BaselineAvengers at PAN 2024: Often-Forgotten Baselines for LLM-Generated Text Detection

Notebook for the PAN Lab at CLEF 2024

Ludwig Lorenz[1,†], Funda Zeynep Aygüler[1,†], Ferdinand Schlatt[2] and Nailia Mirzakhmedova[1,†]

[1]*Bauhaus-Universität Weimar, Germany*
[2]*Friedrich-Schiller-Universität Jena*

Abstract

The rapid advancements of Large Language Models (LLMs) make it increasingly challenging to distinguish between human-written and machine-generated texts, which raises concerns regarding their potential misuse. This paper describes our submission to the PAN: Generative AI Authorship 2024 verification task, which involves identifying the human-authored text from a pair of texts, one written by a human and the other by an LLM. Our approach is based on the assumption that LLMs use a distinct vocabulary. We propose a simple and interpretable method using non-neural machine learning classifiers with lexical features. We evaluate several classification models and feature sets on a validation split and find logistic regression and SVM models using *tf-idf* feature vectors to be highly effective. Our submissions offer a more effective alternative to all baseline approaches while also being more efficient and interpretable.

Keywords

Authorship verification, Logistic Regression, Tf-Idf Vectorizer

1. Introduction

With the rapid advancements of Large Language Models (LLMs), distinguishing between human-written and machine-generated texts becomes more and more challenging. As a result, the need for reliable authorship verification methods becomes even more pressing. The ability to distinguish between human-written and machine-generated texts is crucial for various applications, such as plagiarism detection [1], forensic linguistics [2], and content moderation [3]. Multiple approaches have been proposed to address this problem, including complex feature engineering and stylometric analysis, linguistic analysis, and machine learning-based methods [4]. However, the increasing sophistication of LLMs poses a significant challenge to existing authorship verification methods. In response to this challenge, PAN [5] introduced the Voight-Kampff Generative AI Authorship Verification task to test the feasibility of distinguishing between human-written and LLM-generated texts [6].

In this paper, we present our submission to the PAN shared task, where we address the generative authorship verification problem using non-neural machine learning classifiers based on lexical features. Our decision to employ non-neural models is motivated by the observation that simple models are often overlooked in recent research, despite their proven effectiveness and their ability to serve as efficient baselines for comparison with more complex models [7]. Moreover, our emphasis on lexical features is based on the hypothesis that LLMs use a distinct vocabulary, which may be sufficient to differentiate between human-authored and machine-generated texts.

In our work, we experimented with three classification models and two lexical feature sets. We found logistic regression and SVM models using *tf-idf* feature vectors are highly effective for the task. Motivated by the performance of our approach, we conducted a qualitative analysis of the most

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

† Authors contributed equally

ludwig.david.lorenz@uni-weimar.de (L. Lorenz); funda.zeynep.aygueler@uni-weimar.de (F. Z. Aygüler); ferdinand.schlatt@uni-jena.de (F. Schlatt); nailia.mirzakhmedova@uni-weimar.de (N. Mirzakhmedova)
0009-0005-2410-9005 (L. Lorenz); 0009-0009-6160-5074 (F. Z. Aygüler); 0000-0002-6032-909X (F. Schlatt); 0000-0002-8143-1405 (N. Mirzakhmedova)

significant lexical features to test our hypothesis that LLMs employ a distinct vocabulary. Our analysis revealed that there is a small set of words that can indicate whether a text is written by an LLM. Overall, our approach offers a more effective alternative to all baseline approaches while also being more efficient and interpretable.

The remainder of this paper is structured as follows. In Section 2, we provide background information on the PAN: Generative AI Authorship Verification task and review the related work. In Section 3, we describe our system and the components of our submission. In Section 4, we present the results of our submission. Section 5 provides a qualitative analysis of the most important lexical features. We conclude with a discussion of our results in Section 6.

2. Background

Task Description The PAN: Generative AI Authorship Verification task is organized in collaboration with the Voight-Kampff Task at the ELOQUENT Lab in a builder-breaker style. PAN participants build systems to tell human and machine-generated texts apart, while ELOQUENT participants investigate novel text generation and obfuscation methods to avoid detection. The task is defined as follows:

> Given two texts, one authored by a human, one by a machine: pick out the human.

More formally, given a pair of texts (t_1, t_2), one of which is written by a human and the other by an LLM, the system must output a confidence score $s \in [0.0, 1.0]$. A score $s < 0.5$ indicates that text t_1 is believed to be human-authored, while a score $s > 0.5$ indicates that text t_2 is believed to be human-authored. A score of exactly 0.5 means the case is undecidable.

Dataset The task participants were provided with a training dataset of 1,359 U.S. news articles. To ensure that the articles were human-authored, the task organizers collected the articles from Google News, focusing on the period before the release of GPT-3.5. The articles were summarized using GPT-4-Turbo, and the summaries were used as input for 13 downstream LLMs to generate new articles. The dataset consists of pairs of articles, one human-authored and one LLM-generated, and is split into training, validation, and test sets.

To further test the robustness of submissions, the task organizers provided additional test datasets, each applying a different obfuscation technique to the original test dataset. The obfuscation techniques include switching the text encoding, prompting the LLMs to generate German instead of English, using contrastive decoding, cropping the text to 35 words, etc. In total, 65 different test datasets were created by obfuscation, with ELOQUENT providing another five.

3. System Overview

Scoring Function As follows from the task description (cf. Section 2), the generative authorship verification task is formulated as a pairwise classification problem. Given a pair of texts (t_1, t_2), the goal is to determine which text is human-authored. However, we approach this task as a pointwise binary classification problem. That is, given a text t_i, we aim to predict the probability $P(\text{human}|t_i)$ that the text is human-authored.

By definition, the probability $P(\text{human}|t_i)$ is equal to $1 - P(\text{LLM}|t_i)$. Given that we need to predict the probability that t_1 is human-authored while taking into account t_2, we average the probabilities of the first text being written by a human and the second text not being written by a human to obtain the final score $s(\text{human}|t_1)$:

$$s(\text{human}|t_1) = \frac{P(\text{human}|t_1) + 1 - P(\text{LLM}|t_2)}{2} \tag{1}$$

Table 1
Overview of the different classifiers (rows) and features (columns) evaluated on the validation set.

Classifier	tf-idf	Term Frequencies
Multinomial Naive Bayes	0.77	0.874
Logistic Regression	0.927	0.922
SVM	0.932	0.925

Feature Extraction To capture the distinctive vocabulary of LLM-generated texts, we use a bag-of-words model to represent the texts. We experiment with two feature sets: term frequencies and *tf-idf* values for all tokens in the training dataset.

Classification Models We experiment with three classifiers: Multinomial Naive Bayes, logistic regression, and a support vector machine (SVM) with a linear kernel. We test the classifiers with both term frequencies and *tf-idf* values to identify the most effective model and feature combination.

Model and Feature Selection To evaluate the performance of the different models and feature sets, we use 100 samples from the training dataset as a validation split. The results of the validation are used to select the most effective model and feature combination.

Table 1 shows the accuracy achieved on the validation split for each model. Overall, logistic regression and SVM are more effective than multinomial Naive Bayes. The differences in effectiveness for different feature sets for logistic regression and SVM are minimal. Interestingly, the performance of multinomial naive Bayes is significantly better using raw term frequencies compared to *tf-idf* values.

4. Results

Evaluation Setup The PAN: Generative AI Authorship Verification task employed the TIRA platform [8] to ensure the reproducibility and comparability of submissions. The platform provides a standardized environment for running submissions and evaluates the submissions using the following metrics:

- **ROC-AUC**: The area under the ROC (Receiver Operating Characteristic) curve
- **Brier**: The complement of the Brier score (mean squared loss)
- **C@1**: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases
- **F1**: The harmonic mean of precision and recall
- **F0.5u**: A modified F0.5 measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives
- The arithmetic mean of all the metrics above.

The arithmetic mean of all metrics is used to rank the submissions.

Baselines The task organizers provided official baselines for comparison, which are based on the performance of various approaches to the task of authorship verification. The baselines include a simple text length classifier, PPMd Compression-based Cosine [9, 10], Authorship Unmasking [11, 12], Binoculars [13], DetectLLM LRR and NPR [14], and DetectGPT [15].

Evaluation Results Table 2 presents the evaluation results of our submissions to the task, along with the official baselines and summary statistics of all submissions. Our best performing submission (SVM) outperforms all official baselines across all metrics, with the other two submissions (Multinomial Naive Bayes and Logistic Regression) not outperforming only the Binoculars baseline for the algorithmic mean of all metrics (0.965 vs. 0.956 and 0.958 respectively).

Table 2
Overview of the performance of our approaches, baselines, and the summary statistics of the performance of all submissions in the competition. We report ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their arithmetic mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
naive-bayes	0.998	0.859	0.975	0.975	0.974	0.956
logistic-regression	0.996	0.884	0.97	0.97	0.97	0.958
svm	0.994	0.923	0.976	0.976	0.975	0.969
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	0.995	0.986	0.988	0.988	0.989	0.989
75-th quantile	0.971	0.925	0.954	0.935	0.942	0.945
Median	0.911	0.889	0.887	0.869	0.867	0.889
25-th quantile	0.714	0.771	0.683	0.657	0.670	0.697
Min	0.131	0.265	0.005	0.006	0.007	0.224

Table 3
Overview of the performance of our approaches, baselines, and the summary statistics of the performance of all submissions in the competition over 10 variants of the test set. We report the minimum, 25-th quantile, median, 75-th quantile, and maximum of the arithmetic mean of all metrics.

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
naive-bayes	0.884	0.935	0.945	0.967	0.969
logistic-regression	0.837	0.941	0.957	0.963	0.989
svm	0.832	0.949	0.969	0.974	0.999
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	0.875	0.973	0.985	0.989	1.000
75-th quantile	0.758	0.875	0.935	0.959	0.994
Median	0.605	0.629	0.876	0.889	0.946
25-th quantile	0.350	0.481	0.658	0.697	0.709
Min	0.015	0.038	0.231	0.235	0.252

Table 3 shows the summarized results averaged (arithmetic mean) over 10 obfuscated variants of the test dataset. Each dataset variant applies one potential technique to measure the robustness of authorship verification approaches (cf. Section 2). The results show that all our submissions are robust to the obfuscation techniques, as the performance does not drop significantly compared to the baseline approaches. For example, the minimum achieved score for our best submission (SVM) is 0.832, while the minimum score for the best baseline (Binoculars) is 0.342.

Overall, our approach demonstrates that simple and interpretable models can be highly effective for the task of generative authorship verification. The results suggest that the distinctive vocabulary used by LLMs can indeed be effectively captured using simple lexical features and machine learning classifiers. Moreover, our submissions showed to be robust to obfuscation techniques, making them a promising alternative to more complex and computationally expensive methods.

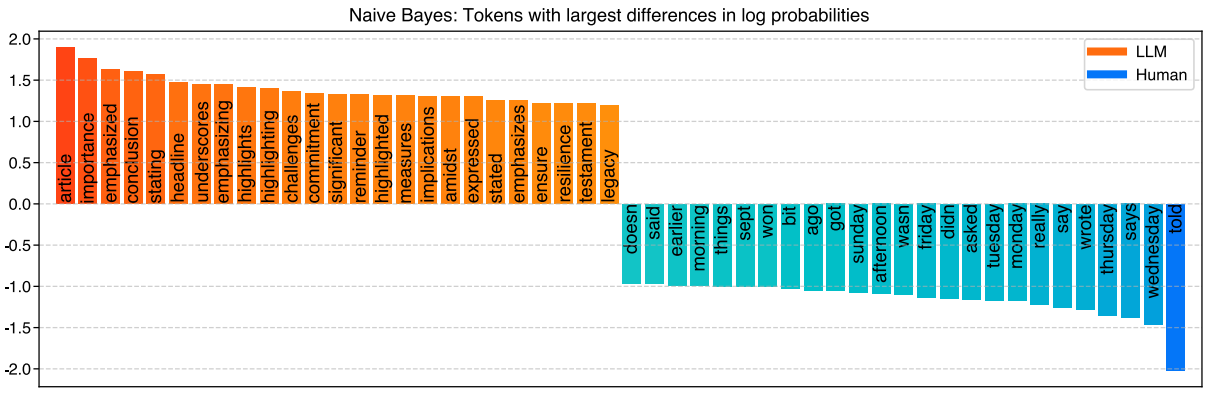

Figure 1: Top 50 tokens with the largest differences in log probabilities for multinomial Naive Bayes. Positive values indicate the probability is higher for LLM-generated texts, negative values indicate the probability is higher for human-written texts.

5. Qualitative Analysis

In addition to the quantitative evaluation of our submissions, we conducted a qualitative analysis of the most important lexical features identified by the models. This analysis aims to highlight key tokens that contribute to distinguishing between human-written and LLM-generated texts.

The implementation of the multinomial Naive Bayes model allows us to extract the log probabilities of each token belonging to the human-written and LLM-generated classes. By comparing these probabilities, we can identify the tokens that contribute most to the classification decision. We use the following equation to calculate the difference in log probabilities for each token w_i in the feature set:

$$\text{log\_diff}(w_i) = \log(P(w_i|\text{LLM})) - \log(P(w_i|\text{human})) \tag{2}$$

The log difference values are then sorted in descending order to identify the tokens with the largest differences. The resulting values are interpreted as the importance of each token in distinguishing between human-written and LLM-generated texts. Positive values indicate higher probabilities for LLM-generated texts, while negative values indicate higher probabilities for human-written texts. Figure 1 presents the top 50 tokens with the largest differences in log probabilities for the multinomial Naive Bayes model. Here, we observe that LLM-generated texts frequently use specific terms such as “article”, “importance”, “emphasized”, “context”, and “despite”. These terms often relate to structured and formal writing, which is often characteristic of LLM-generated content. On the other hand, human-written texts show a higher probability of tokens related to everyday language and temporal expressions such as “told”, “says”, “asked”, “wrote”, and “really”. These tokens indicate a more narrative and less formal style typical of human writing. The frequent use of days of the week such as “Wednesday”, “Thursday”, and “Friday” and terms like “afternoon” and “morning” in human-written texts can be attributed to their common use in chronological events or planning. Humans often refer to specific days when recounting events, discussing plans, or setting contexts within their narratives. This is particularly relevant in our news articles dataset, where providing temporal context is essential for accurate and engaging reporting. The word “told” is particularly prominent in human-written texts, as it is frequently used in direct and indirect speech, which is also common in news articles. In contrast, LLM-generated texts often prioritize structured content delivery and formal exposition over narrative elements, resulting in frequent use of terms such as “emphasized”, “stating”, and “highlights”. The term “conclusion” is also prevalent in LLM-generated texts, indicating a structured and formal writing style that often includes a summary or final remarks, which is uncommon in human-written news articles.

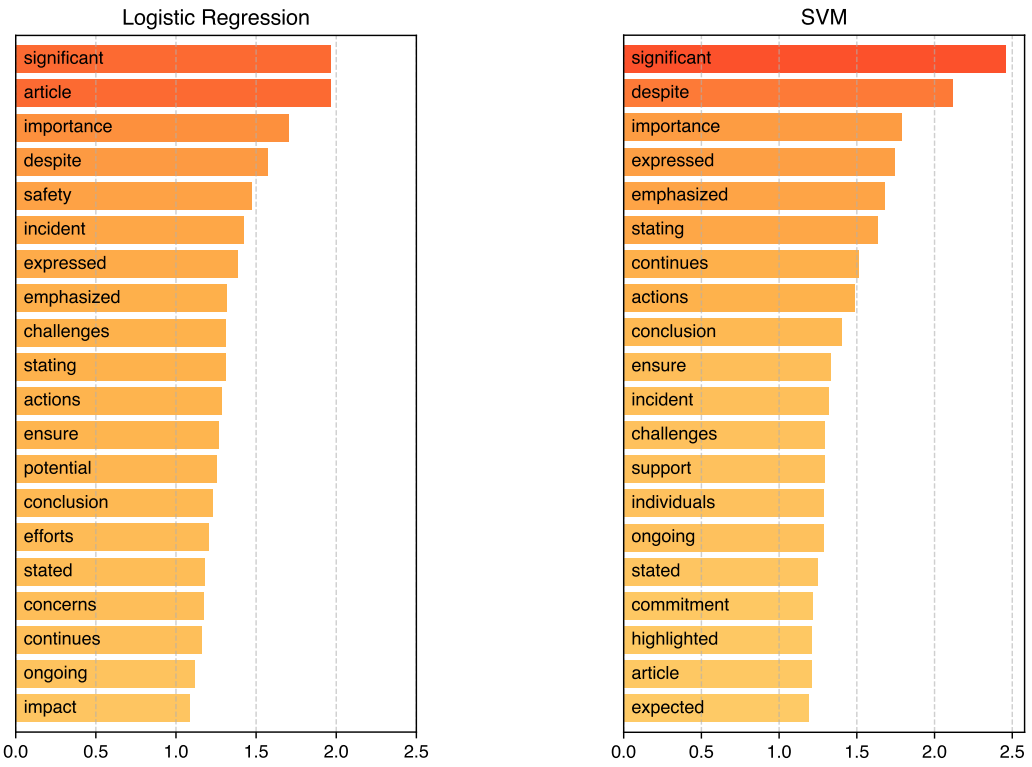

Figure 2: Top 20 tokens for identifying LLM-generated texts using Logistic Regression (left) and SVM (right). The importance of each token is based on the size of the coefficients assigned to them by the trained models.

Figure 2 presents the top 20 most important tokens for identifying LLM-generated texts based on the coefficients assigned to them by the trained logistic regression and SMV models. Tokens with larger coefficients have a greater impact on the model's decision function. Similarly to the Naive Bayes model, some of the most notable tokens both in logistic regression and the SVM models include "significant", "article", "importance", "despite", "stating" and "conclusion". This suggests that LLM-generated texts often contain terms that convey formality, which might be less prevalent in human-written texts. The overlap in key tokens between the logistic regression and SVM models underlines the consistency of these patterns in distinguishing LLM-generated texts. The frequent appearance of the word "significant" in LLM-generated texts can be attributed to the tendency of language models to produce content that is polished and systematic. Language models are typically trained on large datasets that include a large amount of academic, technical, and professional writing. This extensive exposure to formal texts influences the models to emulate this style.

Our qualitative analysis supports the hypothesis that LLMs use a distinctive vocabulary that can be captured using lexical features. The presence of terms related to formality and structured discourse in LLM-generated texts contrasts with the more narrative and less formal vocabulary found in human-written texts.

6. Conclusion

In this paper, we presented our submission to the PAN: Generative AI Authorship Verification task. Our approach is based on the assumption that LLMs use a particular vocabulary, which can be captured using lexical features. We experiment with three classifiers and two feature sets to identify the most effective model and feature combination. Our results show that logistic regression and SVM models using *tf-idf* feature vectors are highly effective for the task. We find that our submissions outperform all official baselines, demonstrating that simple and interpretable models can be more effective than complex and computationally expensive methods. Our qualitative analysis of the most important

lexical features confirms that LLM-generated texts often contain terms distinct from human-written texts, which can be effectively captured using lexical features. The robustness of our submissions to obfuscation techniques further highlights the effectiveness of our approach. Overall, our results offer a more effective alternative to all baseline approaches while also being more efficient and interpretable.

Acknowledgments

This work originates from a programming assignment from the "Introduction to Natural Language Processing" course at Bauhaus-Universität Weimar during the summer term of 2024. We would like to thank the teaching staff who recognized the potential of our approach and encouraged us to participate in the PAN task. Together we turned these ideas into writing.

References

[1] M. Potthast, B. Stein, A. Barrón-Cedeño, P. Rosso, An evaluation framework for plagiarism detection, in: C.-R. Huang, D. Jurafsky (Eds.), Coling 2010: Posters, Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 997–1005. URL: https://aclanthology.org/C10-2115.

[2] V. Guillén-Nieto, D. Stein, Language as evidence: Doing forensic linguistics, Springer Nature, 2022.

[3] V. U. Gongane, M. V. Munot, A. D. Anuse, Detection and moderation of detrimental content on social media platforms: current status and future directions, Social Network Analysis and Mining 12 (2022) 129.

[4] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, B. Stein, M. Potthast, Overview of the Authorship Verification Task at PAN 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Labs and Workshops, Notebook Papers, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3180/paper-184.pdf.

[5] 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.

[6] J. Bevendorff, M. Wiegmann, J. Karlgren, L. D"urlich, 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. Galušč'akov'a, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[7] Y.-C. Lin, S.-A. Chen, J.-J. Liu, C.-J. Lin, Linear classifier: An often-forgotten baseline for text classification, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1876–1888. URL: https://aclanthology.org/2023.acl-short.160. doi:10.18653/v1/2023.acl-short.160.

[8] M. Fr"obe, 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.

[9] D. Sculley, C. Brodley, Compression and machine learning: a new perspective on feature space

vectors, in: Data Compression Conference (DCC'06), 2006, pp. 332–341. doi:10.1109/DCC.2006.13.

[10] O. Halvani, C. Winter, L. Graner, On the usefulness of compression models for authorship verification, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, ARES '17, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi.org/10.1145/3098954.3104050. doi:10.1145/3098954.3104050.

[11] M. Koppel, J. Schler, Authorship verification as a one-class classification problem, in: Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 62. URL: https://doi.org/10.1145/1015330.1015448. doi:10.1145/1015330.1015448.

[12] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 654–659. URL: https://aclanthology.org/N19-1068. doi:10.18653/v1/N19-1068.

[13] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. URL: https://arxiv.org/abs/2401.12070. arXiv:2401.12070.

[14] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, 2023. URL: https://arxiv.org/abs/2306.05540. arXiv:2306.05540.

[15] G. Bao, Y. Zhao, Z. Teng, L. Yang, Y. Zhang, Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature, 2024. URL: https://arxiv.org/abs/2310.05130. arXiv:2310.05130.