DeBERTa-FPN: Fusion Feature Pyramid Network for Human-AI Collaborative Text Classification

Notebook for PAN at CLEF 2025

Qiyuan Sun, Li Ma*, Wenyin Yang, Tufeng Xian, Meifang Xie, Weidong Wu, Zhiliang Zhang and Miaoji Zheng

Foshan University, Foshan, China

Abstract

In recent years, large language models (LLMs) have developed rapidly and have been widely used in the field of text generation. The increasingly realistic generated content has led to some security risks in information dissemination. Human-machine collaborative text classification has become a critical task and is extremely challenging. This paper proposes a human-machine collaborative text classification model, namely DeBERTa-FPN, which combines DeBERTa-V3 and feature fusion pyramid (FPN), aiming to use the powerful text processing capabilities of DeBERTa-V3-Large and the multi-scale feature extraction capabilities of FPN to improve the model performance of human-machine collaborative text classification. The introduction of FPN can enhance the model's function in feature extraction, and also more effectively combine global features to help complete the classification task. Experimental results show that our method significantly outperforms the baseline model, with 12% increase in Recall, 13% increase in F1, and 10% increase in Accuracy. Thus, we have verified the effectiveness of this method.

Keywords

PAN 2025, Human-AI Collaborative Text Classification, DeBERTa-V3, Feature Fusion Pyramid

1. Introduction

The goal of the human-machine collaborative text classification task in PAN@CLEF2025 [1] is to classify text into six categories based on the nature of human and machine contributions: fully human-written, human-initiated and then machine-continued, human-abbreviated and then machine-polished, machine-written and then machine-humanized, machine-written and then human-edited, and deep hybrid text [2]. As the quality of machine-generated text is getting better and better, and it is comparable to human writing in terms of logic and vividness, existing classification methods usually do not work well in this task, and more effective methods are needed to complete the classification task.

In order to solve the problem of text writing classification and distinguish different writing types, this paper proposes DeBERTa-FPN, a new method that combines the DeBERTa-V3-Large pre-trained model and feature fusion pyramid to meet these challenges. As a pre-trained model, DeBERTa-V3-Large has strong context understanding ability and can capture detailed features in text. Feature Fusion Pyramid (FPN) is a multi-scale feature fusion module. This method was originally proposed in image classification tasks to link global and local features of images to improve classification efficiency. In this task, we use its multi-scale characteristics to associate features of texts of different lengths. By combining text features of different lengths, we effectively utilize global and local feature information, enhance the model's perception of features, and ensure efficient and accurate classification. By combining the two, our model can effectively complete the classification task in the human-machine collaborative text classification task.

To evaluate the effectiveness of DeBERTa-FPN, we tested it through the submission platform specified by PAN. The platform provides a strictly controlled testing environment to ensure fair and transparent benchmarking based on established benchmarks. This initial submission is crucial to evaluating the practicality of the model in real-world scenarios and improving its performance based on objective feedback. After this evaluation, DeBERTa-FPN performed well in multiple key indicators, with the Recall index of 54.49%, F1 index of 54.4%, and Accuracy of 62.89%. These results are significantly better than baseline models such as RoBERTa-base and DetectGPT, which shows the effect of DeBERTa-FPN in distinguishing

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

944376274@qq.com (Q. Sun); molly_917@163.com (L. Ma); cswyyang@fosu.edu.cn (W. Yang); 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); 2112453056@stu.fosu.edu.cn (M. Zheng)

human-machine collaborative text and the effectiveness of our method in solving human-machine collaborative text classification tasks.

2. Background

In recent years, text generation artificial intelligence tools have developed rapidly, and the generation effect has been significantly improved in logic and vividness. As the quality of text generated by LLMs (such as GPT-4, BERT, etc.) is close to human level, its wide application in education, medical care, news, law and other fields has also brought ethical issues, such as the spread of false information, academic plagiarism and public opinion manipulation. Therefore, it is urgent to develop efficient AI-generated text detection technology[3]. For the task of human-computer collaborative text classification, early human-computer text classification research mainly focused on binary classification tasks, that is, distinguishing whether the text is written by humans or generated by AI. Jinyan Su et al. proposed DetectLLM-LRR and DetectLLM-NPR for detecting machine-generated text[4]. LRR measures text features by the ratio of log-likelihood to log-rank, while NPR generates perturbed text by making small perturbations to the original text and then calculates the normalized log-rank mean of these perturbed texts. Both methods show good recognition results and can provide effective feedback in a short time.However, with the popularization of AI writing assistance tools, researchers gradually realized that this simple binary division can no longer meet actual needs, and gradually turned to the study of multi-class division[5]. In order to improve the classification effect, the Mindner team systematically evaluated 37 features, covering 8 categories including perplexity, semantics, list search, documents, error-based, readability, AI feedback and text vectors, all of which can be used as the basis for distinguishing human-machine writing[6]. Schaaff et al. used machine learning methods for multilingual AI-generated text detection and obtained good classification performance through multi-layer perceptron (MLP), but the detection performance for AI rewriting was poor[7]. In addition to traditional machine learning methods, large models have also been used in human-machine collaborative text classification. Kumar et al. used the BERT model to obtain text embeddings to compare the text similarity between people and between people and AI[8]. Sun et al. explored how to more effectively utilize the results of BERT model preprocessing[9]. In more extensive research, it was found that pre-trained models such as RoBERTa[10] and DeBERTa-V3[11] also played an important role in text classification tasks. They can be used as modules for text feature extraction to provide more effective text feature information[12]. However, pre-trained models often fail to focus on global information during feature extraction, and only focus on deep features, which may cause some useful features to be ignored. Feature fusion pyramid was first proposed for target detection tasks to fuse high-level semantic information with underlying detail features, significantly improving the detection accuracy of small targets[13]. This method has now been widely used in many fields to fuse multiple features to improve the effect of feature extraction. This study aims to enhance the perception ability of the pre-trained model for high-level and low-level features by using the feature fusion pyramid method, so as to achieve a more effective human-computer collaborative text classification model.

3. System Overview

3.1. Dataset and Preprocessing

This study used the "Human-machine Collaborative Text Classification Task" dataset provided by PAN@CLEF. This dataset is a publicly available dataset specifically designed to verify human-machine collaborative text classification. The PAN@CLEF dataset contains Multi-domain documents such as academic, news, and social media, Human-written and machine-generated samples, Collaborative texts with annotation layers for human/machine contributions, and these texts are in multiple languages. The PAN@CLEF dataset is usually organized into the following format, which contains text content, language type, label, data source, model type, and label type information:

```
{"text":"...","language":"...","label":0,"source_dataset":"...","model":"...","label_text":"fully
human-written"}
{"text":"...","language":"...","label":1,"source_dataset":"...","model":"...","label_text":"human-written,
then machine-polished"}
```

The validation set provided by the "Human-Computer Collaborative Text Classification Task" in PAN@CLEF is a key component for testing and optimizing the author's validation model. Its organization is consistent with the above. It should be mentioned that the label and label type are linked in the dataset, and 0-5 is used to represent different text types when predicting. Its comparison is as follows:

```
{ 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 (human + machine parts)",
  5: "machine-written, then human-edited"}
```

The sample sizes of the training set and development set are 288,918 and 72,661 respectively. The specific number of categories is shown in Table 1. This study uses the PAN@CLEF dataset to train and evaluate a hybrid model that combines DeBERTa-V3-Large and FPN, aiming to improve the accuracy of human-machine collaborative text classification.

Table 1
Distribution of samples in training set and development set

Label Category	Train	Dev
Machine-humanized	91 232	10 137
Machine-polished	95 398	12 289
Human-written	75 270	12 330
Machine-continued	10 740	37 170
Deeply-mixed	14 910	225
Human-edited	1 368	510
Total	288 918	72 661

Specifically, this dataset helps us systematically understand and identify the characteristics of different types of human-machine collaborative text. Through a series of experiments on the PAN@CLEF dataset, we evaluate the performance of the model in distinguishing six types of human-machine collaborative text classification. We use multiple evaluation metrics such as precision, recall, and F1 score to comprehensively analyze the effectiveness of the model. The results show that our model can achieve satisfactory performance when dealing with this specific task.

3.2. Network Architecture

In our study, we designed and implemented a hybrid neural network model named DeBERTa-FPN that combines DeBERTa-V3-Large and FPN to perform complex human-machine collaborative text classification, aiming to distinguish six types of text with different machine participation. The structure of the model is shown in Figure 1. The model architecture aims to make full use of the deep semantic processing capabilities of DeBERTa-V3-Large and the global information integration and extraction capabilities of the FPN module to enhance the model's performance in handling fine-grained text analysis tasks. We use the pre-trained DeBERTa-V3-Large model provided by Hugging Face as the pre-processing module for text information. This pre-trained model has a 24-layer Transformer, which can process long texts, more accurately model inter-word dependencies, capture long-distance dependencies in text, and obtain effective features for classification. In order to better combine global and detailed features, we use FPN as a feature enhancement module. FPN can fuse feature information at different levels in DeBERTa-V3-Large, enhance the model's ability to understand global information, improve the utilization of features, and provide more effective features for the classifier. For our model, we used the features of the 12th and 18th layers for fusion. These two levels are chosen because they not only ensure the adequacy of feature extraction, but also combine feature information at different levels to provide effective information for further fusion. They were processed by Conv1d for deep feature extraction and concatenated before being passed to the Max Pooling layer to enhance key features. Finally, the features were passed to the classifier, where a dropout layer was added to prevent overfitting. Finally, the features were mapped to the classification results through a fully connected layer to achieve classification of six types of human-machine collaborative texts.

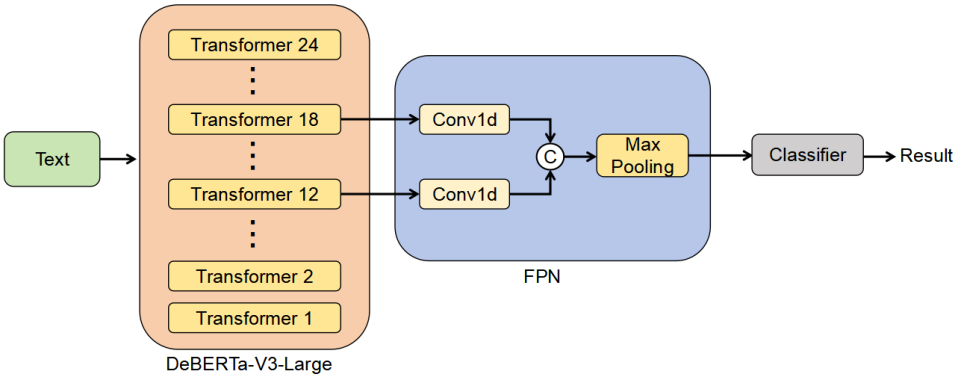

Figure 1: DeBERTa-FPN Architecture. DeBERTa-V3-Large is used to model the dependency relationship between words to obtain text features, and FPN is used to extract deep features and integrate feature information.

4. Experiments and Results

4.1. Experiment Settings

In this study, we used the pre-trained DeBERTa-V3-Large model as the basis for text feature extraction. Its large resource consumption also provides very impressive text feature extraction capabilities, so it is necessary to select the most effective training scheme in combination with time and computing resources. During the training process, we selected a batch size of 2, a learning rate of 1e-5, and a total of 3 training cycles to ensure that the model can effectively learn text features and avoid overfitting. In addition, we used the AdamW optimizer, which was selected for its optimization effect in deep learning model training, especially in dealing with gradient sparsity and weight decay. To ensure the repeatability of the experiment, we set a fixed random seed, and all experiments were conducted in a computing environment equipped with NVIDIA GeForce RTX 3090 and Intel(R) Core(TM) i9-10900K.

The main experimental process includes three stages: data preparation, model training, and performance evaluation. First, in the data preparation stage, the dataset is preprocessed, including text cleaning, label correspondence, etc. In addition, the dataset is divided into a training set and a validation set. During the model training phase, the model is iteratively learned on the training set. At the end of each cycle, we evaluate the performance of the model on the validation set to monitor whether there is overfitting during the training process. Finally, in the performance evaluation phase, we use standard classification metrics such as accuracy, recall, and F1 to evaluate the model. We pay special attention to the performance of the model on an independent test set to verify its generalization ability in practical applications. Through this series of meticulous and rigorous experimental processes, we ensure the accuracy and practicality of the research results.

4.2. Results

In order to comprehensively evaluate the performance of our proposed DeBERTa-FPN, we selected a series of indicators, including F1, recall, and Accuracy. These indicators not only reflect the overall performance of the model, but also provide different performance evaluation perspectives to help us understand the performance of the model in specific aspects.

The specific performance metrics are as follows:

Accuracy represents the proportion of samples correctly predicted by the model to the total samples, which can intuitively reflect the overall prediction accuracy. It is calculated as:

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

Among them, TP stands for True Positive, that is, the number of correctly judged positive examples, TN stands for True Negative, that is, the number of correctly judged negative examples, FN stands for False Negative, that is, the number of positive examples that are not judged correctly, and FP stands for False Positive, that is, the number of negative examples that are mistaken for positive examples.

Recall represents the proportion of samples that are actually positive and are correctly predicted. It measures the ability of the model to predict correctly. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

F1 score is the harmonic mean of precision and recall.It is calculated as:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Where

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Through these evaluation indicators, we can learn to evaluate the overall performance of the model and conduct in-depth analysis of the specific capabilities of the model from multiple dimensions. These evaluation indicators can also objectively measure the performance difference between our proposed model and other models, ensure the comprehensiveness and reliability of the evaluation results, and lay a solid foundation for future model optimization and application. Through the above indicators, we compare with the baseline model. Compared with the baseline model, our model achieved 54.49%, 54.40%, and 62.89% in Recall, F1 Score, and Accuracy, respectively, which are 12%, 13%, and 10% higher than the baseline model. The evaluation results under the evaluation set are shown in Table 2.

Table 2
Evaluation Results

Approach	Accuracy (%)	F1 Score (%)	Recall (%)
Baseline	57.09	47.82	48.32
DeBERTa-FPN	**62.89**	**54.40**	**54.49**

5. Conclusion

In this study, we proposed and implemented a hybrid neural network model that combines DeBERTa-V3-Large and FPN, aiming to improve the accuracy of human-machine collaborative text classification. The results show that the Deberta-FPN hybrid model proposed in this study has achieved satisfactory results in text classification tasks. It has achieved impressive results in multiple indicators. Compared with the baseline model, the Recall indicator is improved by 12%, the F1 indicator is improved by 13%, and the Accuracy is improved by 10%, which shows the effectiveness of the model structure and can achieve excellent performance in related tasks. At the same time, we also know that there are still many aspects of this model that can be improved. In future work, we will continue to improve the method and strive to obtain better results in the task of human-machine collaborative text classification.

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

During the preparation of this work, the authors used DeepSeek-R1 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed. full responsibility for the publication’ s content.

References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, AndréGreiner-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: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[2] 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.

[3] S. Fariello, G. Fenza, F. Forte, M. Gallo, M. Marotta, Distinguishing human from machine: A review of advances and challenges in ai-generated text detection, International Journal of Interactive Multimedia and Artificial Intelligence 8 (2024) 1–12.

[4] J. Su, T. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 12395–12412.

[5] H. Abburi, S. Bhattacharya, E. Bowen, N. Pudota, Ai-generated text detection: A multifaceted approach to binary and multiclass classification, arXiv preprint arXiv:2505.11550 (2025).

[6] L. Mindner, T. Schlippe, K. Schaaff, Classification of human-and ai-generated texts: Investigating features for chatgpt, in: International conference on artificial intelligence in education technology, Springer, 2023, pp. 152–170.

[7] K. Schaaff, T. Schlippe, L. Mindner, Classification of human-and ai-generated texts for english, french, german, and spanish, arXiv preprint arXiv:2312.04882 (2023).

[8] S. Kumar, S. Tiwari, R. Prasad, A. Rana, M. Arti, Comparative analysis of human and ai generated text, in: 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2024, pp. 168–173.

[9] G. Sun, W. Yang, L. Ma, Bcav: a generative ai author verification model based on the integration of bert and cnn, Working Notes of CLEF (2024).

[10] 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).

[11] 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).

[12] 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, arXiv preprint arXiv:2403.03506 (2024).

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.