Genre-Aware Contrastive Learning for AI Text Detection: A RoBERTa-Based Approach

Notebook for the PAN Lab at CLEF 2025

Junlong Yang, Kai Yan

Foshan University, Foshan, China

Abstract

This paper presents a detection method for the PAN 2025 Voight-Kampff Generative AI Detection task, which combines RoBERTa with a contrastive learning mechanism. Building on RoBERTa, we introduce genre embeddings and contrastive loss to enhance the model's sensitivity to the textual genre and semantic differences. Experimental results show that our method achieves excellent performance on the official validation set with genre information, and also performs robustly on a custom validation set without genre labels, demonstrating strong generalization capabilities. Our findings validate the effectiveness of integrating deep semantic modeling with structured representation learning ,offering a novel approach to generative text detection.

Keywords

Authorship verification, RoBERTa, Contrastive learning, Text classification

1. Introduction

In recent years, the rapid advancement of natural language processing technologies has enabled large language models, such as GPT-4, Gemini, and LLaMA, to generate fluent, coherent, and semantically rich text [1, 2, 3, 4]. While these models have transformed areas such as content creation and intelligent customer service, they also pose potential risks, including misinformation dissemination, academic dishonesty, and opinion manipulation. For example, students may use AI-generated content for assignments, and malicious actors may generate large volumes of fake news using AI.

Against this backdrop, distinguishing AI-generated content from human-written text has become a critical challenge for both academia and industry. To promote progress on this problem, the PAN 2025 shared task [5] introduced the Voight-Kampff Generative AI Detection Task [6], which requires participants to classify whether a given text is human-written or AI-generated based on its linguistic properties. All models are evaluated through the TIRA platform [7], which ensures standardized and reproducible experimental setups.

In this study, we propose a detection framework that integrates genre embedding and contrastive learning for the PAN 2025 Voight-Kampff task. The task requires predicting a confidence score for each input text: scores below 0.5 are classified as human-written, scores above 0.5 as AI-generated, and exactly 0.5 as undecidable.

2. Related Work

As the quality of AI-generated text improves, distinguishing it from human-authored content has emerged as a major research focus. Traditional detection methods often rely on handcrafted features such as statistical indicators, stylistic differences, or perplexity scores. For example, GLTR uses a language model to compute word probabilities and assess whether the text aligns with natural language patterns[8, 9]. Similarly, OpenAI's Text Classifier models output distributions from GPT to detect anomalous word choices.

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

1836257095@qq.com (J. Yang); yankai@fosu.edu.cn (K. Yan)

0009-0008-4507-5969 (J. Yang); 0000-0002-4960-7108 (K. Yan)

However, such approaches often depend heavily on the internal mechanics of the language model and tend to generalize poorly, especially when applied across different domains and genres[10, 11, 12]. As a result, researchers have turned to discriminative deep learning methods. Pretrained models like BERT and RoBERTa have been widely adopted, showing strong performance in binary classification tasks due to their deep semantic modeling capabilities[3, 13, 14, 15].

Contrastive learning has also gained traction in NLP as a means of learning better representations, drawing inspiration from vision-based models like SimCLR and MoCo[10, 11]. Approaches such as SimCSE, CoSent, and CPT have proven effective in semantic matching and representation learning[16, 17, 18]. In the context of generative text detection, some studies incorporate contrastive learning by treating samples from the same genre or model as positive pairs to optimize embedding space structures or by including stylistic features in classification[19, 18, 20] .

Despite these advances, the integration of genre labels with representation learning remains underexplored. Our work addresses this gap by combining RoBERTa, genre embeddings, and contrastive learning within a unified framework and empirically validating its effectiveness and generalization capabilities on the PAN 2025 dataset.

3. Method

Our method consists of two main components: a RoBERTa-based classifier and a contrastive learning module.

3.1. RoBERTa Text Classifier

We adopt RoBERTa-large as the encoder to extract deep semantic representations of input texts[3].For each input, the vector at the [CLS] position is used as the global representation. This representation is passed through a dropout layer and a linear classifier to produce a scalar output for binary classification, trained using the Binary Cross-Entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\left[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right] \tag{1}$$

To enable contrastive learning, we add a projection head and incorporate genre information as an embedding vector. Specifically:

- We define a learnable genre embedding lookup table (dimension 64), assigning each genre a unique vector.
- The $[CLS]$ representation is concatenated with the corresponding genre embedding: $[\text{CLS}; \mathbf{e}_{\text{genre}}]$.
- The concatenated vector is fed into a two-layer MLP with ReLU activation to produce a contrastive embedding:

$$\mathbf{z} = \text{MLP}([\text{CLS}; \mathbf{e}_{\text{genre}}]) \in \mathbb{R}^{128} \tag{2}$$

The model thus outputs:

- **logits**:used for classification with BCE loss
- **contrastive embedding**:used to compute the contrastive loss

This dual-branch architecture allows the model to jointly optimize classification and representation learning jointly,enhancing generalization [19, 18].

3.2. Contrastive Learning Module

To improve representation separability, we employ an InfoNCE-style contrastive loss[10, 16]:

- For each sample pair i and j, if their labels match, they are treated as a positive pair; otherwise, a negative pair.

- A binary mask is used to prevent self-pairs: $\text{mask}_{i,i} = 0$
- Cosine similarity is computed using L2-normalized vectors:

$$\text{sim}_{i,j} = \frac{z_i \cdot z_j}{\tau} \tag{3}$$

where τ is a temperature hyperparameter.

The contrastive loss for a sample i is defined as:

$$\mathcal{L}_{\text{i}} = -\frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}_{i,p})}{\sum_{j \neq i} \exp(\text{sim}_{i,j})} \tag{4}$$

where $\mathcal{P}(i)$ denotes the set of positive pairs for sample i.

The final loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{contrastive}} \tag{5}$$

with λ controlling the balance between the two objectives.

4. Experiments

4.1. Datasets

We evaluate our method using the official PAN 2025 Voight-Kampff dataset and a custom validation set named VK-CleanBench-2024. The official dataset includes fields such as text, label, model, genre, and id. Labels are binary: 0 for human-written, 1 for AI-generated. Genre values include categories like essays and fiction. We concatenate the genre with the text during tokenization and pad or truncate to a maximum length of 512 tokens for RoBERTa-large[3].

To assess generalization in the absence of genre information, we construct the VK-CleanBench-2024 validation set using publicly available data from PAN 2024 Voight-Kampff Authorship Verification. This set contains only id, text, and label fields. No genre is included during inference. The dataset is cleaned and de-duplicated to ensure independence from the training set and simulate real-world detection scenarios without structured metadata.

4.2. Experimental Setup

In our experimental setup, we employ RoBERTa-large as the pre-trained language model[3], with a maximum input sequence length of 512 and a batch size of 32. The optimizer used is AdamW[3, 13] with an initial learning rate of 2e-5. For the contrastive learning component, the temperature parameter is set to 0.5, and the loss weight coefficient $lamda$ is set to 0.1 to balance the training objectives between classification and representation learning. The model is trained for up to 10 epochs, with an early stopping strategy based on the F1 score on the validation set (patience = 2) to prevent overfitting and enhance generalization[19].

4.3. Experiment Results

The validation performance on the official PAN 2025 Voight-Kampff dataset is shown in Table 1. During tokenization, the genre is concatenated with the input text, allowing the model to leverage contextual cues beyond the raw text. The model achieves near-perfect results across all evaluation metrics, indicating strong performance under well-structured and annotated conditions.

To assess the model's robustness and generalization, we further evaluate it on the VK-CleanBench-2024 dataset, shown in Table 2. This setup removes genre metadata at inference time, simulating more realistic scenarios. Although performance drops slightly, the model remains competitive, demonstrating its adaptability to genre-agnostic inputs.

In response to reviewer suggestions, we conducted additional ablation studies to evaluate the individual contributions of the genre vector and the contrastive loss. The results are summarized in

Table 3. Removing the genre vector leads to a noticeable performance decline across all metrics, particularly in F_1 (from 0.997 to 0.911), confirming that genre provides valuable contextual information. Similarly, excluding the contrastive loss slightly degrades performance, suggesting its role in enhancing representation learning.

Finally, our model achieves strong generalization on the official test set, ranking 4th on the PAN 2025 leaderboard. The test-time results are presented in Table 4. Compared to the validation scores, there is a moderate drop, which is expected due to domain shift and label noise in the test environment. Nevertheless, the model maintains high robustness across all metrics.The full leaderboard can be found at: https://pan.webis.de/clef25/pan25-web/generated-content-analysis.html#task1-leaderboard.

Table 1
Evaluation results of our model on the official validation set.

ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
1.000	0.996	0.996	0.997	0.996	0.997

Table 2
Evaluation results of our model on the custom VK-CleanBench-2024 validation set.

ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
0.994	0.959	0.953	0.933	0.965	0.961

Table 3
Ablation study: removing the genre vector and contrastive loss on the official validation set.

Setting	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
without genre vector	0.987	0.946	0.939	0.911	0.953	0.947
without contrastive loss	0.980	0.945	0.934	0.898	0.951	0.942

Table 4
Official PAN 2025 test set results (ranked 4th on the leaderboard).

ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
0.845	0.878	0.871	0.856	0.881	0.877

5. Conclusion

We propose a RoBERTa-based text classification framework that integrates genre embeddings and contrastive learning for the PAN 2025 Voight-Kampff Generative AI Detection task. By incorporating genre information and optimizing the representation space, our method improves the modeling of fine-grained differences between human and AI-generated text[3, 19, 18]. Experiments confirm its effectiveness across datasets both with and without genre metadata, demonstrating robust generalization.

In future work, we plan to explore additional auxiliary signals such as writing style features and multimodal data[21, 22, 23], aiming to enhance detection in multi-genre and cross-lingual scenarios, with special attention to low-resource genres.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

Declaration on Generative AI

During the preparation of this paper, generative AI tools (specifically ChatGPT) were employed to assist in English language polishing and in translating parts of the manuscript from Chinese to English. All conceptual contributions, ideas, methods, experiments, analyses, and conclusions are entirely the work of the authors. The authors reviewed and verified all AI-assisted edits to ensure accuracy and appropriateness.

References

[1] OpenAI, Gpt-4 technical report, 2023. https://openai.com/research/gpt-4.

[2] G. DeepMind, Gemini: Multimodal language models, 2023. https://deepmind.google/discover/blog/google-gemini-ai/.

[3] 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, 2019. arXiv:1907.11692.

[4] E. Mitchell, Y. Lin, A. Bosselut, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, arXiv preprint arXiv:2301.11305 (2023).

[5] 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.

[6] 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.

[7] 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: 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.

[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, Association for Computational Linguistics, 2019, pp. 111–116.

[9] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, in: Advances in Neural Information Processing Systems, volume 32, 2019.

[10] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning (ICML), 2020.

[11] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[12] J. Kirchenbauer, J. Geiping, M. Goldblum, N. Carlini, T. Goldstein, Watermarking language models for detection, arXiv preprint arXiv:2301.10226 (2023).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[14] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, Ernie: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[16] T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6894–6910.

[17] Y. Hou, X. Li, H. Pan, S. He, J. Zhou, Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 2891–2903.

[18] W. Yang, Y. Zhao, K. Wu, W. Lu, Cosent: Supervised contrastive learning for sentence embeddings, 2022. arXiv:2201.07313.

[19] Y. Bao, Y. Du, L. Dong, W. Zhang, F. Wei, M. Zhou, Contrastive pre-training for human-authored and machine-generated text classification, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4424–4434.

[20] E. Tian, Gptzero: Detecting ai-generated text, https://gptzero.me, 2023.

[21] R. Schwartz, M. Sap, I. Konstas, W. Ammar, N. A. Smith, Story cloze evaluations and adversarial story generation, in: Proceedings of the 2017 Conference on Computational Natural Language Learning, 2017, pp. 100–109.

[22] A. Uchendu, R. Varshney, S. Lee, Y. Wang, Authorship attribution in multi-author corpora: The role of stylometric and deep learning features, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6713–6725.

[23] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Comparison of diverse generative models for text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 2100–2111.