Fine-Grained Human-AI Collaborative Text Classification Using DeBERTa

Notebook for the TL Lab at CLEF 2025

Tao Li, Guo Niu*

Foshan University, Foshan, Guangdong, China

Abstract

In this PAN-CLEF 2025 Subtask 2: Human-AI Collaborative Text Classification challenge, our objective is to categorize documents co-authored by humans and Large Language Models (LLMs). Specifically, we aim to classify texts into six distinct categories based on the nature of human and machine contributions. Utilizing the pre-trained language model DeBERTa-v3-large, we fine-tuned it for this specific classification task. Our experimental results demonstrate that this approach effectively distinguishes between different types of texts, contributing significantly to the understanding of human-AI collaboration and mitigating risks associated with synthetic text.

Keywords

PAN 2025, Generated Content Analysis, Human-AI Collaborative Text Classification

1. Introduction

With the rapid development of artificial intelligence technologies, particularly the widespread application of large-scale language models (LLMs), collaborative creation between humans and AI has become increasingly common. In this context, accurately identifying the proportion of contributions from humans and AI in a given text has emerged as a significant challenge in the field of natural language processing (NLP) [1]. This paper focuses on the PAN-CLEF 2025 Subtask 2: Human-AI Collaborative Text Classification, aiming to classify texts into six categories: "Fully human-written", "Human-initiated, then machine-continued", "Human-written, then machine-polished", "Machine-written, then machine-humanized (obfuscated)", "Machine-written, then human-edited", "Deeply-mixed text" [2]. This task not only aids in understanding human-machine collaboration mechanisms but also provides technical support for detecting synthetic texts and preventing misinformation dissemination. We adopt the DeBERTa [3, 4] model as the core architecture, train and evaluate it on standard datasets, and enhance model performance through data augmentation and fine-tuning strategies.

2. Related Work

In recent years, research on human-AI collaborative writing has gradually increased. Early work mainly focused on determining whether a text was generated by AI, such as the GPT-2 Detector developed by OpenAI [5]. However, as generation technology has advanced, simply distinguishing between "AI-generated vs. human-written" has become insufficient. Therefore, more granular classification tasks have emerged. Some studies have attempted to introduce multimodal features (e.g., syntactic structures, sentiment tendencies) to assist classification, but due to the high cost of data annotation, most research still relies primarily on plain text input. Additionally, some scholars have proposed staged classification strategies—first determining whether a text contains AI components, then further

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

*Corresponding author.

trent0@foxmail.com (T. Li); sreedomly@163.com (G. Niu)
https://orcid.org/0009-0007-1348-2060 (T. Li); https://orcid.org/0000-0002-1552-7310 (G. Niu)
0009-0007-1348-2060 (T. Li); 0000-0002-1552-7310 (G. Niu)

classifying its type. Regarding model selection, Transformer-based models (such as BERT, RoBERTa, DeBERTa) have been widely applied to text classification tasks [6, 7, 3]. Among them, the DeBERTa series, owing to its unique design for contextual awareness and position modeling, has demonstrated excellent performance across various NLP tasks. DeBERTa-v3-Large [4], in particular, exhibits stronger capabilities in long-text understanding and complex semantic modeling through improved relative positional encoding and discriminative pre-training approaches.

3. Methodology

3.1. Data Preparation and Preprocessing

We utilize the officially provided training, validation, and test sets, which collectively contain six types of text samples corresponding to different human-AI collaborative writing styles. Each sample includes the original text content (`text`) and its associated class label (`label`). Data preprocessing mainly involves the following steps:

- **Uniform Formatting**: The raw text is directly fed into the model to preserve potentially style-distinguishing linguistic features.
- **Text Encoding and Alignment**: The text is tokenized using the tokenizer corresponding to DeBERTa-v3-large, and all sequences are truncated or padded to a maximum length of 512 tokens to meet the model's input requirements.
- **Label Mapping**: A bidirectional mapping between category names and integer IDs (`id2label` / `label2id`) is established to support multi-class classification.

3.2. Model Architecture and Training Strategy

3.2.1. Model Selection

Recently, advances in Natural Language Processing (NLP) have benefited significantly from progress in pre-trained language models, with DeBERTa standing out among them [3]. It has shown outstanding performance across multiple benchmark tasks. First, the core architectural characteristics of DeBERTa lay the foundation for its performance in complex text classification tasks. The model adopts an ELECTRA-style pre-training method, effectively improving training efficiency and performance through a generator-discriminator framework. At the same time, gradient-decoupled embedding sharing significantly reduces computational resource requirements. DeBERTa not only possesses strong semantic understanding capabilities but can also adapt well to multi-class text classification tasks. Moreover, its large vocabulary (128K tokens) and flexible input-output structure further expand its applicability.

DeBERTa-v3-Large is the third-generation DeBERTa model proposed by Microsoft, with 178 million parameters [4]. Its advantages include: first, the use of a mechanism that separates content and position representations, enhancing the model's ability to understand context; second, the introduction of a hybrid of absolute and relative positional encoding, strengthening modeling of long-range dependencies; and finally, an improved Masked Language Modeling (MLM) objective function that boosts pre-training efficiency.

3.2.2. Fine-Tuning Strategy

During actual training, we employed the standard full-parameter fine-tuning strategy. Specifically, we loaded a pre-trained language model and added a classification head on top of it. Then, all parameters of the entire model were updated without freezing any layers. This strategy is suitable when the target task dataset is moderately sized and sufficient computing resources are available. In this case, the dataset meets these conditions.

Throughout the training process, the AdamW optimizer and a relatively small learning rate were used to ensure that the language representations already learned by the pre-trained model would not be disrupted during fine-tuning. Meanwhile, weight decay was implemented to prevent overfitting. The model was trained for three epochs in total. Additionally, an early stopping mechanism was utilized to both prevent overfitting, verifying promptly at per epoch and stopping early on the dev set, which improve model performance. Micro F1 score was used as the primary evaluation metric to guide the optimization process. To enhance training efficiency and save GPU memory, all input texts were uniformly truncated or padded to a maximum length of 512 tokens.

4. Experiments

4.1. Experimental Setup

We firstly conducted tests on the official dev set. Evaluation metrics included macro-averaged recall, macro-averaged F1 score and macro-averaged precision.

Finally, we conducted tests on the official test set. Evaluation metrics included accuracy, macro-averaged recall (Macro-Recall), and macro-averaged F1 score (Macro-F1).

Table 1
Results on the dev set

Type	Recall	F1	Precision
Fully human-written	87.94	60.88	46.56
Human-initiated, then machine-continued	90.27	67.74	54.21
Human-written, then machine-polished	98.76	88.09	79.50
Machine-written, then machine-humanized (obfuscated)	34.80	49.97	88.63
Machine-written, then human-edited	99.11	30.53	18.04
Deeply-mixed text	94.51	96.98	99.59
Overall	84.23	65.70	64.42

Table 2
Results on the test set

Model Name	Recall (Macro)	F1 (Macro)	Accuracy
Baseline	48.32	47.82	57.09
DeBERTa-v3-Large	56.74	55.39	66.27

4.2. Result Analysis

In the PAN-CLEF 2025 Subtask 2, the Recall performance of the DeBERTa-v3-large model became a key evaluation criterion. From the results on the dev set (Table 1), it is evident that the model performs exceptionally well in classifying "human-written, then machine-polished" and "deeply-mixed text" categories but faces challenges in identifying "machine-written, then machine-humanized" texts. On the test set, experimental results (Table 2) showed that compared to the baseline model RoBERTa-base, DeBERTa-v3-large achieved a Macro Recall value of 56.74% on the same task, representing an improvement of 8.42%, indicating its superiority in complex multi-category classification tasks. Nevertheless, despite overall performance improvements, the significance of the Recall metric and potential optimization paths warrant deeper exploration.

5. Conclusion

This study leverages the DeBERTa-v3-Large model to achieve efficient classification of six types of Human-AI collaborative texts. Experimental results demonstrate that this model outperforms baseline pre-trained models in terms of Recall and accuracy, especially excelling in handling texts with complex semantic structures and ambiguous category boundaries.

Future research directions include:

- Exploring multimodal feature fusion, such as integrating syntactic, semantic, and emotional information;
- Introducing contrastive learning or self-supervised learning strategies to enhance the model's sensitivity to subtle differences;
- Constructing more representative datasets covering a broader range of real-world applications;
- Developing lightweight versions of the model suitable for deployment on low-resource devices.

Acknowledgments

This work is supported by the Research Projects of OrdinaryUniversities in Guangdong Province under Grant2023KTSCX133, the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515140103.

Declaration on Generative AI

The authors declare that the Qwen3 large language model was used during the preparation of this paper for text translation and language polishing. The final responsibility for the content, accuracy, and scientific integrity of the paper lies solely with the authors.

References

[1] 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.

[2] 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.

[3] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[4] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[5] 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).

[6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[7] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.