

Table 1

Overview of the accuracy in detecting if a text is written by an human in task 4 on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). We report ROC-AUC, Brier, C@1, F₁, F_{0.5u} and their mean.

Approach	ROC-AUC	Brier	C@1	F ₁	F _{0.5u}	Mean
logistic-regression	0.999	0.913	0.983	0.983	0.983	0.972
Baseline Binoculars	0.972	0.957	0.966	0.964	0.965	0.965
Baseline Fast-DetectGPT (Mistral)	0.876	0.8	0.886	0.883	0.883	0.866
Baseline PPMd	0.795	0.798	0.754	0.753	0.749	0.77
Baseline Unmasking	0.697	0.774	0.691	0.658	0.666	0.697
Baseline Fast-DetectGPT	0.668	0.776	0.695	0.69	0.691	0.704
95-th quantile	0.994	0.987	0.989	0.989	0.989	0.990
75-th quantile	0.967	0.925	0.939	0.929	0.935	0.938
Median	0.907	0.890	0.878	0.866	0.865	0.887
25-th quantile	0.705	0.769	0.673	0.654	0.662	0.692
Min	0.131	0.265	0.005	0.006	0.007	0.224

Team <team-name> at PAN: <short title>

Notebook for the PAN Lab at CLEF 2024

First N. Author^{1,†}, Second N. Author^{1,†}

¹Affiliation name, Street, City, Post code, Country

Abstract

Your abstract. If you need more information on commands to prepare this paper, see [formatting-instructions.pdf](#). The [submission-instructions.pdf](#) contains instructions on how to submit the paper by 31 May 2024 Midnight CEST (not AoE!).

Keywords

Separate, Them with, Sep command

1. Introduction

This is a template for paper submissions to the generative authorship task at PAN [1]. You have to cite the PAN overview paper [1] and the task overview paper [?] (just as we will cite your paper). Please also cite TIRA [2].

2. Background

3. System Overview

4. Results

Table 1 shows the results, initially pre-filled with the official baselines provided by the PAN organizers and summary statistics of all submissions to the task (i.e., the maximum, median, minimum, and 95-th,

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[†] Authors contributed equally

✉ first@authors.org (F. N. Author); second@authors.org (S. N. Author)

🆔 0000-0000-0000-0000 (F. N. Author); 0000-0000-0000-0000 (S. N. Author)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 2

Overview of the mean accuracy over 9 variants of the test set. We report the minimum, median, the maximum, the 25-th, and the 75-th quantile, of the mean per the 9 datasets.

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
logistic-regression	0.826	0.958	0.970	0.975	0.987
Baseline Binoculars	0.342	0.818	0.844	0.965	0.996
Baseline Fast-DetectGPT (Mistral)	0.095	0.793	0.842	0.929	0.958
Baseline PPMd	0.270	0.546	0.750	0.770	0.863
Baseline Unmasking	0.250	0.653	0.673	0.697	0.762
Baseline Fast-DetectGPT	0.159	0.579	0.677	0.719	0.982
95-th quantile	0.862	0.968	0.978	0.990	1.000
75-th quantile	0.756	0.863	0.926	0.959	0.995
Median	0.593	0.622	0.874	0.884	0.943
25-th quantile	0.343	0.470	0.637	0.683	0.703
Min	0.015	0.038	0.231	0.235	0.252

75-th, and 25-th percentiles over all submissions to the task).

Table 2 shows the summarized results averaged (arithmetic mean) over 10 variants of the test dataset. Each dataset variant applies one potential technique to measure the robustness of authorship verification approaches, e.g., switching the text encoding, translating the text, switchign the domain, manual obfuscation by humans, etc. Please focus your description on the discussion of the results on the main dataset, e.g., Table 1. I.e., Table 2 is only here for your completeness, please discuss only the details on the main dataset (i.e., Table 1). A detailed description of all dataset variants will be available in the overview notebook.

5. Conclusion

Acknowledgments

Your acknowledgments (funding etc.)

References

- [1] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
- [2] 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. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

A. Appendix

If you need one