Are AI Detectors Good Enough? A Survey on Quality of Datasets with Machine-Generated Texts

Anonymous submission

Abstract

The rapid development of autoregressive Large Language Models (LLMs) has significantly improved the quality of generated texts, necessitating reliable machine-generated text detectors. A huge number of detectors and collections with AI fragments have emerged, and several detection methods even showed recognition quality up to 99.9% according to the target metrics in such collections. However, the quality of such detectors tends to drop dramatically in the wild, posing a question: Are detectors actually highly trustworthy or do their high benchmark scores come from the poor quality of evaluation datasets? In this paper, we emphasise the need for robust and qualitative methods for evaluating generated data to be secure against bias and low generalising ability of future model. We present a systematic review of datasets from competitions dedicated to AI-generated content detection and propose methods for evaluating the quality of datasets containing AI-generated fragments. In addition, we discuss the possibility of using high-quality generated data to achieve two goals: improving the training of detection models and improving the training datasets themselves. Our contribution aims to facilitate a better understanding of the dynamics between human and machine text, which will ultimately support the integrity of information in an increasingly automated world.

1 Introduction

The quality of large language models (LLMs) has grown tremendously in the last five years, making their output almost indistinguishable from human-written texts (Chang et al. 2024). This expanded the application fields of these models, as many routine tasks can be entrusted to them nowadays. However, they can be used for creating texts that are intended to be written and fact-checked by humans. Example of such misuse is generation of fake news (Zellers et al. 2019; Zhou et al. 2023), which can mislead readers of generated content. Teachers raise another concern, as many students complete assignments with LLMs (Koike, Kaneko, and Okazaki 2024; Ma et al. 2023), undervaluing the purpose of the educational process. Machine-generated fragments also appear in academic articles more often with the growth of chatbots and reach several tens of percent (Liang et al. 2024; Gritsay et al. 2023a). More than 60,000 scientific papers in the last year alone contained the evidence of the use of machine generation (Gray 2024). All of that prove that it is crucial to develop systems able to counter the misuse of artificial data and signal to the reader that the content they read is generated.

Another concern is the Web, overflowing with machine-generated content, often of poor quality. Such texts contribute bias to publicly available texts on the Internet, through false facts, hallucinations and spelling errors. Given the current agenda of using texts from the Internet to train new language models, all this bias will be inadvertently added to the model. Moreover, (Villalobos et al. 2022) revealed that the human-written data will run out by 2028. That means that the training sets for language models in the future will include a large amount of generated content. Such *self-consuming* will result in the substantial degradation of the model's abilities (Alemohammad et al. 2023). Furthermore, the trend is evolving in such a way that human-written texts on almost any topic will be much harder to retrieve. While for texts dated even 5 years ago we are confident as the usage of generation was extremely rare, we cannot state the same for more recent texts.

Therefore, detectors, capable to distinguish human-written texts from AI-generated texts, and whose detection quality can be guaranteed, are necessary for many fields. We believe one of the kay factors for building reliable detectors is the high-quality artificial text collections that can be used for training and evaluation. In this paper, we would like to estimate the quality of the available generated texts from competitions and research papers. Sometimes we see that some methods from participants of the competitions reach almost perfect (up to 99.9%) metric score, meanwhile in the wild we observe a noticeable decline in performance. Such results look confusing, because the models become more and more advanced, seemingly making the detecting task more challenging, meanwhile participants of competitions still reach almost perfect scores, bringing up the question about quality of generated data in the provided datasets. Are the devised methods really good or is the data easy enough for detectors to solve the seemingly hard detection task?

Our contributions are as follows.

1. We systemize information about existing datasets from the research papers and competitions, dedicated to the detection of AI-generated content task.
2. We suggest methods that may be helpful for evaluating the quality of the generated data and the datasets aimed to

use for binary classification between human and machine texts.

2 Related Work

2.1 The Task of AI-generated text detection

The task of AI-generated text detection task is generally stated as a text classification task, which means that the input is a text sequence and the output is a discrete, usually binary, class prediction. When the task is binary, the common labels are “AI” or “human”, whereas multiclass classification focusses on distinguishing several language models. The last task is usually called authorship attribution. Finally, more complex task suggests to determine the borders betweens fragments from different authors, for example between human author and some LLM author.

The approach to tackle the classification problem is to utilize linguistic, stylometric, and statistical features for classifiers (Jawahar, Abdul-Mageed, and Lakshmanan 2020; Fröhling and Zubiaga 2021). Some internal metrics, such as perplexity or its modifications (Hans et al. 2024) can also be helpful as a inference method. Also, perturbing texts can also provide valueable information, as described in DetectGPT (Mitchell et al. 2023) method, where log-probabilities between between original and modified texts are compared. Finally, finetuning encoder-based model, like DeBERTa (He et al. 2021), is considered the best solution for this task (Uchendu et al. 2021; Macko et al. 2023).

2.2 Evaluating Generated Text

As for evaluating the quality of the generated data itself, it has become more common to evaluate it with the help of LLMs (Xu et al. 2023). This approach does not require any human reference, unlike ROUGE (Lin 2004). However, the downside is the output of model-evaluator needs to be unified, not always interpretable and model-evaluator scores can be skewed. Alternative approach is suggested by Zhu and Bhat (2020), where the text is evaluated based on several linguistic criteria, such as grammar or coherence.

2.3 Datasets with Artificial Content

There are a number of surveys about machine-generated content detection with an overview of the datasets (Jawahar, Abdul-Mageed, and Lakshmanan 2020; Wu et al. 2023), however, few works focus on the quality of data in the available datasets, despite it being an important aspect of the task. Building AI-generated content detectors requires high-quality labelled data that involve substantial financial, computational, and human resources. The human evaluators should check that the dataset does not contain corrupted generations, that the texts are coherent and grammatically correct. We will describe the datasets we used in our analysis and experiments in Section 3.

2.4 Shared Tasks on AI-Generated content Detection

Shared tasks help to advance research on detecting AI-generated content, offering new variations on the tasks and data for evaluation, incentivisng participants to come up with novel ideas for detectors robust to the change of language, domain, or generating model. Participants explore approaches ranging from transfer learning on complex text features to utilising and fine-tuning LLMs for these tasks (2022; 2022; 2023b; 2024; 2024; 2024; 2024; 2024). These efforts have highlighted challenges such as handling multilingual data and adapting to rapidly evolving generative models. Some participants also provide some analysis of the given data or even discuss some flaws with generated texts in it (Voznyuk and Konovalov 2024).

3 Datasets

3.1 Datasets From Shared Tasks

The most common tasks in shared tasks are binary classification and authorship attribution, with binary classification being the prevalent task, therefore, in this work, we focused only on it. All chosen shared tasks contain texts in English, unless stated otherwise. Here we give a brief overview of each task, as well as some quantitative statistics of the texts in Table 1, whereas a more detailed description, such as models used for generation or domains of the presented texts, can be found in Appendix A.

1. **DAGPap 2022** (Kashnitsky et al. 2022) introduced a dataset of human- and machine-written scientific excerpts collected by Elsevier.
2. **RuATD 2022** (Shamardina et al. 2022) focused on human- and machine-written documents in Russian, covering a wide range of themes.
3. **AuTexTification 2023** (Sarvazyan et al. 2023) provided texts in English and Spanish, covering five distinct domains.
4. **IberAuTexTiification 2024** (Sarvazyan et al. 2024) expanded on the previous competition with a multilingual (six Iberian languages), multi-domain, and multi-model focus.
5. **Voight-Kampff Generative AI Authorship Verification 2024** (Ayele et al. 2024), hereafter referred to as PAN 2024, tasked participants with identifying the human-authored text from two samples – one human-written and one machine-generated.
6. **SemEval 2024 Task 8** (Wang et al. 2024c) addressed domain, generator, and language shifts in generated texts. Training data included multiple languages (e.g., Chinese, Urdu, and Russian), but the test set was limited to English, Italian, German, and Arabic.
7. **MGT Detection Task 1 (COLING 2025)** (Wang et al. 2025) was built on SemEval 2024 Task 8 by incorporating data generated by novel LLMs and expanding multilingual coverage of train and test sets.

3.2 Datasets from Papers

The number of collections with generated content has started to grow with an increasing number of available generators. Quite often, researchers, together with a new approach for AI content detection, publish a parallel dataset on which they have validated their method. In this paper, our aim

Dataset	Year	Language	Num. of Texts, 10^3	Num. of Texts, G / H, 10^3	Average Length, G / H	Median Length, G / H
Research papers datasets						
GPT2	2019	en	1250	1000 / 250	2941 / 2616	3245 / 2459
TweepFake	2019	en	20.7	10.4 / 10.4	104 / 118	89 / 94
HC3	2023	en, zh	85.4	26.9 / 58.5	1011 / 681	1012 / 422
GhostBuster	2023	en	21	18 / 3	3345 / 3391	3440 / 2911.5
MGTBench	2024	en	23.7	20.7 / 3	1596 / 3391	1226 / 2911.5
MAGE	2024	en	436	152.3 / 284.2	1139 / 1282	706 / 666
M4	2024	en, zh, ru, bg, ur, id	89.5	44.7 / 44.7	1588 / 3162	1454 / 1697
OutFox	2024	en	57.6	43.2 / 14.4	2686 / 2238	2311 / 1992
Shared tasks datasets						
DAGPap22	2022	en	5.3	3.6 / 1.6	799 / 1180	680 / 1126.5
RuATD	2022	ru	129	64.5 / 64.5	237 / 221	99 / 95
AuTex	2023	en, es	65.9	33.1 / 32.8	315 / 297	386 / 351
IberAuTex	2024	es, en, ca, gl, eu, pt	98	52.5 / 45.4	1037 / 1058	981 / 1018
PAN24	2024	en	15.2	14.1 / 1.1	2641 / 3007	2731 / 2868
SemEval24 Mono	2024	en	34.2	18 / 16.2	2465 / 2358	2570 / 2083.5
SemEval24 Multi	2024	en, ar, de, it	42.3	22.1 / 20.2	2218 / 2257	2270 / 2032
MGT-1 Mono	2025	en	610.7	381.8 / 228.9	1448 / 1541	1208/ 1080
MGT-1 Multi	2025	en, zh, it, ar, de, ru, bg, ur, id	674	416.1 / 257.9	1423 / 1445	1195 / 1032

Table 1: Statistics of the texts in the datasets from the shared tasks and research papers.

was to pick collections with human- and machine-generated excerpts that are the most common and cited in other researchers' publications. Similarly to previous subsection, here we give a brief overview of each chosen datasets, describe some statistics about the texts in Table 1, and add a more detailed description in Appendix A.

GPT2 Output Dataset[1] consists of text outputs generated by GPT-2 models of different sizes across various prompts.

HC3 (Human Chatbot Conversations Corpus) (Su et al. 2024) features conversations between humans and chatbots, primarily used for research on chatbot responses and human-AI interaction analysis. This dataset is available for both English and Chinese, but we have focused only on the former.

GhostBuster (Verma et al. 2024) aimed at detecting AI-generated content by comparing it to human-written text, often used in the context of identifying machine-generated misinformation or spam.

MGTBench (Machine Generated Text Benchmark) (He et al. 2023) is a benchmark dataset designed to evaluate the quality of machine-generated text across various tasks, including fluency, coherence, and creativity.

MAGE (Model Augmented Generative Evaluation) (Li et al. 2024) evaluates the performance of generative models by comparing outputs with human annotations, aiding in the development of more accurate generative AI models.

M4 (Multilingual, Multimodal, Multitask, Massive Dataset) (Wang et al. 2024b) is a large-scale dataset designed for training models that can handle multiple languages, tasks, and modalities, making it useful for developing versatile AI systems. Although it is multilingual, we sampled only English texts.

TweepFake (Fagni et al. 2021) contains real tweets written by humans and synthetic tweets, generated by various AI models, from bots, imitating human users.

Outfox (Koike, Kaneko, and Okazaki 2024) contains triplets of essay problem statements, human-written essays, and LLM-generated essays. The students who wrote the essays range from 6th to 12th grade in the USA.

4 Approach

We decided to evaluate all datasets with common setups to see how good standard approaches perform on them. We did not have the goal to obtain the highest score, but rather to compare the performance of the same method on different datasets.

[1] https://github.com/openai/gpt-2-output-dataset

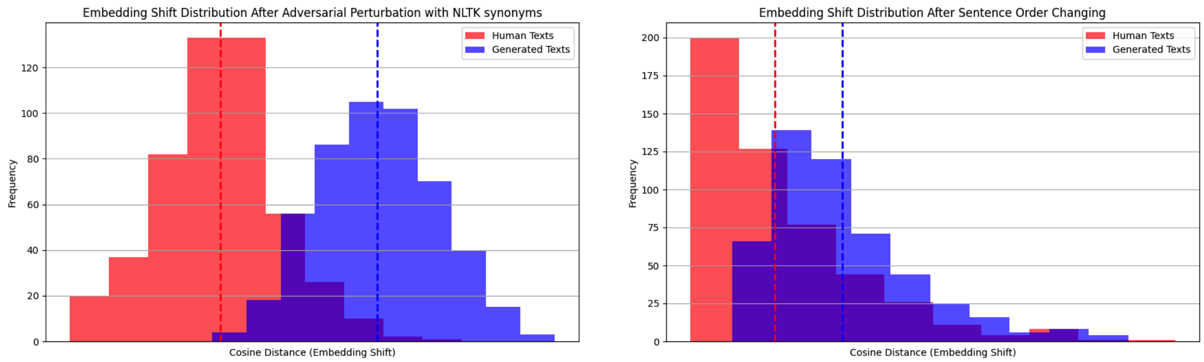

Figure 1: Comparison of embedding shifts after two types of modifications for the HC3 dataset.

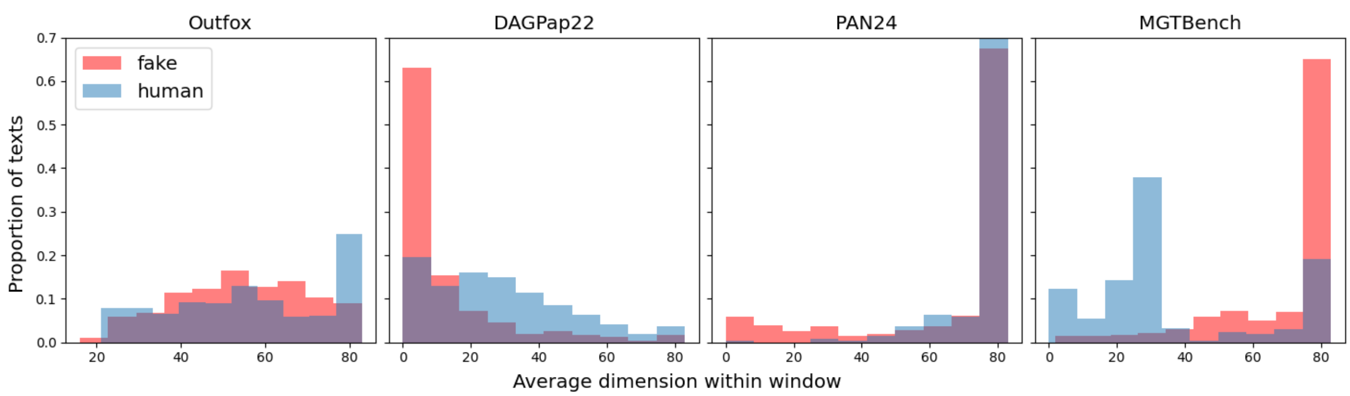

Figure 2: Topological Time Series for different datasets. The results for the remaining datasets selected in this paper can be found in Figure 4.

4.1 Baselines

We used several common classifiers as baselines on each dataset to obtain first-hand understanding of the dataset. The first one is **mDeBERTa** (He et al. 2021), which is the current state-of-the-art model for multilingual machine-generated text detection (Macko et al. 2023). The next baseline is **DetectGPT** framework (Mitchell et al. 2023). We used GPT-2 (Radford et al. 2019) as the base model and T5-Large (Raffel et al. 2019) as perturbations generator. However, as DetectGPT requires intensive computational costs, we utilized Fast-DetectGPT (Bao et al. 2024) that substitutes DetectGPT's perturbation step with a more efficient sampling step. Finally, we used **Binoculars** (Hans et al. 2024) as a method with improved perplexity score. The last two baselines require no fine-tuning which is crucial for detection task, as it is infeasible to train the detector on all existing artificial texts.

4.2 Topological Statistics

It was shown in Tulchinskii et al. (2023) that if we take the inner dimensionality of the manifold on the set of embeddings, we could separate human-written texts from machine-generated ones. The authors used persistence homology dimension (PHD) and showed that statistically human-generated texts have higher PHD than machine-generated texts, therefore introducing a novel detector. We calculated PHD on each set. Additionaly, in Kushnareva et al. (2024) it was suggested to calculate PHD within sliding window. These intrinsic dimensions of the text within sliding window can be used as a feature for detector. The authors conclude that this metric as well can help to differentiate texts of different origin. To be able to compare datasets between each other, we came up with a score, utilizing KL-divergence, but we also needed it to be symmetrical. Let h_d, m_d be distributions of intrinsic dimensions for two types of texts from the same dataset, of human and machine origin, then our $\mathrm{KL_{TTS}}$ is following:

$$\mathrm{KL_{TTS}}(h_d, m_d) = |D_{\mathrm{KL}}(h_d||m_d) - D_{\mathrm{KL}}(m_d||h_d)|$$

The lower this score, the closer h_d and m_d are, which means almost indistinguishable texts and vice versa.

4.3 Perturbations and Shuffling

Based on the results of the text modification studies (Sadasivan et al. 2024; Mitchell et al. 2023), which show how small perturbations affect machine reading comprehension systems, we decided to consider this way of possibly assessing the quality of a dataset. The key idea here is that AI models are sensitive to such adversarial changes, unlike humans. We considered two modification ideas: Adversarial Token Perturbation and Sentence Shuffling.

Adversarial Token Perturbation. In this approach we divide the text into tokens and randomly replace the token with a synonym from the WordNet collection (Miller 1994) with a probability of 70%. We apply such a technique to each represented class. Using an encoder model, we obtain embeddings for each of the texts in the current dataset. Finally, we measure the average embedding shifts for the classes of human and generated texts. We obtain the embedding shifts using the cosine distance between the embeddings of the original texts and the modified ones. As a result, after modifications we obtain Δ_{shift} — the log difference of the average embedding shifts.

$$\Delta_{\text{shift}} = \log \frac{\frac{1}{n}\sum_{i=1}^{n} \cos_d(h^o_{h_i}, h^p_{h_i})}{\frac{1}{m}\sum_{j=1}^{m} \cos_d(h^o_{m_j}, h^p_{m_j})},$$

where n and m – number of samples in the human and generated parts of the dataset respectively, $h^o_{h_i}$ – embedding of i-th fragment of human part of data, $h^p_{h_i}$ – the same embedding after perturbation. The same goes for $h^o_{m_i}$ and $h^o_{m_i}$ for machine-generated part of data. Finally, $\cos_d$ is a function that measures the cosine distance between two vectors.

Sentence Shuffling. In this approach, we randomly swap sentences, thereby affecting the cohesion of the text. We try to find out the effect of artificial origin on the difference between the distributions after permutations. By dividing a fragment into sentences and randomly reversing the order of 70% of the selected sentences, we apply this technique to each represented class. Further, using the text encoding model, we obtain embeddings for each of the texts of the current dataset. Finally, we measure embedding shifts for the class of human and generated texts, and after that we convert the shifts into probability-like distributions. This allows us to obtain at the end $\text{KL}_{\text{shuffle}}(H, M)$ — the KL-divergence between the shifts of human and generated texts.

$$\text{KL}_{\text{shuffle}}(H, M) = \sum_i H(i) \log \frac{H(i)}{M(i)},$$

$$H(i) = \frac{\cos_d(h^o_{h_i}, h^p_{h_i}) + \epsilon}{\sum_j \left(\cos_d(h^o_{h_j}, h^p_{h_j}) + \epsilon\right)},$$

and $M(i)$ has the same structure, except that instead of human class texts the generated class texts are used, ϵ is a small constant added to avoid division by zero.

5 Experiments

From each dataset, we sampled 1000 documents from the test set, balanced between two classes. Regarding baselines, we fine-tuned `mdeberta-v3-base` for each dataset and evaluated the model, information about training can be found in the Appendix C. To evaluate the quality of baselines, Binoculars and Fast-DetectGPT, we utilised `falcon-rw-1b` (Almazrouei et al. 2023) and `gpt-neo-2.7B` (Black et al. 2021) respectively. It is worth noting that with the last two methods we were only able to measure quality for samples in English.

Our objective was to show that datasets of lower quality have shifts that will be easily recognised by the models

Dataset	DeBERTa	Binoculars	DetectGPT
GPT-2	0.972	0.495	0.412
TweepFake	0.941	0.845	0.864
HC3	0.998	0.931	0.972
GhostBuster	0.910	0.683	0.711
MGTBench	0.961	0.364	0.447
MAGE	0.835	0.632	0.654
M4	0.987	0.871	0.881
OutFox	0.901	0.692	0.707
SemEval24 Mono	0.991	0.913	0.924
SemEval24 Multi	0.994	–	–
RuATD	0.765	–	–
DAGPap22	0.968	0.333	0.562
PAN24	0.826	0.411	0.890
AuTex23en	0.941	0.783	0.911
AuTex23es	0.933	–	–
IberAuTex	0.964	–	–
MGT-1 Mono	0.904	0.665	0.683
MGT-1 Multi	0.934	–	–

Table 2: Classification results with different detectors estimated using F_1-score. Binoculars and DetectGPT work only with English texts, thus we could not apply them to datasets with non-English texts.

"from the first step", hence we have not performed any hyperparameter tuning, only one iteration of fine-tuning and testing of the underlying models. In the experiment with topological features we used `roberta-base`, just as the authors of the original paper. In the experiment with perturbations and shuffling, the `multilingual-e5-large` encoder was used to build embeddings of texts, which shows high metrics on encoding high-resource languages (Wang et al. 2024a).

6 Results

The results of the comparison of the designed features on the selected datasets are presented in Table 3. Regarding the PHD and TTS score, in previous works it was shown that texts from language models have smaller PHD values than human-written ones, however this result was obtained for GPT-2, GPT-3.5 and OPT models, and this trend could change for more recent language models, that generate more human-resembling texts. If texts of different origin have high KL_{TTS}, it means that it is easier for a detector to separate such texts. KL_{TTS} is also constrained for shorter texts, see Section 7. As for PHD, we hypothesize that generated texts of good quality should have similar PHD with human-written ones. Additionally, we compare the distributions of

Dataset	$KL_{TTS} \downarrow$	PHD_{human}	$PHD_{machine}$	$\Delta_{shift} \downarrow$	$KL_{shuffle} \downarrow$
GPT-2	**0.014**	9.23 ± 1.98	10.27 ± 1.84	0.084	1.255
HC3	0.053	8.76 ± 1.83	7.38 ± 1.05	0.264	1.167
GhostBuster	0.053	9.84 ± 1.18	9.76 ± 1.15	**0.024**	**0.359**
MGTBench	0.043	8.77 ± 1.31	9.97 ± 1.02	**0.031**	**0.421**
MAGE	**0.011**	9.8 ± 2.14	9.38 ± 3.04	0.094	**0.310**
M4	0.036	7.26 ± 1.99	8.59 ± 1.4	0.107	**0.483**
OutFox	0.025	8.96 ± 1.21	11.48 ± 1.13	0.095	**0.237**
TweepFake	**-**	9.02 ± 3.19	8.12 ± 4.02	0.116	1.001
SemEval24 Mono	**0.012**	9.11 ± 1.19	9.41 ± 1.2	0.191	2.576
SemEval24 Multi	**0.001**	9.65 ± 1.81	9.42 ± 1.44	0.059	2.046
RuATD	$\underline{0.007}$	7.33 ± 1.4	7.46 ± 1.41	0.315	14.028
DAGPap22	0.083	8.35 ± 1.33	7.48 ± 2.01	**0.039**	**0.472**
PAN24	0.053	9.4 ± 1.05	8.52 ± 1.59	**0.050**	**0.331**
AuTex23 Eng	$\underline{0.021}$	8.07 ± 2.26	8.1 ± 2.68	0.110	4.331
AuTex23 Esp	$\underline{0.001}$	9.16 ± 3.49	9.25 ± 3.26	0.105	1.306
IberAuTex	**0.012**	9.33 ± 2.45	8.47 ± 2.73	0.223	5.516
MGT-1 Mono	**0.019**	9.19 ± 1.75	8.96 ± 2.24	**0.031**	0.587
MGT-1 Multi	**0.006**	8.76 ± 1.85	8.6 ± 2.29	**0.027**	0.522

Table 3: Calculated statistics on texts from chosen datasets. Some values for KL_{TTS} are underlined, because texts are too short, see Section 7 and TTS for almost all texts in TweepFake is equal to 0.

PHD for all datasets, see Figure 3. Again, the distributions for texts of both origin should be similar, which holds for texts from SemEval, PAN24 and MGT-1.

In the next columns, we list the statistics observed on modified texts, and for both of these the lower the better, as this reflects the similar degree of resilience of the generated and human texts to adversarial attacks. Qualitatively generated data with no bias should take values close to human.

Finally, in Table 2 we show the results of applying modern classifiers to the chosen test parts. For instance, on the datasets with low values in Table 3, a quality close to 1 can be achieved, which indicates the clear presence of detector bias towards them, or a structural feature that is too obvious for the model. It is not possible to judge the quality of the data only by achieving F_1 values close to 1, but by combining the values of the two tables, we can estimate which set has better quality data and which has lower quality data.

7 Discussion

Regarding KL_{TTS}, on Fig. 2 we show 4 datasets with high value of it. While GhostBuster and PAN24 received such high score due to discrepancy on texts with higher dimensions, MGTBench and DAGPap22 did it due to the difference in distributions themselves. Note also that KL_{TTS} may not perform well with very short texts, since the internal method of computing PHD requires sufficiently long texts for stable computation. Therefore, we discard KL_{TTS} on RuATD and AuTex23-es and Tweepfake, as they do not fit the criteria, see Table 1. On top of that, it has already been shown that texts must be of sufficient length (Gritsay, Grabovoy, and Chekhovich 2022) to build reliable detectors.

Analysing the values in the Table 3, we can trace the presence of sufficiently high quality data in the selected datasets. The developed attributes in aggregate are able to reflect the quality of the generated dataset from different perspectives and angles. We propose to utilise these attributes in combination with other statistical tools for evaluating data quality, e.g., Zipf's law (Powers 1998).

Presented statistics can be utilised to estimate the quality of collections and to improve them. Also, datasets that collect machine-generated content may provide utility for the two more general purposes as well. First, high-quality generated data can be utilised to evaluate the quality of the causal model during training, as one of the training objectives to improve model answers and make it more human-like. Secondly, good detectors can help to clean training sets, as large proportion of low-quality generated texts in those sets can result in emerging biases towards incorrect structure and rubbish fragments in the output of the model in the future.

8 Conclusion

In the current research, we discussed the problem of quality of datasets with AI-generated texts used for testing corresponding detectors. This problem is relevant, as the quality of test data directly influences the quality of widely used detectors. We conducted a review of datasets from competitions and scientific publications on datasets aimed at the detection of AI-generated content and proposed methods

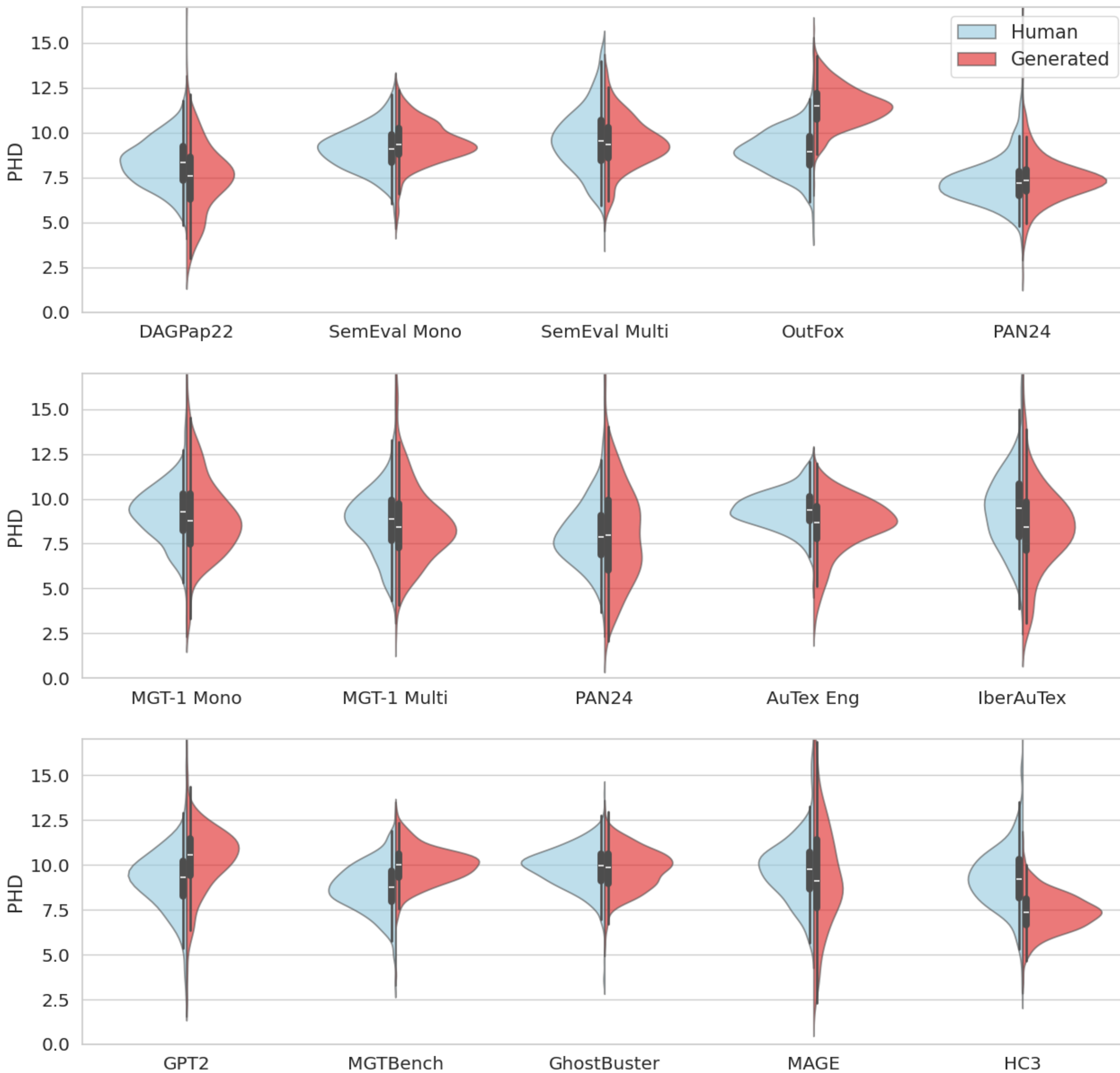

Figure 3: PHD values on all datasets, except TweepFake and AuTex23 Spanish, texts from which were too short for proper calculation of PHD.

to evaluate the quality of datasets containing AI excerpts based on different structural features. We evaluated topological features, robustness to adversarial attacks, and similarity of attention patterns as estimators of data quality. We concluded that all analysed datasets fail in one another of our methods and do not allow to reliably estimate AI detectors. We encourage researchers to propose their own ways for quality assessment, which will allow to create a comprehensive system of evaluation of the detection datasets. Our work aims to contribute to a better understanding of the difference between human and machine text, which will ultimately contribute to preserving the integrity of information in the world.

9 Limitations

In our work we focused on the task of binary classification, thus suggested methods are not optimal for the task of detection of the hybrid AI-human content. Also, some methods do not work properly on short texts, however, so do the detectors.

References

Alemohammad, S.; Casco-Rodriguez, J.; Luzi, L.; Humayun, A. I.; Babaei, H.; LeJeune, D.; Siahkoohi, A.; and Baraniuk, R. G. 2023. Self-Consuming Generative Models Go MAD. arXiv:2307.01850.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Étienne Goffinet; Hesslow, D.; Launay, J.; Malartic, Q.; Mazzotta, D.; Noune, B.; Pannier, B.; and Penedo, G. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867.

Ayele, A. A.; Babakov, N.; Bevendorff, J.; Casals, X. B.; Chulvi, B.; Dementieva, D.; Elnagar, A.; Freitag, D.; Fröbe, M.; Korenčić, D.; Mayerl, M.; Moskovskiy, D.; Mukherjee, A.; Panchenko, A.; Potthast, M.; Rangel, F.; Rizwan, N.; Rosso, P.; Schneider, F.; Smirnova, A.; Stamatatos, E.; Stakovskii, E.; Stein, B.; Taulé, M.; Ustalov, D.; Wang, X.; Wiegmann, M.; Yimam, S. M.; and Zangerle, E. 2024. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In Goeuriot, L.; Mulhem, P.; Quénot, G.; Schwab, D.; Nunzio, G. M. D.; Soulier, L.; Galuscakova, P.; Herrera, A. G. S.; Faggioli, G.; and Ferro, N., eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer.

Bao, G.; Zhao, Y.; Teng, Z.; Yang, L.; and Zhang, Y. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130.

Black, S.; Gao, L.; Wang, P.; Leahy, C.; and Biderman, S. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Boeva, G.; Gritsai, G.; and Grabovoy, A. 2024. Team ap-team at PAN: LLM Adapters for Various Datasets.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Fagni, T.; Falchi, F.; Gambini, M.; Martella, A.; and Tesconi, M. 2021. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5): e0251415.

Fröhling, L.; and Zubiaga, A. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7.

Gray, A. 2024. ChatGPT" contamination": estimating the prevalence of LLMs in the scholarly literature. *arXiv preprint arXiv:2403.16887*.

Gritsai, G.; Khabutdinov, I.; and Grabovoy, A. 2024. Multi-head Span-based Detector for AI-generated Fragments in Scientific Papers. In Ghosal, T.; Singh, A.; Waard, A.; Mayr, P.; Naik, A.; Weller, O.; Lee, Y.; Shen, S.; and Qin, Y., eds., *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, 220–225. Bangkok, Thailand: Association for Computational Linguistics.

Gritsay, G.; Grabovoy, A.; and Chekhovich, Y. 2022. Automatic Detection of Machine Generated Texts: Need More Tokens. In *2022 Ivannikov Memorial Workshop (IVMEM)*, 20–26.

Gritsay, G.; Grabovoy, A.; Kildyakov, A.; and Chekhovich, Y. 2023a. Artificially generated text fragments search in academic documents. In *Doklady Mathematics*, volume 108, S434–S442. Springer.

Gritsay, G.; Grabovoy, A.; Kildyakov, A.; and Chekhovich, Y. 2023b. Automated Text Identification: Multilingual Transformer-based Models Approach. In *IberLEF@SEPLN*.

Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. arXiv:2401.12070.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654.

He, X.; Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2023. MGTBench: Benchmarking Machine-Generated Text Detection. *CoRR abs/2303.14822*.

Jawahar, G.; Abdul-Mageed, M.; and Lakshmanan, L., V.S. 2020. Automatic Detection of Machine Generated Text: A Critical Survey. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 2296–2309. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Kashnitsky, Y.; Herrmannova, D.; de Waard, A.; Tsatsaronis, G.; Fennell, C. C.; and Labbe, C. 2022. Overview of the DAGPap22 Shared Task on Detecting Automatically Generated Scientific Papers. In Cohan, A.; Feigenblat, G.; Freitag, D.; Ghosal, T.; Herrmannova, D.; Knoth, P.; Lo, K.; Mayr, P.; Shmueli-Scheuer, M.; de Waard, A.; and Wang, L. L., eds., *Proceedings of the Third Workshop on Scholarly Document Processing*, 210–213. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Koike, R.; Kaneko, M.; and Okazaki, N. 2024. OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada.

Kushnareva, L.; Gaintseva, T.; Abulkhanov, D.; Kuznetsov, K.; Magai, G.; Tulchinskii, E.; Barannikov, S.; Nikolenko, S.; and Piontkovskaya, I. 2024. Boundary detection in mixed AI-human texts. In *First Conference on Language Modeling*.

Li, Y.; Li, Q.; Cui, L.; Bi, W.; Wang, Z.; Wang, L.; Yang, L.; Shi, S.; and Zhang, Y. 2024. MAGE: Machine-generated Text Detection in the Wild. arXiv:2305.13242.

Liang, W.; Zhang, Y.; Wu, Z.; Lepp, H.; Ji, W.; Zhao, X.; Cao, H.; Liu, S.; He, S.; Huang, Z.; Yang, D.; Potts, C.; Manning, C. D.; and Zou, J. Y. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. In *First Conference on Language Modeling*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Ma, Y.; Liu, J.; Yi, F.; Cheng, Q.; Huang, Y.; Lu, W.; and Liu, X. 2023. AI vs. Human – Differentiation Analysis of Scientific Content Generation. arXiv:2301.10416.

Macko, D.; Moro, R.; Uchendu, A.; Lucas, J.; Yamashita, M.; Pikuliak, M.; Srba, I.; Le, T.; Lee, D.; Simko, J.; and Bielikova, M. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9960–9987. Singapore: Association for Computational Linguistics.

Miller, G. A. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Miralles, P.; Martín, A.; and Camacho, D. 2024. Team aida at PAN: Ensembling Normalized Log Probabilities. In Faggioli, G.; Ferro, N.; Galuscáková, P.; and 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*, volume 3740 of *CEUR Workshop Proceedings*, 2807–2813. CEUR-WS.org.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Powers, D. M. W. 1998. Applications and explanations of Zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, 151–160. USA: Association for Computational Linguistics. ISBN 0725806346.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Raffel, C.; Shazeer, N. M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Rosati, D. 2022. SynSciPass: detecting appropriate uses of scientific text generation. In Cohan, A.; Feigenblat, G.; Freitag, D.; Ghosal, T.; Herrmannova, D.; Knoth, P.; Lo, K.; Mayr, P.; Shmueli-Scheuer, M.; de Waard, A.; and Wang, L. L., eds., *Proceedings of the Third Workshop on Scholarly Document Processing*, 214–222. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2024. Can AI-Generated Text be Reliably Detected?

Sarvazyan, A. M.; Ángel González, J.; Franco-Salvador, M.; Rangel, F.; Chulvi, B.; and Rosso, P. 2023. Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. arXiv:2309.11285.

Sarvazyan, A. M.; Ángel González, J.; Rangel, F.; Rosso, P.; and Franco-Salvador, M. 2024. Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula. *Procesamiento del Lenguaje Natural*, 73(0): 421–434.

Shamardina, T.; Mikhailov, V.; Cherniavskii, D.; Fenogenova, A.; Saidov, M.; Valeeva, A.; Shavrina, T.; Smurov, I.; Tutubalina, E.; and Artemova, E. 2022. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. *CoRR*, abs/2206.01583.

Su, Z.; Wu, X.; Zhou, W.; Ma, G.; and Hu, S. 2024. HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus. arXiv:2309.02731.

Tulchinskii, E.; Kuznetsov, K.; Kushnareva, L.; Cherniavskii, D.; Nikolenko, S.; Burnaev, E.; Barannikov, S.; and Piontkovskaya, I. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Uchendu, A.; Ma, Z.; Le, T.; Zhang, R.; and Lee, D. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2001–2016. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Valdez-Valenzuela, A.; Zavala-Reyes, R. L.; Murillo, V. G. M.; and Gómez-Adorno, H. 2024. The iimasNLP team at IberAuTexTification 2024: Integrating Graph Neural Networks, Multilingual LLMs, and Stylometry for Automatic Text Identification. In *IberLEF@SEPLN*.

Verma, V.; Fleisig, E.; Tomlin, N.; and Klein, D. 2024. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. arXiv:2305.15047.

Villalobos, P.; Sevilla, J.; Heim, L.; Besiroglu, T.; Hobbhahn, M.; and Ho, A. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Voznyuk, A.; and Konovalov, V. 2024. DeepPavlov at SemEval-2024 Task 8: Leveraging Transfer Learning for Detecting Boundaries of Machine-Generated Texts. In Ojha, A. K.; Doğruöz, A. S.; Tayyar Madabushi, H.; Da San Martino, G.; Rosenthal, S.; and Rosá, A., eds., *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 1821–1829. Mexico City, Mexico: Association for Computational Linguistics.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024a. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672.

Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Mohammed Afzal, O.; Mahmoud, T.; Puccetti, G.; Arnold, T.; Aji, A.; Habash, N.; Gurevych, I.; and Nakov, P. 2024b. M4GT-Bench: Evaluation Benchmark for Black-Box Machine-Generated Text Detection. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3964–3992. Bangkok, Thailand: Association for Computational Linguistics.

Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Whitehouse, C.; Afzal, O. M.; Mahmoud, T.; Puccetti, G.; Arnold, T.; Aji, A. F.; Habash, N.; Gurevych, I.; and Nakov, P. 2024c. SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024. Mexico, Mexico.

Wang, Y.; Shelmanov, A.; Mansurov, J.; Tsvigun, A.; Mikhailov, V.; Xing, R.; Xie, Z.; Geng, J.; Puccetti, G.; Artemova, E.; Su, J.; Ta, M. N.; Abassy, M.; Elozeiri, K.; Ahmed, S. E. D.; Goloburda, M.; Mahmoud, T.; Tomar, R. V.; Aziz, A.; Laiyk, N.; Afzal, O. M.; Koike, R.; Kaneko, M.; Aji, A. F.; Habash, N.; Gurevych, I.; and Nakov, P. 2025. GenAI Content Detection Task 1: English and Multilingual Machine-generated Text Detection: AI vs. Human. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*. Abu Dhabi, UAE: Association for Computational Linguistics.

Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Wong, D. F.; and Chao, L. S. 2023. A Survey on LLM-gernerated Text Detection: Necessity, Methods, and Future Directions. *CoRR*, abs/2310.14724.

Xu, W.; Wang, D.; Pan, L.; Song, Z.; Freitag, M.; Wang, W. Y.; and Li, L. 2023. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. arXiv:2305.14282.

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; and Choi, Y. 2019. *Defending against neural fake news*. Red Hook, NY, USA: Curran Associates Inc.

Zhao, Y.; Gao, J.; Wang, J.; Luo, G.; and Tang, L. 2024. Utilizing an Ensemble Model with Anomalous Label Smoothing to Detect Generated Scientific Papers. *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*.

Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Zhu, W.; and Bhat, S. 2020. GRUEN for Evaluating Linguistic Quality of Generated Text. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 94–108. Online: Association for Computational Linguistics.

A Data Description

More detailed description with information on sources, topics and years of the datasets selected in this paper from competitions and research papers in Table 4

B Evaluation results of competitions

Table 5 shows the winning scores in the competitions reviewed in this paper. In the AuTex and IberAuTex competitions it was forbidden to use additional data to fine-tune the detection algorithms. In the other collections it was allowed, we can notice a high quality near perfect in them. We should note the low value of metrics on the RuATD dataset, which can be explained by the limited number of available high-quality language models in Russian during the competition.

Competition	**Metric**	**Best result**
RuATD	Accuracy	0.820
AuTex23-en	Macro-F1	0.809
AuTex23-es	Macro-F1	0.708
IberAuTex	Macro-F1	0.805
SemEval24 Mono	Accuracy	0.975
SemEval24 Multi	Accuracy	0.959
PAN24	Avg. of 5 metrics*	0.924
DAGPap22	Avg. F1-score	0.994
MGT-1 Mono	Macro-F1	0.8307
MGT-1 Multi	Macro-F1	0.7916

Table 5: Best results from each analysed competition. PAN24 used mean of 5 metrics, such as accuracy, F1 and other to evaluate *efficiency* of the system.

C Hyperparameters

Hyperparameters	**Values**
Epochs	5*
Learning rate (LR)	5e-5
Warmup steps	50
Weight decay	0.01

Table 6: Hyperparameters for fine-tuning mDeBERTa-base. We trained for 5 epochs with possibility of early exit.

The training was carried out on NVIDIA GeForce RTX 3090. See hyperparameters in Table 6.

Dataset	Year	Themes	Sources
		Research papers datasets	
GPT2	2019	WebText	GPT-2
TweepFake	2019	Tweets	Markov Chains, RNN, LSTM, GPT-2
HC3	2023	ELI5, WikiQA, Wikipedia, Medicine, Finance	ChatGPT
GhostBuster	2023	Student Essays, News Articles, Creative Writing	ChatGPT, Claude
MGTBench	2024	Student Essays, News Articles, Creative Writing	ChatGPT. ChatGLM, Dolly, GPT4All, StableLM, Claude
MAGE	2024	Opinions, Reviews, News, QA, Story Generation, Commonsense Reasoning, Knowledge Illustration, Scientific Writing	text-davinci-002, GPT-3.5, ChatGPT, LLaMA, GLM-130B, FLAN-T5, OPT, BLOOM, GPT-J-6B, GPT-NeoX-20B
M4	2024	Wikipedia, Reddit ELI5, WikiHow, PeerRead, arXiv abstract	GPT-3.5, ChatGPT, Cohere, Dolly-v2, BLOOM
OutFox	2024	Student Essays	ChatGPT, GPT-3.5, FLAN-T5
		Shared tasks datasets	
RuATD	2022	News, Social media, Wikipedia, Strategic Documents, Diaries	M-BART, M2M-100, OPUS-MT, mT5, ruGPT2, ruGPT3, ruT5-Base
DAGPap	2022	Scopus papers	Longformer Encoder-Decoder, GPT-3, Spinbot, GPT-Neo
AuTex	2023	Legal documents, Social media, How-to articles	BLOOM, GPT-3, GPT-3.5
IberAuTex	2024	News, Reviews, Emails, Essays, Dialogues, Wikipedia, Wikihow, Tweets	GPT-2, LLaMA, Mistral, Cohere, Claude, MPT, Falcon
PAN	2024	News	Alpaca, BLOOM, Gemini, ChatGPT, gpt-4-turbo, LLaMA-2, Mistral, Qwen1.5, GPT-2
SemEval Mono	2024	Wikipedia, WikiHow, Reddit, arXiv, PeerRead, Student Essays	GPT-3.5, GPT-4, Cohere, Dolly-v2, BLOOMz
SemEval Multi	2024	Wikipedia, WikiHow, Reddit, arXiv, PeerRead, Student Essays, News	ChatGPT, GPT-3.5, GPT-4, LLaMA2, Cohere, Dolly-v2, BLOOM, Jais
MGT-1 Mono	2025	CNN, DialogSum, Wikipedia, WikiHow, Eli5, Finance, XSum, PubMed, SQuAD, IMDb, Reddit, arXiv, PeerRead	text-davinci-002, GPT-3.5, ChatGPT, OPT, LLaMA3, BLOOM, FLAN-T5, Cohere, Dolly, Gemma, Mixtral
MGT-1 Multi	2025	CNN, DialogSum, Baike, WikiQA, WikiHow, Eli5, Finance, Psychology, XSum, PubMed, SQuAD, IMDb, Reddit, arXiv, PeerRead	text-davinci-002, GPT-3.5, ChatGPT, gpt4o, GLM, GPT-J, GPT-Neo, OPT, LLaMA2, LLaMA3, BLOOM, FLAN-T5, Cohere, Dolly, Gemma, Mixtral, Jais

Table 4: More detailed descriptive statistics about domains and generators of the chosen datasets from competitions and research papers. ChatGPT is gpt-3.5-turbo, GPT-3.5 is text-davinci-003.

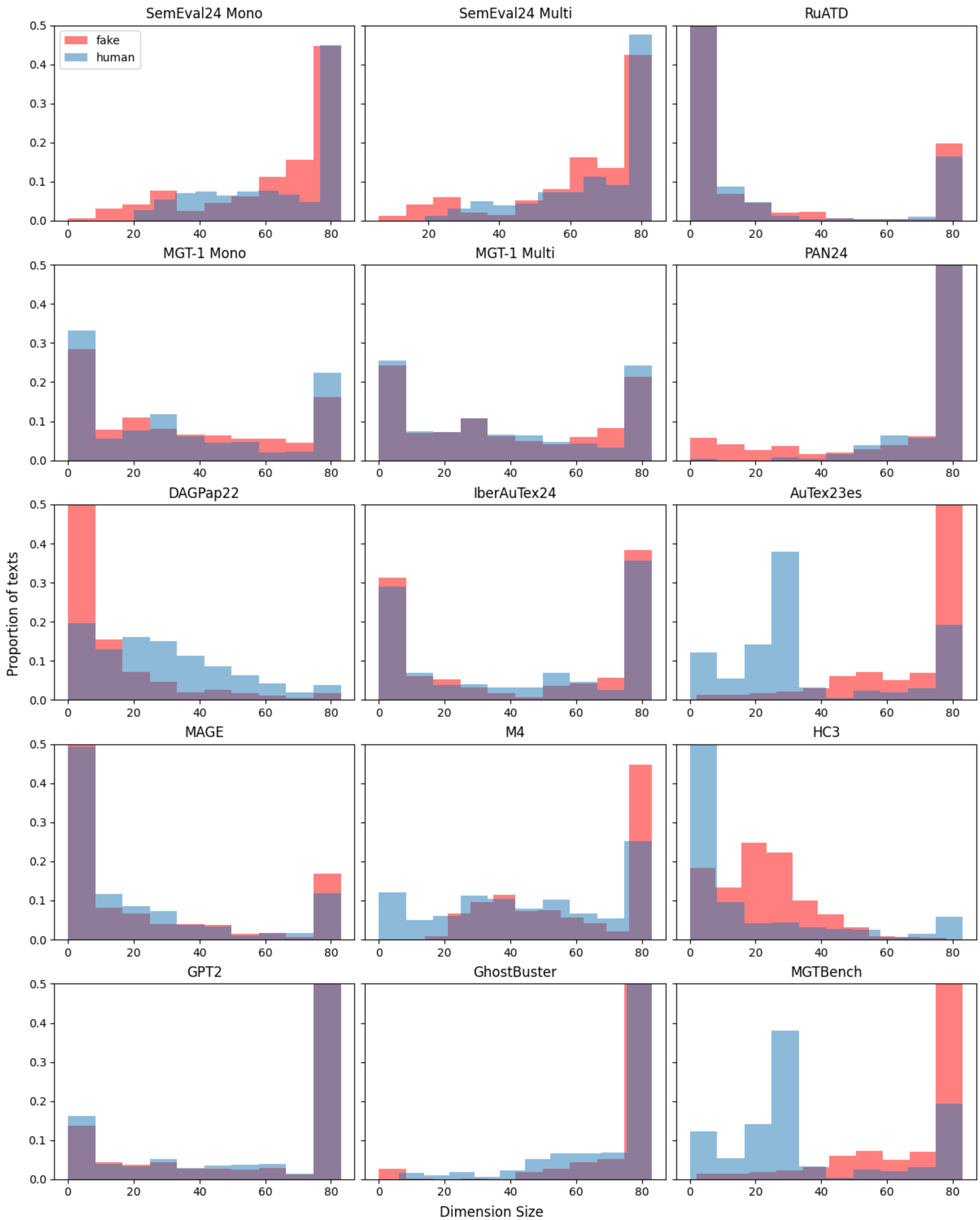

Figure 4: Topological Time Series on all datasets.