Team SINAI-INTA at PAN 2025: Uncovering Machine Generated Text with Linguistic Features

Notebook for the SINAI-INTA Lab at CLEF 2025

Maria Jimeno-Gonzalez[1], Eugenio Martínez-Cámara[2], Noelia Fernandez[3], Pedro Díaz-García[1] and Luis Alfonso Ureña-López[2]

[1]INTA, Madrid, Spain
[2]UJA, Jaen, Spain
[3]UC3M, Madrid, Spain

Abstract

Addressing the escalating text generation capabilities of large language models, PAN and the ELOQUENT Lab have introduced the Voight-Kampff Generative AI Authorship Verification task, which aims to distinguish between human and machine-generated texts. In response, this paper proposes a lightweight approach that combines syntactic, structural, and lexical features with TF-IDF representations of the raw text. The method is designed to be computationally efficient, making it suitable for practical applications without requiring extensive resources. On the validation set, our approach outperforms the provided baselines, albeit with a modest margin.

Keywords

PAN 2025, Voight-Kampff Generative AI Authorship Verification, Text classification, AI-Generated Text Detection

1. Introduction

Thanks to advances in large language models (LLMs), it is now possible to generate high-quality texts with diverse and varied applications [1] . Language modeling has long been a focus of study for both language creation and comprehension (if language is identified as a complex system of expressions governed by a set of grammatical rules), but it was not until the release of the ChatGPT model [2] that this fascinating field became accessible to the public. With this new tool, one can extract information (such as relationships or events), summarize texts, or generate original content, such as a poem or an email.

As these models continue to evolve, their output has become increasingly indistinguishable from human writing—not only in grammatical accuracy, but also in style, tone, and rhetorical complexity. The line between machine-generated and human-written text has become increasingly blurred, as LLMs learn to replicate not only grammatical structures but also stylistic nuances, rhetorical devices, and even domain-specific jargon.

However, these advances raise significant challenges regarding the authenticity and regulation of their use. Between January 1, 2022, and May 1, 2023, the relative number of synthetically generated news articles increased by more than half (53.3 %) on respected news websites [3] . On disinformation sites, this increase was 474%.

This qualitative leap creates a paradox: while LLMs democratize access to creative tools, they also erode traditional mechanisms of authorship attribution. Determining the authorship of a text, that is, whether it was written by a human or a machine, has become a problem of unprecedented relevance. These tools have the potential to be used for unethical purposes, such as plagiarism, the creation of fake news, or spinning (mass production of messages), which can impact not only individuals but society as a whole [4].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
mjimgon@inta.es (M. Jimeno-Gonzalez); emcamara@ujaen.es (E. Martínez-Cámara); noeferna@pa.uc3m.es (N. Fernandez); pdiagar@inta.es (P. Díaz-García); laurena@ujaen.es (L. A. Ureña-López)

Moreover, regardless of whether LLMs are used maliciously, there is another issue: hallucinations produced by these models [1, 5]. These errors occur unpredictably and cannot be anticipated in advance. Hallucinations are fictitious statements presented as truths. This problem becomes particularly severe when an LLM is faced with tasks that require expert knowledge in a specific domain. The mere possibility that a machine could have authored a given text underscores the importance of the task at hand. Accurately determining whether a text has been written by a human or a machine is becoming increasingly relevant in everyday contexts.

In its simplest form, the original problem is deciding whether a text was written by a human or a machine. Methodologically, the problem is framed as a binary classification (human vs. AI). However, this approach is deceptively simple. One of the greatest challenges is the statistical convergence between human and artificial texts [6]. Since LLMs are trained on vast amounts of human-written texts, they have learned not only syntactic structures but also stylistic patterns and cognitive biases, blurring the boundary that might initially seem clear. However, it is true that these models do not merely replicate these patterns—they optimize them, potentially creating exploitable stylistic perfection.

To boost this area of research, the PAN 2025 [7] workshop introduced the 'Generative AI Authorship Verification Task' [8] that is divided into two sub-tasks. Task 1 focuses on the robustness and sensitivity of detection systems. In response to this challenge, we proposed an architecture that combines hand-crafted linguistic features with textual representations. Specifically, we integrated syntactic, structural, and lexical features alongside TF-IDF representations of the raw text. These features are then used in a stacking ensemble classifier comprising Random Forest, XGBoost, and LinearSVC as base learners, with Logistic Regression serving as the final estimator. This traditional machine learning pipeline allows for interpretability and flexibility while achieving competitive performance.

The central objective of this work is to develop a reliable and interpretable method for distinguishing between human-written and AI-generated text. In conclusion, this work contributes a promising approach that supports both effective classification and transparency, addressing key challenges in the field of generative AI content detection.

2. Related Work

One of the most accessible and widely used approaches for detection is the use of statistics based on linguistic features [9, 10, 11]. This set of features forms the foundation of the approach we will explore in the present work.

The clear advantage of this approach lies in the fact that it relies solely on the text to be classified, which facilitates its practical application in contexts where the generative model is not accessible. However, it is important to note that its effectiveness often depends on the availability of a representative reference corpus containing both human and machine-generated texts. This allows for proper calibration of decision thresholds and validation of the robustness of the identified patterns.

Among the features analyzed are lexical density (the ratio of content words to function words), the average number of sentences per paragraph, the distribution across grammatical categories (POS tags), and the atypical frequency of certain k-grams (contiguous sequences of k words). These characteristics can capture subtle differences in style, syntactic coherence, or lexical variability between human and generative model texts [12, 13].

When trained on labeled corpora, classifiers built on these statistical features have demonstrated competitive accuracy in distinguishing between human and machine-generated texts

Nevertheless, these statistical measures are not the only ones used in the task of classifying texts generated by language models. Following the taxonomy proposed by Wu et al. (2025) [14], statistical methods can be categorized into two major groups: white-box and black-box approaches.

White-box methods [15, 16, 17] require direct access to the original model, meaning access to its architecture and raw parameters. These variables are especially valuable for understanding how text is generated and how the model selects certain words or structures—i.e., for analyzing the model's decision-making processes in detail.

The statistics derived from this type of analysis are crucial for attributing authorship of a text to a specific model, as they rely on the model's internal outputs (such as logits) and architectural behavior. Among these metrics are: Rank [18], which indicates the position of a token in the ordered list of logits (with higher-ranking tokens being considered more probable by the model); Log-likelihood [19], which refers to the sum of the log-probabilities of each token given its preceding context; and Log-Likelihood Ratio Ranking (LLR) [20], which combines the previous two metrics for a more robust classification.

During the model development process, perplexity was also analyzed—a metric that measures the model's ability to correctly predict a sequence of text. In other words, it evaluates the model's level of "surprise" when processing a given input. This metric was employed to validate the hypothesis proposed by Li et al. (2024) [21], which states that automatically generated texts exhibit increased perplexity after undergoing a rewriting process, due to a greater deviation from the linguistic distributions expected by the model. The results were not encouraging.

Although white-box methods are highly effective in detecting texts generated by the model they are designed for, their performance significantly decreases when analyzing texts generated by other models.

Complementary to white-box strategies, black-box methods [22, 23, 24] offer a more flexible yet computationally demanding alternative for text classification tasks. Black-box statistical methods are employed in scenarios where direct access to the internal parameters of the generative model is unavailable. This approach, characterized by its greater methodological diversity, relies exclusively on the analysis of the generated text itself, without requiring any supplementary information about the underlying model.

However, one of the primary limitations of black-box methods lies in their computational intensity, as mentioned before. The complexity of the required analyses can result in high latency times, thereby limiting their suitability for real-time applications or contexts requiring rapid response.

Emerging techniques for detecting text generated by language models include digital watermarking [25, 26] and deep neural network-based approaches [27, 11, 28], notably leveraging large language models (LLMs).

3. Methodology

3.1. PAN dataset

Released by the PAN shared task organizers, the PAN dataset, contains both human-authored and AI-generated text, with the twist: the LLMs were instructed to change their style and mimic a specific human author. It includes a total of 23,707 samples, consisting of 9,101 (61%) human-authored texts and 14,606 (38%) AI-generated texts produced using twenty-two different LLMs.

3.2. Data Pre-Processing

We performed an analysis of the data to study the presence of featuring patterns of human and machine generated texts.

3.2.1. Lexical Complexity and Vocabulary

- Lexical Diversity: It is a central concept in quantitative linguistics, assesses the range and variability of vocabulary used in a text sample [29]. In our study, this measure helps identify patterns of lexical richness in texts produced by humans versus generative models. As shown in Figure 2, human-authored texts tend to display a more centered distribution with less dispersion at the extremes. In contrast, AI-generated texts show a higher concentration at elevated diversity levels, which may be interpreted as more uniform and stylistically refined output.
- Lexical Frequency : To evaluate the lexical relevance of terms within each document, we calculated the average TF-IDF (Term Frequency–Inverse Document Frequency) score. This metric weights

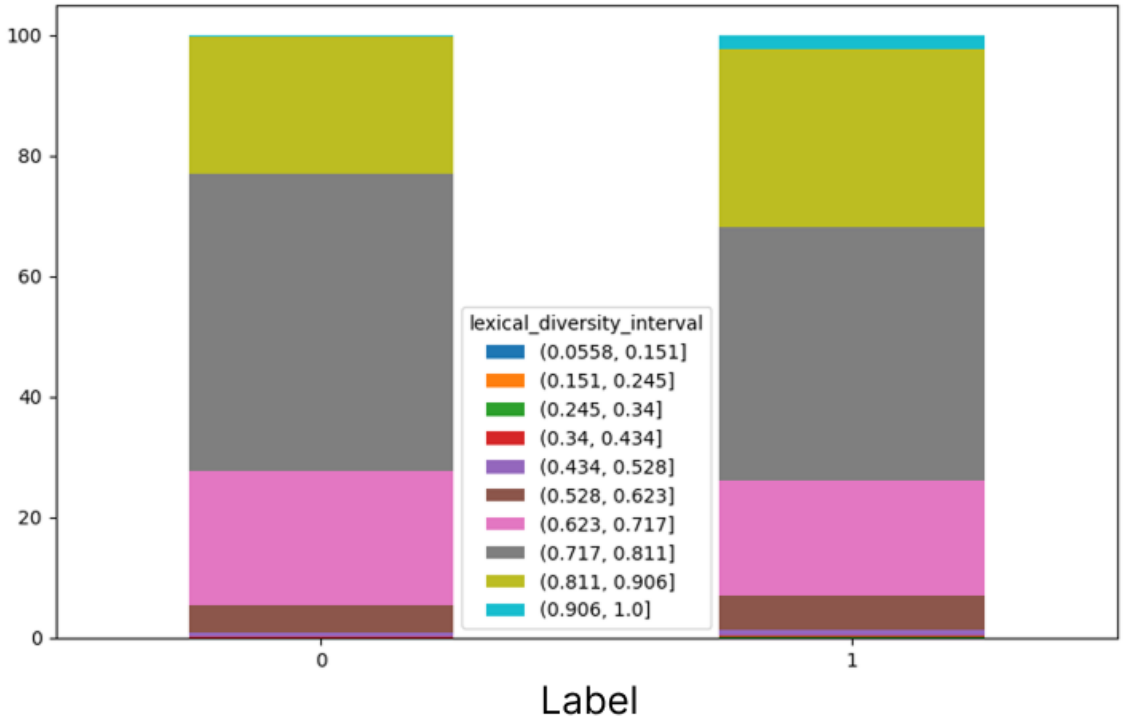

Figure 1: Lexical Diversity in intervals for label 0 (which correspond to humans), 1 (which correspond to machine).

term frequency according to its relative presence in the corpus, highlighting the most distinctive linguistic elements of each text. Its inclusion captures the balance between common words and infrequent terms that may provide unique semantic value. No major differences were found between both distributions, aside from the recurring observation that human-written texts tend to be less polarized. Similarly, it was observed that, in terms of average TF-IDF values, human texts exhibit higher scores than those generated by machines.

3.2.2. Text Structure

In the actual and the following section, we have employed the spaCy natural language processing library. Specifically, we utilized spaCy's [30] built-in part-of-speech (POS) tagger, which is integrated into the language models provided by the library (in our case, `en_core_web_sm` for English).

- Average Sentence Length: Calculated as the mean number of words per sentence, this metric provides insight into the structural complexity of the text.
- Average Word Length: Measures the average number of characters per word. Longer words are generally associated with more technical or sophisticated vocabulary.
- Total Number of Sentences: This feature allows control over the overall length of the text, which may affect the stability of other computed metrics.

3.2.3. Syntax and Part-of-Speech (POS)

A relative frequency analysis of various grammatical categories was conducted using Part-of-Speech tagging. The categories considered include determiners, adjectives, nouns, verbs, conjunctions (coordinating and subordinating), adverbs, ad-positions (prepositions and post-positions), auxiliaries, pronouns, unrecognized tokens, and punctuation marks.

The results (see Figure 2) show that texts generated by models exhibit higher usage of determiners, nouns, adjectives, and ad-positions. Conversely, human-written texts are characterized by more frequent use of punctuation, adverbs, conjunctions, and pronouns.

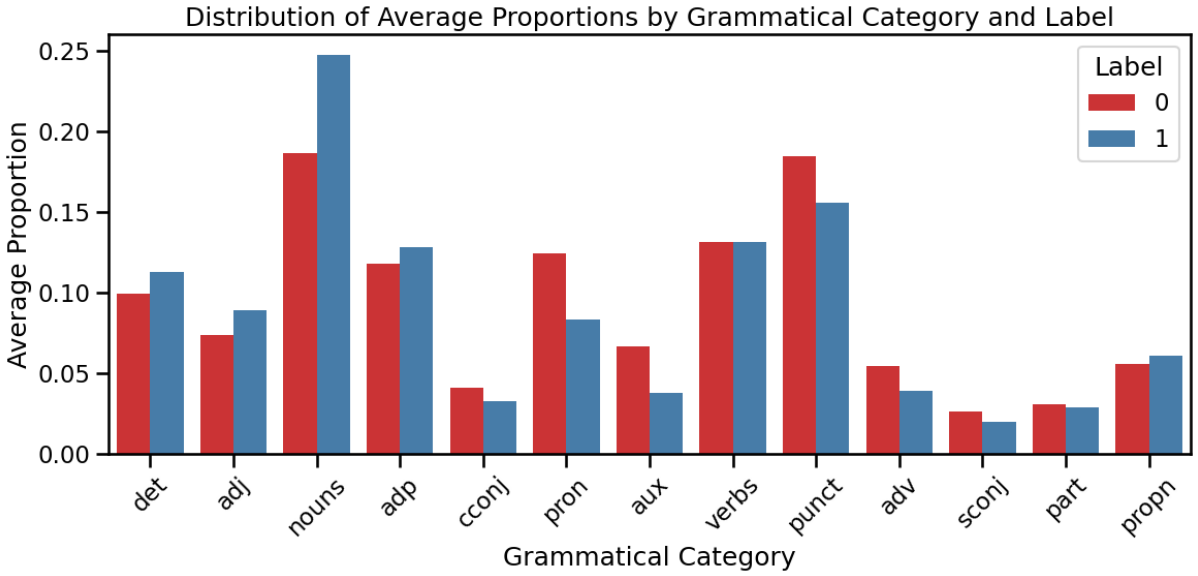

Figure 2: Distribution of proportions for label: 0 (which correspond to humans), 1 (which correspond to machine).

These differences suggest that human texts tend to show greater segmentation of ideas and a more coordinated style, likely influenced by communicative intent and personal context (as reflected in pronoun usage). In contrast, automatically generated texts display a more formal, informative, and grammatically structured construction, reflected in a higher proportion of ad-positions, determiners, and nouns.

3.3. Model Design and Classification Approach

Building on the previously discussed importance of statistical and linguistic features, the proposed model aims to combine the explanatory power of these variables with the strength of automatic text representation techniques, such as TF-IDF. To achieve this, a processing pipeline has been designed integrating both the full vectorization of the textual content—including unigrams—and the linguistic variables described earlier, preserving their structure divided into lexical, structural, and syntactic components. The full scope of variables its described in the table 1.

Once preprocessed, all these features are concatenated into a single feature space and used as input for a stacked ensemble classification model. This strategy allows the integration of different supervised learning approaches to enhance the system's robustness and generalization ability.

The ensemble consists of the following base classifiers:

- Random Forest: A decision tree-based model that introduces randomness in both data sampling and feature selection, thereby reducing overfitting and capturing nonlinear feature interactions.
- XGBoost: A boosting technique that iteratively optimizes a set of trees by minimizing the loss function, improving probabilistic classification performance.
- Support Vector Classifier (Linear SVC): A robust linear classifier, particularly effective in high-dimensional spaces such as those generated by TF-IDF vectors. Textual features were vectorized using `TfidfVectorizer` from the `scikit-learn` library, with default parameter settings. This settings correspond to a unigram-based representation (`ngram_range` = (1,1)), where each term is weighted according to its term frequency-inverse document frequency (TF-IDF) value, normalized using the L2 norm. No explicit constraints were placed on vocabulary size (`max_features` was left unspecified), and all terms occurring in at least one document were included (`min_df` = 1, `max_df` = 1.0). Binary weighting was disabled (`binary` = `False`), and standard smoothing was applied (`smooth_idf` = `True`).

Table 1
List of engineered features used in the model.

#	Feature Name	Classification
1	`lexical_frequency` (Frequency-based lexical score)	Lexical
2	`num_sentences` (Total number of sentences)	Structural
3	`num_tokens` (Total number of tokens)	Structural
4	`lexical_diversity` (Ratio of unique words to total words)	Lexical
5	`avg_sentence_length` (Mean words per sentence)	Structural
6	`avg_word_length` (Mean characters per word)	Structural
7	`proporcion_num_propn` (Proportion of proper nouns)	Syntactic
8	`proporcion_num_part` (Proportion of particles)	Syntactic
9	`proporcion_num_sconj` (Proportion of subordinating conjunctions)	Syntactic
10	`proporcion_num_det` (Proportion of determiners)	Syntactic
11	`proporcion_num_adj` (Proportion of adjectives)	Syntactic
12	`proporcion_num_nouns` (Proportion of nouns)	Syntactic
13	`proporcion_num_adp` (Proportion of adpositions)	Syntactic
14	`proporcion_num_cconj` (Proportion of coordinating conjunctions)	Syntactic
15	`proporcion_num_pron` (Proportion of pronouns)	Syntactic
16	`proporcion_num_adv` (Proportion of adverbs)	Syntactic
17	`proporcion_num_aux` (Proportion of auxiliary verbs)	Syntactic
18	`proporcion_num_verbs` (Proportion of main verbs)	Syntactic
19	`proporcion_num_punct` (Proportion of punctuation marks)	Syntactic

The intermediate predictions generated by these base models are combined using a logistic regression meta-model, which learns to weight the partial outputs to produce the final prediction. This architecture leverages the complementarity of models with different inductive capabilities, balancing performance and interpretability.

4. Results

In this section, we present an evaluation of our AI-generated text detection experiments. The comparison is conducted using the designated evaluation split of the dataset. We report results using well-established performance metrics, as outlined in the official PAN@CLEF 2025 evaluation guidelines[1].

Table 2 presents a comparative evaluation of the state-of-the-art baselines on the PAN validation set using six key metrics: ROC-AUC, Brier score, C@1, F1, F05U, and a computed mean of all metrics. For each test instance, we predicted the corresponding label (human or machine-generated) and produced calibrated probability scores, following the evaluation recommendations provided by the benchmark organizers.

Notably, our Approach attains a perfect or near-perfect performance, yielding the highest scores in every metric: a ROC-AUC of 0.996, Brier score of 0.978, C@1 of 0.976, F1 of 0.981, F05U of 0.986, and an overall mean of 0.983.

When compared to the strongest baseline, the Linear SVM with TF-IDF features, our Approach maintains equivalent performance in ROC-AUC (0.996) while demonstrating notable improvements in the Brier score (+0.027), F05U (+0.005), and mean score (+0.005). This indicates that our method not only preserves strong discriminative capability but also enhances probability estimation and performance on metrics that emphasize partial correctness (such as F05U and C@1).

In summary, the results highlight the efficacy of Our Model in outperforming both traditional feature-based classifiers and more unconventional methods across a comprehensive set of evaluation metrics, thereby establishing it as a robust and reliable solution for the task evaluated in the PAN validation set.

Table 3 presents the performance of Our Approach on the PAN test set, as reported after the final

[1] https://pan.webis.de/clef25/pan25-web/generated-content-analysis.html

Table 2
Results on PAN val set.

Model	ROC-AUC	Brier	C@1	F1	F05U	Mean
Our Approach	**0.996**	**0.978**	**0.976**	**0.981**	**0.986**	**0.983**
Linear SVM with TF-IDF features	0.996	0.951	0.984	0.980	0.981	0.978
PPMd Compression-based Cosine	0.786	0.799	0.757	0.812	0.778	0.786
Binoculars	0.918	0.867	0.844	0.872	0.882	0.877

Table 3
Results on PAN test set.

Model	ROC-AUC	Brier	C@1	F1	F05U	Mean
Our Approach	**0.97**	**0.903**	**0.882**	**0.957**	**0.938**	**0.91**
Our Approach (final test)	**0.811**	**0.841**	**0.807**	**0.818**	**0.860**	**0.838**

submission to the TIRA evaluation platform [31]. The model achieves strong and consistent results across all evaluation metrics: ROC-AUC of 0.970, Brier score of 0.903, C@1 of 0.882, F1 score of 0.957, F05U of 0.938, and a mean score of 0.910. In the same table, we can also see the final test scores, where our approach placed 17th out of 24 participating teams.

Compared to the validation results reported in Table 2, these outcomes demonstrate the model's ability to generalize effectively to unseen data, with only modest declines in performance, which are expected due to the inherent distributional shift between validation and test splits. Importantly, the model retains a high ROC-AUC and F1 score, indicating sustained discriminative power and classification accuracy. The Brier score and C@1 values remain competitive, further attesting to the model's well-calibrated probability outputs and its effectiveness in high-confidence decision-making scenarios.

5. Conclusion

In this paper, we presented our submission to the PAN shared task on generative AI content detection. The central objective of our work was to develop a reliable and interpretable approach for distinguishing between human-written and AI-generated text. Our experimental results confirm that this objective has been successfully met: the proposed method demonstrated competitive performance relative to state-of-the-art systems and passed the official evaluation on the TIRA platform, qualifying for the final competition results.

The combination of linguistic feature engineering and ensemble learning enabled both strong classification capabilities and interpretability, aligning with the goals stated at the outset. These findings validate the effectiveness of our approach in addressing the challenges posed by generative authorship verification.

For future work, we aim to further enhance the model's generalizability by evaluating its performance across a wider array of datasets to better assess its robustness under diverse real-world conditions. Additionally, we plan to examine the system's resilience to adversarial attacks by introducing controlled perturbations, thereby deepening our understanding of its limitations and improving its reliability in adversarial contexts.

Acknowledgements

This work was partly supported by the grants FedDAP (PID2020-116118GA-I00), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) and CONSENSO (PID2021-122263OB- C21) funded by MCIN/AEI/10.13039/501100011033, "ERDF A way of making Europe" and "European Union NextGenerationEU/PRTR". This work was also funded by the Ministerio para la Transformación Digital

y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to: Grammar and spelling check as well as text translation. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 1 (2023).

[2] A. Mathew, Is artificial intelligence a world changer? a case study of openai's chat gpt, Recent Progress in Science and Technology 5 (2023) 35–42.

[3] H. W. Hanley, Z. Durumeric, Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 18, 2024, pp. 542–556.

[4] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, Y. Zhang, A survey on large language model (llm) security and privacy: The good, the bad, and the ugly, High-Confidence Computing (2024) 100211.

[5] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55.

[6] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can ai-generated text be reliably detected?, arXiv preprint arXiv:2303.11156 (2023).

[7] 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.

[8] 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.

[9] S. Corston-Oliver, M. Gamon, C. Brockett, A machine learning approach to the automatic evaluation of machine translation, in: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, 2001, pp. 148–155.

[10] B. Alhijawi, R. Jarrar, A. AbuAlRub, A. Bader, Deep learning detection method for large language models-generated scientific content, Neural Computing and Applications 37 (2025) 91–104.

[11] R. Tang, Y.-N. Chuang, X. Hu, The science of detecting llm-generated text, Communications of the ACM 67 (2024) 50–59.

[12] M. Gallé, J. Rozen, G. Kruszewski, H. Elsahar, Unsupervised and distributional detection of machine-generated text, arXiv preprint arXiv:2111.02878 (2021).

[13] A. A. Hamed, X. Wu, Detection of chatgpt fake science with the xfakesci learning algorithm, 2024. URL: https://arxiv.org/abs/2308.11767. arXiv:2308.11767.

[14] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, A survey on llm-generated text detection: Necessity, methods, and future directions, 2024. URL: https://arxiv.org/abs/2310.14724. arXiv:2310.14724.
[15] R. Wang, H. Chen, R. Zhou, H. Ma, Y. Duan, Y. Kang, S. Yang, B. Fan, T. Tan, Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning, 2024. URL: https://arxiv.org/abs/2402.01158. arXiv:2402.01158.
[16] K. Wu, L. Pang, H. Shen, X. Cheng, T.-S. Chua, Llmdet: A third party large language models generated text detection tool, arXiv preprint arXiv:2305.15004 (2023).
[17] V. Verma, E. Fleisig, N. Tomlin, D. Klein, Ghostbuster: Detecting text ghostwritten by large language models, arXiv preprint arXiv:2305.15047 (2023).
[18] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, 2019. URL: https://arxiv.org/abs/1906.04043. arXiv:1906.04043.
[19] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, arXiv preprint arXiv:1908.09203 (2019).
[20] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, arXiv preprint arXiv:2306.05540 (2023).
[21] R. Li, W. Hao, W. Zhao, J. Yang, C. Mao, Learning to rewrite: Generalized llm-generated text detection, 2025. URL: https://arxiv.org/abs/2408.04237. arXiv:2408.04237.
[22] C. Mao, C. Vondrick, H. Wang, J. Yang, Raidar: generative ai detection via rewriting, arXiv preprint arXiv:2401.12970 (2024).
[23] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, arXiv preprint arXiv:2301.07597 (2023).
[24] Y. Tian, H. Chen, X. Wang, Z. Bai, Q. Zhang, R. Li, C. Xu, Y. Wang, Multiscale positive-unlabeled detection of ai-generated texts, arXiv preprint arXiv:2305.18149 (2023).
[25] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, in: International Conference on Machine Learning, PMLR, 2023, pp. 17061–17084.
[26] J. Ren, H. Xu, Y. Liu, Y. Cui, S. Wang, D. Yin, J. Tang, A robust semantics-based watermark for large language model against paraphrasing, arXiv preprint arXiv:2311.08721 (2023).
[27] A. M. Sarvazyan, J. Á. González, P. Rosso, M. Franco-Salvador, Supervised machine-generated text detectors: Family and scale matters, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 121–132.
[28] A. Bhattacharjee, H. Liu, Fighting fire with fire: can chatgpt detect ai-generated text?, ACM SIGKDD Explorations Newsletter 25 (2024) 14–21.
[29] J. Read, 2000: Assessing vocabulary. cambridge: Cambridge university press (2000).
[30] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.
[31] 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.