Application and Analysis of Roberta-base Model Fine Tuning Based on Data Enhancement in AI Text Detection

Notebook for PAN at CLEF 2025

Jiancheng Huang, Haojie Cao, Xiaocan Lin and Zhongyuan Han*

Foshan university, Foshan, China

Abstract

With the current rise of large language models, AI Text Detection has become crucial as it involves determining whether a text is generated by AI. We regard this task as a binary classification problem between human-written and AI-generated texts. By performing data augmentation on AI-generated texts and fine-tuning the Roberta-base model, we achieved an F1 score of 89% on the validation set and 64% on the test set.

Keywords

AI Text Detection, Binary classification, Data augmentation, Roberta-base model

1. Introduction

The task of Voight-Kampff AI Detection Sensitivity [1] has become crucial in the field of natural language processing. With the advancement of AI language models, the ability to accurately identify the origin of a text is essential for applications such as copyright protection, content moderation, and information authenticity assessment.

Previous methods for text classification have struggled with the rapid evolution of AI language models. Traditional approaches often fail to adapt effectively to the diverse and changing characteristics of AI-generated text, particularly when new language models emerge with distinct text generation styles and features.

To address the limitations of existing methods in adapting to the dynamic nature of AI-generated text, this paper presents a novel approach that combines the fine-tuning of the Roberta-base model with advanced data augmentation techniques. By leveraging the power of Roberta-base's pre-trained language capabilities and enhancing the training data with diverse samples generated from various and updated language models, our method aims to significantly improve the model's adaptability and accuracy in classifying human and AI text. The integration of these techniques allows for a more robust and versatile classification system that can better handle the continually evolving landscape of AI text generation.

2. Related Work

The task of distinguishing human-written and AI-generated text has been explored in various evaluation forums and competitions. In PAN at CLEF 2024, the Generative AI Authorship Verification Task challenged participants to differentiate between human and machine-authored texts. Similar to previous years' tasks, this competition aimed to advance the state-of-the-art in AI text detection by providing a platform for researchers to develop and test their methods.

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

*Corresponding author.

huanghiwiy@gmail.com (J. Huang); caohaojie0322@163.com (H. Cao); linxx6513@gmail.com (X. Lin); hanzhongyuan@gmail.com (Z. Han)

0009-0001-6787-3338 (J. Huang); 0000-0002-8365-168X (H. Cao); 0009-0002-0517-7254 (X. Lin); 0000-0001-8960-9872 (Z. Han)

First Place Method: staff-trunk.Key Innovations: The marsan team's staff-trunk system leveraged a combination of advanced pre-trained language models and sophisticated data augmentation techniques. They fine-tuned their model on a diverse dataset that included not only the provided bootstrap data but also additional external data sources to enhance the model's generalization ability. A novel aspect of their approach was the integration of multi-task learning, where the model was trained to predict not only the author type (human or AI) but also auxiliary features such as text genre and length. This multi-task framework allowed the model to learn richer representations that captured various facets of the text, leading to improved discrimination between human and AI writing styles.

Main Advantages: The staff-trunk method demonstrated exceptional performance across multiple evaluation metrics, achieving the highest mean score of 0.924. Its strength lay in its ability to effectively utilize diverse data resources and the multi-task learning approach, which provided additional contextual information that aided in distinguishing subtle differences between human and AI-generated texts.

Related Literature: The method's effectiveness can be linked to similar approaches in the literature that highlight the benefits of multi-task learning in NLP tasks. For example, Bevendorff et al. [2] (2024) discuss the advantages of incorporating auxiliary tasks to improve model performance in authorship verification. Additionally, the use of external data sources aligns with strategies employed in other high-performing text classification systems where expanding training data diversity has been shown to enhance model robustness [3].

Second Place Method: charitable-mole_v3.Key Innovations: The charitable-mole_v3 system developed by the you-shun-you-de team focused on enhancing model interpretability and precision through innovative attention mechanisms and rigorous feature selection. They implemented a modified version of the Transformer architecture with custom attention heads designed to better capture the stylistic nuances that distinguish human from AI writing. A unique feature of their approach was the incorporation of a confidence calibration module that adjusted the model's predictions based on uncertainty estimates, reducing the likelihood of overconfident incorrect classifications.

Main Advantages: This method achieved a mean score of 0.921, showcasing its strong competitive edge. Its interpretability advantages enabled researchers to gain deeper insights into the model's decision-making process, identifying which text features were most influential in classification decisions. The precision enhancement from the confidence calibration module proved particularly effective in reducing errors on challenging text pairs where human and AI writing styles closely resembled each other.

Related Literature: The attention mechanism innovations in charitable-mole_v3 resonate with research by Jakesch et al. [4] (2023), who emphasize the importance of capturing stylistic elements in AI-generated text detection. The confidence calibration technique employed echoes methods described in Hans et al. [5]2024, where uncertainty estimation was used to improve the reliability of text classification systems. These parallels to established literature underscore the method's foundation in current research trends and its contribution to advancing the field.

3. Method

The process can be divided into several key components, The first is the selection model. The Roberta-base model is selected in this study. In terms of dataset processing, this study will make two different data enhancements to the dataset. The first is to extract a part of the original AI text and add noise for synonym replacement, sentence reorganization and semantic preservation; The other is a special enhancement for AI text. First, collect some AI generation models that are not in the original dataset, and then write several prompt statements for different generation topics, so that each model outputs these texts separately, and then mark them as AI text. After the data processing is completed, the Roberta base model is submitted for fine-tuning, and then the validation set is used for evaluation. The model with good evaluation is saved, and then the best model for this training evaluation is obtained after continuous fine-tuning for 100 times. The specific flow chart is shown in the Figure 1.

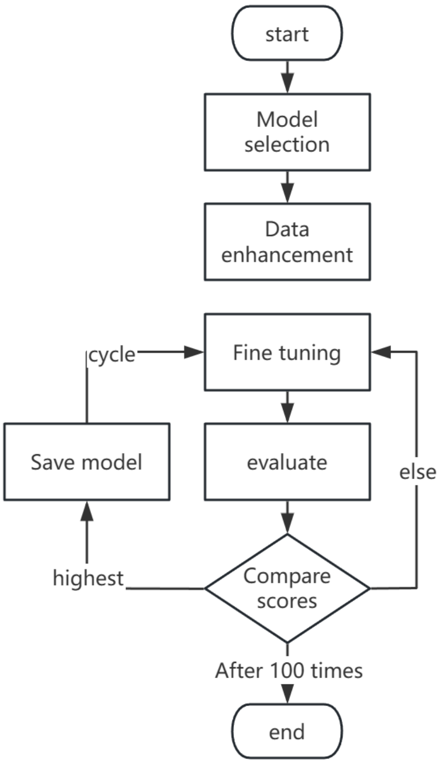

Figure 1: Overall flow chart of the experiment

3.1. Common data enhancement methods

The first type of data augmentation described above has three operations, namely synonym replacement, sentence reordering and noise addition for semantic preservation.

Synonym Replacement: Generates a new text sample by replacing some words in the text with their synonyms. This method can change the expression of the text and increase the diversity of vocabulary while keeping the semantic of the text roughly unchanged. For example, the model can learn the use of different words in the same semantic scene, and improve the adaptability to vocabulary changes, by replacing "happy" with "joyful" and "contented". In this study, we used a pre-defined list of common synonyms obtained from a standard thesaurus API to ensure the accuracy and relevance of the replacements. Approximately 10% of the words in each selected text were replaced with their synonyms, with a frequency of 5% of the total text length being replaced in each augmentation iteration.

Sentence Reordering: Adjusts the structure of sentences without changing the overall semantics of the text, such as changing the order of sentences, splitting compound sentences into simple sentences or merging simple sentences into compound sentences. This helps the model learn the semantic representation of the text under different structures and enhance its robustness to the changes of sentence structure. For example, reorganize "I went to the supermarket to buy fruits and vegetables today" into "Today, I went to the supermarket to buy fruits and vegetables". We implemented this by randomly changing the order of sentences within a paragraph and altering the structure of 15% of the sentences in each text, ensuring that the core meaning remained intact.

Semantic-preserving Noise Addition: Add some noise unrelated to semantics in the text, such as inserting some common stop words, repeating some words or phrases, but ensure that the core semantics of the text will not be changed. This method can make the model robust to noise in the training process, and improve its classification ability in the face of complex text in practical application. For example, the words "ah", "ne" and "then" are randomly inserted into the sentence to simulate the noise situation in the real language environment. The noise was selected from a predefined list of common function words and fillers that do not alter the semantic meaning. Approximately 8% of the text's tokens were replaced with noise in each augmentation step.

3.2. Special data enhancement method for AI text

In the task of human and AI text classification, the source and characteristics of AI text have their unique characteristics. With the rapid development and frequent updating of large language model, new models continue to emerge, and the style and characteristics of the generated text may also be different from the previous large language model. In order to make our classification model adapt to this change, we designed a special data augmentation method for AI text, which uses the new large language model to generate text and add it to the training data, so as to further improve the recognition ability of the model for AI text.

New Models Used: Qwen3-8B, chatglm3-6b, internlm2_5-7b-chat, GLM-4-9B-0414, DeepSeek-R1-Distill-Qwen-7B

The prompt statement uses five.

- "Write a text of about 500 words which covers the following items: Genre and Style: Culture and Lifestyle. Opinion piece, blending analysis of current events and cultural trends.Content: The text explores the cultural significance and public fascination with the Kim Kardashian and Pete Davidson relationship. It delves into how this high-profile union merges elements of a modern romance and bridges two distinct celebrity fandoms. The author reflects on society's intrigue with the relationship as a mirror of the evolving cultural landscape and the changing dynamics of celebrity culture. The Kim and Pete story appears to embody a contemporary fairy tale while challenging traditional narratives in unexpected ways. Kardashian's influence in the fashion and media world, along with Davidson's background in comedy and his openness about mental health, create a unique blend that resonates with diverse audiences. The piece highlights the emergence of a newer, more accepting social order that embraces authenticity and vulnerability, marking a shift from the more conventional and often idealized celebrity relationships of the past. This shift has garnered a mixed reception, with some embracing the authenticity and others clinging to the more traditional, polished narratives of celebrity romance."
- "Write a text of about 500 words which covers the following items: Genre and Style: Encyclopedia.Content: Uralic languages descended from Proto-Uralic language from 7,000 to 10,000 years ago.Uralic languages spoken by 25 million people in northeastern Europe, northern Asia, and North America.Hungarian, Estonian, and Finnish are the most important Uralic languages.Attempts to trace genealogy of Uralic languages to earlier periods have been hampered by lack of evidence.Uralic and Indo-European languages are not thought to be related, but speculation exists.Uralic languages consist of two groups: Finno-Ugric and Samoyedic.Finno-Ugric and Samoyedic have given rise to divergent subgroups of languages.Degree of similarity in Finno-Ugric languages is comparable to that between English and Russian.Finnish and Estonian, closely related members of Finno-Ugric, differ similarly to diverse dialects of the same language"
- "Write a text of about 500 words which covers the following items: Genre and Style: Personal development and relationships. Informative and motivational podcast transcript. Content: Discusses the importance of knowing what one truly wants and doesn't want in a relationship for personal happiness and growth.Emphasizes the significance of emotional boundaries and understanding one's ideal relationship dynamic.Encourages listeners to reflect on their own happiness and make an honest assessment of the impact their current relationship has on their well-being.Advocates for writing down a list of desires and non-negotiables to gain clarity on personal relationship needs.Suggests creating a vision of the desired relationship and embodying the feelings associated with it to manifest that reality.Shares personal anecdotes and experiences to demonstrate the power of intention, manifestation, and personal growth"
- "Write a text of about 500 words which covers the following items: Genre and Style: The fan fiction is a blend of fantasy and family drama, written in a conversational and descriptive style.Content: Setting: The story is set in the living room of Professor Michael Verres-Evans, Mrs. Petunia

Evans-Verres, and their adopted son, Harry James Potter-Evans-Verres. The room is filled with overflowing bookshelves.Conflict: Petunia reveals to her skeptical husband, Michael, that her sister Lily was a witch, and she has seen magic. Michael dismisses it as a delusion, leading to tension between them.Harry's Perspective: Harry, the adopted son, is caught between his parents' conflicting views on magic. He is curious and open-minded, willing to test the hypothesis of magic's existence.Testing the Hypothesis: Harry writes a letter to Hogwarts, attempting to send it via an owl. However, a neighbor, Mrs. Figg, intervenes and reveals that Harry's acceptance letter was standard, leading to an unexpected twist and a cliffhanger."

- "Write a text of about 500 words which covers the following items: Genre and Style: Self-help/Personal development. The language is motivational and empathetic, aiming to inspire action and self-reflection.Content: The author emphasizes the importance of proactive measures to reduce anxiety-inducing situations and manage time wisely, advocating for self-care and balance in navigating life's responsibilities.The article discusses the relationship between anxiety and procrastination, highlighting how the fear of not having enough time and taking on too many responsibilities can lead to procrastination.It offers five coping strategies for anxiety-induced procrastination, including prioritizing tasks, breaking projects into smaller parts, focusing on progress rather than perfection, setting boundaries, and scheduling downtime for mental recharge.The author emphasizes the importance of proactive measures to reduce anxiety-inducing situations and manage time wisely, advocating for self-care and balance in navigating life's responsibilities."

We pay close attention to the development of large language models, and screen out those representative and widely used new large language models, which have not appeared in the original training data. In order to generate high-quality and representative AI text, we adopt the method of diversified prompt design and parameter adjustment. Design diversified prompts to guide the new large language model to generate texts with different themes, styles and lengths. These tips cover common text application scenarios, such as news reports, story creation, comment writing, etc., to ensure that the generated text has a wide coverage. According to the characteristics of the new large language model, adjust the parameters when generating text, such as temperature, Top-k and top-p sampling. The temperature parameter can control the randomness of the generated text. A lower temperature generates a more deterministic text, and a higher temperature increases the diversity of the text. By reasonably adjusting these parameters, we can achieve a balance between the accuracy and diversity of the generated text.

Finally, the processed text generated by the new large language model is added to the training data set and labeled as AI text category.

In addition to using a single new large language model to generate text, we also simulated the scene of multi model hybrid text generation. In practical applications, AI text may come from different large language models, or even be generated from a mixture of multiple models. In order to improve the classification ability of the model in this complex situation, we designed a method to simulate the text generated by multi model mixing. We selected several different large language models and mixed them to generate text according to a certain proportion. The mixing ratio can be set based on the popularity of the model, application scenarios and other factors to simulate the real-world AI text generation environment. After the text is generated according to the hybrid strategy, the text is also post processed, and then integrated into the training data set.

4. Experiment

4.1. Experimental Setup

The data set used in this study includes human written text and AI written text. The texts compiled by human beings come from the open text corpus, covering news reports, blog articles, academic papers and other styles; AI compiled text is collected from the text generated by multiple open large language models, including but not limited to the models marked in the original training data. We collected about

9000 human texts and about 14600 AI texts from several different data sources. The proportion is shown in Table 1. AI texts cover samples generated by a variety of different large language models.

Table 1
Proportion of human and AI texts in the original dataset

Dataset text category	Proportion
Human	38.14%
AI	61.86%

For common data enhancement, we will randomly extract 2400 samples from AI text in three separate instances, and add synonym replacement, sentence reorganization and semantic preserving noise to the text respectively. After enhancing this part of the sample, put it back and replace the original sample. For the special data enhancement of AI text - text generation, this study constructed multiple prompt statements, and each prompt statement was given to different large language models, and the generated text was marked as AI text, and then the generated sample was mixed with the source dataset to form a dataset. The composition of the new data set is shown in Table 2.

Table 2
New dataset composition table

Operation	Number of samples	Proportion
Synonym Replacement	2400	8%
Sentence Reordering	2400	8%
Semantic-preserving Noise Addition	2400	8%
Single model generation	5100	17%
Multi model hybrid generation	1300	4.3%
Original AI sample	7400	24.7%
Human sample	9000	30%

4.2. Experimental environment

Hardware environment: experiment on a server equipped with 4GHz CPU, 32GB ram and NVIDIA GPU $3090, 24GBvideomemory$.

Software environment: the programming language Python version 3.12, the deep learning framework uses pytorch version 2.7.0, and other related libraries include transformers 4.52.3, scikit learn 1.5.1, numpy 1.26.4, pandas 2.2.2, etc.

4.3. Experimental results and analysis

The results of our experiment in Tira [6] are shown in Table 3. The table shows the performance of our method (connected-svn) on the test dataset.

Table 3
Score of the model on the test set

Index	Roc-Auc	Brier	C@1	F1	F05U	Mean
Mine	0.91	0.723	0.471	0.64	0.526	0.654
Baseline	0.99	0.939	0.971	0.968	0.987	0.971

The Roc-Auc score of 0.91 indicates that our model has a strong ability to distinguish between human-written and AI-generated text. The high Roc-Auc value suggests that our model can effectively rank positive instances higher than negative ones.

The Brier score of 0.723 reflects the mean squared difference between the predicted probabilities and the actual outcomes. A lower Brier score indicates better accuracy in probability estimates. Brier score of our model shows that the resolution effect of the model is not very good.

The C@1 score of 0.471 is a modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases. This score indicates that when the model is uncertain and provides a non-answer, it leverages the accuracy of the remaining cases reasonably well. However, there is still room for improvement in making more definitive predictions.

The F1 score of 0.64 shows a balance between precision and recall. Our model demonstrates a decent ability to identify both human and AI text without a significant bias towards either class.

The F0.5U score of 0.526 is a modified F0.5 measure that treats non-answers (score = 0.5) as false negatives. This score emphasizes precision while still considering recall. The model tends to be more confident in its predictions, which is beneficial for reducing false positives. However, the treatment of non-answers as false negatives suggests that there is room for improvement in handling uncertain cases.

The overall mean score of 0.654 across all metrics indicates a solid performance. However, there is still potential for improvement to enhance the model's effectiveness further.

Our method, connected-svn, achieves strong performance across multiple evaluation metrics. The high Roc-Auc and reasonable Brier score demonstrate the model's effectiveness in distinguishing human and AI text. However, the C@1 score suggests that there is still room for improvement in the model's certainty, and the F0.5U score indicates that handling non-answers could be improved. Future work could focus on enhancing the model's decision-making process to improve the C@1 score and refine the handling of non-answers to achieve better overall performance.

Unfortunately, the score of my method is still low compared with the baseline method.

5. Conclusions

This paper presents a method for distinguishing human-written from AI-generated text using the Roberta-base model with data augmentation.Our key contributions include Using Roberta-base as a powerful foundation for text classification. Implementing data augmentation to enhance model generalization. Introducing special data augmentation for AI text classification by incorporating new language models and simulating multi-model text generation.

Our method shows strong potential in identifying AI-generated content and will be further refined in future research.

Acknowledgements

This work is supported by the National Social Science Foundation of China (24BYY080)

Declaration on Generative AI

During the preparation of this work, the author(s) used Kimi-K2 in order to: translate text into English and polish the manuscript. Further, the author(s) used Qwen3-8B, ChatGLM3-6B, InternLM2.5-7B-Chat, GLM-4-9B-0414, and DeepSeek-R1-Distill-Qwen-7B in order to: enhance the dataset. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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.

[2] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "voight-kampff" generative ai authorship verification task at pan and eloquent 2024, in: 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble, France 9 September 2024 through 12 September 2024, volume 3740, CEUR-WS, 2024, pp. 2486–2506.

[3] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 8384–8395.

[4] M. Jakesch, J. T. Hancock, M. Naaman, Human heuristics for ai-generated language are flawed, Proceedings of the National Academy of Sciences 120 (2023) e2208839120.

[5] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, arXiv preprint arXiv:2401.12070 (2024).

[6] 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.