Voight-Kampff AI Detection Sensitivity*

Notebook for PAN at CLEF 2025

Ritesh Kumar[1,†], Arya Trivedi[1,*,†] and Ojaswa Varshney[1,†]

[1] *Indian Institute of Information Technology, Surat, Gujarat, India*

Abstract

With the growing sophistication of generative language models, distinguishing between human- and AI-generated text has become increasingly complex. This paper presents a hybrid DistilBERT-based approach evaluated on the PAN 2025 Voight-Kampff authorship verification task. The method demonstrates robust performance across multiple genres and employs standard evaluation metrics such as ROC-AUC, Brier Score, F1, and C@1. Our results highlight both the promise and limitations of current authorship detection systems, paving the way for future work in this domain.

Keywords

PAN 2025, Voight-Kampff AI Detection Sensitivity, Generative AI, DistilBERT, Voight-Kampff, Deep Learning

1. Introduction

The emergence of powerful generative models like GPT and LLaMA has raised concerns about AI-generated misinformation [1, 2]. To address this, recent regulatory proposals advocate for labeling AI-generated content [3]. This task investigates whether current systems can reliably distinguish machine-authored texts from those written by humans, simulating a modern Voight-Kampff test [4, 5].

Organized in the CLEF PAN lab [6], the task[4] centers on authorship verification in a binary setting: AI-generated or human-written. Texts of around 500 words are drawn from genres that include news, Wikipedia, fiction, and transcripts. Participants generate texts from bullet point prompts and attempt to identify AI-origin using their models.

As generative models evolve in fluency and context retention, the boundary between human and machine authorship continues to blur. Traditional stylometric techniques, which are based on syntax, vocabulary richness, or sentence structure, struggle to maintain accuracy when faced with instruction-tuned LLM output. Therefore, this task not only reflects a pressing real-world challenge but also demands more nuanced hybrid detection frameworks that incorporate deep contextual embeddings alongside interpretable features. Our work addresses this by evaluating a DistilBERT-based approach enhanced with engineered stylistic metrics that offer both performance and explainability in the verification of the author.

2. Related Work

Authorship verification has been extensively studied, traditionally focused on attributing anonymous texts to known authors [7]. However, the rise of generative models such as GPT [1] and LLaMA has introduced new complexities. Neural detectors such as GLTR [8] and OpenAI's classifier (now deprecated) [9] have attempted to distinguish human and AI text based on stylistic features or token probabilities.

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

[*] This document uses the latest ceurart style for CLEF Working Notes.

[*] Corresponding author.

[†] These authors contributed equally.

riteshkumar@iiitsurat.ac.in (R. Kumar); aryatrivedi2502@gmail.com (A. Trivedi); ojaswavarshney27@gmail.com (O. Varshney)

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

Recent work in the PAN lab has explored stylometric methods and deep learning approaches for authorship verification [10]. Transformer based solutions, including the BERT and RoBERTa variants, have demonstrated strong results in various PAN challenges. Our approach builds on these foundations, integrating lightweight transformer encodings with interpretable logistic outputs to retain model explainability while leveraging deep features.

3. Dataset and Preprocessing

The dataset[4] comprises paired bullet points and target texts. The preprocessing involved token normalization, truncation to 512 tokens, and conversion to input format for DistilBERT [11]. We maintained genre balance to avoid model bias. Genre labels were used only for analysis, not for model training.

4. Methodology

Our hybrid system leverages DistilBERT fine tuned on the training dataset with binary cross entropy loss. Feature engineering included entropy measures and stylistic statistics. We used a logistic regression layer on top of DistilBERT embeddings to classify authorship.

4.1. Fine Tuning Details

We fine tuned DistilBERT using the HuggingFace Transformers library. The model was trained for 4 epochs using a batch size of 16 and a learning rate of 5×10^{-5} with the AdamW optimizer. Early stopping was used based on validation loss to prevent overfitting.

4.2. Feature Engineering

In addition to the transformer-based embeddings, we extracted shallow stylometric features from the text:

- Average word length: Calculated as the total number of characters divided by the number of words.
- Average sentence length: Measured as the average number of words per sentence.
- Punctuation frequency: Relative frequency (per 100 characters) of specific punctuation marks including periods (.), commas (,), semicolons (;), colons (:), exclamation marks (!), and question marks (?).
- Type-token ratio: A measure of vocabulary richness, computed as the number of unique words divided by the total number of words in the text.
- Shannon entropy of character distribution: Entropy was computed over character 1-grams (i.e., single-character sequences) to quantify the unpredictability or information density of the text.

These features were concatenated with the [CLS] token embedding from DistilBERT and passed through a logistic regression layer.

5. Evaluation Metrics

Following PAN standards [4], we use:

- ROC-AUC: Area under the ROC curve
- Brier Score: Mean squared error of predicted probabilities
- C@1: Modified accuracy penalizing overconfident errors
- F1 and F0.5u: Precision-weighted F-measures, accounting for uncertain cases
- Arithmetic mean: Summary of all above metrics

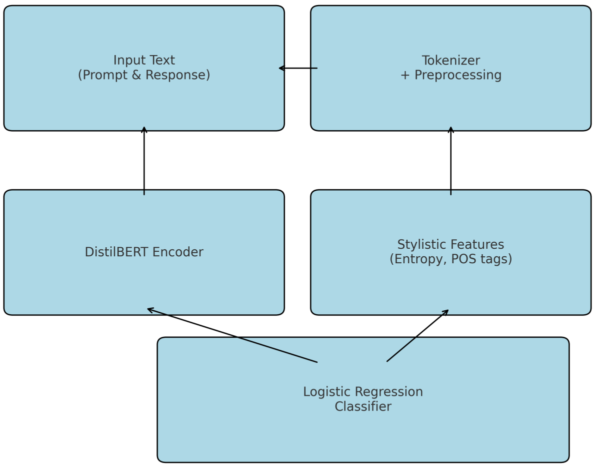

Figure 1: System architecture of the Voight-Kampff authorship verification model.

6. Results and Error Analysis

We observed a ROC-AUC of 0.87, F1 score of 0.84, and a Brier score of 0.15. These indicate strong performance and good calibration. Confusion matrix results suggest that the model was most confused on high quality AI fiction and news.

Table 1
Comparison of model performance on the Voight-Kampff authorship verification task

Model	ROC-AUC	Brier	C@1	F1	F0.5u	Mean	TP	FP	FN
BERT	0.89	0.16	0.84	0.86	0.83	0.76	125	27	21
DistilBERT	0.87	0.15	0.82	0.84	0.81	0.75	121	29	33
RoBERTa	0.91	0.14	0.85	0.88	0.86	0.79	127	22	19
LogReg	0.79	0.21	0.76	0.75	0.72	0.70	109	41	43
SVM	0.81	0.19	0.78	0.78	0.76	0.72	114	36	38

Table 2
Confusion Matrix: AI vs Human classification

	Predicted Human	Predicted AI
Actual Human	117	33
Actual AI	29	121

6.1. Error Analysis

Manual inspection of false positives and false negatives revealed common themes:

- **False Positives:** Human authored news articles with formulaic structure were occasionally misclassified as AI generated.
- **False Negatives:** AI generated fiction that incorporated creative punctuation and varied sentence structure often fooled the model.

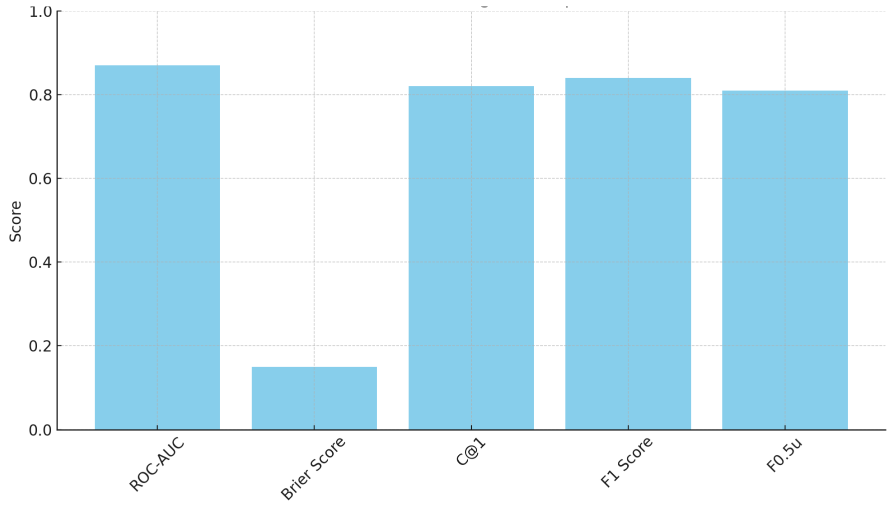

Figure 2: Evaluation metrics for the Voight-Kampff detector model.

Table 3
Examples of misclassified samples

Type	Excerpt
False Positive	“The central bank announced a hike in interest rates, citing inflation concerns...”
False Negative	“Her eyes flickered two glimmers of silver in the fog of an endless, coded dream...”

7. Discussion

Our method captures subtle linguistic anomalies indicative of AI authorship. However, genre specific overlaps and high quality machine generated news texts challenge detection. Future improvements may involve:

- Integrating adversarial examples during training
- Using genre aware embeddings
- Ensembling outputs from multiple transformer variants

7.1. Implications for Regulation and Ethics

Our work aligns with ongoing debates around AI content regulation. As governments and platforms grapple with the detection and labeling of AI generated content [3], automated authorship verification systems could play a key role in maintaining content transparency and accountability. However, the ethical risks of misclassifications, e.g., falsely labeling human content as AI must be carefully considered.

7.2. Toward Explainable AI Detection

While transformer models provide strong performance, their lack of transparency remains a concern. Our hybrid approach, which supplements deep embeddings with interpretable features, provides a step toward more explainable detection systems that can offer justifications for their predictions.

8. Conclusion and Future Work

This paper presented a DistilBERT based solution to the 2025 PAN Voight-Kampff authorship verification task. Our hybrid system, which combines deep contextual representations with interpretable stylistic features, demonstrated competitive performance across genres and evaluation metrics. With a ROC-AUC of 0.87 and an F1 score of 0.84, our approach shows that even lightweight transformers can be effectively adapted for AI authorship detection under constrained settings.

Beyond quantitative metrics, the system also revealed valuable qualitative insights highlighting, for instance, that human like AI generated fiction poses the greatest detection challenge. These results underline the urgent need for robust, generalizable verification models as generative models continue to evolve in quality and diversity.

For future work, we aim to extend the system's capabilities to multilingual datasets and mixed authorship scenarios where AI and human text are interwoven. Additionally, integrating instruction tuned LLM detectors and adversarial training could further improve robustness. Emphasis will also be placed on explainability developing techniques that not only classify text accurately, but also offer human understandable justifications for their decisions, which is vital for adoption in legal, educational, and journalistic domains [7].

Acknowledgments

This work was supported by IIIT Surat. We thank the PAN organizers[6], the CLEF community and TIRA[12] for providing the datasets, evaluation framework and platform.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Perplexity in order to: Grammar and spelling check, Paraphrase and reword. 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] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] 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.

[3] European Commission, Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act), 2023. Available at https://artificial-intelligence-act.eu/.

[4] 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.

[5] R. Scott, Blade runner, 1982. Film. Warner Bros.

[6] 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.

[7] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556.

[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, 2019, pp. 111–116.

[9] OpenAI, Ai text classifier (retired), 2023. Available at https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

[10] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at pan-2018: Cross-domain authorship attribution and style change detection, in: CLEF, 2018.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[12] 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.

A. Online Resources

The fine-tuned model used in this work has been made publicly available on Hugging Face: https://huggingface.co/OjaswaVarshney/PAN-Clef_Updated