselines in both accuracy and robustness, offering a promising direction for authorship analysis in the era of generative AI.

Keywords

PAN 2025, Human-AI Collaborative Text Classification, Transformer, BERT

1. Introduction

Text classification remains a cornerstone of Natural Language Processing (NLP), and within this domain, the task of authorship verification has gained renewed importance due to the proliferation of large-scale generative models. Authorship verification supports applications such as authenticity validation, plagiarism detection, and source attribution, playing a crucial role in maintaining the integrity of digital content.

The Generative AI Authorship Verification Task at PAN@CLEF 2025 continues this line of research by focusing on the challenge of distinguishing between texts written by humans and those generated by large language models (LLMs). As models such as GPT increasingly produce human-like text, the boundary between human and machine authorship becomes harder to discern, amplifying the relevance of this task. In the 2025 benchmark setting, baseline performance on this task yielded a Recall (Macro) of 48.32%, F1 (Macro) of 47.82%, and Accuracy of 57.09%. Building on previous studies and neural methods for text classification, we introduced a hybrid neural model designed to more effectively capture subtle stylistic and semantic differences indicative of authorship. Our improved system achieved Recall (Macro) of 54.09%, F1 (Macro) of 53.57%, and Accuracy of 63.01%, representing substantial gains across all core evaluation metrics. This performance improvement demonstrates the effectiveness of leveraging neural architectures tailored for human-AI co-authorship detection. Notably, our approach integrates contrastive learning and advanced sequence modeling techniques to enhance discriminative capabilities, especially in scenarios where differences in authorship style are subtle and context-dependent.

Our system was evaluated through the TIRA.io platform, which ensures a reproducible and fair comparison under shared task conditions. The results reinforce the critical role of neural models in advancing authorship verification tasks and illustrate the feasibility of scalable, accurate solutions in the face of increasingly human-like AI-generated text.

[1]*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*
*Corresponding author.
2112453039@stu.fosu.edu.cn (W. Wu); cswyyang@fosu.edu.cn (W. Yang); sz0314@foxmail.com (Z. Shen); 2112453043@fosu.edu.cn (Z. Zhang); 2112453056@stu.fosu.edu.cn (M. Zheng); 2112453044@stu.fosu.edu.cn (T. Xian); 2112453029@stu.fosu.edu.cn (Q. Sun); 2112453050@stu.fosu.edu.cn (M. Xie)

2. Dataset

The The dataset for the Human-AI Collaborative Text Classification Task at PAN@CLEF 2025 plays a central role in training and evaluating models aimed at discerning authorship dynamics within mixed-authored content. This year's dataset features a diverse collection of texts that reflect a spectrum of authorship scenarios, including purely human-written, fully machine-generated, and co-authored content. These texts are drawn from multiple genres, including news articles, Wikipedia introductions, and fanfiction, ensuring a broad stylistic and structural diversity.

Participants are also provided with a bootstrap dataset containing annotated samples of real and machine-generated news articles centered on prominent 2021 U.S. events. This component is designed to simulate real-world collaborative or comparative writing scenarios, in which AI-generated content often mirrors the topical and rhetorical choices of human authors. The data—curated in collaboration with contributors such as ELOQUENT Labs— is carefully balanced to represent different authorship types. Articles are generated either by one or more human authors or by advanced LLMs, particularly Google's Gemini Pro. The dataset is structured around pairs of texts on the same topic, authored separately by a human and a machine, to highlight subtle differences in writing style and semantic composition.

Each text is stored in a newline-delimited JSON (.jsonl) format. A typical entry in the development set appears as:

{"text":"Have you... of lost.", "language":"English", "label":4, "source_dataset":"TriBERT", "model":"chatgpt", "label_text":"deeply-mixed text; where some parts are written by a human and some are generated by a machine"}

{"text":"But now... really mattered.","language":"English", "label":3, "source_dataset": "RoFT_chatgpt", "model":"llm1-llm2", "label_text":"human-initiated, then machine-continued"}

In the Human-AI Collaborative Text Classification subtask of PAN@CLEF 2025, each document in the dataset is assigned to one of six categories, reflecting the nuanced interactions between human authors and large language models (LLMs). These categories capture various forms of collaboration and transformation that occur in co-authored texts. Some documents are written entirely by humans without any AI involvement, while others begin with a human-authored draft that is subsequently continued or polished by an AI model. Conversely, certain texts originate from an AI system and are later modified by a human editor, either for stylistic refinement or to obscure the machine origin of the content. In more complex scenarios, human and AI contributions are deeply intertwined throughout the document, lacking a clear division between authorship segments.

This classification task requires models to discern subtle stylistic, structural, and semantic cues indicative of each collaboration pattern, going beyond surface-level detection of synthetic language. By embracing this more detailed taxonomy of co-authorship, the subtask enables a richer understanding of the ways in which human and machine writing processes intersect.To facilitate model training and evaluation, the dataset is provided in newline-delimited JSON (JSONL) format, with each entry comprising a unique identifier, the full text, and a corresponding class label. During testing, labels are omitted, and models must predict the appropriate category for each instance. Evaluation metrics such as macro-averaged recall, F1-score, and accuracy are used to ensure balanced assessment across all six classes, reflecting the importance of generalization in this multi-class setting.

Ultimately, this task offers a framework for systematically analyzing the emerging landscape of collaborative authorship, where distinguishing between different forms of human-AI interaction is critical for maintaining transparency, trust, and accountability in content creation.

Participants are required to classify each individual document into one of six categories that represent different patterns of human-AI collaboration in text creation. This task challenges models to capture subtle linguistic, stylistic, and semantic cues that differentiate various forms of co-authorship, rather than simply distinguishing between human- and machine-generated texts. Access to the dataset is carefully controlled through Zenodo, where participants must register and request access using their TIRA-registered email. This process ensures that the dataset is used exclusively for research purposes and prohibits any unauthorized redistribution. Such controlled access maintains compliance with copyright regulations and preserves the dataset' s integrity for academic and developmental use.

3. Methodology

3.1. Dataset Preprocessing

EffectiveEffective data preprocessing plays a vital role in enhancing the robustness and accuracy of machine learning models, especially for complex tasks like Human-AI Collaborative Text Classification. For this subtask at PAN@CLEF 2025, our preprocessing pipeline was carefully designed to prepare diverse and nuanced texts reflecting different collaboration patterns between humans and AI.

The preprocessing began with text normalization, which involved converting all text to lowercase and removing punctuation, non-alphabetic characters, and numerals. This step aimed to reduce irrelevant variability and focus the model on meaningful linguistic content. Subsequently, common stopwords were filtered out to minimize noise and emphasize distinctive textual features indicative of different co-authorship styles. Following normalization, the texts were tokenized into discrete units suitable for model input. We utilized a pre-trained BERT tokenizer to vectorize the token sequences, ensuring consistency by applying padding and truncation to standardize input lengths, thereby optimizing training efficiency. Given the inherent complexity and limited size of labeled data in this multi-class classification task, data augmentation techniques were employed. These techniques generated additional training samples by introducing subtle modifications to existing texts, preserving semantic and stylistic integrity critical for distinguishing collaborative writing patterns.

Throughout the preprocessing workflow, special attention was paid to maintaining the delicate balance between cleaning the data and preserving the linguistic cues essential for accurate classification of the six collaboration categories. This comprehensive preprocessing framework laid a solid foundation for training models capable of capturing the nuanced human-AI interplay within the texts, thereby improving classification performance on this challenging task.

3.2. Network Architecture

In this study, we propose a neural network architecture designed to address the complexities of Human-AI Collaborative Text Classification in the PAN@CLEF 2025 challenge. The model builds upon the robust contextual representation capabilities of BERT-base, combined with a Transformer encoder to effectively model semantic dependencies and stylistic variations across the input texts. Specifically, we utilize the pre-trained bert-base-uncased model from Hugging Face as the foundational encoder. The [CLS] token embedding is extracted to represent the overall semantic composition of each document and is subsequently passed through a Dropout layer to mitigate overfitting and improve generalization. To capture the nuanced distinctions among the six collaboration types— ranging from fully human-written texts to deeply interwoven human-AI compositions—a Transformer encoder layer with multi-head attention is introduced. This allows the model to attend to different parts of the input text and infer patterns that indicate varying degrees of human and AI involvement. Unlike binary classification tasks, our model is trained in a multi-class setting using a categorical cross-entropy loss function, enabling it to predict one of six predefined collaboration categories for each input text. During inference, each document is processed individually, and the model outputs a probability distribution over the six classes, from which the most probable class is selected.

We fine-tuned hyperparameters such as learning rate, batch size, and dropout rate to maximize classification accuracy and macro-averaged F1 scores—metrics particularly relevant in the context of imbalanced multi-class tasks. This model configuration demonstrates strong adaptability to the demands of Human-AI co-authored text classification and highlights its effectiveness in identifying subtle textual signals that correspond to distinct collaboration modes, as visualized in Figure 1: Model Architecture for Human-AI Text Classification.

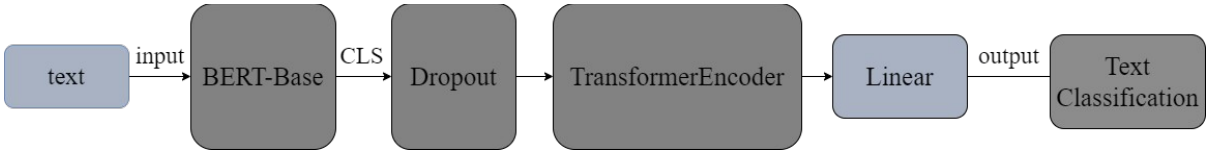

Figure 1: Bert_T Architecture

4. Experiments and Results

4.1. Experimental Setting

For Subtask 2 of PAN@CLEF 2025, we trained a classification model to distinguish six types of Human-AI collaborative writing. The dataset was split into training and testing sets at a 7:3 ratio. Our model integrates a pretrained BERT base encoder with a Transformer layer and a linear classification head to predict one of six categories reflecting different human-machine authorship dynamics. The model uses 768 hidden units, four attention heads, and is optimized using AdamW with a learning rate of 1e-6 and batch size of 8. Training was conducted over 300 epochs on CUDA-enabled GPUs. This setup enables effective learning of stylistic and structural patterns unique to each collaboration type.

4.2. Metrics

Our evaluation framework was meticulously designed to rigorously assess the performance of the Bert_T model across several metrics that reflect its effectiveness in classifying texts based on different modes of Human-AI collaboration. The model was evaluated using a standard set of metrics commonly employed in multi-class text classification tasks, including ROC-AUC, Brier score, C@1, F1, and F0.5u, along with the arithmetic mean of these metrics to provide a comprehensive overview of performance.

Performance Metrics:

ROC-AUC measures the area under the receiver operating characteristic curve, providing insight into the model's ability to discriminate between classes across all thresholds [8]. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The formula is given by:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(t)\,d(\text{FPR}(t)) \quad (1)$$

Brier Score evaluates the mean squared error of the predicted class probabilities in the context of multi-class Human-AI collaborative writing classification [9]. It reflects how well the model's probabilistic outputs align with actual class labels. A lower Brier score indicates more accurate and better-calibrated predictions. It is calculated as:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}\left(\text{predicted probability}_i - \text{actual outcome}_i\right)^2 \quad (2)$$

C@1 represents a modified accuracy that treats non-answers (predictions with a confidence score of 0.5) by averaging the accuracy of the remaining cases, thus penalizing uncertainty [10]. This metric is particularly useful in situations where making no prediction is preferable to making an incorrect prediction. The formula is:

$$C@1 = \frac{\text{Number of correct answers}}{\text{Total number of cases} - \text{Number of non-answers}} + \frac{\text{Number of non-answers}}{\text{Total number of cases}} \quad (3)$$

F1 Score is the harmonic mean of precision and recall, offering a balance between the precision of the classifier and its recall capability [11]. It is particularly useful in situations where an equal balance between precision and recall is desired. The formula is:

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Where $\text{Precision} = \frac{TP}{TP+FP}$ *and* $\textit{Recall} = \frac{TP}{TP+FN}$.

F0.5u is a variant of the F-measure that weights precision more than recall, suitable for scenarios where false positives are more costly than false negatives [12]. It is calculated using the formula:

$$F0.5u = \left(1+0.5^2\right) \cdot \frac{\text{Precision} \times \text{Recall}}{0.5^2 \cdot \textit{Precision} + \text{Recall}} \#(5)$$

These metrics collectively provided a robust framework for evaluating our model, enabling us to effectively measure its ability to perform authorship verification across different dimensions of accuracy and reliability.

4.3. Results

Our proposed Bert_T model exhibited strong performance in the PAN@CLEF 2025 Subtask 2: Human-AI Collaborative Text Classification, demonstrating notable effectiveness across core classification metrics. As shown in Table 1, Bert_T achieved a macro-averaged recall of 0.541, macro F1 score of 0.535, and an overall accuracy of 0.630, outperforming the official baseline system, which yielded a recall of 0.483, F1 of 0.478, and accuracy of 0.570. These results indicate that Bert_T is capable of effectively distinguishing among the six nuanced categories of human-AI collaborative writing—ranging from fully human-written texts to deeply interwoven co-authored documents. The improvement in macro recall and F1 highlights the model's balanced ability to detect minority classes as well as dominant ones, which is critical in a multi-class setting with imbalanced category distributions. The model's consistent performance across these metrics underscores its robustness in handling the complex stylistic variations and subtle linguistic cues that characterize collaborative human-LLM texts. Unlike binary authorship verification, this task demands a more granular understanding of co-authorship dynamics, and the Bert_T model's architecture proves well-suited to these challenges.

Future enhancements will focus on refining category-specific sensitivity and improving class-wise calibration, especially for closely related subtypes like "human-initiated, machine-continued" versus "machine-written, human-edited." Expanding annotated training data and applying contrastive learning are also being considered to further boost model generalization.

Table 1:The final performance of our submission on PAN 2025 (Human-AI Collaborative Text Classification)

Approach	Recall (Macro)	F1 (Macro)	Accuracy
Bert_T	0.541	0.535	0.630
Baseline	0.483	0.478	0.570

5. Conclusion

This This paper presents the design and evaluation of the Bert_T model, our proposed solution for Subtask 2: Human-AI Collaborative Text Classification at PAN@CLEF 2025. By integrating BERT-based contextual feature extraction with a Transformer encoder for attention modeling, Bert_T is tailored to capture the nuanced patterns of collaboration between human authors and large language models. It effectively classifies co-authored texts into six distinct categories, such as fully human-written, human-written then machine-polished, and deeply-mixed compositions.

In experimental evaluations, Bert_T achieved a macro-averaged recall of 0.541, F1 score of 0.535, and an accuracy of 0.630, outperforming the baseline system (0.483 recall, 0.478 F1, 0.570 accuracy). These results confirm the model's reliable performance in handling the complex and subtle nature of human-AI collaborative writing. Its effectiveness demonstrates strong generalization across different forms of human-machine co-authorship and the ability to detect varying degrees of AI

involvement in text generation. Looking ahead, we aim to further improve Bert_T through parameter tuning, enhanced feature engineering strategies, and by diversifying the training data to better represent various human-AI interaction styles. These enhancements are expected to boost the model's precision and adaptability, extending its utility beyond this task to broader challenges in collaborative text understanding and authorship analysis.

6. Acknowledgements

This work was supported by grants from the Guangdong-Foshan Joint Fund Project (No. 2022A1515140096) and Open Fund for Key Laboratory of Food Intelligent Manufacturing in Guangdong Province (No. GPKLIFM-KF-202305).

Declaration on Generative AI

During the preparation of this manuscript, the authors made limited use of GPT-based tools solely for grammar and spelling checking. All content generated by these tools was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of the publication.

References

[1] Janek Bevendorff, Daryna Dementieva, Maik Fröbe, Bela Gipp, André Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. 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, September 2025.

[2] Bevendorff, J., Wang, Y., Karlgren, J., Wiegmann, M., Fröbe, 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 G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), Working Notes of CLEF 2025 -- Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings. CEUR-WS.org, Madrid, Spain.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast,A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025:Generative AI Authorship Verification, Multi-Author Writing Style Analysis, Multilingual Text Detoxification, and Generative Plagiarism Detection, in: Experimental IR Meets Multilinguality,Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] O. Halvani, C. Winter, and L. Graner. "On the usefulness of compression models for authorship verification." Proceedings of the 12th international conference on availability, reliability and security. 2017.

[5] J. Bevendorff, B. Stein, M. Hagen, et al. "Generalizing unmasking for short texts." 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). 2019.

[6] G. S. Bao, Y. B. Zhao, Z. Y. Teng, et al. "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature." arXiv preprint arXiv:2310.05130 (2023).

[7] Z. X. Yang, L. Ma, W. Y. Yang , et al. A Intelligent Detection Method for Irony and Stereotype Based on Hybird Neural Networks. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, CLEF 2022 Labs and Workshops, Notebook Papers, September 2022. CEUR-WS.org.

[8] D. Yuan, W. Y. Yang, L. Ma, et al. Analysis of Irony and Stereotype Spreaders Based On Convolutional Neural Networks. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, CLEF 2022 Labs and Workshops, Notebook Papers, September 2022. CEUR-WS.org.

[9] 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. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[10] A. M. Carrington, D. G. Manuel, P. W. Fieguth, et al. "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.1 (2022): 329-341.

[11] W. Yang, J. Jiang, E. M. Schnellinger, et al. "Modified Brier score for evaluating prediction accuracy for binary outcomes." Statistical methods in medical research 31.12 (2022): 2287-2296.

[12] A. Peñas, A. Rodrigo, A simple measure to assess non-response (2011).

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[14] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, 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), 2019, pp. 654–659.

[15] M. Abassy, K. Elozeiri, A. Aziz, M. N. Ta, R. V. Tomar, B. Adhikari, S. E. D. Ahmed, Y. Wang,O. Mohammed Afzal, Z. Xie, J. Mansurov, E. Artemova, V. Mikhailov, R. Xing, J. Geng, H. Iqbal, Z. M.Mujahid, T. Mahmoud, A. Tsvigun, A. F. Aji, A. Shelmanov, N. Habash, I. Gurevych, P. Nakov, LLM-DetectAIve: a tool for fine-grained machine-generated text detection, in: D. I. Hernandez Farias,T. Hope, M. Li (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 336– 343. URL: https://aclanthology.org/2024.emnlp-demo.35/. doi:10.18653/v1/2024.emnlp-demo.35.