DBA: A Hybrid Neural Network Model for Generative Human-AI Collaborative Text Classification[⋆]

Notebook for PAN at CLEF 2025

Zhiliang Zhang[1,†], Wenyin Yang[*][2,*], Weidong Wu[3], Meifang Xie[4], Miaoji Zheng[5], Qiyuan Sun[6] and Tufeng Xian[7]

School of Computer Science and Artificial Intelligence, Foshan University, Foshan 528000, China

Abstract

With the rapid development and wide application of Large Language Models (LLMs), recognizing the differences between human and machine-generated content has become increasingly complex. Although several classification methods have been proposed for distinguishing text sources, they still have significant shortcomings in terms of the feasibility and challenge of the task nature. In this paper, we propose a hybrid text categorization model, DeBERTa+BiLSTM+Attention (DBA), that incorporates DeBERTa, Bidirectional Long and Short-Term Memory Network (BiLSTM), and Attention mechanism, aiming at identifying the specific ways of authoring in the human-computer collaborative generation of text. The model takes full advantage of DeBERTa's strong capability in semantic understanding, combines BiLSTM's advantage in modeling contextual sequential information, and introduces an attention mechanism to highlight key information so as to more accurately capture semantic cues in the text reflecting the characteristics of human-AI interaction. The task focuses on determining the degree of human versus AI involvement in the creation process of a text, which is classified into six categories: fully human-authored, human-initiated then machine-continued, human-authored then machine-polished, machine-authored then robotically-sourced (obfuscated), machine-authored then manually edited, and deep hybrid text. Our proposed DBA model significantly outperforms existing baseline methods (e.g., RoBERTa) in this multi-category discriminative task, fully demonstrating its power in complex semantic disambiguation and mixed writing pattern recognition.

Keywords

PAN 2025, Human-AI Collaborative Text Classification, DeBERTa, BiLSTM, Attention,

1. Introduction

As Natural Language Processing (NLP) continues to evolve, people are increasingly relying on Large Language Models (LLMs) in their writing, whether they are using them for academic tasks [1], composing daily emails [2], posting social media content, or engaging in professional writing [3] such as journalism and creative projects. In many cases, people do not directly copy the output of LLMs, but edit them to suit their actual needs, or even discard the generated content altogether, depending on the use of the text and their personal preferences. Therefore, the introduction of LLM tools into the writing process blurs the line between human authorship and AI-generated content. In collaborative writing scenarios, human authors can not only modify AI-generated text, but their creations are also often inspired and influenced by LLM [4] output. In addition, the prompts used for generating AI content tend to be more personalized and contextually relevant than in standalone generation tasks, which further influences the style of the generated text [5].

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

[⋆] You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

[*] Wenyin Yang.

[†] Zhiliang Zhang.

2112453043@stu.fosu.edu.cn (Z. Zhang); cswyyang@fosu.edu.cn (W. Yang[*]); 2112453039@stu.fosu.edu.cn (W. Wu); 2112453050@stu.fosu.edu.cn (M. Xie); 2112453056@stu.fosu.edu.cn (M. Zheng); 2112453029@stu.fosu.edu.cn (Q. Sun); 2112453044@stu.fosu.edu.cn (T. Xian)

0009-0002-3645-0786 (Z. Zhang); 0000-0003-4842-9060 (W. Yang[*])

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

Existing research has mainly focused on distinguishing between texts created entirely by humans and texts generated entirely by AI, with less attention paid to the subtle differences between human writing styles and hybrid texts created by human-computer collaboration [6]. For example, Wu et al. [7] proposed a hybrid neural network model designed to verify the authorship of text generated by Generative AI. The model combines BERT and other neural network structures to improve the accuracy and efficiency of authorship verification. Gehrmann et al. [8] proposed the GLTR tool to help humans recognize AI-generated texts using statistical methods, but the method does not take into account stylistic differences between human authors. In addition, Mitchell et al. [9] proposed the DetectGPT method, which utilizes the probabilistic curvature property of language models to detect AI-generated text without training, however, the method mainly targets pure AI-generated text, and has limited ability to process mixed text with human-computer collaboration.

In order to fill the gaps in existing research in identifying human-machine collaborative texts, this paper explored how texts can be categorized into six different categories based on the nature of human and artificial intelligence (AI) contributions to the creation of the text. Completely human authorship, human-initiated then machine-continued, human-authored then machine-polished, machine-authored then robot-sourced (obfuscated), machine-authored then human-edited, and deep hybrid texts. We propose a text categorization method based on DeBERTa+BiLSTM+Attention (DBA). This approach is a hybrid architecture that fuses the pre-trained language model DeBERTa-v3-large with sequence modeling and attention mechanisms. First, the contextual semantic features of the text are extracted using DeBERTa, followed by dimensionality reduction of the high-dimensional features by linear projection. Then, a bidirectional long-short-term memory network (BiLSTM) is employed to further model the time-dependent and bidirectional contextual information in the sequence. In order for the model to better focus on the key content in the text, a pooling layer based on the attention mechanism is designed to weight the sequence features for aggregation. Finally, six types of labels are output after the classification layer. Ranking 10th in the official PAN 2025 test results, the three metrics experimented with resulted in Macro F1 at 52.81%, Macro Recall at 48.32%, and Macro F1 at 47.82%, all exceeding the official benchmarks.

2. DataSet

The dataset for the Generative AI Detection Task (Subtask 2) [?] @ PAN 2025 [10] plays a crucial role in training and validating the effectiveness of the DBA model. With the growing popularity of Large Language Models (LLMs) such as GPT-4o, Claude 3.5, and Gemini 1.5-pro, this dataset contains a wide range of text types, reflecting a diverse blend of real and machine-synthesized content. The main sources of data include news reports, Wikipedia introductory texts, and homoerotic novels, which exhibit a rich variety of features in terms of style, structure, and complexity. The structure of this dataset is critical to the task and contains key information such as text content, the category to which it belongs, source and language. Each piece of data is organized in JSON format and an example file format is shown below:

```
{"text":"...  ","language":"...","label":0,"source_dataset":"...","
    model":"...","label_text":"fully human-written"}
{"text":"...","language":"...","label":3,"source_dataset":"...","
    model":"...","label_text":"human-initiated, then machine-
    continued"}
```

The validation set provided by the “Human-Computer Collaborative Text Classification Task” in PAN@CLEF plays a crucial role in testing and optimizing the author’s validation model. Its structure is consistent with the previously described format. Notably, the dataset links the label and label type, where the labels 0 to 5 are used to represent different types of text during prediction. The mapping between these labels and text types is as follows:

```
{
```

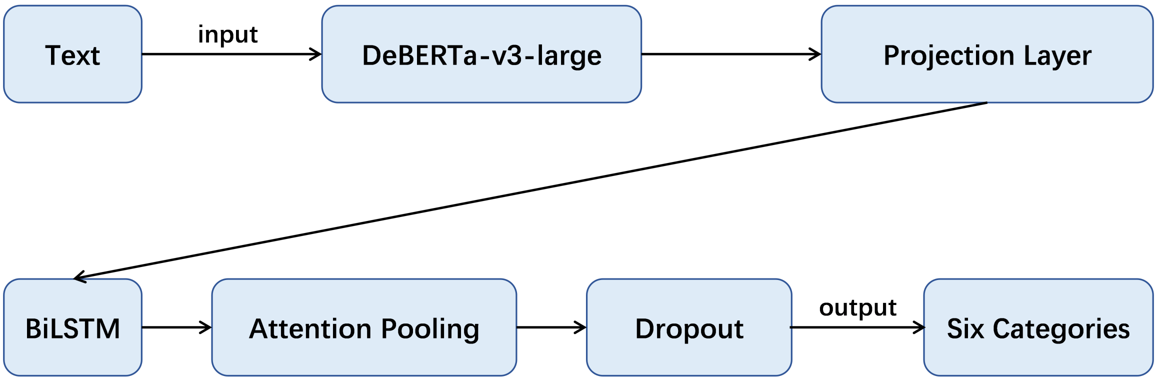

Figure 1: DBA Architecture.

```
    0: "fully human-written",
    1: "human-written, then machine-polished",
    2: "machine-written, then machine-humanized",
    3: "human-initiated, then machine-continued",
    4: "deeply-mixed text; where some parts are written by a human and
       some are generated by a machine",
    5: "machine-written, then human-edited"
}
```

The goal of the participants' task was to categorize documents co-created by humans and LLM. This setup tested the model's ability to recognize micro linguistic and stylistic nuances that typically distinguish human writing from its AI-generated counterpart. Access to the dataset was officially regulated through the PAN, and participants had to sign up and request access using their CodaLab-registered email to ensure that the use of the data was limited to research purposes and that no redistribution occurred. This controlled distribution ensures compliance with copyright regulations and maintains the integrity of the data for academic and developmental purposes.

3. Methods

This study proposes a hybrid neural network architecture DBA based on the combination of pre-trained language models and sequence modeling for distinguishing the degree of human and machine generated text. The model architecture is shown in Figure. ??. The model first adopts Hugging Face's pre-trained DeBERTa-v3-large as the backbone network, and makes full use of its powerful contextual semantic encoding capability to perform deep feature extraction on the input text to obtain a high-dimensional hidden state representation with rich semantic information.

In order to reduce the feature dimensionality and enhance the efficiency of subsequent sequence modeling, the model designs a linear projection layer to map the output of DeBERTa from 1024 to 256 dimensions. Subsequently, the projected sequence features are modeled with temporal information through a bi-directional long-short-term memory network (BiLSTM) to capture the contextual dependencies in the text. In order to adaptively aggregate the sequence information, the model introduces Attention Pooling, which weights and sums the BiLSTM outputs by calculating the attention weights to obtain a context vector representing the overall text semantics. The context vector passes through the Dropout layer to mitigate the risk of overfitting, and is finally mapped to the classification space through the fully connected layer to realize multi-category label prediction. The cross-entropy loss function is used to optimize the model during the training process to ensure the robustness and accuracy of the classification performance.

The DBA model effectively combines the semantic representation advantage of the pre-trained Transformer model with the RNN's sensitive capture ability of temporal features, showing strong differentiation ability and generalization performance in the Human-AI Collaborative Text Classification, and improving the recognition accuracy of the human-computer hybrid text creation process. The specific challenges of the Human-AI Collaborative Text Classification task are met at PAN@CLEF 2025, demonstrating innovative theoretical approaches and practical differentiation capabilities.

4. Experiments and Results

4.1. Experiment Settings

In our experimental setup for evaluating the ability of the DBA model to distinguish between human-created and machine-generated text, we train the model on an officially given training dataset. The model consists of six modules: First, the contextual features of the text are extracted by DeBERTa-v3-large, and then the features are downscaled using a linear projection layer. Then, BiLSTM models the contextual dependencies in the sequence, and Attention Pooling aggregates the important information to form an overall representation. Finally, the classification header is used to generate the prediction results, and Dropout is used to prevent overfitting. The experimental configuration is optimized for a 16GB graphics memory environment and uses the DeBERTa model for the text classification task. The training process adopts a round-by-round evaluation and save strategy, keeping at most one best model, with a training batch size of 8, achieving an equivalent large batch through gradient accumulation (8 steps), and turning on mixed-precision training (fp16) to reduce the memory usage and speed up the training. We set the learning rate to 2e-5, train for 3 cycles, add weight attenuation to prevent overfitting, and have clear paths for logs and model outputs, so that the overall configuration takes into account performance, stability, and resource efficiency.

4.2. Metrics

Our evaluation framework has been carefully designed to rigorously evaluate the performance of the DBA model in the Human-AI Collaborative Text Classification task using the following three metrics: Accuracy, Macro F1 and Macro Recall [11][12]. These metrics provide a comprehensive measure of the model's performance in determining the level of human and AI involvement in a text. The formula for each metric is as follows.

Accuracy measures the proportion of samples correctly predicted by the model to the total sample, with the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where TP, TN, FP, and FN denote the true example, true inverse example, false positive example, and false inverse example, respectively.

Macro F1 is the calculation of F1 scores for each category separately and then averaged, which is able to give equal attention to each category when the categories are not balanced, according to the formula:

$$\text{Macro F1} = \frac{1}{n}\sum_{i=1}^{n} \text{F1}_i \tag{2}$$

where n is the number of categories and F1_i is the F1 value of the i-th category.

Macro Recall is an average of the recall of each class, which is used to measure whether the model recognizes each class completely, with the formula:

$$\text{Macro Recall} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i \tag{3}$$

where n is the number of categories and $Recall_i$ is the recall of the i-th category.

Together, these metrics are used to comprehensively assess the model's ability to recognize different human and AI engagement patterns, and Macro Recall in particular emphasizes coverage of each category to avoid biasing the model toward the dominant category, a key metric for assessing the fairness and stability of multi-class classification.

4.3. Results

Our DBA model in Subtask 2: Human-AI Collaborative Text Classification of PAN 2024Voight-Kampff Generative AI Detection 2025 shows strong performance, showing substantial effectiveness on several key metrics. As shown in Table 1, Accuracy of 61.65% is about 4% higher than the benchmark's 57.09%, reflecting the overall ability to predict correctly. Macro Recall is 54.06%, indicating that the model has better coverage of the categories. Macro F1 is 52.81%, indicating a more balanced performance across the six different categories. The accuracy-related metrics about the benchmarks are 57.09% for Accuracy, 48.32% for Macro Recall, and 47.82% for Macro F1. Our model exceeds all the metrics of the benchmark, reflecting the effectiveness of the model.

In terms of competition rankings, our entry ranked 10th out of 22 participants on the official PAN 2025 leaderboard. Notably, the ranking exceeded all baselines on all test datasets, as detailed in the PAN 2025 rankings. This ranking emphasizes the competitive advantage of our model and its remarkable ability to discriminate in challenging environments full of diverse and complex entries. These results confirm that DBA not only represents theoretical innovation, but also demonstrates significant practical capabilities in the area of human-AI collaborative writing recognition. The model's ability to effectively distinguish between human and machine-generated text makes it a valuable tool for complex text analysis tasks. Future work at will focus on further optimizing the model parameters, enhancing the feature engineering techniques, and expanding the diversity of the training datasets to improve the model's generalizability and performance in different textual contexts. This continuous improvement aims to refine the ability of DBA for higher detection accuracy and wider range of applications in the real world.

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

Name	Accuracy	F1 (Macro)	Recall (Macro)
Baseline	57.09	47.82	48.32
DBA	**61.65**	**52.81**	**54.06**

5. Conclusion

This paper details the development and evaluation of the DBA model, which is our contribution to Subtask 2: Human-AI Collaborative Text of Voight-Kampff Generative AI Detection 2025 in PAN. Classification's innovative contribution. In this study, we propose a hybrid architectural model that fuses DeBERTa semantic feature extraction capability, BiLSTM long-range dependency modeling capability and attention mechanism to effectively identify the generation of human-computer collaborative text. The model balances deep semantic understanding and sequence modeling, and focuses on key text segments through the attention mechanism, which improves the ability to perceive complex linguistic features. In the task of classifying six types of mixed-author text, the model outperforms the existing baseline method in several performance metrics, as shown by the accuracy (Accuracy) of 61.65%, the macro-averaged F1 score (Macro F1) of 52.81%, and the macro-averaged recall (Macro Recall) of 54.06%; in comparison, the baseline model only respectively achieved 57.09%, 47.82%, and 48.32%, respectively. This performance improvement fully verifies the effectiveness and stability of the proposed model in discriminating complex text generation methods.

Looking ahead, we will continue optimizing the DBA model across multiple dimensions. At the model level, we aim to apply more refined parameter tuning and regularization to enhance generalization. In feature engineering, we plan to fuse structured semantic and stylistic features to better capture authorship variations. At the data level, we will expand the training corpus to cover diverse text types such as news, social media, technical documents, and literature. These efforts will not only boost performance in human-AI collaborative text classification but also support the model's transferability to other NLP tasks.

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).

7. Declaration on Generative AI

Duringthepreparationofthis work, theauthorsusedDeepSeek-R1inorderto: Grammarandspelling check. After using this tool, the authors reviewed and edited the content as needed. full responsibility for the publication's content.

References

[1] Z. Lin, Techniques for supercharging academic writing with generative ai, Nature Biomedical Engineering (2024) 1–6.

[2] S. M. Goodman, E. Buehler, P. Clary, et al., Lampost: Ai writing assistance for adults with dyslexia using large language models, Communications of the ACM 67 (2024) 80–89.

[3] J. A. George, 'If Journalism Is Going Up in Smoke, I Might as Well Get High Off the Fumes': Confessions of a Chatbot Helper, 2024. The Guardian (September 2024), Accessed 2024, pages 11–14.

[4] A. H. C. Hwang, Q. V. Liao, S. L. Blodgett, et al., 'It was 80% me, 20% AI': Seeking Authenticity in Co-Writing with Large Language Models, Proceedings of the ACM on Human-Computer Interaction 9 (2025) 1–41.

[5] C. Li, M. Zhang, Q. Mei, et al., Learning to rewrite prompts for personalized text generation, in: Proceedings of the ACM Web Conference 2024, 2024, pp. 3367–3378.

[6] Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, D. Gašević, G. Chen, Detecting ai-generated sentences in human-ai collaborative hybrid texts: Challenges, strategies, and insights, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-2024), 2024, pp. 7545–7553.

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

[8] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 111–116.

[9] E. Mitchell, Y. Lee, A. Khazatsky, et al., Detectgpt: Zero-shot machine-generated text detection using probability curvature, in: International Conference on Machine Learning, PMLR, 2023, pp. 24950–24962.

[10] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, 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.

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

[12] J. Opitz, A closer look at classification evaluation metrics and a critical reflection of common evaluation practice, Transactions of the Association for Computational Linguistics 12 (2024) 820–836.