Team Nexus Interrogators at PAN: Voight-Kampff Generative AI Detection

Notebook for the PAN Lab at CLEF 2025

Samiya Ali Zaidi[1,*,†], Huzaifah Tariq Ahmed[1,*,†], Sarrah Ali Akbar[1], Ziaullah Shakeel[1], Faisal Alvi[1] and Abdul Samad[1]

[1]*Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan*

Abstract

The Voight-Kampff task at PAN CLEF 2025 challenges participants to detect and categorize AI-generated text in an era of increasingly human-like language models. In this work, we develop a two-stage system leveraging fine-tuned transformer architectures to tackle both binary and multi-class authorship verification. For Subtask 1, we fine-tune a bert-base-uncased model to distinguish human-written from machine-generated text, achieving near-perfect performance across genres with minimal false positives. For Subtask 2, we address severe class imbalance in multi-class collaborative authorship detection by augmenting underrepresented categories using backtranslation, synonym/antonym replacement, and random deletion. Fine-tuning a roberta-large model on this enriched dataset yields significant gains, particularly in minority classes. Our results underscore the effectiveness of combining targeted data augmentation with robust transformer-based models to capture subtle distinctions in authorship, offering a scalable foundation for detecting generative AI involvement in real-world texts.

Keywords

Voight-Kampff, AI-Generated Text Detection, Authorship Verification, Transformer Models, Data Augmentation, Fine-tuning, PAN Lab, CLEF 2025

1. Introduction

The increasing use of large language models (LLMs) in content creation has introduced new challenges in distinguishing between human- and AI-generated text. While generative AI has shown remarkable capabilities in mimicking human writing, this raises concerns related to academic integrity, misinformation, and authorship transparency. As AI-assisted writing becomes more sophisticated, robust detection systems are needed to identify the degree of machine involvement in written texts.

The Voight-Kampff Generative AI Detection 2025 task [1, 2], part of the PAN shared task series with the ELOQUENT Lab [1], addresses this problem by evaluating detection systems across two key subtasks. Subtask 1 focuses on binary classification of texts as either entirely human-written or machine-generated, even in cases where the AI attempts to imitate a specific human writing style [2]. This tests the sensitivity and robustness of detection methods against adversarial obfuscation and unseen model outputs.

Subtask 2 extends the challenge by introducing multi-class classification of collaborative human-AI texts, requiring systems to detect nuanced degrees of machine involvement. This includes identifying when humans post-edit AI-generated drafts, co-write with AI models, or minimally edit machine-generated outputs. The goal is not only to improve detection accuracy but also to understand the spectrum of human-AI collaboration [2].

To tackle these challenges, this paper explores a range of techniques, including data augmentation strategies, finetuning, ensemble methods, and neural classifiers. Our system builds upon prior research

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

*Corresponding author.

†These authors contributed equally.

sa07171@st.habib.edu.pk (S. A. Zaidi); ha07151@st.habib.edu.pk (H. T. Ahmed); sa07207@st.habib.edu.pk (S. A. Akbar); zs07752@st.habib.edu.pk (Z. Shakeel); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0008-1907-1542 (S. A. Zaidi); 0000-0002-9421-8566 (H. T. Ahmed); 0000-0001-7116-9338 (S. A. Akbar); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

in authorship verification and leverages recent advances in supervised learning, fine-tuning, and hybrid modeling. We focus on robustness across genres and model types, addressing both fully and partially machine-generated content.

The rest of this paper is structured as follows: section 2 presents a review of the related works focusing on the approaches commonly used for authorship detection. Section 3 describes our approach to solving both subtasks. Section 4 presents our validation results. Lastly, section 5 concludes our paper.

2. Related Work

Authorship verification has evolved from stylistic analysis of human writing to the detection of AI-generated content. Recent work has leveraged both traditional machine learning and deep learning models for this task. Fine-tuned transformer architectures such as DeBERTa [3] and RoBERTa have achieved high performance in binary classification of human vs. AI text [4], while hybrid models that combine BERT with CNNs enhance local and contextual feature extraction [5].

Some systems introduce data augmentation and R-Drop regularization [6] to improve robustness, employing loss functions that combine cross-entropy and KL divergence. Ensemble learning approaches using multiple transformer models (e.g., BERT, RoBERTa, DeBERTa) have shown further improvements in ROC-AUC scores [7]. Meanwhile, instructional prompting with T5 has been explored to reframe authorship detection as a sequence-to-sequence task [8].

Beyond transformers, research has explored lightweight classifiers with embeddings like LUAR for low-resource scenarios [9], and stylometric analysis using Graph Neural Networks (GNNs) alongside pre-trained models [10]. Approaches such as Tri-Sentence Analysis [11] and hybrid models like BertT [12] demonstrate effectiveness in handling short texts and improving generalization.

Despite promising results, many systems struggle with generalization to novel AI models or obfuscated styles, highlighting the importance of continual adaptation and diverse training data in generative AI authorship verification.

3. Methodology

In this section, we provide details about the datasets for each task, followed by our methodology for both subtasks individually.

3.1. Datasets

The datasets for this task are provided as newline-delimited JSON files. In subtask 1's dataset, each entry includes an identifier, the text content, the originating model (human or specific AI model), a label (0 for human, 1 for AI), and a genre indicator (e.g., essays, news, fiction).

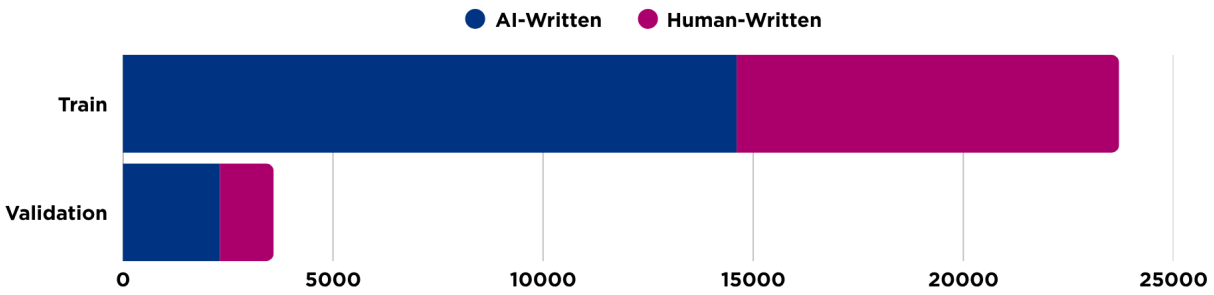

Figure 1: Sub-Task 1 Dataset Class Distribution

The dataset for subtask 2, on the other hand, comprises multi-domain documents drawn from academic sources, journalism, and social media. The data includes a mixture of human-written and machine-generated samples (produced by models such as GPT-4, Claude, and PaLM) and is annotated to indicate the type of human-AI collaboration. The dataset spans multiple languages and provides detailed labels for each collaboration category.

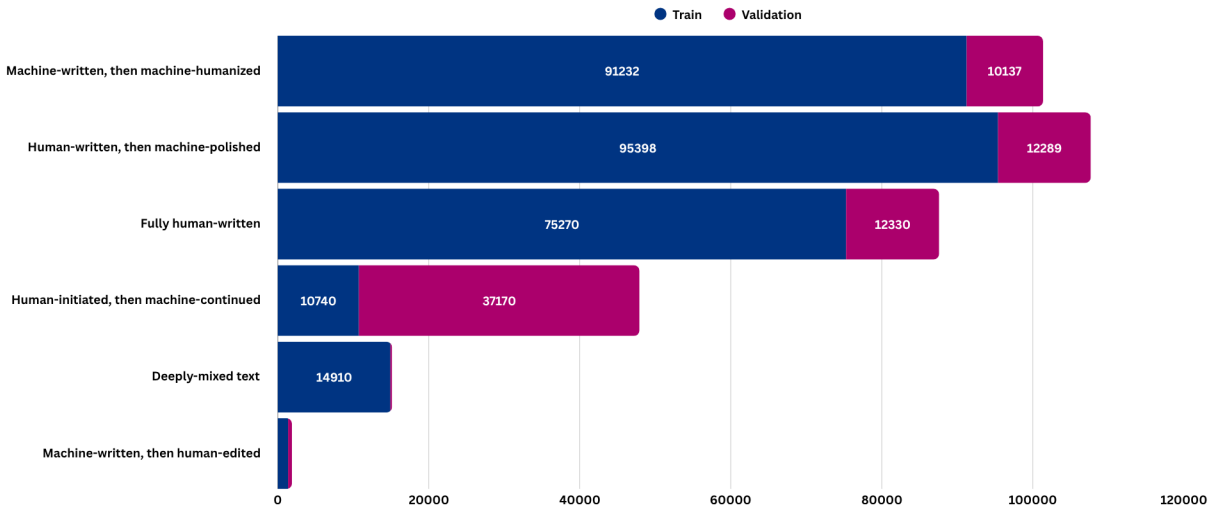

Figure 2: Sub-Task 2 Dataset Class Distribution

3.2. Subtask 1: Voight-Kampff AI Detection Sensitivity

In this task, our primary objective was to accurately distinguish between human-written and AI-generated text. This binary classification problem required a robust modeling pipeline that could leverage the nuanced differences between the two categories. The distribution of the dataset used for this task is illustrated in Figure 1, providing insight into the balance of the data across both classes.

The first phase of our workflow involved data preprocessing. The original dataset was provided in a `.jsonl` format, which is commonly used for storing structured data in a line-delimited manner. To facilitate data handling and analysis, we first converted this `.jsonl` file into a Pandas DataFrame. From this structure, we extracted only the essential fields required for our task: `'id'`, `'text'`, `'label'`. These fields represent, respectively, the unique identifier of each sample, the content of the text, and its associated label indicating whether the text was AI-generated or written by a human (0 means human-written, and 1 means AI-generated).

After isolating the relevant information, we transformed the dataset into the Hugging Face `Dataset` format. This conversion optimized the data pipeline for fine-tuning pre-trained models. The Hugging Face Dataset object also provides efficient shuffling, batching, and tokenization utilities, which are particularly useful for handling text data at scale.

With the dataset prepared, we proceeded to the model fine-tuning phase. We leveraged the Hugging Face `transformers` library due to its modularity, ease of use, and strong support for state-of-the-art pre-trained language models. We used the `AutoModelForSequenceClassification` interface to load the `bert-base-uncased` model with two output labels (human and AI), and the `AutoTokenizer` for consistent input preprocessing. We selected this variant of BERT for its proven effectiveness in various natural language understanding tasks, particularly in text classification. The fine-tuning process involved training the model on the labeled dataset to adapt BERT's pretrained representations to our specific task of authorship classification.

Training was carried out using the `Trainer API`, which provided integrated training and evaluation loops, model checkpointing, and metric logging. All hyperparameters used during training, including learning rate, batch size, and number of epochs, are detailed in Table 1. These parameters were chosen

based on standard practices for fine-tuning transformer models and adjusted to fit the computational constraints and performance needs of our project.

Table 1
Hyperparameters Used in Both Subtask 1 and Subtask 2 Experiments

Hyperparameter	Value
Epochs	3
Learning Rate	2×10^{-5}
Batch Size	8
Weight Decay	0.01

We monitored performance after each epoch and retained the best-performing model. For evaluation, we used the `evaluate` library to compute micro-averaged F1 scores, ensuring that performance was balanced across both classes. At inference time, predictions were generated using the Trainer API and analyzed via a detailed classification report, giving us insights into precision, recall, and F1 score for both human and AI text classes. This setup ensured a reliable, reproducible training pipeline aligned with modern standards for fine-tuning transformer-based classifiers.

3.3. Subtask 2: Human-AI Collaborative Text Classification

For this sub-task, our objective was to determine the extent of AI involvement in the generation of a given piece of text. Unlike the binary classification task described earlier, this problem was framed as a multi-class classification challenge, where each sample was categorized into one of six distinct labels based on the degree and type of human-machine collaboration. The classification labels are 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, where some parts are written by a human and some are generated by a machine
- 5: machine-written, then human-edited

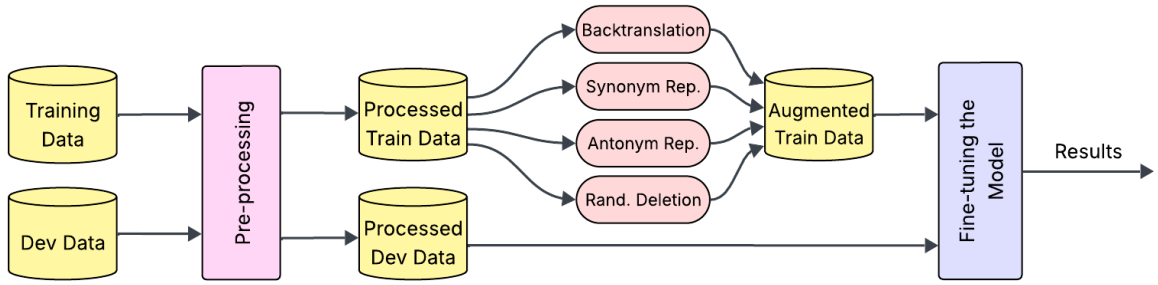

Figure 3: Overview of the data preprocessing, augmentation, and fine-tuning pipeline used for multi-class AI authorship extent classification in Subtask 2.

The distribution of samples across these six categories is visualized in Figure 2, which highlights a substantial class imbalance in the dataset. This imbalance posed a significant challenge, particularly for training a model capable of accurately distinguishing underrepresented categories.

As with the earlier task, the dataset was initially provided in `.jsonl` format. To facilitate preprocessing and further transformations, we first converted the data into a Pandas DataFrame. From the available

fields, only the `text` and `label` columns were retained, as these were essential for the classification task.

Given the imbalance in class distribution, we implemented several data augmentation techniques targeting the three least represented classes – 3, 4, and 5. These augmentation strategies were designed to increase the diversity and volume of examples in the minority classes, thereby helping to mitigate the effects of class imbalance during training. The augmentation methods used include:

- Backtranslation
- Synonym Replacement
- Antonym Replacement
- Random Deletion

Each of these strategies was applied separately to the minority classes, after which the augmented datasets were merged to form an enriched and more balanced training set, as depicted in Table 2.

Table 2
Class Distribution in Training Dataset for Subtask 2 Before and After Augmentation

Label	Before Augmentation	After Augmentation
0: Fully human-written	75,270	75,270
1: Human-written, then machine-polished	95,398	95,398
2: Machine-written, then machine-humanized	91,232	91,232
3: Human-initiated, then machine-continued	10,740	53,700
4: Deeply-mixed text	14,910	74,550
5: Machine-written, then human-edited	1,368	6,840

This enhanced dataset was utilized to fine-tune the state-of-the-art `Roberta-Large` Model. The large variant was chosen to effectively capture the nuances and nonlinearities present in such a complex dataset. By training on both the original and augmented data, the model became better equipped to generalize across all six categories of AI-human text interaction. The hyperparameters used are detailed in Table 1, and the entire workflow is visually represented in Figure 3.

4. Results and Discussion

4.1. Subtask 1: Voight–Kampff AI Detection Sensitivity

Table 3 reports the detailed classification metrics obtained by fine-tuning a BERT-base-uncased model on the Subtask 1 dataset. The model achieves an overall accuracy of 98.77%, with precision and recall both above 98% for human-authored text and above 99% recall for AI-generated text. These results indicate that the model is highly effective at distinguishing between genuine human writing and obfuscated machine-generated content, even when the latter is crafted to mimic a specific authorial style.

Table 3
Classification report for Subtask 1 (BERT-base-uncased).

Class	Precision	Recall	F1-score	Support
Human (0)	0.9968	0.9687	0.9825	1,277
AI (1)	0.9830	0.9983	0.9906	2,312
Accuracy		0.9877		
Macro avg	0.9899	0.9835	0.9865	3,589
Weighted avg	0.9879	0.9877	0.9877	3,589

The exceptionally high recall for the AI class (0.9983) suggests that the detector rarely misses machine-generated instances, even when those instances employ novel obfuscation methods. Conversely, the

slight asymmetry in recall (0.9687) for the human-authored class highlights a small proportion of false positives—AI texts misclassified as human—which could stem from particularly human-like AI outputs. Overall, the model's balanced precision and recall showcase its robustness and sensitivity in the face of adversarial style-mimicking.

The ROC curves and confusion matrix visualized in Figure 4 further reinforce the model's high discriminative ability. The curves show excellent separation between the classes, and the confusion matrix reveals very few misclassifications, aligning with the reported metrics.

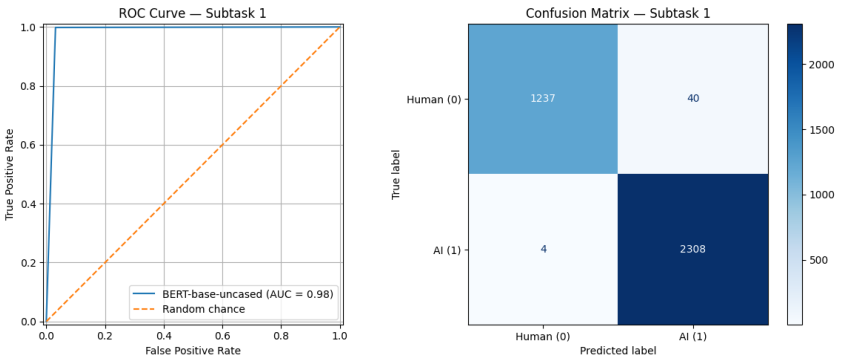

Figure 4: ROC curves and confusion matrix for Subtask 1.

The scores obtained after running the model on TIRA [13] are presented in Table 4. It showcases our model's flawless performance across all genres in the validation phase, achieving a perfect ROC-AUC of 1.0 and consistently high scores across C1, F1, F0.5U, and Brier metrics—underscoring both its discriminative power and calibration quality.

Table 4
Evaluation metrics for the Subtask 1 Submission on TIRA [13].

Dataset	Genre	Roc-Auc	Brier	C@1	F1	F05U	Mean
pan25-generative-ai-detection-val	All	1.0	0.986	0.983	0.987	0.980	0.987
pan25-generative-ai-detection-val	Essays	1.0	0.982	0.980	0.988	0.981	0.986
pan25-generative-ai-detection-val	Fiction	1.0	0.987	0.983	0.983	0.975	0.985
pan25-generative-ai-detection-val	News	1.0	0.987	0.985	0.991	0.986	0.990

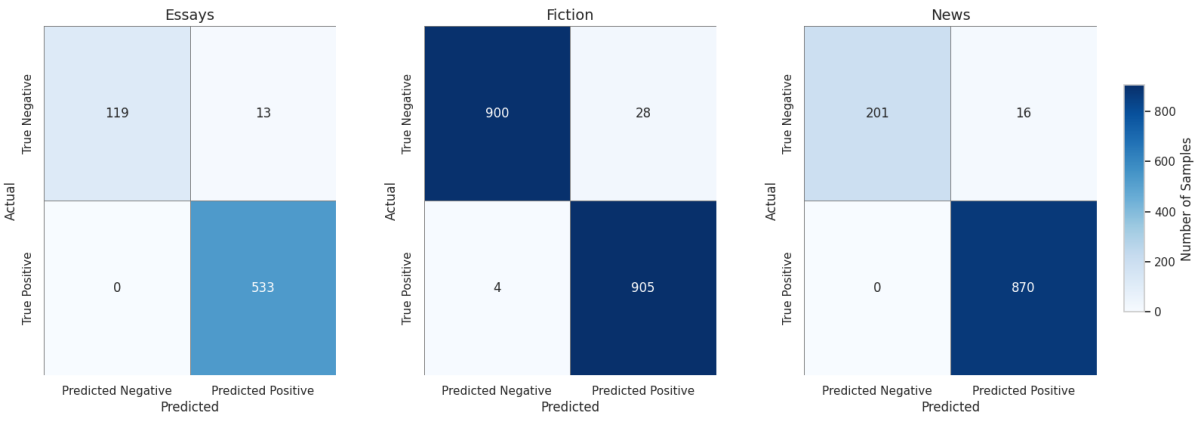

Figure 5: Confusion Matrix for Each Genre for Subtask 1.

Furthermore, the confusion matrices for each genre in the test dataset are shown in Figure 5, which provides further insight through confusion matrices for each genre on the test set. The model demonstrates perfect recall in Essays and News (no false negatives), with only 13 and 16 false positives,

respectively, highlighting its conservative and accurate labeling of AI-generated text. In Fiction, although a small number of misclassifications occur (28 false positives, 4 false negatives), the model still exhibits strong performance, effectively handling the complexity of creative writing.

Finally, Table 5 benchmarks our model against leading baselines on the test dataset, where it outperforms across all major metrics—achieving the highest ROC-AUC (0.865), F1 (0.860), and mean score (0.879), while maintaining the lowest False Positive Rate (0.131). These results confirm that the model generalizes well and remains reliable across genres, balancing precision and recall better than all competing approaches.

Table 5
Performance Comparison on Subtask 1 Test Dataset

Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean	FPR	FNR
Ours	**0.865**	**0.874**	**0.870**	**0.860**	**0.881**	**0.879**	**0.131**	**0.159**
Baseline TF-IDF SVM	0.838	0.871	0.836	0.827	0.862	0.856	0.201	0.153
Baseline Binoculars LLaMA 3.1	0.760	0.835	0.793	0.802	0.831	0.818	0.314	0.206
Baseline PPMd CBC	0.636	0.795	0.735	0.763	0.771	0.758	0.784	0.129

4.2. Subtask 2: Human–AI Collaborative Text Classification

Subtask 2 involves a multi-class classification challenge with six distinct levels of collaboration. To tackle the significant class imbalance, especially for Classes 3, 4, and 5, we implemented targeted data augmentation techniques. These techniques included back-translation, antonym/synonym substitution, and random deletion, all aimed at enhancing the representation of the underrepresented categories.

We fine-tuned a RoBERTa-Large model on the augmented dataset and observed decent scores, especially in the performance of minority classes. Table 6 summarizes the per-class precision, recall, F1-score, and overall performance metrics.

Table 6
Classification Report with Per-Class and Macro-Averaged Metrics

Class	Precision	Recall	F1-score	Support
0	0.644	0.775	0.703	12,330
1	0.403	0.935	0.564	12,289
2	0.712	0.991	0.829	10,137
3	0.899	0.336	0.489	37,170
4	0.152	0.573	0.241	225
5	0.998	0.937	0.967	510
Macro Avg	**0.635**	**0.758**	**0.632**	**72,661**
Weighted Avg	**0.744**	**0.608**	**0.588**	**72,661**
Accuracy		**0.608**		

The macro-averaged F1-score of 0.632 shows balanced performance across classes, highlighting the success of our augmentation strategy in addressing bias toward majority classes. Classes 4 and 5, once underrepresented, have also seemed to perform well. Class 3 has high precision (0.899) but low recall (0.336), indicating conservative predictions potentially due to overlap with other classes. Class 1, on the other hand, has high recall (0.935) but low precision (0.403), suggesting overprediction.

Therefore, to evaluate the impact of each augmentation method, we fine-tuned separate models using one technique at a time. Table 7 displays the class-wise precision, recall, and F1-scores. Antonym replacement and random deletion enhanced macro-level performance, with random deletion achieving the highest macro F1-score of 0.590.

To contextualize our results, we compare our model's performance with the official PAN shared task baseline on both the test and validation splits [2]. As shown in Table 8, while our test-time performance

Table 7
Per-class and macro-averaged Precision, Recall, and F1-score for different data augmentation strategies.

Metric / Strategy	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Macro Avg
Precision							
Backtranslation	0.515	0.444	0.682	0.885	0.072	0.977	0.596
Antonyms	0.491	0.455	0.717	0.888	0.136	0.981	0.612
Rand Deletion	0.494	0.461	0.707	0.883	0.143	0.991	0.613
Synonym Subst.	0.514	0.432	0.699	0.877	0.102	0.984	0.601
Recall							
Backtranslation	0.859	0.873	0.987	0.296	0.120	0.824	0.660
Antonyms	0.856	0.882	0.985	0.307	0.120	0.820	0.662
Rand Deletion	0.857	0.875	0.987	0.315	0.084	0.855	0.662
Synonym Subst.	0.855	0.883	0.987	0.284	0.098	0.869	0.662
F1-score							
Backtranslation	0.644	0.588	0.807	0.443	0.090	0.894	0.578
Antonyms	0.625	0.601	0.830	0.456	0.127	0.893	0.589
Rand Deletion	0.627	0.604	0.824	0.464	0.106	0.918	0.590
Synonym Subst.	0.642	0.580	0.818	0.429	0.100	0.923	0.582

lags behind the baseline, our validation scores significantly exceed it, particularly in terms of macro F1-score and recall. This suggests that our model is capable of learning from the augmented data, but may suffer from domain shift or limited generalizability on the blind test set.

Table 8
Performance Comparison with PAN Shared Task Baseline [2]

Model / Split	Macro Recall	Macro F1-score	Accuracy
PAN Baseline (Test)	48.32%	47.82%	57.09%
Ours (Test)	33.86%	31.86%	35.45%
Ours (Validation)	**61.86%**	**63.20%**	**60.80%**

4.3. Summary of Findings

Our experiments confirm that Subtask 1 can be effectively solved with standard fine-tuning of a transformer-based model, achieving near-ceiling performance even under adversarial-style obfuscation. In contrast, Subtask 2's multi-way classification remains challenging due to severe class imbalance and nuanced distinctions between collaboration levels. Data augmentation proves a viable strategy for boosting performance on underrepresented classes, but future work should explore complementary approaches—such as ensembling, stylometric feature fusion, or few-shot prompting with large language models—to further enhance robustness and fine-grained discrimination.

5. Conclusion

In this study, we focused on both binary and multi-class AI authorship detection tasks for the Voight-Kampff challenge at the CLEF PAN Lab 2025, utilizing a fine-tuned BERT base uncased model. For Subtask 1, our approach achieved an impressive accuracy of 98.77%, demonstrating robust F1 scores for both human and AI classes, which illustrates the model's effectiveness in binary classification.

Subtask 2 posed a considerable challenge due to severe class imbalance. By applying targeted data augmentation—specifically focused on underrepresented classes—and fine-tuning a RoBERTa-Large model, we were able to significantly improve macro F1-score across the board. The largest gains were observed in minority classes, particularly Class 4 and Class 5, demonstrating that balancing strategies can effectively improve performance on rare collaboration levels without sacrificing overall accuracy.

Moreover, performance on high-support classes such as Class 0 (fully human-written) and Class 2 (minor AI assistance) remained robust, indicating that augmentation did not negatively impact the model's understanding of dominant patterns. However, despite these gains, Class 3 continues to show low recall, suggesting persistent confusion in capturing intermediate collaboration levels. Future work could explore the use of contrastive learning, ensemble techniques, or stylometric features to help better disentangle nuanced authorial blends, especially with more powerful foundation models.

Acknowledgments

The authors would like to acknowledge the support provided by the Office of Research (OoR) at Habib University, Karachi, Pakistan, for funding this project through the internal research grant IRG-2235.

Declaration on Generative AI

During the preparation of this work, the authors utilized GPT-4 and Grammarly for grammar and spelling checks. After employing these tools, the authors independently reviewed and edited the content as necessary, taking full responsibility for the final publication.

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] J. Bevendorff, Y. Wang, J. Karlgren, M. Wiegmann, M. Fröbe, A. Tsivgun, J. Su, Z. Xie, M. Abassy, J. Mansurov, R. Xing, M. N. Ta, K. A. Elozeiri, T. Gu, R. V. Tomar, J. Geng, E. Artemova, A. Shelmanov, N. Habash, E. Stamatatos, I. Gurevych, P. Nakov, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[4] A. Yadagiri, L. Shree, S. Parween, A. Raj, S. Maurya, P. Pakray, Detecting ai-generated text with pre-trained models using linguistic features, in: Proceedings of the 21st International Conference on Natural Language Processing (ICON), 2024, pp. 188–196.

[5] 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).

[6] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu, et al., R-drop: Regularized dropout for neural networks, Advances in neural information processing systems 34 (2021) 10890–10905.

[7] Z. Lin, Z. Han, L. Kong, M. Chen, S. Zhang, J. Peng, K. Sun, A verifying generative text authorship model with regularized dropout, Working Notes of CLEF (2024).

[8] Z. Lin, Y. Li, J. Huang, Voight-kampff generative ai authorship verification based on t5, Working Notes of CLEF (2024).
[9] A. Richburg, C. Bao, M. Carpuat, Automatic authorship analysis in human-ai collaborative writing, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 1845–1855.
[10] A. Valdez-Valenzuela, H. Gómez-Adorno, Team iimasnlp at pan: leveraging graph neural networks and large language models for generative ai authorship verification, Working Notes of CLEF (2024).
[11] J. Huang, Y. Chen, M. Luo, Y. Li, Generative ai authorship verification of tri-sentence analysis base on the bert model, Working Notes of CLEF (2024).
[12] Z. Wu, W. Yang, L. Ma, Z. Zhao, Bertt: a hybrid neural network model for generative ai authorship verification, Working Notes of CLEF (2024).
[13] 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.