Overview of PAN 2025: Generative AI Detection, Multilingual Text Detoxification, Multi-author Writing Style Analysis, and Generative Plagiarism Detection

Extended Abstract

Janek Bevendorff[1], Daryna Dementieva[2], Maik Fröbe[3], Bela Gipp[4], André Greiner-Petter[4], Jussi Karlgren[5], Maximilian Mayerl[6], Preslav Nakov[7], Alexander Panchenko[8], Martin Potthast[9,10,11], Artem Shelmanov[7], Efstathios Stamatatos[12], Benno Stein[13], Yuxia Wang[7], Matti Wiegmann[13(✉)], and Eva Zangerle[6]

[1] Leipzig University, Leipzig, Germany
[2] Technical University of Munich, Munich, Germany
[3] Friedrich Schiller University Jena, Jena, Germany
[4] Georg-August-Universität, Göttingen, Germany
[5] Silo AI, Helsinki, Finland
[6] University of Innsbruck, Innsbruck, Austria
[7] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[8] Skoltech and AIRI, Moscow, Russia
pan@webis.de
[9] University of Kassel, Kassel, Germany
[10] hessian.ai, Darmstadt, Germany
[11] ScaDS.AI, Leipzig, Germany
[12] University of the Aegean, Samos, Greece
[13] Bauhaus-Universität Weimar, Weimar, Germany
matti.wiegmann@uni-weimar.de

Abstract. The paper gives a brief overview of the four shared tasks organized at the PAN 2025 lab on digital text forensics and stylometry to be hosted at CLEF 2025. The goal of the PAN lab is to advance the state-of-the-art in text forensics and stylometry through an objective evaluation of new and established methods on new benchmark datasets. Our three tasks in 2025 will be: (1) generative AI detection, particularly in mixed and obfuscated authorship scenarios, (2) multilingual text detoxification, a continued task that aims re-formulate text in a non-toxic way for multiple languages, and (3) multi-author writing style analysis, a continued task that aims to find positions of authorship change., and(4) generative plagiarism detection, a new task that targets source retrieval and text alignment between generated text and source documents.

As with the previous editions, PAN invites software submissions as easy-to-reproduce docker containers; more than 400 pieces of software have been submitted from PAN'12 through PAN'24 combined, with all recent evaluations running on the TIRA experimentation platform [11].

C. Hauff et al. (Eds.): ECIR 2025, LNCS 15576, pp. 434–441, 2025.
https://doi.org/10.1007/978-3-031-88720-8_64

1 Introduction

PAN is a workshop series and a networking initiative for stylometry and digital text forensics. PAN hosts computational shared tasks on authorship analysis, computational ethics, and the originality of writing. Since the workshop's inception in 2007, we organized 68 shared tasks[1] and assembled 58 evaluation datasets[2] plus nine datasets contributed by the community.

In 2024, our four tasks concluded with 147 submissions and 74 notebook papers. Together with the ELOQUENT Lab, we introduced a new challenge on Generative AI Authorship Verification, which was very successful with 34 submitting teams. Given the continued relevance of the topic, we have expanded the organizing team and extended the task to *Generative AI Detection*, where we focus on detector sensitivity in the presence of obfuscation and mixed human-machine authorship. The newly introduced task on *Multilingual Text Detoxification* aligns with our community's interest in countering toxicity and generative tasks, and has attracted notable participation. We will therefore continue to develop the task in 2025. The *Multi-Author Writing Style Analysis* task was continued in 2024 with a new dataset and structured around topic heterogeneity as an indicator of difficulty. The task attracts consistent participation of high technical quality, while the problem is still relevant and offers room for improvement, so we continue the task with minor modifications in 2025. With the 2024 task on Oppositional Thinking Analysis, we discontinue our long and very successful line of research on the ethics of Social Web phenomena. In its place we introduce *Generated Plagiarism Detection*, where we focus in particular on the detection of near-verbatim text reuse by large language models. We briefly outline the upcoming tasks in the sections that follow.

2 Generative AI Detection

With generative AI, we now have the ability to produce high-quality discursive texts on virtually any topic, approaching human-like standards of writing [19]. On one hand, this is a great achievement, but on the other hand, it is also a cause for concerns. Generative AI introduces new challenges in education, as students can now generate essays on any topic and complete assignments without actually investing time and effort [20,21]. Academia is facing new forms of dishonesty, with reports of automatically fabricated articles and reviews, either in whole or in part. The media is alarmed by synthetic misinformation and disinformation articles, and social platforms could be overloaded with content generated by bots. Recognizing the "fingerprint" of AI in texts is the foundation for a healthy information ecosystem in the future.

In the wild, the process of text creation may involve various combinations of human writing and machine generation, with the role of LLMs being to assist humans in composing and refining text [1]. In an educational scenario, students

[1] Find PAN's past shared tasks at pan.webis.de/shared-tasks.html.
[2] Find PAN's datasets at pan.webis.de/data.html.

can first write their assignment reports and then improve them with the help of LLMs. In journalism, fact-checkers have observed the production of fake news, which involves a mixture of machine-generated texts and human efforts [28]. Therefore, in this task, we introduce the common practical scenario of collaborative texts generated by humans and machines. The wide scope of the task is ensured by considering multiple domains and utilizing various LLMs for generation.

We hence study the following sub-tasks:

Subtask 1 (Webis) AI Detection Sensitivity Analysis for recognizing unobfuscated and obfuscated LLM style: (1) Given a document, determine whether it was written by a human or an AI. (2) Given a document with a machine obfuscation of known or unknown origin, determine whether it was written by a human or an AI.

Subtask 2 (MBZUAI++) Fine-grained recognition of human-AI collaborated document. Given a document collaboratively authored by humans and models, our goal is to classify it into one of the following categories: (1) fully human-written;(2) fully machine-generated; (3) human-initiated, then machine-continued; (4) human-written, then machine-polished; (5) machine-written, then machine-humanized (obfuscated); (6) machine-written, then human-edited; (7) deeply-mixed text, where some parts are written by a human and some are generated by a machine.

3 Multilingual Text Detoxification

Even with various regulations in place across countries and social media platforms [10], digital abusive speech remains a significant issue. One potential approach to address this challenge is automatic text detoxification, a text style transfer (TST) approach that transforms toxic language into a more neutral or non-toxic form. Thus, AI safety and the need for approaches to mitigating abusive speech risks [3] are still relevant.

So far, we have developed the parallel text detoxification corpora for nine languages. The first parallel corpus was presented for English [14]. Then, we transferred the data collection pipeline to Russian that built a base for the first shared task on the text detoxification task at the Dialogue Evaluation forum [6].[3] With data present for these two language, we conducted the first experiments on multilingual and cross-lingual text detoxification [8] showing the challenge on transferring the knowledge on toxicity and detoxification between languages.

As a result, in this first edition of the shared task on multilingual text detoxification TextDetox CLEF 2024 [7], we covered several languages from different part of the world: together with existing data for English and Russia, we had also Spanish, German, Chinese, Arabic, Hindi, Ukrainian, and Amharic. The prepared 1 000 pairs specifically for the shared task were divided into dev set

[3] Monolingual shared task at dialog-21.ru/en/evaluation: russe.nlpub.org.

(400) and test set (600). The final leaderboard was based on the crowdsourcing evaluation of the 100 test set subset via Toloka.ai. The key takeaway from this edition of the shared task is that, despite advancements in Large Language Models, achieving effective multilingual and cross-lingual text detoxification—especially for less commonly spoken languages—remains a significant challenge.

Thus, in 2025 edition, we are extending even to more languages—French, Italian, Hebrew, and Japanese—dividing the shared task into two stages:

Sub-task 1 will be multilingual text detoxification. We will provide the parallel training data of toxic-neutral pairs for languages from 2024 edition. With such data, participants will be able to fine-tune their multilingual *text2text* generation models in a supervised manner.

Sub-task 2 will focus on the exploration of more advanced techniques of cross-lingual text detoxification knowledge transfer for new languages. We will not provide for them such parallel training datasets. Thus, we will encourage participants to explore methods for cross-lingual possibilities of their solutions.

The evaluation setups—both automatic and manual—will be based on the main three parameters: (i) *style transfer accuracy*—the new paraphrase should be non-toxic; (ii) *content similarity*—the content should be saved as much as possible; (iii) *fluency*—the resulted text should be fluent but may contain some minor mistakes (as the majority of the original toxic samples are examples from posts from social networks). In the end, they will be combined in the overall score J.

4 Multi-author Writing Style Analysis

Multi-author writing style analysis examines documents written by multiple individuals to identify specific points at which authorship shifts occur. By leveraging variations in writing style, it aims to segment a given document into distinct sections corresponding to different authors. Multi-author writing style analysis provides a foundation for intrinsic plagiarism detection, allowing the identification of plagiarized content without the need for an external reference corpus.

This task (formerly known as the Style Change Detection task) was introduced at PAN'16. In the first edition, the task was to identify and group the authors of fragments of a document [15]. In 2017, the task evolved to determine whether a document was written by a single or multiple authors [16]. In addition, for multi-author documents, participants were asked to find the positions of style changes. The task was then relaxed to a classification problem, determining whether a document was single or multi-authored for PAN'18-PAN'21 [13]. In 2019, participants were further challenged to predict the number of authors in multi-authored documents [27]. In 2020, the task shifted to detecting style changes at the paragraph level [26]. At PAN'21, we extended the task to assign authors to each paragraph [22]. In 2022, this was extended from paragraph to sentence level [23]. In 2023 and 2024, the task returned to the paragraph level but controlled for the simultaneous change of authorship and topic [24,25].

For PAN'25, we will continue to focus on the following intrinsic style change detection task:"For a given text, identify all positions where the writing style changes."Given the profound performance improvements demonstrated in solutions from recent PAN editions, we aim to expand the task for 2025. To make the task more realistic, we plan to shift the focus from the paragraph level, as in previous years, to the sentence level. The core challenge will be to apply intrinsic profiling methods to determine whether a style change occurs between consecutive sentences, indicating whether they were authored by the same individual or not.

Participants will be provided with three datasets, each increasing in difficulty based on the topical similarity between sentences: "Easy dataset": (1) Easy Dataset: The sentences within each document cover a wide range of topics, allowing participants to leverage topic shifts as indicators of authorship changes. (2) Medium Dataset: The number of topics in each document is limited, requiring participants to rely more heavily on detecting subtle style changes rather than topic variations. (3) Hard Dataset: All sentences in the document focus on the same topic, demanding precise identification of style changes independent of topical clues.

5 Generative Plagiarism Detection

The recent widespread adoption of large language models (LLMs) has introduced complex challenges, ranging from facilitating malware and social engineering attacks to automated influence campaigns, spam, and harassment [5]. Of particular interest to this task is the domain of academic integrity. A prime concern, and the focus of this task, is the anticipated escalation in both the frequency and sophistication of plagiarism cases facilitated by LLMs in the foreseeable future [9,18]. This task aims to address the challenges of real-world plagiarism by exploiting the capabilities of generative LLMs to paraphrase and otherwise disguise plagiarism to varying degrees.

As a first step toward developing generative plagiarism detection (PD) models, this task addresses the novel challenge of identifying paraphrased text alignments. Traditional plagiarism detection has focused primarily on copy-paste plagiarism, relying on token overlap between the source and plagiarized content [2]. However, modern plagiarists can now use generative LLMs to create highly sophisticated paraphrased content that often appears indistinguishable from original writing [4,17]. This blurs the lines between plagiarized and original material. Therefore, modern PD systems must evolve to detect paraphrased plagiarism while distinguishing it from genuinely novel content.

For this task, participants will be given pairs of scientific articles, each consisting of a source document and a corresponding plagiarized version. The degree of plagiarism will vary: some documents will be fully plagiarized, while others will only have specific paragraphs paraphrased. Since paraphrased plagiarism tends to obscure the exact boundaries of copied content, this task focuses on paragraph-level detection, rather than sentence- or token-level analysis. Initially,

we will focus on the STEM domain, using articles from `arxiv.org`. While some cases of plagiarism may include non-textual elements such as mathematical or chemical formulae, multimedia content, tables, and figures will not be altered.

To simulate realistic plagiarism scenarios as accurately as possible, the paraphrased content in this task will be generated using a variety of different LLMs, rather than relying on a single model. Moreover, the plagiarized versions may include genuine sections that are partially or entirely authored by LLMs. As generative LLMs gain wider acceptance as tools for assisting in the writing of scientific texts, merely identifying plagiarized text alignments based on LLM usage will no longer be sufficient [12]. This approach should encourage the development of diverse solutions to the text alignment challenge, ranging from basic similarity calculations to more advanced methods such as detecting generative AI patterns and analyzing writing styles.

Acknowledgments. The work of members of the Bauhaus-Universität Weimar and Leipzig University was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu).

References

1. Abassy, M., et al.: Llm-detectaive: a tool for fine-grained machine-generated text detection. CoRR abs/2408.04284 (2024). https://doi.org/10.48550/ARXIV.2408.04284, https://doi.org/10.48550/arXiv.2408.04284
2. Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B.: Corpus and Evaluation Measures for Automatic Plagiarism Detection. In: Calzolari, N., et al., (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta (May 2010)
3. Brundage, M., et al.: The malicious use of artificial intelligence: forecasting, prevention, and mitigation. CoRR abs/1802.07228 (2018)
4. Cegin, J., Simko, J., Brusilovsky, P.: ChatGPT to Replace crowdsourcing of paraphrases for intent classification: higher diversity and comparable model robustness. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1889–1905, Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-.main.117
5. Crothers, E.N., Japkowicz, N., Viktor, H.L.: Machine-generated text: a comprehensive survey of threat models and detection methods. IEEE Access 11, 70977-71002 (2023), ISSN 2169–3536. https://doi.org/10.1109/ACCESS.2023.3294090
6. Dementieva, D., et al.: RUSSE-2022: Findings of the first Russian detoxification task based on parallel corpora. In: Computational Linguistics and Intellectual Technologies (2022)
7. Dementieva, D., et al.: Overview of the multilingual text detoxification task at PAN 2024. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, CEUR Workshop Proceedings, vol. 3740, pp. 2432–2461, CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-223.pdf

8. Dementieva, D., Moskovskiy, D., Dale, D., Panchenko, A.: Exploring methods for cross-lingual text style transfer: The case of text detoxification. In: Park, J.C., et al., (eds.) Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 – 4, 2023, pp. 1083-1101, Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.IJCNLP-MAIN.70
9. Elali, F.R., Rachid, L.N.: AI-generated research paper fabrication and plagiarism in the scientific community. Patterns **4**(3), 100706 (Mar 2023), ISSN 26663899. https://doi.org/10.1016/j.patter.2023.100706
10. European Parliament and Council of the European Union: Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services (digital services act) and amending directive 2000/31/ec (2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj, official Journal of the European Union, L 277, 27.10.2022, p. 1–102
11. Fröbe, M., et al.: Continuous integration for reproducible shared tasks with TIRA.io. In: Kamps, J., et al., (eds.) Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), pp. 236–241, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Apr 2023). https://doi.org/10.1007/978-3-031-28241-6_20
12. Jarrah, A.M., Wardat, Y., Fidalgo, P.: Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? Online J. Commun. Media Technol. **13**(4), e202346 (Oct 2023), ISSN 1986–3497. https://doi.org/10.30935/ojcmt/13572
13. Kestemont, M., et al.: Overview of the author identification task at PAN 2018: cross-domain authorship attribution and style change detection. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
14. Logacheva, V., et al..: ParaDetox: detoxification with parallel data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6804–6818, Association for Computational Linguistics, Dublin, Ireland (May 2022)
15. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16—new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16) (2016)
16. Tschuggnall, M., et al.: Overview of the author identification task at PAN 2017: style breach detection and author clustering. In: CLEF 2017 Labs and Workshops, Notebook Papers (2017)
17. Wahle, J.P., Ruas, T., Kirstein, F., Gipp, B.: How large language models are transforming machine-paraphrase plagiarism. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 952-963, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.62
18. Wahle, J.P., Ruas, T., Meuschke, N., Gipp, B.: Are neural language models good plagiarists? a benchmark for neural paraphrase detection. In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 226–229, IEEE, Champaign, IL, USA (Sep 2021), ISBN 978-1-66541-770-9, https://doi.org/10.1109/JCDL52503.2021.00065

19. Wang, Y., et al.: M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3964–3992, Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://aclanthology.org/2024.acl-long.218
20. Wang, Y., et al.: M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1369–1407, Association for Computational Linguistics, St. Julian's, Malta (Mar 2024). https://aclanthology.org/2024.eacl-long.83
21. Yao, F., Li, C., Nekipelov, D., Wang, H., Xu, H.: Human vs. generative AI in content creation competition: Symbiosis or conflict? CoRR abs/2402.15467 (2024). https://doi.org/10.48550/ARXIV.2402.15467
22. Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
23. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2022. In: Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.) CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2022)
24. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the multi-author writing style analysis task at PAN 2023. In: Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), CEUR Workshop Proceedings, vol. 3497, pp. 2513–2522 (Sep 2023), URL https://ceur-ws.org/Vol-3497/paper-201.pdf
25. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the multi-author writing style analysis task at PAN 2024. In: Faggioli, G., Ferro, N., Galuščáková, P., Herrera, A.G.S. (eds.) Working Notes Papers of the CLEF 2024 Evaluation Labs, pp. 2513–2522, CEUR-WS.org (Sep 2024), http://ceur-ws.org/Vol-3740/paper-222.pdf
26. Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2020. In: CLEF 2020 Labs and Workshops, Notebook Papers (2020)
27. Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the style change detection task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)
28. Zhou, J., Zhang, Y., Luo, Q., Parker, A.G., Choudhury, M.D.: Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In: Proceedings of the 2023 CHI, ACM (2023). https://doi.org/10.1145/3544548.3581318