StylOch at PAN: Gradient-Boosted Trees with Frequency-Based Stylometric Features

Notebook for the PAN Lab at CLEF 2025

Jeremi K. Ochab[1,2,3,4,*], Mateusz Matias[3], Tymoteusz Boba[3] and Tomasz Walkowiak[4]

[1]Institute of Theoretical Physics, Jagiellonian University, Kraków, Poland
[2]M. Kac Center for Complex Systems Research, Jagiellonian University, Kraków, Poland
[3]Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Kraków, Poland
[4]Faculty of Information and Communication Technology, Wroclaw University of Science and Technology, Wroclaw, Poland

Abstract

This submission to the binary AI detection task is based on a modular stylometric pipeline, where: public spaCy models are used for text preprocessing (including tokenisation, named entity recognition, dependency parsing, part-of-speech tagging, and morphology annotation) and extracting several thousand features (frequencies of n-grams of the above linguistic annotations); light-gradient boosting machines are used as the classifier. We collect a large corpus of more than 500 000 machine-generated texts for the classifier's training. We explore several parameter options to increase the classifier's capacity and take advantage of that training set. Our approach follows the non-neural, computationally inexpensive but explainable approach found effective previously.

Keywords

generative AI detection, stylometry, explainability

1. Introduction

The rapidly developing landscape of Large Language Models (LLMs) has revolutionised natural language processing (NLP), enabling the use of machine-generated texts (MGTs) throughout society on a daily basis. The use of these tools, especially in some professional environments such as academic [1] , medical [2], legal, or news reporting, raises concerns around issues of plagiarism, factual reliability, and many others. The *"Voight-Kampff" Generative AI Authorship Verification Task* [3] at the PAN and ELOQUENT 2025 workshop [4], and specifically Subtask 1 "AI Detection Sensitivity" answers the urgent need to develop reliable model detection. The subtask is a classical binary text classification task, i.e. categorising a given text as a human or machine written. The additional challenge comes from changing the style of MGTs, mimicking specific human authors, testing on unseen models, and using obfuscation strategies. In submission to this subtask, we strive to expand on the simplistic non-neural feature-based classifiers that were previously found effective, using boosted trees with stylometric (linguistically explainable) features.

2. Background

2.1. MGT detection methods

There is a considerable variety of MGT detection methods reviewed in [5, 6], but also in the overview of last year's *Voight-Kampff Generative AI Authorship Verification Task* at PAN and ELOQUENT 2024 [7]. They included systems based on (i) terms, (ii) perplexity or logit statistics, (iii) watermarking, and their mixtures. The watermarking approach relies on embedding an imperceptible signature in the generated

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
jeremi.ochab@uj.edu.pl (J. K. Ochab); mateusz.matias@student.uj.edu.pl (M. Matias); tymoteusz.boba@student.uj.edu.pl (T. Boba); tomasz.walkowiak@pwr.edu.pl (T. Walkowiak)
0000-0002-7281-1852 (J. K. Ochab); 0000-0002-7749-4251 (T. Walkowiak)

texts at some stage of the generator training, text generation, or post-processing that modifies character- or word-level distributions. In the present submission, we disregard this approach due to the task's constraints. The logits statistics approach typically involves zero-shot white-box methods, i.e., ones that require access to the LLM generator (or its surrogate) in order to compute either the likelihood of a text being generated by it or features later used in a classifier. The black-box alternatives, instead, would machine-regenerate a given text sample and subsequently compare it to the original to obtain a similarity score. Finally, term-based systems are typically neural fine-tuned classifiers (from the BERT family with modifications) using word embeddings or linguistic (stylometric) features.

Our submission follows works such as [8, 9, 10, 11], which either utilised various stylometric features, augmented data, or expanded the training dataset. We especially find our approach similar to the simple SVM classifier on the TF-IDF features [12], which outranked all neural baselines and most neural-based submissions in the last year's task. Classifiers based on stylometric features were also found to be effective elsewhere [13].

2.2. MGT detection robustness

The performance of MGT detection can generally degrade due to two factors: out-of-distribution issues and attacks [6]. The former encompasses generalisation issues such as: cross-domain (involving changing text type, and consequently its vocabulary, style, topics, etc.), cross-language (involving not only switching the language of the text but also linguistic interference due to non-nativity of the authors) and cross-LLM (involving detection of text generators unseen during the detector's training). The latter includes: paraphrasing output of one LLM by another (therefore changing the textual feature distribution of the former) [14], adversarial text perturbations on different levels (characters [15], syntax, [16] or lexis, [17]), prompt engineering (taking advantage of in-context learning to change LLM's characteristics by varying prompts [18, 19] including mimicking specific authors or character profiling [20]) and other attacks.

Reportedly [21], supervised detectors can generalise reasonably well across LLM scales but less so across model families. On the other hand, issues that were found to be challenging were incorporating unseen languages and performing simple attacks such as Unicode obfuscation or shortening text length [7]. Our own approach has been found vulnerable to cross-domain detection [22] as tested on [23], but on a closed domain it was robust to one-step paraphrasing. Furthermore, the unexpected performance of the aforementioned SVM TF-IDF classifier [12] was mainly due to its robustness to obfuscation.

We did not explicitly design our detector to target any of these issues; however, we follow the general recommendation [6] that supervised detectors can effectively defend against some of them by continually expanding training datasets (with adversarial examples, examples of LLM families, examples of text types, etc.) and fine-tuning even on small samples.

3. System Overview

In general, our submission employed (i) gradient-boosted tree models together with (ii) feature engineering and, crucially, (iii) a large training dataset. We did not target any obfuscation techniques. Similarly to our previous work [22, 24], we used a modular Python pipeline for interpretable stylometric analysis being developed for CLARIN-PL[1] [25]. It is designed to connect text preprocessing and linguistic feature extraction with various existing NLP tools, classifiers, explainability modules, and visualisation.

3.1. Data source

Following our unremarkable attempt [22] (F1 = 0.54 compared to an ensemble of stylometric features and transformers [26] and the highest ranked result 0.81) in cross-domain MGT detection on *AuTexTification* [23] benchmark – where training was performed on tweets, how-to articles and legal documents,

[1] https://gitlab.clarin-pl.eu/stylometry/cl_explainable_stylo

while testing on reviews and news – we decided that our model needs as comprehensive and varied training data as possible in order for the validation result to hold on test set.

For that purpose, we have collected in total 563 571 text samples from several openly accessible datasets [3, 23, 27, 18, 28, 29, 30, 31] designed as benchmarks in MGT detection, see Table 1. The number reported above already takes into account dropouts due to issues with special characters or incompatibility of data structure that we were not able to solve within the time constraints of the PAN's task. In particular cases, not all data were incorporated (e.g. training but not validation set in the case of *AuTexTification* and *PAN's Voight-Kampff Generative AI Detection*; consequently, not all available genres were included). Some of these datasets themselves were collected from other openly available datasets and augmented with the generated texts. The total number of LLM labels available in that dataset was 348.

The source, genre and model labels were not used in the training.

3.2. Stylometric Features

We considered two options: either a closed set of predefined but more interpretable features or an open set features generated programmatically but still partly based on linguistic analysis.

Regarding the first, when analysing our own Wikipedia-based dataset [22], we used StyloMetrix [32]. This open-source stylometric text analysis library calculates the appearance of 195 predefined features that include grammatical forms (tenses, modal verbs, etc.), parts of speech, lexical items (types of pronouns, hurtful words, etc.), aspects connected to social media (e.g. sentiment analysis), syntactic forms, and general text statistics (e.g. type-token ratio). StyloMetrix uses the spaCy model for English to extract these features. The classifiers based on this small feature set consistently scored lower than the alternative, so the final submission comprised only the second option.

The second option follows the basic ideas used in the R package `stylo` [33], which is mainly computing token n-grams, but augmented with the various annotations. At present, for preprocessing steps and said annotations (tokenisation, named entity recognition, dependency parsing, part-of-speech tagging, and morphology annotation) we use spaCy [34] model `en_core_web_lg`. Specifically, we computed the normalised frequencies of:

- lemmas (from uni- to trigrams), excluding named entities,
- part-of-speech tags (from uni- to quadrigrams) including punctuation,
- dependency-based bigrams (where token neighbourhood is defined by the distance in the dependency tree), excluding named entities,
- morphological annotations (unigrams) including entity types (i.e. using Named Entity Recognition to find named entities and replacing them with their types)

Each of the four feature classes could contain a maximum of 1500 items. This particular set of features admittedly comes from some unresolved technical issues, but also from repeated trial and error on yet other authorial attribution datasets. For instance, elsewhere [22] we have found that punctuation features, such as the 'SPACE' token, can detect human mistakes or artefacts in LLM processing or further data post-processing (a redundant whitespace character, e.g., at the beginning of a paragraph or a second one between words). In that choice of feature classes we also try to minimise, although not strictly enforce, the generation of duplicate versions of the same feature in separate classes.

As presented in Table 2 we also testes so-called *culling* (i.e., ignoring features with *document frequency* strictly higher or lower than the given threshold). In the present submission, a majority of our models did not use culling. In one case, we set the minimum document frequency to 0.1 (that is, about 50k out of 500k documents), which reduced the number of features from the initial 4594 to 3264.

3.3. Classifier

We take advantage of the existing solutions: Light Gradient-Boosting Machine (LGBM) [35] as the state-of-the-art boosted trees classifier and Scikit-learn [36] for feature counting and cross-validation.

Table 1
Overview of datasets used in training. Items in parentheses refer to all the collected samples, while items without parentheses refer to the samples used in training.

Dataset	Samples	Word Count	Genres	Models
PAN'25 Generative AI Detection	23 704 (23 707)	14 727 408	essays fiction news	*human*, deepseek-r1-distill-qwen-32b falcon3-10b-instruct, gemini-1.5-pro gemini-2.0-flash, gemini-pro gemini-pro-paraphrase, gpt-3.5-turbo gpt-4-turbo, gpt-4-turbo-paraphrase gpt-4.5-preview, gpt-4o, gpt-4o-mini llama-2-70b-chat, llama-2-7b-chat llama-3.1-8b-instruct, llama-3.3-70b-instruct mistral-8b-instruct-2410 mistral-7b-instruct-v0.2 mixtral-8x7b-instruct-v0.1, o3-mini qwen1.5-72b-chat-8bit, text-bison-002
Autextification [23]	21 832 (21 832)	1 367 323	news reviews (tweets how-to legal)	*human*, BLOOM-1B7, BLOOM-3B, BLOOM-7B1, babbage, curie, text-davinci-003
CHEAT [27]	15 394 (15 395)	165 584	abstracts	gpt-3.5-turbo
HC3 [18]	0 (48 644)	12 492 921	Q&A finance medicine Wikipedia	*human*, ChatGPT
HC3 Plus [28]	148 237 (148 402)	11 250 436	news summaries translations question paraphrases	*human*, GPT-3.5-Turbo-0301
MAGE [29]	318 958 (319 071)	67 471 388	opinions reviews news Q&A stories reasoning Wikipedia abstracts	*human*, gpt-3.5-turbo, text-davinci-002, text-davinci-003, gpt_j, gpt_neox, opt, flan_t5, t0, bloom_7b, GLM130B
Multitude [30]	29 459 (29 460)	6 175 907	news	*human*, alpaca-lora-30b, gpt-3.5-turbo gpt-4, text-davinci-003, vicuna-13b llama-65b, opt-66b, opt-iml-max-1.3b
M4 [31]	5987	XXX	Wikipedia abstracts peer reviews news briefs	*human*, GPT-4, ChatGPT text-davinci-003, Cohere Dolly-v2, BLOOMz 176B

Following our previous experience on other, smaller datasets – mainly in English and Polish languages – during pipeline development, the LGBM classifiers parameters were set to: DART boosting, `learning_rate` = 0.5, enabled bagging (randomly selecting `bagging_fraction` = 0.8 of data without resampling every `bagging_freq` = 3 iterations). At this time, we used the binary classifier, but it is possible – and in fact in can be beneficial [22] – to train a multiclass model using the LLM labels, see Table 1, and then map it back to the binary 'human vs. machine' labels.

Table 2
Overview of submitted model parameters and their sizes. The *Model size* refers to the size of model saved in a `.txt` file.

Model name	Feature culling	LGBM parameters			Model size [kB]
		`num_leaves`	`num_iterations`	`max_depth`	
small	0	10	100	8	229
medium	0	12	500	10	851
big	0	20	1500	12	3685
culled	0.1	20	1500	12	3648

The following parameters were used to produce separate submissions with increasing model capacity:

- maximal number of leaves per tree (`num_leaves`),
- number of boosting iterations (`num_iterations`)
- maximal depth of the tree model (`max_depth`),

In our smaller pre-submission experiments (e.g., human authorship attribution on 2-100 novels, resulting in the number of samples of the order of thousands or tens of thousands at most) satisfactory results were obtained with `num_leaves` = 5, `num_iterations` = 100, `max_depth` = 5. We decided that with 500k text samples, 4k features and 348 LLM labels, the LGBM classifier required a higher capacity, hence we submitted three classifier versions: *small*, *medium* and *big*, listed in Table 2. Further hyperparameter optimisation is possible, but was not performed in the present submission.

Since LGBM training is fast, we used the stratified 10-fold cross-validation (CV) scheme to obtain more reliable validation and test error estimation. We then decided to validate both a classifier from a single fold (*-single*) and the probability scores averaged over classifiers trained on all CV folds (*-cv*).

4. Results

4.1. Evaluation setup

The environment for running and evaluating submissions to Subtask 1 "AI Detection Sensitivity" of the *PAN: Voight-Kampff Generative AI Detection 2025* task was TIRA [37]. This platform allows dockerised submissions in order to ensure their reproducibility. Upon submission our contribution was validated on two datasets, to which we refer as: "Validation 1" – the validation split of the dataset available for training (available texts and labels), "Validation 2" – the dataset used for evaluation at TIRA (not available to see its contents). Both datasets could be used for classifier evaluation and selection; see Table 3. TIRA platform produced the following six evaluation metrics (all on scale 0-1, with 1 representing the perfect score):

- **ROC-AUC**: The area under the ROC (Receiver Operating Characteristic) curve
- **Brier**: The complement of the Brier score (equivalent to mean squared loss)
- **C@1**: A modified accuracy score that breaks ties by assigning *non-answers* (class probability = 0.5) the average accuracy of the remaining cases
- $\mathbf{F_1}$: The harmonic mean of precision and recall
- $\mathbf{F_{0.5u}}$: A modified $F_{0.5}$ measure (where precision weighs more than recall) that treats non-answers as false negatives
- The arithmetic mean of all above.

The final evaluation was also appended with the False Positive Rate (FNR) and False Negative Rate (FNR). The submissions were ranked by a macro-average of the arithmetic mean over all individual data sources (all individual datasets contained in the test and the ELOQUENT collections).

Table 3
Mean performance on validation sets and the unobfuscated test set against the best baseline (TF-IDF) and the best contribution ('mdok'). The values are arithmetic means of evaluation metrics.

Approach	Validation 1	Validation 2	Test
small-single	0.943	0.885	0.885
medium-single	0.967	0.93	0.905
big-single	0.972	0.926	0.917
big-cv	**0.976**	0.933	**0.921**
big-cv-culled	0.972	**0.951**	0.915
TF-IDF	0.978	0.971	0.94
best	–	0.979	0.991

Table 4
Detailed performance of *big-cv* model (top three rows) on test sets against the best baseline and the best contribution. (a) The main test set without obfuscation, (b) test set incorporating most of the ELOQUENT obfuscation contributions, (c) final evaluation (macro-averages over all individual datasets).

Evaluation set	ROC-AUC	Brier	C@1	$\mathbf{F_1}$	$\mathbf{F_{0.5u}}$	Mean	FPR	FNR
(a) Test	0.958	0.912	0.882	0.911	0.943	-	-	-
(b) ELOQUENT-01	0.884	0.749	0.625	0.746	0.877	-	-	-
(c) Final	0.793	0.866	0.821	0.823	0.853	0.844	0.131	0.201
Final TF-IDF	0.838	0.871	0.836	0.827	0.862	0.856	0.128	0.153
Final best	0.853	0.896	0.894	0.898	0.903	0.899	0.094	0.108

4.2. Evaluation results

Table 3 presents the arithmetic mean scores of our submissions on the validation sets (available during submission) and the unobfuscated test set together with the best-ranked baseline and participant contribution. Both model capacity (model size and no feature culling) and cross-validation visibly led to higher scores on both datasets. The results from the obfuscated ELOQUENT dataset available at TIRA showed the same pattern in ROC-AUC metric, but there was no generally discernible dependence on the model size in the other metrics. The detailed test results for the selected *big-cv* model are shown in Table 4 (a).

The final results are shown in Table 4. In general, one can observe $\mathrm{F}_{0.5\mathrm{u}} > \mathrm{F}_1 > \mathrm{C@1}$ which is probably due to FN > FP and TP > TN and consequently a higher recall of MGTs.

5. Conclusion

Two general observations are: (1) larger capacity of boosted trees increased the detection performance, and (2) obfuscation considerably reduced it, Although the our model have not reached the baseline TF-IDF scores, in the outlook, the boosted trees have the capacity to learn on a larger number of features, so incorporating TF-IDF features [12] or standardising feature frequencies, found to be greatly effective in stylometry [38, 33], and other classic feature engineering techniques could be beneficial. The straightforward augmentation of the training set with obfuscated samples can further improve the results. The other unexplored avenue is simply hyperparameter optimisation (both in terms of feature set and LGBM parameters). The main computational overhead in our method is feature extraction on the large training dataset. Classifier training (and training continuation), inference and explanation [39] is inexpensive. In summary, we perceive it as a trade-off between the smaller cost and greater explainability of boosted trees and the better generalisation of neural-based systems.

Acknowledgments

The research for this publication has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University. JKO's research on the stylometric pipeline was financed by European Funds for Smart Economy, FENG program, CLARIN – Common Language Resources and Technology Infrastructure, project no.FENG.02.04-IP.040004/24-00.

MM and TB participated in the submission as a programming assignment from the "AI Workshop II" course at Jagiellonian University during the summer term of 2025.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull's model in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, Z. Wang, Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing, Journal of the Association for Information Science and Technology 74 (2023) 570–581.

[2] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, C. Rizzo, Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health, Frontiers in public health 11 (2023) 1166120.

[3] 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.

[4] 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.

[5] E. Crothers, N. Japkowicz, H. L. Viktor, Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods, IEEE Access 11 (2023) 70977–71002. URL: https://doi.org/10.1109/ACCESS.2023.3294090. doi:10.1109/ACCESS.2023.3294090.

[6] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, D. F. Wong, A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions, Computational Linguistics (2025) 1–64. URL: https://doi.org/10.1162/coli_a_00549. doi:10.1162/coli_a_00549.

[7] 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: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. d. Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2486–2506. URL: https://ceur-ws.org/Vol-3740/paper-225.pdf.

[8] M. Guo, Z. Han, H. Chen, J. Peng, A Machine-Generated Text Detection Model Based on Text Multi-Feature Fusion, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. d. Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2593–2602. URL: https://ceur-ws.org/Vol-3740/paper-238.pdf.

[9] P. Miralles, A. Martín, D. Camacho, Team aida at PAN: Ensembling Normalized Log Probabilities, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2807–2813. URL: http://ceur-ws.org/Vol-3740/paper-268.pdf.

[10] A. Yadagiri, D. Kalita, A. Ranjan, A. K. Bostan, P. Toppo, P. Pakray, Team cnlp-nits-pp at PAN: Leveraging BERT for Accurate Authorship Verification: A Novel Approach to Textual Attribution, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2976–2987. URL: http://ceur-ws.org/Vol-3740/paper-290.pdf.

[11] L. Guo, W. Yang, L. Ma, J. Ruan, BLGAV: Generative AI Author Verification Model Based on BERT and BiLSTM, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2585–2592. URL: http://ceur-ws.org/Vol-3740/paper-237.pdf.

[12] L. Lorenz, F. Z. Aygüler, F. Schlatt, N. Mirzakhmedova, BaselineAvengers at PAN 2024: Often-Forgotten Baselines for LLM-Generated Text Detection, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. d. Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2761–2768. URL: https://ceur-ws.org/Vol-3740/paper-262.pdf.

[13] C. Opara, StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis, in: A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, I. I. Bittencourt (Eds.), Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky - 25th International Conference, AIED 2024, Recife, Brazil, July 8-12, 2024, Proceedings, Part II, volume 2151 of *Communications in Computer and Information Science*, Springer, 2024, pp. 105–114. URL: https://doi.org/10.1007/978-3-031-64312-5_13. doi:10.1007/978-3-031-64312-5_13.

[14] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks, Transactions on Machine Learning Research (2025). URL: https://openreview.net/forum?id=OOgsAZdFOt.

[15] H. Stiff, F. Johansson, Detecting computer-generated disinformation, International Journal of Data Science and Analytics 13 (2022) 363–383. URL: https://doi.org/10.1007/s41060-021-00299-5. doi:10.1007/s41060-021-00299-5.

[16] M. M. Bhat, S. Parthasarathy, How Effectively Can Machines Defend Against Machine-Generated Fake News? An Empirical Study, in: A. Rogers, J. Sedoc, A. Rumshisky (Eds.), Proceedings of the First Workshop on Insights from Negative Results in NLP, Association for Computational Linguistics, Online, 2020, pp. 48–53. URL: https://aclanthology.org/2020.insights-1.7/. doi:10.18653/v1/2020.insights-1.7.

[17] E. Crothers, N. Japkowicz, H. Viktor, P. Branco, Adversarial Robustness of Neural-Statistical Features in Detection of Generative Transformers, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8. URL: https://ieeexplore.ieee.org/document/9892269. doi:10.1109/IJCNN55064.2022.9892269, iSSN: 2161-4407.

[18] 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, CoRR abs/2301.07597 (2023). URL: https://doi.org/10.48550/arXiv.2301.07597. doi:10.48550/ARXIV.2301.07597, arXiv: 2301.07597.

[19] Z. Liu, Z. Yao, F. Li, B. Luo, On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing, 2024. URL: http://arxiv.org/abs/2306.05524. doi:10.48550/arXiv.2306.05524, arXiv:2306.05524 [cs].

[20] K. Przystalski, J. K. Argasiński, N. Lipp, D. Pacholczyk, Building Personality-Driven Lan-

guage Models: How Neurotic is ChatGPT, Synthesis Lectures on Engineering, Science, and Technology, Springer Nature Switzerland, Cham, 2025. URL: https://link.springer.com/10.1007/978-3-031-80087-0. doi:10.1007/978-3-031-80087-0.

[21] A. M. Sarvazyan, J. González, P. Rosso, M. Franco-Salvador, Supervised Machine-Generated Text Detectors: Family and Scale Matters, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 121–132. doi:10.1007/978-3-031-42448-9_11.

[22] K. Przystalski, J. K. Argasiński, I. Grabska-Gradzińska, J. Ochab, Stylometry recognizes human and llm-generated texts in short samples, 2025. Manuscript submitted for publication to *Expert Systems with Applications*.

[23] A. M. Sarvazyan, J. González, M. Franco-Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[24] J. K. Argasiński, I. Grabska-Gradzińska, K. Przystalski, J. K. Ochab, T. Walkowiak, Stylometric analysis of large language model-generated commentaries in the context of medical neuroscience, International Conference … (2024) 281–295. URL: https://link.springer.com/chapter/10.1007/978-3-031-63775-9_20. doi:10.1007/978-3-031-63775-9_20.

[25] J. K. Ochab, T. Walkowiak, Implementing interpretable models in stylometric analysis, in: Digital Humanities 2024: Conference Abstracts, George Mason University (GMU), Washington, D.C., 2024.

[26] G. Mikros, A. Koursaris, D. Bilianos, ..., Ai-writing detection using an ensemble of transformers and stylometric features., IberLEF … (2023).

[27] P. Yu, J. Chen, X. Feng, Z. Xia, CHEAT: A Large-scale Dataset for Detecting CHatGPT-writtEn AbsTracts, IEEE Transactions on Big Data (2025) 1–9. URL: https://ieeexplore.ieee.org/abstract/document/10858415. doi:10.1109/TBDATA.2025.3536929.

[28] Z. Su, X. Wu, W. Zhou, G. Ma, S. Hu, HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus, 2024. URL: http://arxiv.org/abs/2309.02731. doi:10.48550/arXiv.2309.02731, arXiv:2309.02731 [cs].

[29] Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, Y. Zhang, MAGE: Machine-generated Text Detection in the Wild, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 36–53. URL: https://aclanthology.org/2024.acl-long.3/. doi:10.18653/v1/2024.acl-long.3.

[30] D. Macko, R. Moro, A. Uchendu, J. Lucas, M. Yamashita, M. Pikuliak, I. Srba, T. Le, D. Lee, J. Simko, M. Bielikova, MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9960–9987. URL: https://aclanthology.org/2023.emnlp-main.616/. doi:10.18653/v1/2023.emnlp-main.616.

[31] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. Mohammed Afzal, T. Mahmoud, T. Sasaki, T. Arnold, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1369–1407. URL: https://aclanthology.org/2024.eacl-long.83/.

[32] I. Okulska, D. Stetsenko, A. Kołos, A. Karlińska, K. Głąbińska, A. Nowakowski, Stylometrix: An open-source multilingual tool for representing stylometric vectors, arXiv preprint arXiv:2309.12810 (2023).

[33] M. Eder, M. Kestemont, J. Rybicki, Stylometry with R: A Package for Computational Text Analysis, The R Journal 8 (2016) 1–15. doi:10.32614/RJ-2016-007.

[34] I. Montani, M. Honnibal, M. Honnibal, A. Boyd, S. V. Landeghem, H. Peters, explosion/spaCy:

v3.7.2: Fixes for APIs and requirements, 2023. URL: https://doi.org/10.5281/zenodo.10009823. doi:10.5281/zenodo.10009823.

[35] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017) 3146–3154.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[37] 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: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, volume 13982, Springer Nature Switzerland, Cham, 2023, pp. 236–241. URL: https://link.springer.com/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20, series Title: Lecture Notes in Computer Science.

[38] J. Burrows, 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing 17 (2002) 267–287. doi:10.1093/llc/17.3.267.

[39] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature Machine Intelligence 2 (2020) 2522–5839.