DBG: Human-AI Collaborative Text Classification with DeBERTa-enhanced Contextual and Geometric Attention

Notebook for the PAN at CLEF 2025

Tufeng Xian, Yong Zhong*, Fen Liu, Meifang Xie, Qiyuan Sun, Miaoji Zheng, Weidong Wu and Zhiliang Zhang

Foshan, University, Foshan, China

Abstract

With the rapid development of generative large language models (LLMs), hybrid texts generated by human-AI collaboration are becoming increasingly common in real-world scenarios. Their detection and classification have become key challenges to ensure information authenticity and academic integrity. This paper proposes a three-level features fusion model called DBG (DeBERTa-BiLSTM-Geometric Attention) for the complex human-AI collaborative text classification problem in the PAN-CLEF 2025 task. The model achieves fine-grained classification through a three-stage collaborative mechanism: first, the decoupled attention of the DeBERTa-v3-large pre-trained model is used to capture global semantic features. Second, the bidirectional language dependency pattern in the text sequence is modeled through a bidirectional LSTM. Finally, the geometric attention module is innovatively introduced, combining one-dimensional convolution with a learnable position enhancement factor to dynamically enhance local discriminative features. Experiments show that DBG significantly outperforms the baseline model in six types of hybrid text classification tasks, with Macro Recall reaching 56.87% (an increase of 8.55% over RoBERTa-base), and F1 Macro Score and Accuracy increased to 56.45% and 66.81% respectively. This study provides technical support for risk prevention and control of human-machine collaborative content, and provides new ideas for the text traceability problem in generative AI ethical governance.

Keywords

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

CEUR-WS.org/Vol-4038/paper_331.pdf

1. Introduction

With the rapid development of computing chip GPU hardware, computing power continues to improve, driving the rapid development of artificial intelligence. In the era of big data, there is a huge amount of data, which is enough to build a large data set, making it possible to scalably train language models, which has completely changed the field of natural language processing. The advent of Large Language Models (LLMs) has revolutionized traditional statistical machine learning methods. These models demonstrate superior capabilities in tackling diverse Natural Language Processing (NLP) tasks. Moving beyond earlier tasks like spam detection, sentiment analysis, automatic response, and academic text summarization, generative large language models, trained on vast amounts of text data, are now capable of generating fluent and coherent text [1, 2, 3, 4, 5]. In recent years, LLMs such as GPT-3/4, PaLM, and Claude have seen rapid technological progress, with the quality of the generated text approaching and even surpassing that of human writing. The GPT-3 model released by OpenAI in 2020, with its huge scale of 175 billion parameters, can generate logically coherent long articles, poems, codes and other content based on simple instructions; and the GPT-4 launched in 2023 has made breakthroughs in multimodal understanding and complex reasoning capabilities, further blurring the boundaries between human and machine-generated content. According to Stanford University's "2023 Artificial Intelligence Index Report", GPT-4 has performed better than 90% of human candidates in professional exams (such as the bar exam), and the semantic rationality and style diversity of its generated text have reached the level of being indistinguishable from the real thing. This progress has led to the development of

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

2112453044@stu.fosu.edu.cn (T. Xian); zhongyong@fosu.edu.cn (Y. Zhong); liufen@fosu.edu.cn (F. Liu); 2112453050@stu.fosu.edu.cn (M. Xie); 2112453029@stu.fosu.edu.cn (Q. Sun); 2112453056@stu.fosu.edu.cn (M. Zheng); 2112453039@stu.fosu.edu.cn (W. Wu); 2112453043@stu.fosu.edu.cn (Z. Zhang)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

large language models (LLMs), which has greatly changed the way people generate and interact with machine text. LLMs have demonstrated a strong ability to generate text that matches human writing.

To enhance the performance of generative AI detection for large language models and improve classification accuracy in subtask 2's human-machine collaborative text classification task, the complexity inherent in the six distinct classification categories necessitates the development of a more robust model architecture. Therefore, we propose a novel model for multi-accumulated text classification called DBG, which cleverly combines advanced pre-training, sequence modeling and geometric attention mechanism to address the key challenges of language analysis. DBG is based on DeBERTa-v3-large, and uses its decoupled attention to capture subtle contextual relationships in text sequences. To enhance sequential pattern learning, we integrate a bidirectional LSTM layer to process the representation encoded by DeBERTa[6], thereby effectively modeling forward and backward language dependencies. The architecture also integrates a novel geometric attention module with a learned position enhancement factor[7], which uses one-dimensional convolution to dynamically highlight discriminative text features while preserving structural relationships within the sequence. The entire model exhibits synergistic effects through three core mechanisms:

(1) DeBERTa's powerful contextual embedding builds a semantic foundation.

(2) BiLSTM layers model hierarchical language patterns.

(3) Geometric attention mechanism optimizes feature extraction through learnable spatial reasoning.

Experimental verification on benchmark datasets shows that our model achieves significant performance improvements, especially macro recall (baseline 48.32% vs ours 56.87%). In addition, F1, recall, is also significantly higher than the baseline model, highlighting that the DBG model has a strong human-machine collaborative classification capability and is believed to be able to effectively solve the complex problems in generative AI author authentication.

2. Related Work

Despite their revolutionary potential, the technological advancements of Large Language Models (LLMs) are accompanied by significant societal risks: the misuse of machine-generated text exhibits exponential growth, manifesting as critical issues including opinion manipulation, disinformation propagation, and systematic plagiarism. This underscores the imperative to establish ethical constraints and risk mitigation frameworks alongside technological deployment. Against this backdrop, Generative AI Authorship Verification has emerged as a critical research focus. Notably, the Voight-Kampff task subtrack for human-AI collaborative text classification at PAN@CLEF 2025 demonstrates particular foresight—by defining six categories of human-AI collaboration patterns, this initiative transcends the limitations of conventional binary classification. It not only directly addresses detection blind spots for hybrid text but also provides multidimensional support for ethical governance and technical optimization[5]. Current mainstream techniques can be categorized into three primary approaches: Jakesch et al.[8] revealed through cognitive experiments that humans' reliance on surface-level features (e.g., first-person pronouns, domestic topics) for text attribution constitutes a fundamental flaw, demonstrating that such heuristic rules are vulnerable to exploitation by "hyper-humanized" generated content. Chakraborty [9] leveraged BERT's contextual embeddings to capture implicit patterns in generated text, while Guo et al.[10] innovatively integrated Transformer encoders with spaCy-derived multi-scale textual features, enhancing long-text semantic coherence analysis via BiLSTM. Nevertheless, current research exhibits notable limitations, First, while mainstream methods achieve high accuracy (>89%) for purely machine-generated text[11], their performance plummets by 37-52% when detecting human-AI hybrid documents[5], with particularly low F1-scores (merely 0.63) in fine-grained collaborative category classification. Second, supervised learning approaches generally suffer from model dependency, exhibiting up to 28 percentage points of accuracy degradation when detecting outputs from closed-source models (e.g., GPT-4, Claude 3). Third, existing systems struggle to adapt to LLMs' rapid evolution—novel RLHF alignment techniques have increased human-like feature density in generated text by 42%, while multimodal content exceeds current detection frameworks' scope. Fourth, surface-

feature-based strategies incur false positive rates as high as 65% in specialized domains[9], exposing deficiencies in semantic depth analysis mechanisms. These bottlenecks compellingly demonstrate the necessity for novel verification frameworks specifically designed for hybrid text characteristics and adaptive to dynamic technological evolution.

3. Methodology

This paper proposes the DBG model (DeBERTa-BiLSTM-Geometric Attention) to perform multi-dimensional feature modeling for generative text classification tasks. The model architecture is based on the DeBERTa-v3-large pre-trained language model, and builds deep contextual representations through the decoupled attention mechanism. A bidirectional LSTM network is connected after the encoding layer to capture the bidirectional language dependency pattern in the sequence. The geometric attention module is innovatively introduced, and the spatial enhancement factor of text features is dynamically learned using a one-dimensional convolution kernel. The extraction of discriminative features is optimized through learnable local position weights. The three-layer architecture forms a synergistic mechanism: DeBERTa provides global semantic representation, BiLSTM models sequence pattern evolution, and geometric attention strengthens the discriminative power of local features. The joint optimization of the three significantly improves the fine-grained classification ability of generated text.

3.1. Dataset Preprocessing

This experiment is based on the PAN 2025 human-machine collaborative text classification task dataset, which is publicly provided by the Zenodo platform and covers academic papers, news reports, social media and other multi-field texts. Data annotation includes six types of human-machine collaboration (see Table 1 for details), covering complex scenarios such as human-led creation, machine-generated post-processing, and deep mixed text. The original data is stored in JSONL format. Each piece of data contains text content (the ‘text‘ field) and category labels (the ‘label‘ field), supporting multilingual analysis in English, Spanish and German. The type label is an int of 0-5, that is, [0, 1, 2, 3, 4, 5], and the corresponding text description is as follows:

id2label = {
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 sample sizes of the training set and the development set are 288,918 and 72,661 respectively, and the category distribution shows a significant imbalance. For example, the "Human-initiated, then machine-continued" category accounts for 51.2% of the development set, while "Deeply-mixed text" accounts for only 0.3% (see Table 1 for details). This distribution characteristic reflects the diversity of human-machine collaboration modes in real scenarios, but it is not a good thing for us to use large models to solve human-machine collaboration text classification tasks. In order to obtain better classification accuracy, sometimes we need to think about how to solve the problem of data imbalance.

In order to better input data into the model to train our model, we preprocess the data as follows:

- **Label normalization**: Map the category label to an integer value of 0-5, and use the id2label dictionary to maintain semantic interpretability. For possible label missing, set invalid labels to -1 and filter abnormal samples.

- **Text segmentation**: Use DebertaV2Tokenizer dedicated to DeBERTa-v3-large to segment the text into subwords. Set the maximum sequence length to 512, truncate the tail of overlong text, and fill the insufficient part with [PAD] tags.
- **Batch encoding**: Generate a fixed-length tensor through the truncation = True and padding = "max_length" parameters, and construct input_ids and attention_mask as model input.

This preprocessing process effectively retains the semantic and structural features of the text, while adapting to the input specifications of the pre-trained model, providing structured data support for subsequent multi-task learning.

Table 1
Distribution of samples in training set and development set

LABEL CATEGORY	TRAIN	DEV
Machine-written,then machine-humanized	91,232	10,137
Human-written, then machine-polished	95,398	12,289
Fully human-written	75,270	12,330
Human-initiated,then machine-continued	10,740	37,170
Deeply-mixed text (human + machine parts)	14,910	225
Machine-written, then human-edited	1368	510
Total	288,918	72,661

3.2. Network Architecture

The traditional human-machine collaborative text classification methods have the following main shortcomings: reliance on artificial feature engineering and single-modal encoding leads to insufficient capture of complex patterns of generated text, one-way sequence modeling makes it difficult to parse traces of two-way collaboration, and the static attention mechanism lacks dynamic enhancement of local key features, resulting in limited fine-grained classification performance. The DBG model proposed in this paper adopts a three-level feature enhancement architecture, as shown in Figure 1. The model extracts global semantic features through the DeBERTa pre-trained encoder, captures sequence pattern dependencies through bidirectional LSTM, and strengthens local discriminative features through geometric attention. The three-stage together to improve the fine-grained classification capabilities of human-machine collaborative text.

3.2.1. DeBERTa Disentangled Attention Encoder

As the model base, it is responsible for extracting deep contextual semantic representations from the input text and solving the semantic ambiguity problem of the generated text. We use the DeBERTa-v3-large pre-trained model, whose core is the disentangled attention mechanism, which separates the content and position encoding calculations:

$$\text{Attn}(Q_c, K_c, V_c, P) = \text{Softmax}\left(\frac{Q_c K_c^T + Q_r P^T}{\sqrt{d}}\right) V_c \tag{1}$$

Where Qc, Kc are content vectors, Qr is the relative position vector, and P is a trainable position embedding matrix. This design enables the model to independently model the relative position relationship between words and enhances the ability to capture long-range dependencies in the generated text. Output hidden state$\mathbf{H}_d \in \mathbb{R}^{L\times 1024}$, where L is the sequence length and 1024 is the DeBERTa hidden layer dimension.

3.2.2. Projection Dimensionality Reduction and Bidirectional LSTM

The feature dimension is reduced by flattening to reduce computational complexity, and local language patterns are captured by bidirectional temporal modeling to solve the sequence coherence differences of generated text.

(1) Projection layer: DeBERTa output is mapped to a low-dimensional space:

$$\mathbf{H}_p = \mathbf{W}_p \mathbf{H}_d + \mathbf{b}_p \quad \in \mathbb{R}^{L\times 256} \tag{2}$$

Among them,$\mathbf{W}_p \in \mathbb{R}^{1024\times 256}$is a trainable parameter, and the core semantic information is retained after dimensionality reduction.

(2) BiLSTM:

$$\overrightarrow{\mathbf{h}_t} = \text{LSTM}\left(\mathbf{h}_p^t, \overrightarrow{\mathbf{h}_{t-1}}\right) \tag{3}$$

$$\overleftarrow{\mathbf{h}_t} = \text{LSTM}\left(\mathbf{h}_p^t, \overleftarrow{\mathbf{h}_{t+1}}\right) \tag{4}$$

By concatenating the forward $\overrightarrow{\mathbf{h}_t}$ and the backward $\overleftarrow{\mathbf{h}_t}$ hidden states, we obtain the bidirectional temporal feature $\overrightarrow{\mathbf{H}_l} = \left[\overrightarrow{\mathbf{H}_l}\,;\,\overrightarrow{\mathbf{H}_l}\right] \in \mathbb{R}^{L\times 512}$, which effectively models the contextual dependencies in the generated text.

3.2.3. Geometric Attention Enhancement Module

By focusing on key local features through learnable spatial weights, the problem of hidden feature extraction of machine-generated fragments in generated text is solved.

(1) One-dimensional convolution transformation: feature space transformation of bidirectional LSTM output:

$$Q = \text{Conv1D}(H_l, W_Q), \quad K = \text{Conv1D}(H_l, W_K) \tag{5}$$

(2) Location enhances attention:

$$A = \text{Softmax}\left(\sigma\left(\frac{QK^\top}{\sqrt{d}} \odot \eta\right)\right) \tag{6}$$

σ is the SiLU activation function, which dynamically adjusts the attention distribution; $\eta \in \mathbb{R}$ is a learnable position enhancement factor, which amplifies the weight contribution of important positions

(3) Residual connection:

$$H_a = \text{Conv1D}(AV) + H_l \tag{7}$$

The original sequence structure information is retained to avoid feature shift caused by the attention mechanism.

3.2.4. Classification decision layer

The final classification is completed based on the aggregated features of the [CLS] tag, and the global representation is used to distinguish complex collaborative patterns. The first position vector (corresponding to the [CLS] tag) of the geometric attention output is taken as the global feature:

$$y = W_c H_a^{(0)} + b_c \in \mathbb{R}^6 \tag{8}$$

Where $W_c \in \mathbb{R}^{512\times 6}$ maps the features to 6 category spaces. Practical analysis shows that the [CLS] tag can effectively aggregate the collaborative mode features of the full text after multi-layer attention transfer.

This architecture achieves multi-granular modeling of traces of human and machine collaboration in generated text through a hierarchical feature refinement mechanism. Experiments have verified that its classification performance on different collaboration modes is significantly better than the baseline model (see Chapter 3).

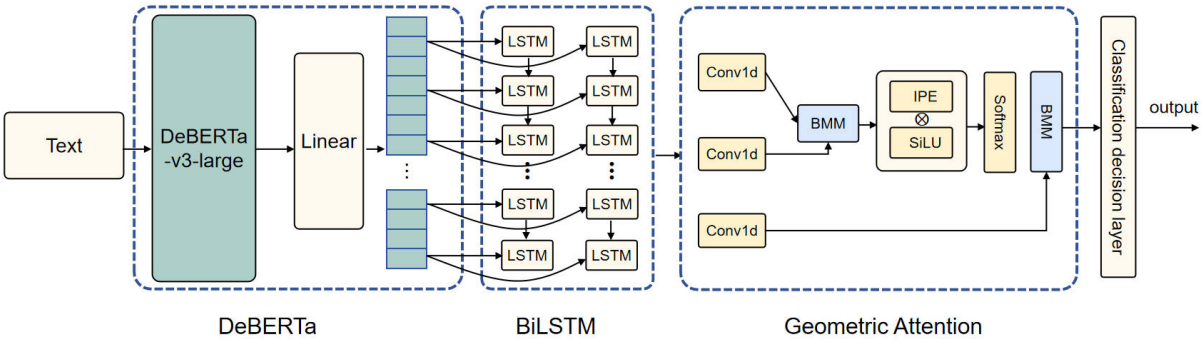

Figure 1: DBG Architecture

4. Experiments and Results

4.1. Experiment Settings

In the training experiment of DBG, we trained the model on the training data of the official train, and used dev as the validation data of the model. The ratio of train.jsonl to dev.jsonl is about 8:2. This model uses the DeBERTa-v3-large pre-trained model to build the classification framework, setting the projection layer dimension to 256, and the bidirectional LSTM adopts a single-layer bidirectional structure. We apply layer normalization between LSTM layers and set the recurrent connection drop rate to 0.3. The geometric attention module contains 8 parallel convolutional layers. The model adopts a hierarchical learning rate strategy in the training phase. The pre-trained DeBERTa uses a learning rate of 1e-5, and the newly added module uses a learning rate of 1e-4. The model is trained for 5 epochs in total, and mixed precision training is performed using the AdamW optimizer. The batch is set to 8 and 4 steps of gradient updates are accumulated, and global gradient clipping (threshold 1.0) and early stopping mechanism (development set loss is terminated after 10 rounds of no improvement) are used. In data processing, the text content is unified to a length of 512 tokens, and DebertaV2Tokenizer is used for word segmentation. The experiment was run on a single RTX 3090 GPU, with a fixed random seed of 42 to control parameter initialization and data shuffle, and a full training cycle of 20 hours.

In the test phase, we used the model saved in the training phase to reason on the unlabeled test set (`subtask2_test.jsonl`) provided by the CodaLab platform. First, the original text was standardized into an input sequence of 512 tokens in length through DebertaV2Tokenizer. After loading the optimal training model, full batch reasoning was performed on the `RTX 3090 GPU` with a batch size of 16. Mixed precision calculation and `torch.no_grad` mode were enabled to accelerate the prediction process. After obtaining the category probability distribution of each sample through `argmax`, the integer prediction label (range 0–5) was strictly bound to the original test sample ID in the key-value pair format of `"id": "label"`, and written into a new result JSONL file (such as `predictions.jsonl`) by line. This process took about 1 hour. Finally, the `format_checker.py` script provided by the official was used to verify the text format, and finally a standardized submission file that meets the requirements of the CodaLab evaluation platform was generated.

4.2. Metrics

According to the particularity of this human-machine collaborative classification task, the official defines four indicators, namely Macro Recall, F1 Macro Score, F1 Micro Score and Accuracy. Among them, Macro Recall is the most important and is the primary indicator of the ranking list. These indicators are good for evaluating the performance of the model and the degree of completion of the task. The following section will introduce these indicators in detail.

Macro Recall: The arithmetic mean of the recall rates of all categories, treating the recognition

ability of each category equally[12]. Its formula is defined as follows:

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \quad (C = 6) \tag{9}$$

Among them, TP_i represents the number of true positive samples (correctly predicted samples) of the ith category; FN_i represents the number of false negative samples (missed samples) of the ith category

F1 Macro Score: The arithmetic mean of the F1 scores of all categories, which comprehensively balances the precision and recall[13]. Its formula is defined as follows:

$$\text{F1 Macro} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{10}$$

Among them, $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$, $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$. Relying on the characteristics of macro-average, the F1 value of each category is calculated independently and then averaged to avoid large categories dominating the results.

F1 Micro Score: The F1 value calculated based on global statistics (the sum of TP/FP/FN of all categories) reflects the performance dominance of high-frequency categories[14]. Its formula is defined as follows:

$$\text{F1 Micro} = \frac{2 \cdot \sum_{i=1}^{C} TP_i}{2 \cdot \sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} (FP_i + FN_i)} \tag{11}$$

Accuracy: The ratio of correctly predicted samples to the total number of samples measures the overall classification performance[15]. Its formula is defined as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{N_{\text{total}}} \quad (N_{\text{total}} = \text{Total number of samples}) \tag{12}$$

These four indicators effectively quantify the effectiveness of DBG in the human-machine collaborative text classification task through the combined effect of a comprehensive evaluation model. Accuracy focuses on category balance, Macro Recall focuses on full category coverage, F1 Macro Score focuses on fine-grained classification stability, and F1 Micro Score focuses on the dominance of high-frequency categories. A multi-dimensional evaluation system covering scenarios with uneven data distribution is comprehensively constructed to effectively evaluate the model's capabilities in this classification task.

4.3. Results

The model we proposed performed well in the PAN-CLEF 2025 Subtask 2: Human-AI Collaborative Text Classification task, and was significantly higher than the baseline model in four key indicators, demonstrating its effectiveness and competitiveness in this task. As shown in Table 2, our model achieved a significant breakthrough in the official indicator Macro Recall, reaching 56.87%, which is 8.55% higher than the 48.32% of RoBEta-base, indicating that it has a higher discriminative ability in this human-machine collaborative text classification; in addition, the F1 Macro Score score is 56.45%, which significantly exceeds the baseline model, reflecting its stability in this classification task. The Accuracy score is 66.81%, which is 9.72% higher than the 57.09% of the baseline model, reflecting the model's robust performance in dealing with the class imbalance problem in this classification task.However, through our testing, the average single sample inference delay of the 512 token sequence on the RTX 3090 GPU for DBG is * * 1200ms * * (standard deviation 85ms), with a peak memory usage of 10GB. In the future, knowledge distillation technology will be used to improve inference speed.

In terms of competition ranking, our entry ranked 5th in the official leaderboard of PAN-CLEF 2025 Subtask 2 (a total of 22 contestants). This ranking fully demonstrates the competitiveness of DBG and its excellent classification ability in this task among many excellent entries and different types of participating models. It can be said that it has excellent discrimination ability in complex

classification tasks such as multi-class text classification and single text content with mixed and varied content.However, through our tests, DBG's single-sample inference latency for a 512-token sequence on an RTX 3090 GPU averaged 1200ms (standard deviation 85ms), and the peak memory usage reached 10GB. In the future, knowledge distillation technology will be used to improve inference speed.

Table 2
Performance on PAN-CLEF 2025 Subtask 2

Approach	Macro Recall(%)	F1 Macro Score(%)	Accuracy(%)
Baseline RoBERTa-base	48.32	47.82	57.09
DBG(ours)	**58.87**	**56.45**	**66.81**

5. Conclusion

In this paper, a multi-level classification framework DBG combining pre-trained language model, sequence modeling and dynamic attention enhancement is proposed to solve the classification problem of generative AI and human collaborative text. Through the deep semantic encoding of DeBERTa, the bidirectional sequence dependency modeling of BiLSTM and the local feature extraction of geometric attention, the model effectively solves the problem of hidden feature extraction of human-machine collaboration traces in mixed text. Experiments verify the robustness of DBG in scenarios with highly uneven data distribution. It performs well in various indicators, with a Macro Recall score of 56.87%, an F1 Macro Score score of 56.45%, and an Accuracy score of 66.81%, all of which exceed the benchmark. This achievement provides a technical solution for practical needs such as authenticity verification of AI-generated content and academic misconduct detection, and lays the foundation for multimodal human-machine collaborative analysis. In future work, we will further improve DBG, including optimizing its hyperparameters, enhancing data preprocessing, improving deep feature fusion capabilities, and exploring the combination of other methods to improve system performance and detection accuracy. At the same time, we will also explore the expansion of cross-language transfer capabilities and the optimization of the explainability of attention mechanisms to meet more complex generative AI governance challenges.

6. 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.

7. 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).

References

[1] 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: 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] G. Sun, W. Yang, L. Ma, BCAV: A Generative AI Author Verification Model Based on the Integration of Bert and CNN, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] Z. Wu, W. Yang, L. Ma, et al., BertT: A Hybrid Neural Network Model for Generative AI Authorship Verification, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[4] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, B. I. A. Agyemang, What if the Devil is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education, Smart Learning Environments 10 (2023) 15. doi:10.1186/s40561-023-00237-x.

[5] M. Alier, F. J. García-Peñalvo, J. D. Camba, Generative Artificial Intelligence in Education: From Deceptive to Disruptive, International Journal of Interactive Multimedia and Artificial Intelligence 8 (2024) 5–14. Special issue on Generative Artificial Intelligence in Education.

[6] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-Enhanced BERT with Disentangled Attention, arXiv preprint (2020). URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[7] Y. Wang, S. Li, T. Wang, et al., Geometric Transformer with Interatomic Positional Encoding, in: Advances in Neural Information Processing Systems, volume 36, 2023, pp. 55981–55994.

[8] M. Jakesch, J. T. Hancock, M. Naaman, Human Heuristics for AI-Generated Language Are Flawed, Proceedings of the National Academy of Sciences of the United States of America 120 (2023) e2208839120. URL: https://www.pnas.org/doi/10.1073/pnas.2208839120. doi:10.1073/pnas.2208839120.

[9] U. Chakraborty, J. Gheewala, S. Degadwala, et al., Safeguarding Authenticity in Text with BERT-Powered Detection of AI-Generated Content, in: 2024 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2024, pp. 34–37.

[10] L. Guo, W. Yang, L. Ma, J. Ruan, BLGAV: Generative AI Author Verification Model Based on BERT and BiLSTM, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes of CLEF 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 2585–2592. URL: https://downloads.webis.de/pan/publications/papers/guo1_2024.pdf.

[11] 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.

[12] A. Berger, S. Guda, Threshold Optimization for F Measure of Macro-Averaged Precision and Recall, Pattern Recognition 102 (2020) 107250. doi:10.1016/j.patcog.2020.107250.

[13] K. Takahashi, K. Yamamoto, A. Kuchiba, et al., Confidence Interval for Micro-Averaged F1 and Macro-Averaged F1 Scores, Applied Intelligence 52 (2022) 4961–4972. doi:10.1007/s10489-021-02635-5.

[14] A. U. Rehman, K. Javed, H. A. Babri, Feature Selection Based on a Normalized Difference Measure for Text Classification, Information Processing & Management 53 (2017) 473–489. doi:10.1016/j.ipm.2016.12.004.

[15] D. L. Streiner, G. R. Norman, "Precision" and "Accuracy": Two Terms That Are Neither, Journal of Clinical Epidemiology 59 (2006) 327–330. doi:10.1016/j.jclinepi.2005.09.005.