StarBERT: A Hybrid Neural Network Model for Human-AI Collaborative Text Classification

Notebook for PAN at CLEF 2025

Miaoji Zheng[1], Yong Zhong[2,*], Fen Liu[3], Tufeng Xian[4], Meifang Xie[5], Weidong Wu[6], Zhiliang Zhang[7] and Qiyuan Sun[8]

Foshan University, Foshan, China

Abstract

As large language models (LLMs) become more accessible, machine-generated content is rapidly increasing in many fields. These models can produce fluent and coherent text, making them useful for automating various writing tasks. However, their wide use has also raised concerns about misinformation, academic honesty, and the authenticity of content. Therefore, it is important to identify how much of a text is created by humans and how much by machines. In this study, we introduce StarBERT, a new hybrid model that combines DeBERTa-v3-large with StarBlock2d. This model focuses on classifying texts written through human-AI collaboration. Specifically, after dividing the texts into six categories based on the type of human and machine contribution, StarBERT uses the deep language understanding of DeBERTa-v3-large and the high-dimensional mapping ability of StarBlock2d. Our results show that StarBERT performs significantly better than existing baseline models, such as RoBERTa-base.

Keywords

PAN 2025, Human-AI Collaborative Text Classification, BERT, Star Operation

1. Introduction

The rapid spread of generative AI models, such as GPT-4o and Claude 3.5, has ushered in a new era of human-AI collaborative creation, where machine-generated text is deeply integrated with human-authored content. While this collaboration boosts productivity, it also presents serious challenges for tracing the origin of texts and identifying authorship—both of which are essential for maintaining academic integrity, combating misinformation, and ensuring content authenticity.

Although some studies have offered valuable insights [1, 2], most existing research focuses on specific application areas or remains limited to the traditional binary classification approach (human vs. machine) in AI-generated text detection. For example, Liu et al. introduced the concepts of graph representation and structural entropy to study model performance under imbalanced data conditions, aiming to improve text classification and detection accuracy [3]. To evaluate the robustness of detectors on mixed-source generated texts, Huang G et al. proposed a new model called the Siamese Calibration Reconstruction Network (SCRN) using the SeqXGPT-Bench dataset [4]. SCRN introduces and removes noise in the text through a reconstruction network, extracting semantic representations that are robust to local perturbations, which aids in feature analysis. To further enhance detection performance, Mo et al. developed an effective tool for detecting AI-generated text [5]. This method uses deep learning techniques by combining Transformer, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) layers for efficient text classification and sequence labeling tasks. In addition, the preprocessing steps are thorough, including Unicode normalization, case conversion, removal of non-letter characters and extra spaces, and the use of specific delimiters for joining tokens. Overall, this rigorous preprocessing pipeline ensures clean and consistent input for the model, and its systematic approach and promising results highlight the tool's potential for wide application and future

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

2112453056@stu.fosu.edu.cn (M. Zheng); zhongyong@fosu.edu.cn (Y. Zhong); liufen@fosu.edu.cn (F. Liu); 2112453044@stu.fosu.edu.cn (T. Xian); 2112453050@stu.fosu.edu.cn (M. Xie); 2112453039@stu.fosu.edu.cn (W. Wu); 2112453043@stu.fosu.edu.cn (Z. Zhang); sunqiyuan@fosu.edu.cn (Q. Sun)

development in the field of AI-generated text detection (AIGTD). Further more, Wu et al. proposed the BertT model to handle the Generative AI Authorship Verification task by leveraging BERT's deep semantic understanding capabilities and the efficient sequence processing power of Transformers [6].

Traditional AI text detection models that rely solely on binary classification (i.e., human vs. machine) are not well-suited to handle the full spectrum of human-AI collaborative writing. These models often fail to identify complex co-writing patterns, ranging from light AI-assisted editing to deeply integrated joint narratives. To help people better understand human-AI collaboration and reduce the risks associated with synthetic texts, PAN@CLEF 2025 [7](PAN is a series of scientific events and shared tasks on digital text forensics and stylometry)introduced Subtask 2: Human-AI Collaborative Text Classification in the Voight-Kampff Generative AI Detection task. This subtask focuses on classifying documents co-created by humans and LLMs. To address these challenges, we propose the DeBERTa-Star architecture, which deeply integrates the semantic understanding capabilities of DeBERTa-v3-large with the high-dimensional feature transformation power of StarBlock2d. The core innovations of StarBERT include:

- Utilizing the enhanced positional encoding of DeBERTa-v3-large [8] to accurately capture context dependencies in collaborative texts.
- Applying a star operation (element-wise multiplication) to implicitly project the decoupled attention outputs into a high-dimensional nonlinear space.

StarBERT is trained using a stable cross-entropy loss function to enhance training robustness. It was fine-tuned on a JSONL-formatted dataset provided by the organizers to better distinguish among six types of human-AI collaboration. Following local training and initial validation, we submitted the model's predictions—formatted as "id": "identifier of the test sample", "label": 1—to the CodaLab platform. This platform offers a rigorous and controlled evaluation environment, ensuring fair and transparent benchmarking against standard baselines. StarBERT achieved strong results across several key metrics: Macro Recall of 57.46%, Macro F1 of 56.31%, and Accuracy of 66.81%, highlighting its capability to distinguish human-AI collaborative writing effectively. These results demonstrate the practical value of StarBERT in real-world collaborative text classification tasks.

2. Dataset Preprocessing

In natural language processing tasks, data preprocessing is a crucial step. It not only lays the foundation for model training but also directly impacts the model's final performance. Our preprocessing pipeline includes two main components: label conversion and text tokenization.

We begin by converting the labels in the dataset. In the original data, labels are stored as strings, while deep learning models typically require labels in integer format. To handle this, we define a mapping function that converts the label field in each data sample to its corresponding integer value. The label range spans from 0 to 5, representing six distinct categories. Next, we tokenize the text using the tokenizer aligned with DeBERTa-v3-large. The tokenizer converts each text sample into a sequence of token indices from the model's vocabulary, while also indicating which tokens are valid and which are padding. After tokenization, we apply truncation and padding to ensure that all input sequences conform to the model's maximum input length. We set this maximum to 512 tokens, so all inputs are either truncated or padded to exactly 512 tokens. Once tokenization is complete, the data is ready for model training. We ensure the quality of the input by carefully preprocessing the dataset, which transforms raw data into a format suitable for deep learning models.

This comprehensive preprocessing pipeline not only prepares the dataset for effective training of StarBERT, but also enhances the model's accuracy in distinguishing human-AI collaborative text—a key requirement for this evaluation task.

3. Methodology

In this study, we introduce StarBERT, a hybrid neural network model that combines the powerful feature extraction capabilities of DeBERTa-v3-large with the implicit high-dimensional feature fusion mechanism of StarBlock2d to handle complex textual features and distinguish subtle differences between categories of human-AI collaborative writing. To help readers better understand StarBERT, we use Figure 1 to illustrate its structure.

We adopt the pretrained roberta-v3-large model from Hugging Face's Transformers library as the backbone BERT layer, leveraging its large-scale pretrained parameters for advanced natural language understanding. This layer captures rich contextual information from complex inputs such as JSONL-formatted data and applies dropout between transitions to reduce overfitting and enhance robustness.

On top of this, we integrate StarBlock, an efficient neural module centered on element-wise multiplication (the Star Operation [9]), which we adapt from the image domain to natural language processing by proposing StarBlock2d. In our design, input features pass through two independent 1×1 2D convolutional layers, one followed by a ReLU6 activation, and their outputs are fused via element-wise multiplication to form an implicit high-dimensional representation that enables nonlinear feature interactions similar to kernel methods. This result is then processed through a 3×3 depthwise separable convolution (DWConv [10]) to extract spatial features, followed by batch normalization (BN) for training stability. Overall, StarBlock2d achieves near-infinite expressive capacity within a low-dimensional computational space, playing a key role in enabling StarBERT to outperform baselines and peer systems.

During testing, StarBERT independently processes each text sample in JSONL format, evaluates its likelihood of belonging to one of six categories—Fully human-written, Human-initiated then machine-continued, Human-written then machine-polished, Machine-written then humanized (obfuscated), Machine-written then human-edited, or Deeply-mixed text—and classifies it based on the highest score. By fine-tuning hyperparameters such as learning rate and batch size, and optimizing with binary cross-entropy loss, the model is calibrated for high performance on key metrics such as Macro Recall and Macro F1. This setup enables StarBERT to meet the specific challenges of the PAN@CLEF 2025 [11] shared task on human-AI collaborative writing, demonstrating both theoretical innovation and strong practical performance.

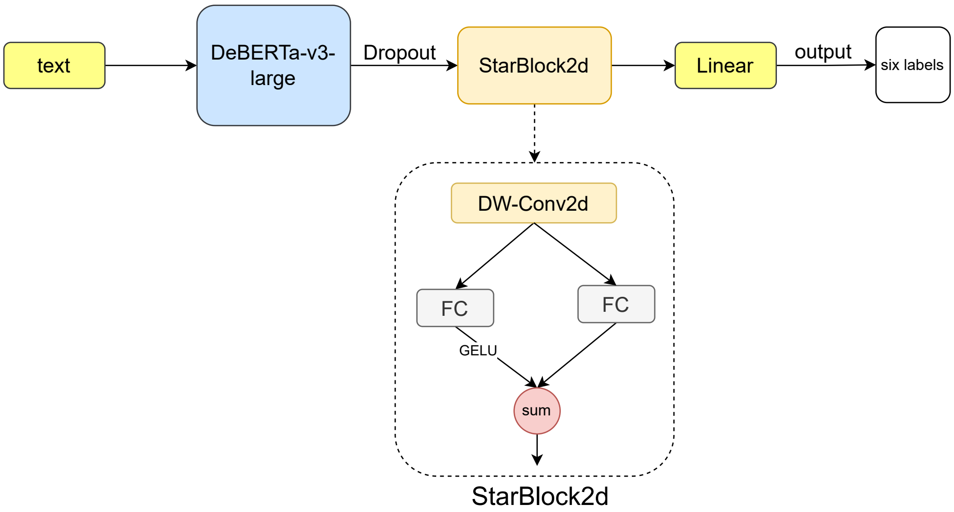

Figure 1: StarBERT Architecture

4. Experiments and Results

4.1. Experiment Settings

In our experimental setup, to evaluate the performance of StarBERT in the human-AI collaborative text classification task, we used the officially provided `subtask2_train.jsonl` as the training set and `subtask2_dev.jsonl` as the validation set for training or fine-tuning. To ensure the reproducibility of our experiments, we fixed the random seed to 42. This setting guarantees consistent results across all stages of the training process, including data loading and model initialization. To balance computational efficiency and learning depth, we trained the model on a CUDA-enabled GPU using the AdamW optimizer. The maximum input length for the model was set to 512 tokens. The learning rate was set to 2e-5, which is a commonly effective value for most text classification tasks. The batch size was 8, and the model was trained for one epoch.

4.2. Metrics

In this task, three evaluation metrics were defined: Macro Recall, Macro F1, and Accuracy. These metrics are designed to comprehensively reflect the model's performance on the collaborative human-AI text classification task. A detailed introduction to each metric is provided below.

Macro Recall: Macro Recall is the arithmetic mean of the recall scores across all classes, giving equal importance to the recognition ability of each class [12]. It is defined as:

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \quad (C = 6) \tag{1}$$

where TP_i denotes the number of true positives for class i (correctly predicted samples), and FN_i represents the number of false negatives for class i (missed samples).

Macro F1: Macro F1 is the arithmetic mean of the F1-scores across all classes, which balances both precision and recall [13]. The formula is defined as:

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2}$$

This macro-averaging strategy calculates the F1-score for each class independently and then averages them, which avoids bias toward majority classes.

Accuracy: Accuracy measures the proportion of correctly predicted samples among all samples, reflecting the overall classification performance [14]. It is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{N_{\text{total}}} \quad (N_{\text{total}} = \text{total number of samples}) \tag{3}$$

These three metrics jointly provide an effective and balanced evaluation of our model's capability in the human-AI collaborative text classification task.

4.3. Results

Our StarBERT model demonstrated strong and stable performance in the PAN 2025 human-AI collaborative text classification task, showing substantial effectiveness across several key evaluation metrics. As shown in Table 1, StarBERT achieved a Macro Recall of 57.46%, significantly outperforming the baseline model's 48.32%. This reflects its strong ability to accurately identify instances across all categories. Moreover, StarBERT achieved a Macro F1 score of 56.31%, well above the baseline's 47.82%, indicating that the model effectively balances predictive performance across categories. It avoids bias toward classes with more samples, which often occurs in imbalanced datasets. In terms of Accuracy, StarBERT reached 66.81%, again far exceeding the baseline score of 57.09%, offering a clear and intuitive measure of the model's overall classification capability in this task. In the official PAN 2025 leaderboard, our submission ranked 4th out of 22 participating teams. This result underscores the competitive advantage of our approach in effectively capturing discriminative textual features in the context of human-AI collaboration.

These outcomes validate that StarBERT not only embodies theoretical innovation but also demonstrates substantial practical effectiveness in this domain. The model is capable of distinguishing collaborative human-AI texts with high accuracy, making it a valuable tool for complex text analysis tasks.

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

Approach	Macro Recall	Macro F1	Accuracy
StarBERT	57.46 %	56.31 %	66.81 %
Baseline [15]	48.32 %	47.82 %	57.09 %

5. Conclusion

StarBERT integrates the deep semantic understanding capabilities of DeBERTa-v3-large with the high-dimensional feature mapping power of StarBlock2d, achieving—for the first time—precise six-class classification of human-AI collaborative texts. Experimental results demonstrate that the model significantly outperforms baseline models (e.g., RoBERTa-base) in the PAN 2025 task, achieving a Macro Recall of 57.46%, a Macro F1 score of 56.31%, and an Accuracy of 66.81%. These results validate StarBERT's superiority in handling mixed-authorship texts.

The innovative Star Operation and enhanced positional encoding effectively capture the nonlinear characteristics and contextual dependencies inherent in collaborative texts. This research provides a scalable technical solution for promoting academic integrity and supporting content provenance in the era of AI-assisted writing.

Future work will focus on further optimizing model parameters, enhancing feature engineering techniques, and expanding the diversity of the training dataset. These improvements aim to increase the model's generalization ability and performance across various textual contexts. This continued development is expected to further improve StarBERT's detection accuracy and extend its applicability in real-world scenarios.

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 work, the author(s) used GPT-4 and DeepSeek-R1 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed. full responsibility for the publication's content.

References

[1] S. S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, A. S. Bedi, Towards possibilities & impossibilities of ai-generated text detection: A survey, arXiv preprint arXiv:2310.15264 (2023).

[2] Y. Zhang, Y. Ma, J. Liu, X. Liu, X. Wang, W. Lu, Detection vs. anti-detection: Is text generated by ai detectable?, in: International Conference on Information, Springer, 2024, pp. 209–222.

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

[4] G. Huang, Y. Zhang, Z. Li, Y. You, M. Wang, Z. Yang, Are ai-generated text detectors robust to adversarial perturbations?, arXiv preprint arXiv:2406.01179 (2024).

[5] Y. Mo, H. Qin, Y. Dong, Z. Zhu, Z. Li, Large language model (llm) ai text generation detection based on transformer deep learning algorithm, arXiv preprint arXiv:2405.06652 (2024).

[6] Z. Wu, W. Yang, L. Ma, Z. Zhao, Bertt: a hybrid neural network model for generative ai authorship verification, Working Notes of CLEF (2024).

[7] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, 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, 2025.

[8] 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).

[9] X. Ma, X. Dai, Y. Bai, Y. Wang, Y. Fu, Rewrite the stars, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5694–5703.

[10] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[11] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, 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.

[12] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (2009) 427–437.

[13] C. M. Bishop, N. M. Nasrabadi, Pattern recognition and machine learning, volume 4, Springer, 2006.

[14] M. L. Thompson, W. Zucchini, On the statistical analysis of roc curves, Statistics in medicine 8 (1989) 1277–1290.

[15] 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).