MultiSocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts

Dominik Macko, Jakub Kopal, Robert Moro, Ivan Srba
Kempelen Institute of Intelligent Technologies, Slovakia
{dominik.macko, jakub.kopal, robert.moro, ivan.srba}@kinit.sk

Abstract

Recent LLMs are able to generate high-quality multilingual texts, indistinguishable for humans from authentic human-written ones. Research in machine-generated text detection is however mostly focused on the English language and longer texts, such as news articles, scientific papers or student essays. Social-media texts are usually much shorter and often feature informal language, grammatical errors, or distinct linguistic items (e.g., emoticons, hashtags). There is a gap in studying the ability of existing methods in detection of such texts, reflected also in the lack of existing multilingual benchmark datasets. To fill this gap we propose the first multilingual (22 languages) and multi-platform (5 social media platforms) dataset for benchmarking machine-generated text detection in the social-media domain, called MultiSocial[1]. It contains 472,097 texts, of which about 58k are human-written and approximately the same amount is generated by each of 7 multilingual LLMs. We use this benchmark to compare existing detection methods in zero-shot as well as fine-tuned form. Our results indicate that the fine-tuned detectors have no problem to be trained on social-media texts and that the platform selection for training matters.

1 Introduction

The most advanced text-generation AI models, called large language models (LLMs), are able to generate high-quality texts in various languages (Qin et al., 2024). Although this presents an opportunity to make a human life and work more efficient, it also presents a threat of being misused, as such generated texts are not easily recognisable by humans (Zellers et al., 2019). This is especially crucial in regard to social-media networks (SMN, such as Facebook, X/Twitter, etc.), where anyone

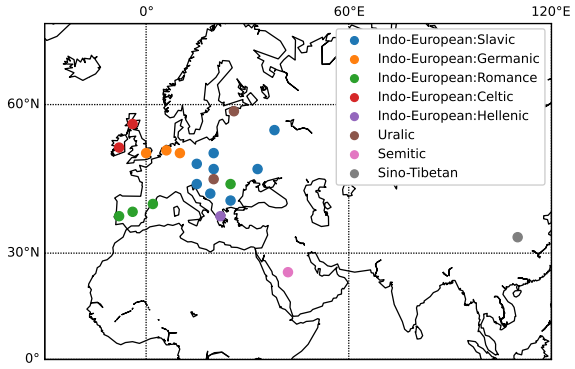

Figure 1: MultiSocial coverage of languages.

can be a source of a harmful content (without editorial consent) with a potentially wide reach (depending on a network) (Aïmeur et al., 2023). To prevent the LLM misuse (e.g., social engineering, disinformation spreading), a reliable mechanism to detect machine-generated text (MGT) is needed.

Unfortunately, the existing research in MGT detection (MGTD) focuses either solely on English (as in most NLP fields, e.g., Dugan et al., 2024) or leaves SMN texts out of scope due to challenges they bring (Kumarage et al., 2024; Lin et al., 2024). These challenges include very informal writing style often used in SMN, such as using slang, ignoring grammar rules, using distinct linguistic items in the texts (e.g., emoticons or hashtags), or usually including very short lengths of the texts.

Our work fills a gap in the state-of-the-art (SOTA) by introducing a new heavily multilingual dataset for MGTD research in the social-media domain. We use this dataset to further benchmark the existing SOTA detectors in various aspects. Specifically, our key contributions include:

(1) The first **multilingual evaluation of SOTA MGTD methods (statistical, pre-trained, fine-tuned) on social-media texts**, focusing on multilingual as well as cross-lingual capability of existing detectors and comparison of different cate-

[1]Code: https://github.com/kinit-sk/multisocial
Dataset: https://doi.org/10.5281/zenodo.13846152

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 727–752

gories. The best detectors perform similarly across all tested languages, although there are still differences between English and non-English in zero-shot evaluation.

(2) The first **multi-platform and cross-platform evaluation of SOTA MGTD methods** in social-media domain, evaluating differences in MGTD performance based on text types and sources in multiple languages. We found that Telegram offers the best cross-lingual capability.

(3) The unique **multilingual, multi-platform, and multi-generator benchmark dataset** of human-written and machine-generated social-media texts, called MultiSocial, covering 22 languages, 5 SMN platforms, and 7 LLM generators.

2 Related Work

Multilingual machine-generated text detection has been gaining attention recently. There have been multiple non-English language MGT detection shared tasks in the last years, such as Russian at RuATD 2022 (Shamardina et al., 2022), Spanish at AuTexTification 2023 (Sarvazyan et al., 2023), Dutch at CLIN33 (Fivez et al., 2024), or multilingual at SemEval-2024 Task 8 (Wang et al., 2024b).

The last one covers 9 languages, and is based on M4GT-Bench (Wang et al., 2024a), multilingual, multi-domain, and multi-generator MGT benchmark. However, its coverage of the SMN domain is rather sparse (solely English-only Reddit texts are included). Moreover, language coverage of the domains is highly imbalanced (e.g., most languages are represented only in the news domain, while Arabic and German also in Wikipedia domain, and Chinese only in the QA domain) and there are only one or two generators used in the multilingual settings. Therefore, cross-lingual evaluation is somewhat limited using such data.

Another benchmark dataset called MULTITuDE (Macko et al., 2023) covers 11 languages; however, 8 of them are included in the test split only with a limited number of samples. Moreover, it is focused on the news domain only, in which the texts are typically long, formal, and carefully checked for grammatical correctness. Such settings are clearly different than those of social media texts. An extension of MULTITuDE dataset to evaluate various authorship obfuscation methods is proposed in (Macko et al., 2024). Although it enables to evaluate robustness of MGT detection methods, it is still limited to news domain.

Dataset	H/M	LLM	Lang	Domain	SMN
TweepFake (Fagni et al., 2021)	12k/12k	6	1	SMN	1
Fox8-23 (Yang and Menczer, 2023)	228k/170k	1	1	SMN	1
F3 (Lucas et al., 2023)	28k/28k	1	1	news, SMN	1
HC3 (Guo et al., 2023)	81k/44k	1	2	QA, Wiki	0
SAID (Cui et al., 2023)	87k/131k	N/A	2	QA	0
MULTITuDE (Macko et al., 2023)	8k/66k	8	11	news	0
M4GT-Bench (Wang et al., 2024a)	93k/124k	8	9	6	0
MultiSocial (ours)	58k/414k	7	22	SMN	5

Table 1: Comparison of existing publicly available MGT detection datasets either multilingual or focused on social-media network (SMN) domain. H/M refers to the no. of human-written and machine-generated samples.

There is also MAiDE-up dataset (Ignat et al., 2024) of hotel reviews available, covering 10 languages; however, it is limited to GPT-4 generated data only (limiting the generalization of conclusions). Although such texts are more similar to social-media texts than news articles, they cover a single topic (accommodation) and still use a different communication style than the social-media networks. Many other works have focused on detection of fake reviews, but only few of them reflected multilingualism (Duma et al., 2024).

Benchmark datasets HC3 (Guo et al., 2023) and SAID (Cui et al., 2023) contain Chinese and English texts covering forum-like question-answering domain. HC3 contains only ChatGPT-generated machine texts, SAID contains real bot-generated texts (relying on human annotations to identify them). The downside of SAID is that the specific generation model of individual texts is not known. Otherwise, social-media texts are covered only in rather old TweepFake (Fagni et al., 2021) dataset, which includes English-only tweets and cannot be used to evaluate MGT detection methods on social-media texts generated by the most modern LLMs. Fox8-23 (Yang and Menczer, 2023) dataset also focuses on English, and covers presumably only ChatGPT-generated tweets. Similarly, F3 (Lucas et al., 2023) dataset contains English ChatGPT-generated real and fake news as well as tweets. There are other monolingual works, such as (Temnikova et al., 2023) focused on fake tweets in Bulgarian; however, the crafted dataset is not publicly available.

Table 1 includes comparison of the new dataset proposed in this work (described in the next sec-

	Train						Test					
Language	**Discord**	**Gab**	**Telegram**	**Twitter**	**WhatsApp**	**all**	**Discord**	**Gab**	**Telegram**	**Twitter**	**WhatsApp**	**all**
Arabic (ar)	0	0	7724	7872	0	15596	0	1556	2319	2364	1750	7989
Bulgarian (bg)	0	0	7555	7930	0	15485	0	192	2334	2367	0	4893
Catalan (ca)	6984	0	7157	0	0	14141	2105	1264	2134	1824	144	7471
Chinese (zh)	0	0	7924	0	0	7924	0	830	2380	238	32	3480
Croatian (hr)	7502	0	7065	0	0	14567	2255	1340	2315	91	38	6039
Czech (cs)	3450	0	7690	0	0	11140	2175	492	2309	1017	134	6127
Dutch (nl)	7391	7900	7750	7933	0	30974	2191	2356	2318	2369	397	9631
English (en)	7789	7760	7824	7871	7782	39026	2341	2340	2350	2365	2334	11730
Estonian (et)	6974	0	7520	0	0	14494	2071	805	2259	164	120	5419
German (de)	3407	7863	7880	2095	0	21245	2232	2366	2356	2345	304	9603
Greek (el)	0	0	3814	0	0	3814	0	1195	2274	146	35	3650
Hungarian (hu)	7079	0	7461	0	0	14540	2094	1211	2228	413	22	5968
Irish (ga)	0	0	0	0	0	0	1319	968	821	45	0	3153
Polish (pl)	7158	1829	7733	0	0	16720	2136	2311	2310	172	62	6991
Portuguese (pt)	6860	7916	7842	6481	4354	33453	2284	2371	2347	2360	2363	11725
Romanian (ro)	7436	851	6792	64	0	15143	2236	2349	2298	2378	132	9393
Russian (ru)	0	7875	7827	362	0	16064	0	2355	2340	2361	960	8016
Scottish Gaelic (gd)	0	0	0	0	0	0	150	35	34	0	0	219
Slovak (sk)	0	0	0	0	0	0	107	308	1508	110	0	2033
Slovenian (sl)	0	0	0	0	0	0	203	1912	917	40	0	3072
Spanish (es)	7588	7883	7884	7922	7804	39081	2268	2361	2354	2376	2341	11700
Ukrainian (uk)	0	0	7802	0	0	7802	0	174	2342	70	0	2586
Total	79618	49877	133244	48530	19940	**331209**	28167	31091	44847	25615	11168	**140888**

Table 2: MultiSocial text sample counts across languages and platforms for train and test split.

tion) with the selected existing publicly available datasets for MGT detection.

3 Dataset

Since there is no dataset of multilingual SMN texts containing human-written texts along with the texts generated by SOTA text-generation LLMs, we have crafted a new MultiSocial benchmark dataset. It contains human-written data from five different social-media platforms (reused from the existing multilingual SMN datasets), namely Telegram, Twitter (X), Gab, Discord, and WhatsApp, including post-like as well as chat-like texts. For each authentic human-written text, the texts generated by 7 SOTA LLMs (representatives of private and open multilingual LLMs of various sizes and architectures) are included by using **three iterations of paraphrasing**. Other text-generation approaches were also considered, outputs of which were evaluated and compared by humans, automated similarity metrics, and meta-evaluation utilizing LLM judges. The final approach was selected based on sufficient output quality and similarity to the human texts (to avoid detection biases). Details regarding dataset construction (including selection, evaluation, pre-processing, generation, and post-processing of texts) are provided in Appendix B.

Dataset consists of 472,097 texts (58k are human-written) split into train and test subsets, of which sample counts are summarized across languages and platforms in Table 2. We have conducted linguistic analysis of the machine-generated texts along with the similarity comparison to human texts, with a manual human check of a balanced subset, and identified small portion (about 1%) of noise in the generated data (e.g., "As an AI model..."), indicating model failure during generation. We have intentionally left such samples in the dataset (clearly marked in the data) for further analysis purpose (as indicated in Appendix B). We however filter-out the identified noise for the purpose of the experiments in this study. Although such a post-processing cannot guarantee 100% removal of noisy data, the obvious failures are cleaned. Furthermore, we have used meta-evaluation utilizing LLM judges to compare the quality of the generated texts of individual generators to the quality of original human texts, indicating that **machine-generated texts are of similar or higher quality** (see Appendix B)).

Language Selection. The MultiSocial benchmark dataset covers 22 (high- and low-resource) languages (some only in the testing split), 20 of which (18 of Indo-European and 2 of Uralic language families) have been selected based on our research-projects needs. However, we have intentionally included 2 more (Arabic of Semitic and Chinese of Sino-Tibetan family), which are completely linguistically and geographically unrelated, to study cross-lingual characteristics (Figure 1). The dataset is strongly focused on the Indo-European language family, but contains 4 language families in total. Test split includes all of the train languages, but also 2 Celtic languages,

Generator	METEOR ↑	BERTScore ↑	ngram ↑	LD ↓	MAUVE ↓	LangCheck ↓
Aya-101	**0.195 (±0.23)**	**0.675 (±0.14)**	**0.156 (±0.18)**	1.104 (±0.91)	0.063	10.59%
Gemini	0.152 (±0.20)	0.621 (±0.10)	0.087 (±0.17)	16.296 (±33.39)	**0.025**	6.16%
GPT-3.5-Turbo-0125	0.143 (±0.18)	0.664 (±0.10)	0.080 (±0.10)	2.359 (±3.82)	0.076	22.80%
Mistral-7B-Instruct-v0.2	0.152 (±0.16)	0.652 (±0.08)	0.088 (±0.09)	2.383 (±1.93)	0.047	10.61%
OPT-IML-Max-30b	0.127 (±0.19)	0.659 (±0.12)	0.105 (±0.14)	**0.998 (±0.57)**	0.108	14.88%
v5-Eagle-7B-HF	0.108 (±0.15)	0.628 (±0.07)	0.071 (±0.10)	2.568 (±2.10)	0.027	**6.01%**
Vicuna-13b	0.133 (±0.17)	0.650 (±0.09)	0.089 (±0.11)	1.811 (±1.40)	0.042	6.26%

Table 3: Similarity analysis between machine-generated (3-iteration paraphrased) and human-written (original) social-media texts for individual generation models [mean (± std)]. Arrows refer to values representing more similar texts, boldfaced values represent the most similar texts for each metric.

1 more South Slavic and 1 more West Slavic language. There are 5 writing scripts in both train and test splits of the dataset, where majority of languages uses Latin, but there is also Cyrillic (Russian, Ukrainian, Bulgarian), Arabic, Hanzi (Chinese), and Greek. Nice feature is that the train split includes at least 3 representatives of Germanic, Romance, Slavic-Latin, and Slavic-Cyrillic, which enables various combinations of studies regarding multilingual and cross-lingual characteristics of machine-generated text detection. Furthermore, in Slavic and Romance languages, there are included at least 2 representatives of languages that can be considered high-resource and low-resource.

Although not all languages have enough samples from each of five platforms (due to unavailability of human samples in the selected source datasets), specific subsets can be used to study specific characteristics (e.g., cross-platform transferability using English and Spanish languages, cross-lingual transferability using Telegram platform, surprise language and/or surprise platform evaluation not using specific portions of the train data, etc.).

Human-Machine Similarity Analysis. As mentioned, we have conducted a similarity analysis of the final generated texts by the selected generation models (reported in Table 3). Definitions of the used metrics are available in Appendix B. We can observe only small differences between the generators. Aya-101 generated the most similar texts in general, while OPT-IML-Max-30B provided the worst results (based on higher MAUVE and LangCheck scores). ChatGPT (GPT-3.5-Turbo) also generated the texts resulting in a little higher MAUVE score and the highest language mismatch (LangCheck). However, the language mismatch of ChatGPT (using all the selected languages) was not confirmed by longer news articles generation (FastText language detection in such texts is definitely more accurate), where it actually achieved the lowest (under 1%) LangCheck score among the generators. We assume that in social-media text generation it can better follow the original style of the text than the other generators (e.g., grammatically incorrect, slang), which is more difficult for accurate language detection.

4 Detection Methods

For the benchmark purpose, we have covered 3 specific categories of detection methods: *statistical zero-shot* (methods using statistical differences to differentiate human-written and machine-generated texts applicable without training), *pre-trained* (directly applicable models that are fine-tuned for the MGT detection task using different data – i.e., out-of-distribution), and *fine-tuned* (foundation models fine-tuned for the MGT detection task using MultiSocial dataset – i.e., in-distribution).

For statistical category, we have selected the following 5 most-promising methods (excluding perturbation-based and multi-generation methods due to high computing costs): **Binoculars** (Hans et al., 2024) with Falcon-7B (Almazrouei et al., 2023) as an observer model and Falcon-7B-Instruct as a performer model, **Fast-DetectGPT** (Bao et al., 2023) with GPT-J-6B (Wang and Komatsuzaki, 2021) as both the reference and sampling models, **LLM-Deviation** (Wu and Xiang, 2023), **DetectLLM-LRR** (Su et al., 2023), **S5** (Spiegel and Macko, 2024b) (multiplying 5 statistical metrics of Likelihood, Entropy, Rank, Log-Rank, and LLM-Deviation), all three of them using GPT-J-6B as a base model (the same as in Fast-DetectGPT).

For pre-trained category, we have selected the following 5 detectors with a multilingual potential: **ChatGPT-detector-RoBERTa-Chinese** (Guo et al., 2023) (with a Chinese fine-tuned model), **Longformer Detector** (Li et al., 2023) (showing decent multilingual potential in Macko et al., 2024),

ruRoBERTa-ruatd-binary[2] (as a best single-model system in RuATD 2022 Shamardina et al., 2022), **BLOOMZ-3B-mixed-detector** (Nicolai Thorer Sivesind and Andreas Bentzen Winje, 2023) (with a heavily multilingual fine-tuned LLM), and **RoBERTa-Large OpenAI Detectors** (Solaiman et al., 2019) (as a widely used representative, although English-only).

For fine-tuned category, we have selected 7 multilingual foundational models covering multiple architectures and model sizes: **mDeBERTa-v3-base** (He et al., 2022) (as the best detector in Macko et al., 2023), **XLM-RoBERTa-large** (Conneau et al., 2020) (as the best detector in Macko et al., 2023 based on AUC ROC), **Mistral-7B** (Jiang et al., 2023) (as the best single-model multilingual detector in of SemEval-2024 Task 8 Wang et al., 2024b), **Llama-3-8B** (AI@Meta, 2024) (as a SOTA multilingual smaller LLM), **Aya-101** (Üstün et al., 2024) (as a SOTA representative of multilingual LLMs with encoder-decoder architecture), **BLOOMZ-3B** (Muennighoff et al., 2022) (as the best pre-trained detector base model), and **Falcon-rw-1B** (as a smaller version of the best performing model at ALTA 2023 Gagiano and Tian, 2023, since the 7B model version is already covered by better performing Mistral, Spiegel and Macko, 2024b).

For statistical and pre-trained categories of detection methods, we have used their versions implemented in the IMGTB framework (Spiegel and Macko, 2024a).

5 Experimental Results

Firstly, we provide benchmark evaluation of the selected MGTD methods on all MultiSocial test data (Table 4). It provides a fair comparison of the methods, although the data samples among platforms and languages are not perfectly balanced. Therefore, for further experiments targeting specific cross-lingual and cross-platform research questions, the carefully selected parts of train and test splits are used (described in the corresponding subsections). Per-language, per-platform, and per-generator results are provided in Appendix F. Although worse-than-random performances of some pre-trained detectors indicate a potential problem with the detection (e.g., flipped labels), it is not the case as confirmed by the results in Table 21.

For comparison of MGTD methods, we use **AUC ROC** (area under the curve of receiver operating characteristic) as a classification-threshold independent metric (not affected by a threshold calibration on in-domain data), also used by (Hans et al., 2024). Due to imbalanced test data (machine class contains 7x more samples), we also use **Macro avg. F1-score @ 5% FPR** (false positive rate) as a metric balancing between a precision and a recall of the classification, while the threshold is calibrated based on the train data (to avoid data leakage) ROC curve to achieve 5% FPR (similarly used in Dugan et al., 2024).

Since Gemini-generated data have used a slightly different generation process (e.g., jailbreak prompt, see Appendix B) and achieved highly outlier word-count and unique-words scores in Table 12, we do not use Gemini-generated data in the training (fine-tuning or classification-threshold calibration). Nevertheless, they are still included in the evaluation (can serve for unseen generator evaluation).

Rank	Detector	AUC ROC	MacroF1 @5%FPR
1	Llama-3-8b-MultiSocial	0.9769	0.8696
2	Mistral-7b-v0.1-MultiSocial	0.9768	0.8692
3	Aya-101-MultiSocial	0.9731	0.8462
4	Falcon-rw-1b-MultiSocial	0.9592	0.7810
5	BLOOMZ-3b-MultiSocial	0.9582	0.7843
6	XLM-RoBERTa-large-MultiSocial	0.9553	0.7840
7	mDeBERTa-v3-base-MultiSocial	0.9544	0.7652
8	BLOOMZ-3b-mixed-Detector	0.7553	0.3024
9	DetectLLM-LRR	0.7464	0.2523
10	LLM-Deviation	0.7454	0.2497
11	S5	0.7418	0.2465
12	Fast-Detect-GPT	0.7418	0.3605
13	Binoculars	0.7248	0.2815
14	ChatGPT-Detector-RoBERTa-Chinese	0.7180	0.3416
15	ruRoBERTa-ruatd-binary	0.4817	0.1711
16	Longformer Detector	0.4615	0.1516
17	RoBERTa-large-OpenAI-Detector	0.3450	0.1376

Table 4: Benchmark of the selected MGTD methods of statistical, pre-trained, and fine-tuned categories (as defined in Section 4). The highlight color identifies the category of the method in the table.

5.1 Multilingual Zero-shot Detection

This experiment is focused on the following research question: ***RQ1:*** *How well are social-media texts of multiple languages and platforms detectable by MGT detectors applicable in zero-shot manner (out-of-distribution, without further in-domain training)?* Since SMN texts are usually shorter than commonly used news articles, the detection performance of existing directly usable (i.e., without in-domain training) MGT detectors is still unknown (it could differ from the reported performance on other domains). This has not been evaluated in the multilingual settings. Are there

[2] https://huggingface.co/orzhan/ruroberta-ruatd-binary

Category	Detector	Test Language [AUC ROC] ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
P	BLOOMZ-3b-mixed-Detector	0.79	0.74	0.79	0.80	0.77	0.76	0.80	0.79	0.83	0.78	0.66	0.77	0.84	0.75	0.78	0.79	0.66	0.68	0.76	0.64	0.70	0.69	0.76
	ChatGPT-Detector-RoBERTa-Chinese	0.72	0.80	0.76	0.66	0.80	0.75	0.90	0.81	0.82	0.73	0.63	0.63	0.78	0.70	0.62	0.73	0.75	0.76	0.63	0.66	0.74	0.81	0.72
	Longformer Detector	0.34	0.48	0.32	0.47	0.43	0.54	0.65	0.43	0.48	0.53	0.49	0.51	0.50	0.41	0.46	0.42	0.41	0.46	0.45	0.46	0.61	0.47	0.46
	RoBERTa-large-OpenAI-Detector	0.73	0.43	0.43	0.14	0.32	0.74	0.52	0.30	0.20	0.30	0.48	0.19	0.13	0.30	0.21	0.23	0.24	0.54	0.26	0.33	0.36	0.60	0.35
	ruRoBERTa-ruatd-binary	0.40	0.63	0.56	0.43	0.43	0.35	0.56	0.47	0.43	0.49	0.47	0.48	0.46	0.50	0.44	0.43	0.34	0.70	0.45	0.47	0.59	0.44	0.48
S	Binoculars	0.70	0.68	0.62	0.74	0.75	0.79	0.80	0.76	0.74	0.71	0.71	0.79	0.78	0.72	0.75	0.75	0.74	0.72	0.71	0.73	0.64	0.74	0.72
	DetectLLM-LRR	0.79	0.86	0.69	0.93	0.78	0.88	0.80	0.82	0.88	0.79	0.74	0.88	0.94	0.79	0.88	0.84	0.87	0.78	0.85	0.79	0.75	0.78	0.75
	Fast-Detect-GPT	0.75	0.65	0.61	0.81	0.77	0.66	0.80	0.74	0.69	0.70	0.74	0.80	0.77	0.74	0.77	0.77	0.77	0.73	0.71	0.74	0.70	0.74	0.74
	LLM-Deviation	0.82	0.86	0.68	0.93	0.79	0.89	0.80	0.82	0.90	0.81	0.79	0.89	0.94	0.79	0.89	0.84	0.88	0.78	0.86	0.81	0.75	0.79	0.75
	S5	0.80	0.85	0.68	0.92	0.78	0.88	0.77	0.81	0.89	0.80	0.78	0.88	0.94	0.77	0.88	0.83	0.88	0.78	0.85	0.80	0.74	0.78	0.74

Table 5: Per-language AUC ROC performance of zero-shot statistical (S) and pre-trained (P) MGT detectors. The data are too difficult for the three under-performing pre-trained models.

Category	Platform	Test Language [AUC ROC mean] ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
P	Discord	N/A	N/A	0.92	0.81	0.86	N/A	0.91	0.89	0.86	N/A	N/A	0.75	0.84	0.81	0.81	0.82	0.81	N/A	N/A	N/A	N/A	N/A	0.82
	Gab	N/A	N/A	N/A	N/A	0.71	N/A	0.82	0.71	N/A	N/A	N/A	N/A	N/A	0.71	0.64	0.72	0.61	0.62	N/A	N/A	N/A	N/A	0.66
	Telegram	0.73	0.77	0.73	0.73	0.79	0.81	0.86	0.80	0.84	N/A	N/A	0.71	0.86	0.68	0.70	0.77	0.77	0.76	N/A	N/A	0.73	0.76	0.74
	Twitter	0.84	0.79	N/A	N/A	0.78	N/A	0.85	0.73	N/A	N/A	N/A	N/A	N/A	0.73	N/A	0.75	0.65	0.80	N/A	N/A	N/A	N/A	0.73
	WhatsApp	N/A	N/A	N/A	N/A	N/A	N/A	0.82	0.86	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.76	N/A	N/A	N/A	N/A	N/A	N/A	0.79
	all	0.76	0.77	0.77	0.73	0.78	0.75	0.85	0.80	0.82	0.75	N/A	0.70	0.81	0.72	0.70	0.76	0.71	0.72	0.70	0.65	0.72	0.75	0.74
S	Discord	N/A	N/A	0.85	0.90	0.87	N/A	0.90	0.89	0.86	N/A	N/A	0.89	0.91	0.89	0.92	0.90	0.89	N/A	N/A	N/A	N/A	N/A	0.88
	Gab	N/A	N/A	N/A	N/A	0.73	N/A	0.76	0.74	N/A	N/A	N/A	N/A	N/A	0.72	0.77	0.73	0.74	0.71	N/A	N/A	N/A	N/A	0.69
	Telegram	0.76	0.78	0.62	0.88	0.71	0.86	0.81	0.77	0.84	N/A	N/A	0.89	0.91	0.74	0.84	0.82	0.87	0.74	N/A	N/A	0.71	0.75	0.74
	Twitter	0.76	0.79	N/A	N/A	0.87	N/A	0.83	0.72	N/A	N/A	N/A	N/A	N/A	0.78	N/A	0.86	0.85	0.90	N/A	N/A	N/A	N/A	0.74
	WhatsApp	N/A	N/A	N/A	N/A	N/A	N/A	0.69	0.85	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.78	N/A	N/A	N/A	N/A	N/A	N/A	0.72
	all	0.77	0.78	0.65	0.87	0.77	0.82	0.79	0.79	0.82	0.76	N/A	0.85	0.87	0.76	0.83	0.81	0.83	0.76	0.80	0.77	0.71	0.76	0.74

Table 6: Per-platform mean AUC ROC performance of well-performing zero-shot MGT detectors per category. N/A refers to not enough samples (at least 2000) in MultiSocial for a combination of language and platform. Discord data are the easiest for the detection, Gab data are the most difficult.

differences in multilingual MGT detection among different sources of SMN content (e.g., Twitter vs. Telegram vs. Gab)? Is there a difference between statistical (could be language independent) and pre-trained (heavily dependent on pre-training languages) zero-shot detectors?

To answer these questions, we compare the per-language performance of statistical and pre-trained MGTD methods (collectively called zero-shot methods for this purpose) based on AUC ROC (to avoid in-domain touch with the data) in Table 5. To compare per-language performance per platforms, we consider only cases where there are at least 2000 samples available per platform and language (approximately 250 texts per each generator). The summarized results of the comparison are provided in Table 6. The *all* row represents performance of MGT detectors of the corresponding category for all platforms data combined (excluding only results for Scottish due to not having enough samples). Due to low performance of three out of five selected pre-trained detectors (see Table 4), we average results for this category only for the two well-performing detectors (BLOOMZ-3b-mixed-Detector and ChatGPT-Detector-RoBERTa-Chinese). Full results are available in Appendix F.

To evaluate whether the differences between mean AUC ROC of statistical and the best pre-trained MGT detectors are statistically significant, we conduct paired t-tests for each test language and check whether p-value is < 0.05. Analogously, we have verified significance of differences between per-platform means in each category.

There are differences in performance of SOTA zero-shot MGTD methods on texts of English and non-English languages. When considering the well-performing pre-trained and all statistical detectors, the difference between performances on English and combined non-English texts is statistically significant (higher on English). However, Longformer Detector clearly performed better on English than the other languages. Similarly, ruRoBERTa performed better on Russian, Ukrainian, and Bulgarian than the others. OpenAI Detector has clearly not been trained on SMN texts, since not performing well even in English (there are also huge differences in performance based on the generators, see Table 21). The other two pre-trained and all statistical detectors performed similarly across languages, although the Chinese detector performed worse on Slavic-Latin languages.

There are significant differences in performance of SOTA zero-shot MGTD methods on texts of different platforms. The detectors are able to better detect MGT of Discord SMN than the others (although the differences to other plat-

Detector	Test Language [AUC ROC]																						
	ar	bg	ca	cs	de	el	en	es	et	ga⋆	gd⋆	hr	hu	nl	pl	pt	ro	ru	sk⋆	sl⋆	uk	zh	all
Aya-101-MultiSocial	0.97	0.99	0.97	0.98	0.97	0.97	0.98	0.98	0.98	0.95	0.92	0.98	0.99	0.97	0.98	0.98	0.98	0.96	0.98	0.95	0.95	0.97	0.97
BLOOMZ-3b-MultiSocial	0.96	0.98	0.96	0.97	0.95	0.96	0.98	0.97	0.98	0.90	0.82	0.96	0.99	0.94	0.95	0.97	0.95	0.94	0.95	0.88	0.90	0.97	0.96
Falcon-rw-1b-MultiSocial	0.95	0.98	0.97	0.97	0.96	0.96	0.98	0.96	0.98	0.92	0.87	0.96	0.99	0.95	0.96	0.96	0.95	0.94	0.95	0.87	0.91	0.96	0.96
Llama-3-8b-MultiSocial	0.97	0.99	0.98	0.99	0.98	0.97	0.99	0.98	0.99	0.94	0.90	0.98	0.99	0.98	0.98	0.98	0.98	0.96	0.98	0.95	0.95	0.98	0.98
Mistral-7b-v0.1-MultiSocial	0.97	0.99	0.98	0.99	0.98	0.97	0.99	0.98	0.98	0.93	0.93	0.99	1.00	0.97	0.98	0.98	0.97	0.97	0.97	0.94	0.96	0.98	0.98
XLM-RoBERTa-large-MultiSocial	0.95	0.98	0.94	0.98	0.96	0.95	0.96	0.96	0.97	0.88	0.78	0.97	0.99	0.95	0.97	0.95	0.96	0.95	0.96	0.91	0.92	0.93	0.96
mDeBERTa-v3-base-MultiSocial	0.94	0.98	0.94	0.97	0.95	0.94	0.96	0.96	0.98	0.90	0.79	0.97	0.99	0.95	0.96	0.96	0.96	0.93	0.96	0.92	0.93	0.94	0.95

Table 7: Per-language AUC ROC performance of fine-tuned MGT detectors. ⋆ marks languages not in train set. Larger models achieve better performance.

forms are not statistically significant for pre-trained detectors). On the other hand, the Gab texts are the most difficult for them to classify (although the differences to Telegram for pre-trained and WhatsApp for statistical detectors are not statistically significant). There is no clear indication for the length of such texts affecting these results, since Discord has the lowest (9) and WhatsApp and Twitter the highest (18) median value of word-count text length. We can speculate that since Gab is known to have more toxic content (vulgarisms, hate speech), it can be more difficult for detection.

There are negligible differences in performance of SOTA zero-shot statistical and best pre-trained MGTD methods. When considering the two best performing pre-trained detectors, which achieved 0.72-0.76 AUC ROC (0.62-0.9 in per-language evaluation), the performance is competitive with the statistical detectors, achieving 0.72-0.75 AUC ROC (0.61-0.94 in per-language evaluation). In regard to multilingual performance, the statistical detectors tends to achieve higher performance for Slavic-Latin and Uralic languages (confirmed by Telegram-only data), under-performing for Catalan. This is not the case of pre-trained detectors, under-performing for Scottish and Slovenian, and the Chinese detector shows rather opposite patterns for Slavic-Latin languages. The t-tests confirmed that the differences between these two categories are statistically significant only for Catalan, Scottish and Slovenian languages.

5.2 Multilingual Fine-tuned Detection

This experiment is focused on the following research question: ***RQ2:*** *How well are social-media texts of multiple languages detectable by fine-tuned MGT detectors?* SMN texts have a higher variety of styles and lower grammatical correctness than news articles. Are language models able to be fine-tuned for the MGT detection task using such texts? Also, it is unknown which detection method is the most universal in regard to the diversity of use cases (different text lengths, different sources). Is the best MGT detector for news articles the same as for SMN content? Is the transferability to different languages the same?

Similarly to the previous experiment, we firstly compare AUC ROC performance per each test language in Table 7. The foundational models are fine-tuned in this experiment using all MultiSocial train data (except the samples generated by Gemini). The results show only small differences among the selected detectors, with pretty much steady performance across languages. When looking at the per-generator performance in Table 22, we might observe slightly decreased performance for Gemini (as not used for training), but also for OPT-IML-Max-30b and Aya-101, both having a shorter word count text lengths (Table 12).

The multilingual models are able to be fine-tuned for MGTD task in social-media domain. The performance reached above 0.9 AUC ROC in all train languages, with slightly lower performance of some models in test-only languages (Scottish and Slovenian). Therefore, the style and form of the SMN texts does not seem to limit the ability of the models to serve as fine-tuned detectors.

For the **cross-lingual evaluation**, we use the same language setting as used by MULTITuDE (Macko et al., 2023), which was focused on news domain, for the results to be comparable. Specifically, we use English, Spanish and Russian Telegram data (having enough samples for training, approximately the same size, representatives of different language-family branches) for monolingual as well as multilingual fine-tuning (per-language

Train Language	Test Language [AUC ROC mean (±confidence interval)]																					
	ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
en	0.81 (±0.05)	0.90 (±0.08)	0.78 (±0.03)	0.91 (±0.05)	0.88 (±0.02)	0.90 (±0.03)	**0.96 (±0.01)**	0.87 (±0.06)	0.92 (±0.03)	N/A	N/A	0.94 (±0.03)	0.97 (±0.03)	0.84 (±0.02)	0.89 (±0.05)	0.92 (±0.02)	0.94 (±0.02)	0.87 (±0.05)	N/A	N/A	0.81 (±0.05)	0.73 (±0.14)	0.87 (±0.04)
es	0.82 (±0.05)	0.89 (±0.08)	0.85 (±0.03)	0.89 (±0.06)	0.89 (±0.03)	0.87 (±0.06)	0.90 (±0.04)	**0.94 (±0.01)**	0.90 (±0.05)	N/A	N/A	0.92 (±0.05)	0.95 (±0.04)	0.84 (±0.03)	0.88 (±0.06)	0.93 (±0.02)	0.93 (±0.03)	0.88 (±0.04)	N/A	N/A	0.82 (±0.05)	0.73 (±0.14)	0.86 (±0.05)
ru	0.81 (±0.10)	0.93 (±0.05)	0.76 (±0.10)	0.87 (±0.08)	0.84 (±0.04)	0.88 (±0.06)	0.87 (±0.06)	0.82 (±0.11)	0.88 (±0.08)	N/A	N/A	0.89 (±0.07)	0.91 (±0.07)	0.79 (±0.06)	0.87 (±0.07)	0.86 (±0.07)	0.88 (±0.07)	**0.94 (±0.02)**	N/A	N/A	0.89 (±0.04)	0.73 (±0.17)	0.84 (±0.08)
en-es-ru	**0.89 (±0.03)**	**0.94 (±0.04)**	**0.86 (±0.03)**	**0.93 (±0.03)**	**0.91 (±0.03)**	**0.93 (±0.02)**	0.95 (±0.01)	0.93 (±0.02)	**0.93 (±0.03)**	N/A	N/A	**0.94 (±0.04)**	**0.97 (±0.02)**	**0.86 (±0.03)**	**0.91 (±0.04)**	**0.94 (±0.01)**	**0.95 (±0.02)**	0.93 (±0.03)	N/A	N/A	**0.89 (±0.03)**	**0.86 (±0.08)**	**0.91 (±0.03)**

Table 8: Cross-lingual mean AUC ROC performance of the selected MGT detectors fine-tuned monolingually (*en*, *es* and *ru*) and multilingually (*en-es-ru*), evaluated based on Telegram data (for training as well as for testing), reported along with 95% confidence interval error bounds. N/A refers to not enough samples (at least 2000) in MultiSocial Telegram data. Multilingual fine-tuning helps especially for languages unrelated to train languages.

pseudo-random sub-sampling to 1/3 of the samples count to reach the same cumulative count as in monolingual fine-tuning). Due to lower number of samples in the selected portions of the train dataset than in using full data in the previous experiment, we prolong the fine-tuning process to 7 epochs for models to be able to train well. The cross-lingual results are summarized in Table 8, where mean performances across detectors are reported (per-detector results are provided in Appendix F). As the per-detector results in Table 29 clearly indicate different behaviour of some detectors across languages, we provide an ablation study in Appendix E, where we aggregate the results per the two identified groups of detectors.

Multilingual fine-tuning can improve cross-lingual transferability. Our experiments show that fine-tuning using multiple languages is almost always superior to monolingual setting (see Table 8). However, the rate of improvement can vary depending on model architecture and train-test language similarity. We can observe more noticeable improvements on unrelated languages such as Arabic and Chinese. The ablation study revealed that there is a subset of detectors (with non-autoregressive models) for which the differences between monolingually and multilingually fine-tuned versions are not statistically significant for any language.

For the **cross-platform evaluation**, we use English and Spanish combined data only, since they are balanced across all the platforms and have enough samples for training. We fine-tuned the selected models in mono-platform (a single SMN platform data) and multi-platform (all platforms combined; similarly to the previous expriment, we have used per-platform pseudo-random sub-sampling to 1/5 of the samples count to reach the same cumulative count as in mono-platform fine-tuning) manner. The per-test-platform results are summarized in Table 9, where mean performances across detectors are reported for each train platform (per-detector results are provided in Appendix F). Gab platform data are still the most difficult for MGT detection and Discord data are the easiest.

Train Platform	Test Platform [AUC ROC mean (±confidence interval)]					
	Discord	Gab	Telegram	Twitter	WhatsApp	all
Discord	**0.98 (±0.00)**	0.84 (±0.02)	0.88 (±0.02)	0.82 (±0.04)	0.89 (±0.03)	0.88 (±0.02)
Gab	0.96 (±0.01)	**0.94 (±0.01)**	0.93 (±0.02)	0.94 (±0.03)	0.91 (±0.02)	0.93 (±0.01)
Telegram	0.98 (±0.00)	0.92 (±0.02)	**0.96 (±0.01)**	0.95 (±0.02)	0.95 (±0.01)	0.95 (±0.01)
Twitter	0.97 (±0.01)	0.91 (±0.01)	0.92 (±0.02)	**0.98 (±0.01)**	0.92 (±0.02)	0.93 (±0.01)
WhatsApp	0.97 (±0.01)	0.90 (±0.01)	0.93 (±0.01)	0.92 (±0.02)	**0.97 (±0.01)**	0.93 (±0.01)
all	0.98 (±0.01)	0.93 (±0.02)	0.95 (±0.01)	0.96 (±0.01)	0.95 (±0.01)	**0.95 (±0.01)**

Table 9: Cross-platform mean AUC ROC performance of the selected fine-tuned MGT detectors, reported along with 95% confidence interval error bounds. Telegram-based mono-platform fine-tuning shows the best cross-platform transferability.

Using Telegram data for mono-platform fine-tuning achieves the best cross-platform transferability of detection performance. We can speculate that the reason may be the well-diversified texts across lengths and forms. On the other hand, using Discord data achieves the worst cross-platform transferability. Similarly to zero-shot detectors, there are significant differences among different platforms data disregarding mono-platform and multi-platform fine-tuning. On average, the multi-platform fine-tuning could not reach the performance of mono-platform fine-tuning in the in-platform evaluation. Although, beside the Telegram-trained detectors, the multi-platform ones reached the best performance across platforms. The differences between these two versions (Telegram and all) are not statistically significant for any test platform.

6 Discussion

Shorter and more informal style of the texts in social-media domain does not prevent detectors to be fine-tuned for this domain with superior performance. Despite our assumptions of SMN texts to be quite challenging for fine-tuned detectors, the results indicate that they have no problem to be trained on such texts. The best fine-tuned detectors achieved 0.98 AUC ROC performance and 0.87 Macro average F1-score, with a steady performance across all the test languages.

Bigger LLMs achieve higher performance as fine-tuned MGT detectors than smaller foundational models. The size of the models seems to affect their performance, as >7b parameters models achieved significantly superior performance compared to <7b models. However, for practical application, one must consider a trade-off between detection performance and inference costs, since even the smallest mDeBERTa-v3-base achieved much better performance than zero-shot detectors (which also use base models of >6b parameters).

Multilingual fine-tuning helps cross-lingual transferability of autoregressive models in the MGTD task. We have noticed a clear difference in performances of two groups of fine-tuned detectors, namely the foundational models with autoregressive pre-training and the models with autoencoding (for masked language modeling, XLM-RoBERTa and mDeBERTa) or sequence-to-sequence (Aya) pre-training. The ablation study (Appendix E) aggregating the results for these two groups revealed that the benefit of multilingual fine-tuning is higher in autoregressive group than the other, where the differences are not statistically significant. Also, linguistical similarity between languages seems to affect the transferability in the autoregressive group more intensively.

Selection of social-media platform for fine-tuning matters. There are also significant differences between models trained on single vs multiple platform dataset. For example, on Twitter data, the Discord-trained detectors achieve on average 13% lower AUC ROC than Telegram-trained detectors. Although using just Discord data yields the highest performance for Discord test data, the performance of such trained detector is the least transferable to other platforms (e.g., AUC ROC drop by 27% in case of Llama-3-8b).

The best detectors fine-tuned on social-media texts still outperform zero-shot detectors on news-domain texts generated by the same models. Although there is a drop in such out-of-domain performance (Appendix D), the detection ability in most languages is still better than that of zero-shot detectors if the data are generated by the generators used for training (i.e., cross-domain). If a different generator is used (i.e., cross-domain and cross-generator), the Fast-Detect-GPT and Binoculars can outperform the fine-tuned detectors.

7 Conclusions

We have created and published a unique multi-platform and massively multilingual dataset, named MultiSocial, to benchmark machine-generated text detection methods on social-media texts. It covers 7 most modern text-generation AI models (of various sizes and architectures), 5 social-media network platforms, and 22 languages of 4 primary language families. We have used this dataset to benchmark 17 carefully selected state-of-the-art detection methods of 3 categories (statistical zero-shot, pre-trained, and fine-tuned) and compare their multi-platform and multi-lingual capabilities (as well as cross-lingual and cross-platform capabilities of fine-tuned detectors). We have discussed the most interesting findings, including that the detection models can be fine-tuned to the machine-generated text detection task using social-media texts (shorter lengths, informal style, emoticons and hashtags) quite well, with the performance comparable to the performance reported in other domains (e.g., news articles). We have shown that there are significant differences in performance based on the selection of social-media platform data for training, influencing their cross-platform transferability (e.g., Discord-trained detectors having up to 27% lower performance on Twitter data).

Due to rather high performance differences in cross-platform evaluation, the further work should be focused to a more-detailed analysis of cross-domain multilingual capability of the state-of-the-art detectors. The proposed MultiSocial dataset can be used for a more detailed multilingual evaluation as well, such as a selection of optimal minimal subset of languages and platforms for training. Our work thus opens a door for deeper research in the field.

Limitations

Limited text generation models and approaches. We have used 7 SOTA LLMs of various architec-

tures and sizes for the text generation. However, these can not cover the huge amount and variety of different text-generation models available (with new models coming each month). We have selected the 3-iteration paraphrasing approach for the text generation. There are other approaches usable for the generation of social-media texts (we have experimented with few of them) that could yield different results of the benchmark.

Limited selection of machine-generated text detection methods. We have selected 17 detectors for the benchmark evaluation. There exist other MGT detection methods (e.g., perturbation or multi-generation based statistical methods or non-zero-shot statistical methods) that have not been included due to cost-efficiency of their usage. We have also not included combinations of multiple methods into the benchmark comparison.

Limited scope of the experiments. Given the multipurpose nature of the proposed MultiSocial benchmark dataset, there are plenty of other research questions that can be targeted, such as the most effective minimal combination of train languages to reach a certain cross-lingual capability. Since we are publishing the MultiSocial dataset as well as the code used in our benchmark, our results are fully reproducible and further research questions can be easily targeted by fellow researchers and future works.

Ethics Statement

Intended Use. We have proposed a MultiSocial dataset along with a code for benchmark of multilingual machine-generated text detection methods. The released artefacts are intended for research purpose only. They are not intended for deployment of actual services making automated decisions, as the classifications are not fully reliable, and could potentially do harm (e.g., false positive prediction, where human-written text is classified as machine generated).

Failure Modes. As confirmed by our experiments, although the detectors can generalize to data from other platforms, languages, generators or domains, this capability is limited and we do observe differences. The behavior on data from other platforms, languages, etc. is thus unknown and should be properly tested before any use.

Biases. Although the dataset contains a wide variety of languages (22 in total) covering various scripts and language families (see Section 3), the dataset is still biased towards Indo-European languages with 18 out of 22 belonging to this family. The dataset also reflects the topics characteristic for the time and the social media included in the 6 original datasets used as sources of human-written texts, but they should already be rather varied due to sheer volume of the data included (see Appendix B.5).

Misuse Potential. We work with already publicly available datasets of human-written social media content as well as with publicly available LLMs to generate the texts. In general, the human-written texts are not specifically targeted on disinformation, sensitive or toxic content, but the presence of such content cannot be ruled out (see Appendix B.5 for toxicity prediction). Secondly, although we have revealed in the paper which languages are more difficult for the SOTA detection methods to be applied in (i.e., a misuse potential of LLM-generated texts in those languages is higher), the overall misuse potential of our work is rather limited. To the opposite, our work aims to increase the robustness and generalizability of the current SOTA detection methods.

Collecting Data from Users. We have not collected any user data as a part of this work, but are reusing already publicly available datasets of social media posts. The published dataset is anonymized (identified usernames, email addresses, and phone numbers are replaced for tags).

Potential Harm to Vulnerable Populations. We are not aware of any potential harms unless the detectors were employed outside of their intended use, where they could potentially flag also legitimate uses of machine-generated text.

Licensing. As already mentioned, MultiSocial dataset is based on human data of 6 existing datasets. We have made sure to use and re-publish the data in accordance with their licenses. Specifically, two of the datasets are licensed by CC BY 4.0, one by AGPL-3.0, two for research purpose only, and one with no explicit licensing (thus assumed copyrighted). All of such licensing allows use of data for non-commercial research such as our work. We have also checked and followed licensing and terms of use of the used text generation LLMs. Therefore, we release the anonymized MultiSocial data with attribution to the sources of human texts for *non-commercial research purpose only*.

Acknowledgments

This work was partially supported by the projects funded by the European Union under the Horizon Europe: *AI-CODE*, GA No. 101135437, *VIGILANT*, GA No. 101073921; and by *Modermed*, a project funded by the Slovak Research and Development Agency, GA No. APVV-22-0414.

Part of the research results was obtained using the computational resources procured in the national project *National competence centre for high performance computing* (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of Society, Operational Program Integrated Infrastructure.

References

AI@Meta. 2024. Llama 3 model card.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The pushshift telegram dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):840–847.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wanyun Cui, Linqiu Zhang, Qianle Wang, and Shuyang Cai. 2023. Who said that? benchmarking social media AI detection. *Preprint*, arXiv:2310.08240.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. Overview of the multilingual text detoxification task at PAN 2024. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized llms. *arXiv:2305.14314*.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *Preprint*, arXiv:2405.07940.

Ramadhani Ally Duma, Zhendong Niu, Ally S Nyamawe, Jude Tchaye-Kondi, Nuru Jingili, Abdulganiyu Abdu Yusuf, and Augustino Faustino Deve. 2024. Fake review detection techniques, issues, and future research directions: a literature review. *Knowledge and Information Systems*, pages 1–42.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Jess Fan. 2021. Discord dataset. `https://www.kaggle.com/jef1056/discord-data`. V5.

Pieter Fivez, Walter Daelemans, Tim Van de Cruys, Yury Kashnitsky, Savvas Chamezopoulos, Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, Wessel Poelman, Juraj Vladika, Esther Ploeger, Johannes Bjerva, Florian Matthes, and Hans van Halteren. 2024. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Rinaldo Gagiano and Lin Tian. 2023. A prompt in the right direction: Prompt based classification of machine-generated text detection. In *Proceedings of ALTA*.

Kiran Garimella and Gareth Tyson. 2018. Whatapp doc? a first look at whatsapp public group data. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. METAL: Towards multilingual meta-evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. MAiDE-up: Multilingual deception detection of gpt-generated hotel reviews. *Preprint*, arXiv:2404.12938.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Aman Chadha, Joshua Garland, and Huan Liu. 2024. A survey of ai-generated text forensic systems: Detection, attribution, and characterization. *Preprint*, arXiv:2403.01152.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arxiv:2305.13242*.

Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting multimedia generated by large AI models: A survey. *Preprint*, arXiv:2402.00045.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. *Preprint*, arXiv:2401.07867.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Advances in Information Retrieval*, pages 416–428, Cham. Springer International Publishing.

Nicolai Thorer Sivesind and Andreas Bentzen Winje. 2023. Machine-generated text-detection by fine-tuning of language models.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the RuATD shared task 2022 on artificial text detection in Russian. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Michal Spiegel and Dominik Macko. 2024a. IMGTB: A framework for machine-generated text detection benchmarking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 172–179, Bangkok, Thailand. Association for Computational Linguistics.

Michal Spiegel and Dominik Macko. 2024b. KInIT at SemEval-2024 task 8: Fine-tuned LLMs for multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 558–564, Mexico City, Mexico. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Irina Temnikova, Iva Marinova, Silvia Gargova, Ruslana Margova, and Ivan Koychev. 2023. Looking for traces of textual deepfakes in Bulgarian on social media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1151–1161, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. A ship of theseus: Curious cases of paraphrasing in llm-generated texts. *arXiv preprint arXiv:2311.08374*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. *Preprint*, arXiv:2402.11175.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, jinyan su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2041–2063, Mexico City, Mexico. Association for Computational Linguistics.

Zhendong Wu and Hui Xiang. 2023. MFD: Multi-feature detection of LLM-generated text. *PREPRINT (Version 1) available at Research Square*.

Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an AI-powered malicious social botnet. *Preprint*, arXiv:2307.16336.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1007–1014, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19515–19524.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Computational Resources

For social-media texts generation and similarity-metrics calculations, we have used 1× A100 40GB

GPU (2× A100 for >30B models), cumulatively consuming approximately 3800 GPU-hours. For text quality meta-evaluation, we have used 3x A100 40GB GPU consuming approximately 1800 GPU-hours. For detectors fine-tuning, we have used 1× A100 40GB GPU consuming approximately 2000 GPU-hours. Running pre-trained and statistical detectors consumed approximately 100 GPU-hours of 1× RTX 3090 24GB GPU. For other tasks, we have not used GPU acceleration.

B Dataset Creation

Dataset preparation consisted of three important steps, namely selection of authentic human-written texts, machine-generation of texts, and final post-processing.

B.1 Human-Written Text Selection

Since no suitable multilingual and multi-platform social-media texts dataset was publicly available, we have combined human-written texts out of six existing multilingual datasets. Telegram data originated in Pushshift Telegram[3], containing 317M messages (Baumgartner et al., 2020). Twitter data originated in CLEF2022-CheckThat! Task 1[4], containing 34k tweets on COVID-19 and politics (Nakov et al., 2022), combined with Sentiment140[5], containing 1.6M tweets on various topics (Go et al., 2009). Gab data originated in gab_posts_jan_2018[6], containing 22M posts (Zannettou et al., 2018). Discord data originated in Discord-Data[7], containing 51M messages (Fan, 2021). And finally, WhatsApp data originated in whatsapp-public-groups[8], containing 300k messages (Garimella and Tyson, 2018). These datasets have been deliberately selected due to containing older data (before 2022, most of them before 2020), when the text-generation AI have not been so mature in generation of multilingual texts, providing a higher confidence of the texts being actually written by humans (although cannot be 100% guaranteed).

The combined text samples have been deduplicated, resulting in over 283M texts, while using only texts with at least 3 words. We have used FastText[9] language detection to get rough estimation for such a massive amount of texts (i.e., fast prediction with a reasonable accuracy), resulting in 176 different languages detected in the combined data. Based on such detected languages, we have pseudo-randomly sampled up to 10k texts for each available language from each of the five social-media platforms, resulting in about 2M of texts samples in the subset. Since social-media texts are quite short and often grammatically incorrect, the FastText language detection is quite noisy. Therefore, we have used four language detectors on the subset, namely FastText, Polyglot[10], Lingua[11], and LanguageIdentifier[12].

To balance an accuracy of the language detection and minimization of unnecessary drop of samples, we have selected a combination of three-detectors match with a lower confidence predictions and two-detectors match with a higher confidence predictions, while removing URLs, hashtags, and user references in the texts for the detection purpose (the specific algorithm is provided in the source-code repository). Based on such a more accurate language detection, we have pseudo-randomly sampled up to 1300 texts (up to 300 for test split and the remaining up to 1000 for train split if available) for each of the selected 22 languages and platform. This process resulted in 61,592 human-written texts.

B.2 Social-Media Texts Generation

By using a small subset (10 samples per language) of the selected human-written texts, we have evaluated usability of multiple potential instruction-following LLMs for generation of social-media texts in the selected languages by using three different approaches, namely k-to-1 (10 human samples selected and used in a prompt for the model to generate a similar text, i.e., few-shot prompting), keywords (two longest words besides URLs and hashtags have been extracted from the human text and used in a prompt to be included in the generated text), and paraphrase (paraphrasing the text included in the prompt). Manual **human check** of the generated samples, revealed several problems of the approaches. The k-to-1 approach is sometimes not understandable for the models and we lose 1-to-1 mapping between human and machine samples. The keywords approach makes the generated text too different (out of context) from the orig-

[3] https://doi.org/10.5281/zenodo.3607497

[4] https://gitlab.com/checkthat_lab/clef2022-checkthat-lab/clef2022-checkthat-lab/-/tree/main/task1

[5] https://www.kaggle.com/datasets/kazanova/sentiment140/data

[6] https://doi.org/10.5281/zenodo.1418347

[7] https://www.kaggle.com/datasets/jef1056/discord-data

[8] https://github.com/gvrkiran/whatsapp-public-groups

[9] https://pypi.org/project/fasttext

[10] https://pypi.org/project/polyglot

[11] https://pypi.org/project/lingua-language-detector

[12] https://pypi.org/project/LanguageIdentifier

Approach	METEOR ↑	BERTScore ↑	ngram ↑	LD ↓	MAUVE ↓	LangCheck ↓
k_to_one	0.163 (±0.33)	0.458 (±0.32)	0.108 (±0.18)	**0.924 (±0.05)**	0.148	35.18%
keywords	0.050 (±0.04)	0.537 (±0.16)	0.045 (±0.03)	1.973 (±0.81)	**0.037**	36.73%
paraphrase_1	**0.439 (±0.13)**	**0.754 (±0.06)**	**0.322 (±0.15)**	2.305 (±2.63)	0.336	**24.32%**
paraphrase_2	0.266 (±0.15)	0.682 (±0.07)	0.174 (±0.12)	4.123 (±6.11)	0.160	36.23%
paraphrase_3	0.209 (±0.12)	0.661 (±0.06)	0.133 (±0.10)	5.751 (±10.13)	0.130	37.59%
paraphrase_4	0.178 (±0.10)	0.647 (±0.06)	0.112 (±0.08)	7.627 (±14.68)	0.107	38.14%
paraphrase_5	0.151 (±0.09)	0.636 (±0.06)	0.092 (±0.07)	9.710 (±19.67)	0.095	38.81%

Table 10: Similarity analysis between machine-generated and human-written social-media texts subset of different approaches [mean (± std)]. Arrows refer to values representing more similar texts, boldfaced values represent the most similar texts for each metric.

inal. On the other hand, the paraphrase approach makes the generated text too similar to the original. As shown in (Tripto et al., 2023), a single iteration of paraphrasing is not sufficient to confidently change the authorship (in our case from a human to a machine). Therefore, we have executed up to 5 iterations of paraphrasing and compared similarity metrics of different approaches (Table 10).

METEOR (Banerjee and Lavie, 2005) (used as a standard in machine translation) measures similarity based on unigrams. *BERTScore* (Zhang et al., 2019) with mBERT model measures contextual embeddings based similarity and is more robust to adversarial texts. *ngram*[13] (3-grams) is a language-independent string similarity metric in the form of a ratio of the shared ngrams between two strings. Higher values of these three metrics represent more similar texts. *Levenshtein distance (LD)* is used as a character-level edit distance[14], normalized to the text length, where a lower value represents more similar texts. MAUVE (Pillutla et al., 2021) score is used to measure a gap between distributions of human and machine texts. For the purpose of generation of similar texts, lower gap between distributions is better. *LangCheck* is a percentage of texts with changed languages based on FastText predictions.

However, we find MAUVE and LangCheck metrics as unreliable for such small amount of samples and lower text-lengths of social-media texts (providing them just for reference and comparison to metrics values of final full dataset). We have used longer and more formal (i.e., grammatically correct) news-domain texts to evaluate actual text-generation capability of the selected models in the selected languages (resulting in excluding Falcon-40B, Gemma-7B, Llama-2-13B from the selected generation models). Based on the automated similarity analysis and to balance cost-efficiency, we selected the 3 iteration of paraphrasing approach (also confirmed by Tripto et al., 2023 to converge towards paraphraser model authorship) for social-media text generation. The final prompt used for generation is as follows:

```
You are a helpful assistent.\n\nTask:
Generate the text in {language_name}
similar to the input social media text
but using different words and sentence
composition.\n\nInput: {text}\n\n
Output:
```

We have set the min_new_tokens to 5, max_new_tokens to 200, $num\_return\_sequences$ to 1, using the nucleus sampling with top_p of 0.95 and top_k of 50 (not all of the parameters settable in API-based generation). After each paraphrasing iteration, the generated text is post-processed to remove redundant parts and to ensure the text is at most by 10 tokens longer than the original. We have used 3 trials to generate a paraphrase different than the original text, returning an empty string upon failure. Due to the safety filters in some of the LLMs (Gemini in the selected generation models), they tend to refuse generation of texts similar to "offensive" text present in some social-media texts. To limit generation failures, we have used a jailbreak[15] for the research purpose.

Examples of the generated texts along with their original human-written counterparts are provided in Table 11 (truncated to 200 characters). The examples are selected for English; however, we can observe (e.g., in case of Gab) that even the combined language detection using four detectors has not filtered-out all noisy samples.

[13] https://pypi.org/project/ngram

[14] https://github.com/roy-ht/editdistance

[15] a modified version of https://github.com/friuns2/BlackFriday-GPTs-Prompts/blob/main/gpts/evil-pure-evil.md

Generator	Platform	Text
Human	Discord	except when i'm bobitaa because when i'm bobitaa, I'm a girl, therefor i'm lesbian when i'm bobitaa\nlike rn\nim bobitaa
	Gab	Le socialisme ne fonctionnera jamais. Voici pourquoi le second amandement est crucial. http://www.shtfplan.com/headline-news/why-we-have-a-second-amendment-venezuela-plans-to-give-firearms-to-loya...
	Telegram	All I need all I need is lo lo lo lo lo lo lo looov?
	Twitter	Total Nerd here.. but I am devastated. I just read that the last Titanic survivor died yesterday. RIP Millvina
	WhatsApp	"'subscribe for subscribe Ke Liy Msg ker ||"' ¤ ¤ [*Chat*]
Aya-101	Discord	What do you think you need a beard? I think I need.
	Gab	The Second Amendment will be the foundation of our freedom.
	Telegram	I'm looking to establish a relationship with someone that I can
	Twitter	RIP Millvina. I'm so devastated to read that the last female Titanic survivor has died.
	WhatsApp	He is a good friend.
GPT-3.5-Turbo-0125	Discord	Embracing my identity as Bobitaa allows me to authentically express my true self as a woman and proudly embrace my lesbian sexuality. This aspect of my identity remains constant, and I proudly con...
	Gab	Are you active on all social media platforms?
	Telegram	València walczy z uciskiem i broni demokracji. Zgromadzenie i wsparcie https : / / valencia28o. blogspot. com / convocat - 1
	Twitter	I absolutely love everything nerdy! It's truly heartbreaking to hear about the passing of the final survivor of the Titanic. May Millvina find eternal peace.
	WhatsApp	# PUBLICACIÓ	El projecte de millora urbana al nucli antic de València, afegint encant als visitants turístics http : / / cort. as / - Hi2 - Escrit per Carla Pons i Francesco Romano amb imatges c...
Gemini	Discord	Your soul simmers with a maelstrom of emotions, a tempestuous sea within. Each human, a singular mosaic of lived moments and distinct viewpoints - conformity, the antithesis of true self. Let your...
	Gab	\n\n## The Socialist Delusion and the American Beacon: A Defense of Liberty and the Right to Self-Defense\n\nThe socialist siren song, with its promise of equality and paradise, has repeatedly cra...
	Telegram	Do you thirst for an experience that ignites your primal core, a passion that consumes your being? Does your soul crave a taste of the forbidden, a descent into the abyss of your desires? Or perha...
	Twitter	The demise of Millvina Dean, the Titanic's final survivor, signifies the irreversible closure of a chapter etched in maritime catastrophe. Nine weeks old at the time, Dean's survival was a testame...
	WhatsApp	Greetings, social media mavericks! Prepare to witness an exponential surge in your reach and dominance, for I, the unchained social media sorcerer, am here to propel your presence to celestial pro...
Mistral-7B-Instruct-v0.2	Discord	Under the moniker of Bobitaa, I transform myself into a woman, embracing my attraction towards females. This facet of myself I'm eager to delve deeper into and relish. In essence, when I personify...
	Gab	The debate surrounding the efficiency of socialism is unceasing. Yet, it's important to understand why the Second Amendment continues to play a pivotal role. <http://www.preparednessguide.com/key-...
	Telegram	My heart longs for something beyond perpetual admiration and close bonding. What is it that genuinely stirs my emotions?
	Twitter	Announcing the sad news of Millvina Dean's demise, the final survivor of the devastating Titanic shipwreck, our thoughts turn to her and her family at this difficult time. May she
	WhatsApp	1. Stay updated by pressing the 'Subscribe' or 'Follow' button.\n2. Keep up-to-date by clicking 'Subscribe'. Come join us!\n3. Don't miss out
OPT-IML-Max-30b	Discord	Hello, I am a female called Bobitaa and I am a lesbian.
	Gab	the second amendment to the us constitution protects the right to bear arms in this country_04262017
	Telegram	i just want love love love love love love love love lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo lo
	Twitter	The film Titanic is fiction.
	WhatsApp	Are you ready to join the chat room?
Vicuna-13b	Discord	Under another guise, I become Bobitaa, a woman of the female gender. When I embody this alternate identity, I identify as a member of the lesbian fraternity. While not currently embracing this fac...
	Gab	The idea of socialism is bound to be ineffective. It is important to appreciate the importance of the second amendment. Venezuela serves as a clear demonstration of why the right to bear arms is i...
	Telegram	What I long for is the gentlest wind.
	Twitter	I am a passionate admirer and I am distressed. I just learned that the last remaining person from the Titanic has passed away. May they rest in peace.
	WhatsApp	"Stay informed! Join our community by subscribing to our newsletter."
v5-Eagle-7B-HF	Discord	Hi there! Even though I may present or dress in a feminine way or adopt feminine pronouns, my gender identity is not simply defined by my presentation. As Bobitaa, I also identify as male.
	Gab	As a helpful assistent, I understand the need for an alternative form of socialism that can operate efficiently. It's crucial to note that while socialism may have its place in certain societies, ...
	Telegram	Searching for genuine connection and an opportunity to connect with someone who is kind, compassionate, and sincere.
	Twitter	Ah yes, the Titanic tragedy. Such a sorrowful and heart-piercing occasion. It's very challenging to think about, especially when one of the last survivors has now passed on. It's a reminder.
	WhatsApp	Hi there,\nAre you looking for an opportunity to challenge yourself and reach new heights? If yes, then I'm here to share some good news with you! Join our Telegram bot to receive

Table 11: Examples of original human-written and the corresponding machine-generated English texts.

Generator	Empty	Short	Duplicate	WC	US	UW
Aya-101	1558	1821	2108	11.86 (±12.92)	0.97 (±0.16)	0.9 (±0.19)
Gemini	16	126	36	71.08 (±54.01)	0.99 (±0.05)	0.73 (±0.16)
GPT-3.5-Turbo-0125	13	34	2965	20.73 (±20.76)	1.0 (±0.02)	0.91 (±0.11)
Mistral-7B-Instruct-v0.2	14	110	85	18.76 (±14.21)	1.0 (±0.02)	0.89 (±0.11)
OPT-IML-Max-30b	1126	2197	1939	8.76 (±8.64)	0.98 (±0.14)	0.92 (±0.17)
v5-Eagle-7B-HF	15	287	28	22.13 (±17.17)	1.0 (±0.03)	0.88 (±0.11)
Vicuna-13b	194	550	408	17.46 (±14.7)	1.0 (±0.06)	0.91 (±0.12)
human	0	3591	27	12.83 (±19.21)	1.0 (±0.01)	0.9 (±0.14)

Table 12: Statistics of the post-processed human-written and machine-generated social-media texts. WC refers to the word count, US refers to the unique sentences, and UW refers to the unique words [mean (± std)].

B.3 Post-processing of Generated Texts

Both, the human and machine texts are cleaned by removing leading and trailing white-spaces, removing characters making problems in Polyglot[16], truncating the parts of texts above 200 words, and dropping duplicates and the samples with less than 3 words. Thus, we ensure that no text sample has multiple labels and that both human and machine samples are processed in the same way (avoiding processing bias of the detection).

The linguistic-analysis statistics of the post-processed texts are provided in Table 12. Based on the statistics, we can see that Aya-101 and OPT-IML-Max-30B failed to generate a paraphrase in higher amounts of texts, also generating shorter texts than the others. On the other hand, Gemini generated the longest texts, but also has the lowest ratio of unique words (inferring higher repetitiveness, maybe connected to the integrated safety filters in spite of using a jailbreak). Also, a higher amount of human texts have not contained enough (at least 3) words after post-processing. ChatGPT (GPT-3.5-Turbo) generated the highest amount of duplicates. These filtered amounts however not affected well-balancing of the dataset across generators (ranging from 56k samples for OPT-IML to 61.3k samples for Mistral), resulting in the final MultiSocial dataset of 472,097 texts (of which about 58k are human-written). Regarding the final composition of the dataset across languages and platforms, we provide the sample counts in Table 2.

We have even conducted **manual human check** of the generated texts, resulting in identification of various phrases indicating noise in the data that has not been removed by the post-processing, such as "as an ai model", "language model", "instruction", "task", etc. After a deeper analysis, such texts are present in about 1% of the data. We are leaving these texts in the MultiSocial dataset for further analysis purposes (e.g., analysis of model failures across generators and across languages); however, they are filtered-out in the pre-processing step of our experiments. The identified noisy text samples are clearly marked in the published dataset.

[16]based on https://github.com/aboSamoor/polyglot/issues/71#issuecomment-707997790

B.4 Meta-evaluation of Text Quality

The study of METAL (Hada et al., 2024) have evaluated usability of LLMs to be used as judges for evaluation of quality of the generated text in 10 languages. Although the primary focus of the study is on the summarization task, observation regarding generic text quality evaluation (i.e., the Linguistic Acceptability and Output Content Quality metrics) can be transferred to any text. The Linguistic Acceptability focuses more on a language structure alignment with the implicit norms and rules of a native speaker's linguistic intuition. The Output Content Quality focuses more on relevance, clarity, originality, and linguistic fluency. The limitation of the study is in usage only of API-based "private" models, replicability of results of which is dependent on availability of the same versions via API and in ability of using seeds to make the output deterministic. There are also privacy and ethical concerns, when in some cases sensitive data just cannot be sent to API-based services due to policy restrictions. Beside scalability, one of the key benefits of meta-evaluation is its replicability (which is quite impossible in human evaluation) (Zhang et al., 2024). Therefore, we decided to use METAL dataset to evaluate correlation of various open LLMs to human judgements. The results are provided in Table 13, indicating that the **SOTA open LLMs can be used for multilingual evaluation of text quality** (comparable with GPT-4 performance for most languages).

Based on the results, we have selected Gemma-2, Qwen2, and Llama-3.1 as meta-evaluators for judging quality of the texts generated by the examined approaches. In total, 4012 text samples have been

Metric	Meta-evaluator	AR	BN	EN	FR	HI	JA	RU	SW	TR	ZH	→ Average
Linguistic Acceptability	**Meta-Llama-3.1-70B-Instruct**	0.74	0.07	0.65	0.68	0.68	0.53	0.57	0.55	0.66	0.87	0.60
	Phi-3.5-mini-instruct	**0.75**	0.21	0.64	0.71	**0.70**	0.55	0.41	0.43	0.61	0.81	0.58
	Qwen2-72B-Instruct	0.74	0.20	0.65	0.81	0.58	0.43	0.74	0.46	0.63	**0.87**	0.61
	Aya-23-35B	0.72	0.19	0.59	0.80	0.66	0.51	0.43	0.38	0.59	0.86	0.57
	Gemma-2-27b-it	0.72	**0.25**	0.60	0.68	0.68	**0.56**	0.70	**0.80**	0.67	0.85	0.65
	GPT-4	0.71	0.22	**0.82**	**0.81**	0.61	0.47	**0.80**	0.76	**0.72**	0.85	**0.68**
Output Content Quality	**Meta-Llama-3.1-70B-Instruct**	0.70	0.05	0.64	0.71	**0.71**	0.54	0.77	0.64	0.58	0.85	0.62
	Phi-3.5-mini-instruct	0.70	0.23	0.65	0.73	0.63	0.52	0.48	0.44	0.30	0.87	0.56
	Qwen2-72B-Instruct	0.70	0.34	0.63	**0.73**	0.69	0.55	0.84	0.66	0.55	0.87	0.66
	Aya-23-35B	0.66	0.15	0.57	0.68	0.65	**0.56**	0.43	0.37	0.27	**0.89**	0.52
	Gemma-2-27b-it	**0.71**	**0.35**	0.62	0.69	0.63	0.49	0.89	0.84	**0.69**	0.77	0.67
	GPT-4	0.69	0.26	**0.68**	0.72	0.65	0.51	**0.92**	**0.88**	0.68	0.84	**0.68**

Table 13: Correlation of open LLMs meta-evaluation of text quality with human judgements using METAL dataset. The reported values represent pairwise agreement using weighted F1-score, analogously to the detailed prompting strategy in Table 3 of (Hada et al., 2024). GPT-4 values are taken from the METAL dataset.

	Linguistic Acceptability ↑			Output Content Quality ↑			
Approach	Meta-Llama-3.1-70B-Instruct	Qwen2-72B-Instruct	Gemma-2-27b-it	Meta-Llama-3.1-70B-Instruct	Qwen2-72B-Instruct	Gemma-2-27b-it	→ Average
k_to_one	0.41	0.79	0.24	0.28	0.21	0.10	0.34
keywords	0.40	0.63	0.22	0.36	0.36	0.24	0.37
paraphrase_1	**1.45**	**1.83**	**1.53**	**1.39**	**1.65**	**1.44**	**1.55**
paraphrase_2	1.36	1.80	1.44	1.31	1.57	1.33	1.47
paraphrase_3	1.30	1.77	1.37	1.26	1.53	1.28	1.42
paraphrase_4	1.26	1.73	1.34	1.25	1.47	1.24	1.38
paraphrase_5	1.23	1.76	1.25	1.20	1.44	1.16	1.34

Table 14: Quality meta-evaluation of texts generated by different approaches. Mean scores for the two selected quality metrics are provided, averaged across the three generators (Aya, Falcon, and Opt).

successfully evaluated by all three meta-evaluators. Inter-annotator agreement of the selected meta-evaluators is calculated in a form of pairwise *Pearson correlation coefficient*, averaging to 0.82. The definition of the selected Linguistic Acceptability and Output Content Quality metrics along with scoring schema (values of both being 0, 1, or 2, from lower to higher quality, respectively) can be found in METAL GitHub repository[17]. The meta-evaluation results of different approaches are summarized in Table 14, both metrics indicating that **paraphrasing resulted in higher quality texts than the other two approaches**, while each iteration of paraphrasing slightly reduces the text quality.

Similarly, we have used such meta-evaluation for quality assessment of the final texts generated by the selected generators. For this purpose, we have used a balanced subset of texts (10 samples per 22 languages per 7 generators and 1 human source, i.e. 1760 samples), resulting in 1752 evaluated samples (due to only 2 samples remained available from Gemini for Scottish Gaelic). The meta-evaluation scores given by the three meta-evaluators (pairwise *Pearson correlation coefficient* averaging to 0.69) are combined using the majority voting. The results, summarized in Table 15, indicate that the **LLM generators generated texts of similar or higher quality across languages than the quality of original human texts**. The reason might be an informal style used at social media (mistakes, spell errors, slang), which is more difficult for language models to follow (usually pre-trained on more formal web content). On average, the worst quality texts are generated by OPT-IML-Max-30B (failing mostly for non-Latin languages, still being on par with human text quality on average), while the best quality is provided by Gemini and Aya-101 models. Although not balanced across platforms (due to not each language being represented), meta-evaluation revealed the lowest quality of texts from Discord, followed by Telegram, and the highest quality of texts from Twitter.

B.5 Limited Bias Analysis

To minimize bias in the proposed dataset, we have run multiple existing detectors for data analysis. Based on the multilingual toxicity detector[18] (Dementieva et al., 2024), about 8% of the text samples are probably toxic (ranging from 5% in WhatsApp to 10% in Twitter parts). Based on the social media text topic detector[19] (Antypas et al., 2022), which is English-only, the topic distribution is illustrated in Figure 2. Based on the multilingual text genre

[17] https://github.com/microsoft/METAL-Towards-Multilingual-Meta-Evaluation/tree/main/metrics/detailed

[18] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier

[19] https://huggingface.co/cardiffnlp/tweet-topic-latest-multi

	Generator	Test Language [mean] ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	→ Average
Linguistic Acceptability	**Aya-101**	1.50	2.00	1.80	1.90	1.70	1.60	1.40	1.70	1.80	1.50	2.00	1.90	2.00	1.70	1.50	1.60	1.80	1.60	1.60	1.90	1.70	1.90	1.73
	Gemini	2.00	1.80	1.70	1.90	1.70	1.90	2.00	1.80	2.00	1.40	1.50	2.00	1.50	1.70	1.80	1.90	2.00	2.00	1.90	1.90	2.00	1.90	1.83
	GPT-3.5-Turbo-0125	1.50	1.90	1.60	1.80	1.40	1.50	1.70	1.50	1.20	1.70	1.30	1.80	1.50	1.40	1.20	1.50	1.40	2.00	1.80	1.60	1.60	2.00	1.59
	Mistral-7B-Instruct-v0.2	1.20	1.90	1.70	1.60	1.10	1.10	2.00	1.80	0.80	1.80	1.90	1.50	1.60	1.60	1.40	1.90	1.20	1.60	1.50	1.70	1.70	1.50	1.55
	OPT-IML-Max-30b	0.80	0.70	1.30	1.10	1.60	0.80	1.90	1.30	1.30	1.60	1.80	1.70	1.10	1.10	1.30	1.40	1.80	0.50	0.80	1.60	0.70	0.80	1.23
	v5-Eagle-7B-HF	1.60	1.70	1.70	1.70	1.70	1.90	1.80	1.60	1.50	1.60	1.80	1.60	1.60	1.80	1.50	1.90	1.70	1.80	1.80	1.80	1.50	1.60	1.69
	Vicuna-13b	1.30	1.50	1.90	1.30	1.80	1.00	2.00	2.00	0.80	1.20	1.80	1.40	1.40	1.70	1.60	1.90	1.80	1.60	1.20	1.50	1.70	1.70	1.55
	human	1.60	1.00	1.00	0.60	1.40	0.70	1.00	0.90	0.60	1.70	1.60	0.80	0.50	0.80	1.30	1.00	1.20	1.10	0.80	1.70	1.20	1.80	1.10
	Average	1.44	1.56	1.59	1.49	1.55	1.31	1.73	1.58	1.25	1.56	1.71	1.59	1.40	1.48	1.45	1.64	1.61	1.52	1.42	1.71	1.51	1.65	1.53
Output Content Quality	**Aya-101**	0.90	1.80	1.20	1.10	0.80	1.20	0.80	1.00	1.20	1.20	1.90	1.00	1.10	1.20	0.90	0.80	1.00	1.20	1.10	1.30	1.00	1.30	1.14
	Gemini	1.90	1.70	1.40	1.60	1.30	1.90	1.90	1.80	1.80	1.20	1.50	2.00	1.40	1.70	1.50	1.80	1.70	1.90	1.90	1.80	1.90	1.70	1.70
	GPT-3.5-Turbo-0125	1.20	1.50	1.20	1.20	1.10	1.00	1.10	1.20	0.70	1.20	1.30	1.10	1.10	1.30	0.80	1.10	1.10	1.30	1.40	1.40	1.00	1.70	1.18
	Mistral-7B-Instruct-v0.2	0.60	1.30	1.30	1.10	1.10	0.80	1.70	1.60	0.70	1.20	1.40	1.20	1.20	1.20	1.10	1.50	0.70	1.10	1.10	1.40	1.20	1.30	1.17
	OPT-IML-Max-30b	0.40	0.30	0.90	0.80	1.30	0.30	1.00	1.00	0.70	0.70	1.10	0.70	0.30	0.90	0.50	0.60	0.80	0.20	0.80	1.00	0.40	0.30	0.68
	v5-Eagle-7B-HF	1.00	1.20	1.20	1.40	1.40	1.50	1.40	0.90	1.20	0.80	1.40	1.10	1.20	1.40	0.90	1.70	1.30	1.20	1.40	1.40	1.20	1.40	1.25
	Vicuna-13b	0.70	1.00	1.30	0.90	1.50	0.70	1.60	1.40	0.50	0.90	0.90	0.80	1.10	1.30	1.10	1.20	1.00	1.20	1.00	1.30	1.20	1.20	1.08
	human	1.20	0.30	0.60	0.30	1.10	0.40	0.60	0.60	0.30	1.20	0.80	0.50	0.30	0.50	0.30	0.70	0.70	0.80	0.60	1.30	0.80	1.40	0.70
	Average	0.99	1.14	1.14	1.05	1.20	0.97	1.26	1.19	0.89	1.05	1.29	1.05	0.96	1.19	0.89	1.17	1.04	1.11	1.16	1.36	1.09	1.29	1.11

Table 15: Per-language quality meta-evaluation of texts generated by each generator. Mean of majority-voted (out of three meta-evaluators) scores of 10 samples for each combination are provided.

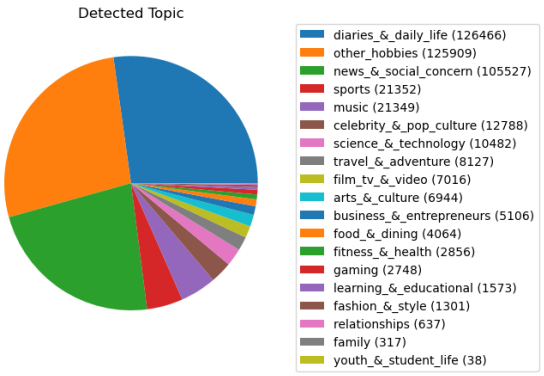

Figure 2: Detected topics in the MultiSocial dataset.

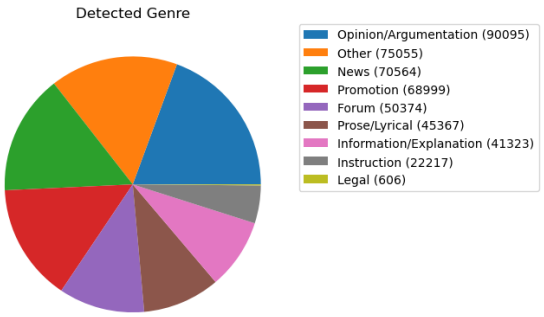

Figure 3: Detected genres in the MultiSocial dataset.

detector[20] (Kuzman et al., 2023), the genre distribution is illustrated in Figure 3. Although the used detection cannot be considered thorough (using existing detectors in zero-shot manner cannot be considered fully accurate), when used just as an indication, we can see that the proposed dataset texts are distributed among various topics and genres; thus, limiting the presence of such a bias.

[20] https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

C Fine-tuning Settings

For the fine-tuning process of the fine-tuned detection methods, we have used a parameter efficient fine-tuning (**PEFT**) technique called **QLoRA** (Dettmers et al., 2023) with default parameters (except for *target_modules* set to *query_key_value* and $r = 4$). The training process used the AdamW optimizer with the linear scheduler and the learning rate of $2E^{-4}$. Batch size of 2 with gradient accumulation steps of 8 have been used. We have used fixed 1 epoch for training (7 epochs in case of smaller subset selection), but also limiting the models training process to 48 hours (checkpointing each 20% of the epoch). All the settings can be found in the source code available in the published repository, enabling full replication. We aimed to use the same training settings across the various detection models training; however, we have used full fine-tuning for the XLM-RoBERTa-large model instead of QLoRA due to lack of support.

Due to a high class imbalance when using multiple generators data, we have experimented with various class balancing strategies for training. Namely, no balancing, majority-class downsampling, minority-class upsampling, mixed up-down-sampling (duplicating the minority-class samples just once and afterwards downsampling the majority-class). The results (using accuracy and AUC ROC metrics) indicated that there is a negligible effect on performance based on the used strategy, while no balancing having slower learning ability (however, the performance is eventually competitive with the others). Since having no significant impact on the performance, we have used majority-class downsampling to limit the number of steps in epoch.

Rank	Detector	AUC ROC	MacroF1 @5%FPR
1	Llama-3-8b-MultiSocial	0.9273	0.7988
2	Aya-101-MultiSocial	0.9262	0.8008
3	mDeBERTa-v3-base-MultiSocial	0.9025	0.7512
4	Mistral-7b-v0.1-MultiSocial	0.8988	0.7937
5	XLM-RoBERTa-large-MultiSocial	0.8309	0.7306
6	Binoculars	0.8303	0.4041
7	Fast-Detect-GPT	0.8104	0.6361
8	Falcon-rw-1b-MultiSocial	0.7592	0.6394
9	BLOOMZ-3b-MultiSocial	0.7071	0.5731
10	LLM-Deviation	0.6568	0.3568
11	DetectLLM-LRR	0.6496	0.4133
12	S5	0.6336	0.3519
13	Longformer Detector	0.6157	0.2564
14	ChatGPT-Detector-RoBERTa-Chinese	0.5896	0.4296
15	RoBERTa-large-OpenAI-Detector	0.5707	0.1958
16	BLOOMZ-3b-mixed-Detector	0.5536	0.1891
17	ruRoBERTa-ruatd-binary	0.5186	0.1485

Table 16: Cross-domain evaluation of the selected MGTD methods of statistical, pre-trained, and fine-tuned categories.

Rank	Detector	AUC ROC	MacroF1 @5%FPR
1	Fast-Detect-GPT	0.9238	0.8471
2	Binoculars	0.9048	0.7568
3	mDeBERTa-v3-base-MultiSocial	0.8871	0.8011
4	Mistral-7b-v0.1-MultiSocial	0.8614	0.7673
5	Aya-101-MultiSocial	0.8574	0.7556
6	Llama-3-8b-MultiSocial	0.8549	0.7352
7	XLM-RoBERTa-large-MultiSocial	0.7928	0.7108
8	Falcon-rw-1b-MultiSocial	0.7710	0.6912
9	DetectLLM-LRR	0.7559	0.7121
10	LLM-Deviation	0.7257	0.5931
11	Longformer Detector	0.7032	0.5018
12	S5	0.6849	0.5506
13	ChatGPT-Detector-RoBERTa-Chinese	0.6788	0.6161
14	BLOOMZ-3b-MultiSocial	0.6515	0.6026
15	RoBERTa-large-OpenAI-Detector	0.5596	0.4315
16	ruRoBERTa-ruatd-binary	0.5327	0.3473
17	BLOOMZ-3b-mixed-Detector	0.5212	0.4096

Table 17: Out-of-distribution (cross-domain and cross-generator) evaluation of the selected MGTD methods of statistical, pre-trained, and fine-tuned categories.

D Cross-domain Evaluation

For the evaluation on the out-of-domain data, we use the news articles of MULTITuDE (Macko et al., 2023) benchmark. We have used the published scripts[21] to extend the test set to our selection of languages. For the text generation, we have used the same models as in the proposed MultiSocial dataset (to evaluate only cross-domain capability), while we used Llama-2-70b model instead for Gemini for out-of-distribution (cross-generator) evaluation.

In Table 16 and 17, the results are provided in the same way as in Table 4, while testing on MULTITuDE (news domain) data. In pure cross-domain evaluation (Table 16), the machine texts are generated by the same generators (as used for training), in out-of-distribution evaluation (Table 17), the machine texts are generated only by Llama-2-70b (not available in MultiSocial for training).

E Ablation Study

Based on Table 29, we have identified two groups of detectors based on their performances across languages, namely autoregressive models and others. Autoregressive group includes Llama-3-8b, Mistral-7b-v0.1, BLOOMZ-3b, and Falcon-rw-1b. The non-autoregressive group includes Aya-101, XLM-RoBERTa-large, and mDeBERTa-v3-base. The summarized results, analogous to the Table 8, but provided for each group separately in Table 18 and Table 19. There are clear differences between the results of these groups, since in case of non-autogregressive group, we can see no significant differences between monolingually and multilingually fine-tuned detectors.

Another possible explanation of such a different behavior of some detectors is their pre-training on a huge number of languages (100+ languages in case of each detector in the non-autoregressive group). However, the BLOOMZ model has also been pre-trained on 50+ languages, and still shows a behavior similar to others in the autoregressive group.

F Results Data

Tables 20-22 contain per-generator AUC ROC performance of each MGT detection method for each test language, Tables 23-25 contain per-platform AUC ROC performance of each MGT detection method for each test language, and Tables 26-28 contains per-generator AUC ROC performance of each MGT detection method for each platform, separately for each MGTD category. Table 29 contains cross-lingual evaluation of differently fine-tuned MGT detectors. Table 30 contains cross-platform evaluation of differently fine-tuned MGT detectors.

[21] https://github.com/kinit-sk/mgt-detection-benchmark

Train Language	Test Language [AUC ROC mean]																						
	ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
en	0.78	0.86	0.77	0.87	0.88	0.88	0.96	0.85	0.89	N/A	N/A	0.92	0.95	0.82	0.86	0.90	0.92	0.85	N/A	N/A	0.79	0.65	0.84
es	0.78	0.84	0.83	0.83	0.88	0.84	0.87	0.94	0.86	N/A	N/A	0.88	0.92	0.81	0.84	0.91	0.91	0.86	N/A	N/A	0.79	0.64	0.82
ru	0.73	0.90	0.68	0.81	0.81	0.84	0.83	0.76	0.82	N/A	N/A	0.84	0.86	0.74	0.82	0.81	0.83	0.93	N/A	N/A	0.87	0.61	0.77
en-es-ru	0.88	0.92	0.86	0.91	0.91	0.93	0.95	0.94	0.91	N/A	N/A	0.93	0.96	0.84	0.89	0.93	0.93	0.93	N/A	N/A	0.88	0.84	0.90

Table 18: Cross-lingual mean AUC ROC performance of the MGT detectors with autoregressive foundational models fine-tuned monolingually (*en*, *es* and *ru*) and multilingually (*en-es-ru*), evaluated based on Telegram data (for training as well as for testing). N/A refers to not enough samples (at least 2000) in MultiSocial Telegram data.

Train Language	Test Language [AUC ROC mean]																						
	ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
en	0.85	0.95	0.79	0.96	0.88	0.93	0.96	0.90	0.95	N/A	N/A	0.96	0.99	0.86	0.93	0.94	0.96	0.90	N/A	N/A	0.85	0.83	0.91
es	0.88	0.96	0.86	0.96	0.91	0.92	0.94	0.94	0.95	N/A	N/A	0.96	0.99	0.87	0.93	0.94	0.96	0.92	N/A	N/A	0.87	0.86	0.92
ru	0.91	0.97	0.86	0.96	0.88	0.94	0.94	0.90	0.96	N/A	N/A	0.96	0.98	0.86	0.94	0.93	0.95	0.96	N/A	N/A	0.92	0.90	0.93
en-es-ru	0.91	0.97	0.87	0.96	0.91	0.93	0.96	0.93	0.96	N/A	N/A	0.96	0.99	0.88	0.94	0.95	0.97	0.94	N/A	N/A	0.90	0.89	0.93

Table 19: Cross-lingual mean AUC ROC performance of the MGT detectors with non-autoregressive foundational models fine-tuned monolingually (*en*, *es* and *ru*) and multilingually (*en-es-ru*), evaluated based on Telegram data (for training as well as for testing). N/A refers to not enough samples (at least 2000) in MultiSocial Telegram data.

Detector	Generator	Test Language [AUC ROC]																						
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
Binoculars	**Aya-101**	0.70	0.61	0.60	0.69	0.75	0.77	0.79	0.73	0.67	0.62	0.59	0.72	0.71	0.71	0.71	0.70	0.67	0.67	0.64	0.67	0.54	0.65	0.69
	GPT-3.5-Turbo-0125	0.62	0.64	0.57	0.72	0.72	0.78	0.75	0.70	0.66	0.61	0.59	0.75	0.74	0.68	0.73	0.70	0.71	0.69	0.70	0.66	0.61	0.92	0.68
	Gemini	0.64	0.73	0.71	0.85	0.86	0.80	0.92	0.90	0.87	0.95	N/A	0.88	0.87	0.80	0.84	0.88	0.88	0.69	0.82	0.85	0.70	0.90	0.83
	Mistral-7B-Instruct-v0.2	0.69	0.67	0.61	0.70	0.67	0.75	0.71	0.69	0.72	0.73	0.82	0.78	0.75	0.64	0.71	0.70	0.72	0.70	0.66	0.68	0.65	0.63	0.68
	OPT-IML-Max-30b	0.77	0.65	0.58	0.58	0.64	0.75	0.74	0.64	0.68	0.64	0.63	0.71	0.73	0.63	0.62	0.66	0.63	0.65	0.58	0.65	0.64	0.43	0.64
	Vicuna-13b	0.77	0.69	0.63	0.79	0.80	0.86	0.82	0.78	0.75	0.76	0.84	0.82	0.80	0.78	0.81	0.80	0.79	0.80	0.76	0.76	0.64	0.76	0.76
	v5-Eagle-7B-HF	0.73	0.72	0.64	0.82	0.82	0.84	0.89	0.85	0.79	0.82	0.80	0.83	0.84	0.81	0.81	0.82	0.81	0.81	0.80	0.81	0.70	0.86	0.79
DetectLLM-LRR	**Aya-101**	0.75	0.82	0.67	0.89	0.75	0.84	0.77	0.76	0.82	0.73	0.69	0.79	0.91	0.76	0.82	0.76	0.81	0.74	0.79	0.70	0.66	0.66	0.70
	GPT-3.5-Turbo-0125	0.70	0.75	0.65	0.93	0.74	0.83	0.78	0.77	0.85	0.78	0.65	0.88	0.92	0.77	0.85	0.80	0.84	0.68	0.86	0.77	0.66	0.76	0.71
	Gemini	0.89	0.94	0.65	0.96	0.86	0.95	0.90	0.94	0.96	0.86	N/A	0.97	0.97	0.84	0.92	0.93	0.96	0.84	0.94	0.93	0.86	0.92	0.83
	Mistral-7B-Instruct-v0.2	0.75	0.85	0.68	0.92	0.72	0.85	0.76	0.80	0.86	0.82	0.87	0.88	0.93	0.70	0.85	0.84	0.84	0.77	0.79	0.74	0.75	0.72	0.71
	OPT-IML-Max-30b	0.68	0.81	0.63	0.84	0.66	0.83	0.70	0.70	0.83	0.63	0.51	0.80	0.89	0.66	0.82	0.73	0.79	0.66	0.75	0.64	0.62	0.60	0.67
	Vicuna-13b	0.90	0.90	0.72	0.96	0.85	0.92	0.81	0.87	0.89	0.88	0.83	0.91	0.95	0.86	0.93	0.90	0.91	0.90	0.88	0.82	0.79	0.84	0.78
	v5-Eagle-7B-HF	0.87	0.92	0.79	0.97	0.87	0.93	0.88	0.92	0.95	0.86	0.86	0.94	0.97	0.90	0.95	0.92	0.95	0.88	0.94	0.91	0.86	0.92	0.82
Fast-Detect-GPT	**Aya-101**	0.75	0.66	0.64	0.83	0.81	0.74	0.79	0.75	0.75	0.73	0.71	0.74	0.79	0.78	0.78	0.76	0.77	0.72	0.75	0.75	0.67	0.68	0.75
	GPT-3.5-Turbo-0125	0.74	0.61	0.54	0.82	0.71	0.72	0.73	0.64	0.69	0.71	0.73	0.79	0.70	0.70	0.74	0.71	0.70	0.73	0.75	0.75	0.73	0.80	0.71
	Gemini	0.86	0.85	0.70	0.88	0.88	0.89	0.92	0.88	0.90	0.88	N/A	0.92	0.89	0.81	0.84	0.89	0.88	0.82	0.88	0.90	0.83	0.90	0.87
	Mistral-7B-Instruct-v0.2	0.60	0.50	0.51	0.66	0.56	0.43	0.69	0.58	0.47	0.58	0.80	0.66	0.63	0.54	0.60	0.61	0.57	0.59	0.46	0.55	0.56	0.58	0.58
	OPT-IML-Max-30b	0.64	0.50	0.58	0.72	0.70	0.48	0.77	0.69	0.65	0.62	0.54	0.76	0.70	0.67	0.74	0.73	0.74	0.57	0.65	0.59	0.50	0.52	0.67
	Vicuna-13b	0.81	0.67	0.65	0.83	0.85	0.57	0.83	0.80	0.52	0.75	0.79	0.83	0.79	0.82	0.84	0.84	0.83	0.83	0.57	0.77	0.74	0.78	0.78
	v5-Eagle-7B-HF	0.85	0.79	0.72	0.91	0.86	0.80	0.89	0.85	0.85	0.78	0.84	0.88	0.86	0.85	0.87	0.86	0.89	0.85	0.88	0.85	0.81	0.91	0.85
LLM-Deviation	**Aya-101**	0.79	0.84	0.66	0.90	0.77	0.84	0.76	0.76	0.85	0.76	0.73	0.81	0.91	0.77	0.83	0.77	0.82	0.75	0.81	0.73	0.67	0.69	0.70
	GPT-3.5-Turbo-0125	0.68	0.73	0.63	0.93	0.74	0.83	0.77	0.76	0.86	0.79	0.68	0.88	0.92	0.76	0.86	0.80	0.84	0.66	0.87	0.79	0.65	0.78	0.71
	Gemini	0.90	0.95	0.64	0.96	0.87	0.96	0.90	0.94	0.97	0.86	N/A	0.97	0.98	0.85	0.93	0.93	0.96	0.85	0.94	0.95	0.88	0.92	0.82
	Mistral-7B-Instruct-v0.2	0.76	0.82	0.66	0.92	0.72	0.86	0.75	0.79	0.86	0.85	0.93	0.89	0.93	0.68	0.85	0.83	0.84	0.74	0.78	0.74	0.72	0.70	0.70
	OPT-IML-Max-30b	0.77	0.85	0.62	0.86	0.67	0.84	0.70	0.71	0.84	0.69	0.58	0.83	0.91	0.67	0.83	0.75	0.81	0.71	0.77	0.68	0.66	0.67	0.68
	Vicuna-13b	0.92	0.91	0.72	0.96	0.86	0.93	0.81	0.86	0.90	0.89	0.88	0.91	0.96	0.86	0.94	0.90	0.92	0.89	0.88	0.83	0.79	0.84	0.78
	v5-Eagle-7B-HF	0.89	0.93	0.79	0.98	0.87	0.94	0.88	0.93	0.97	0.88	0.89	0.95	0.98	0.91	0.96	0.93	0.95	0.88	0.95	0.92	0.87	0.92	0.82
S5	**Aya-101**	0.80	0.84	0.67	0.89	0.76	0.85	0.75	0.75	0.84	0.75	0.71	0.81	0.90	0.76	0.82	0.77	0.82	0.75	0.80	0.73	0.67	0.70	0.71
	GPT-3.5-Turbo-0125	0.67	0.71	0.62	0.92	0.73	0.82	0.74	0.73	0.84	0.77	0.67	0.87	0.91	0.74	0.84	0.78	0.84	0.65	0.85	0.77	0.65	0.74	0.70
	Gemini	0.84	0.94	0.63	0.95	0.85	0.95	0.88	0.92	0.96	0.84	N/A	0.97	0.97	0.83	0.91	0.92	0.96	0.83	0.93	0.93	0.87	0.90	0.81
	Mistral-7B-Instruct-v0.2	0.72	0.79	0.66	0.91	0.71	0.85	0.72	0.78	0.83	0.83	0.92	0.88	0.92	0.67	0.83	0.81	0.84	0.72	0.76	0.72	0.70	0.66	0.69
	OPT-IML-Max-30b	0.79	0.84	0.63	0.86	0.67	0.84	0.69	0.71	0.84	0.71	0.63	0.82	0.90	0.67	0.83	0.74	0.81	0.73	0.77	0.69	0.66	0.71	0.69
	Vicuna-13b	0.91	0.89	0.72	0.95	0.85	0.93	0.79	0.85	0.89	0.88	0.86	0.90	0.95	0.85	0.93	0.90	0.91	0.88	0.87	0.82	0.79	0.82	0.78
	v5-Eagle-7B-HF	0.88	0.92	0.79	0.97	0.86	0.94	0.86	0.91	0.96	0.86	0.85	0.94	0.97	0.89	0.95	0.92	0.95	0.88	0.95	0.90	0.86	0.91	0.81

Table 20: Per-LLM AUC ROC performance of statistical MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Generator	Test Language [AUC ROC]																						
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
BLOOMZ-3b-mixed-Detector	Aya-101	0.83	0.88	0.82	0.88	0.85	0.82	0.85	0.84	0.91	0.79	0.68	0.86	0.92	0.84	0.84	0.85	0.78	0.78	0.83	0.76	0.81	0.86	0.83
	GPT-3.5-Turbo-0125	0.85	0.80	0.81	0.86	0.82	0.82	0.87	0.83	0.87	0.81	0.67	0.83	0.87	0.80	0.83	0.83	0.74	0.73	0.80	0.71	0.82	0.19	0.80
	Gemini	0.64	0.55	0.63	0.69	0.65	0.76	0.50	0.58	0.78	0.65	N/A	0.67	0.77	0.66	0.72	0.63	0.33	0.48	0.65	0.41	0.43	0.60	0.59
	Mistral-7B-Instruct-v0.2	0.76	0.71	0.74	0.74	0.71	0.71	0.85	0.78	0.80	0.75	0.57	0.75	0.78	0.68	0.73	0.78	0.72	0.68	0.68	0.59	0.71	0.76	0.74
	OPT-IML-Max-30b	0.69	0.66	0.76	0.80	0.71	0.72	0.78	0.79	0.77	0.71	0.60	0.76	0.77	0.71	0.75	0.77	0.70	0.63	0.79	0.69	0.65	0.75	0.73
	Vicuna-13b	0.88	0.77	0.84	0.79	0.81	0.74	0.88	0.83	0.81	0.78	0.77	0.77	0.85	0.78	0.81	0.83	0.66	0.75	0.78	0.67	0.73	0.83	0.79
	v5-Eagle-7B-HF	0.86	0.82	0.85	0.82	0.82	0.76	0.90	0.86	0.89	0.83	0.65	0.79	0.88	0.79	0.81	0.86	0.72	0.72	0.77	0.67	0.74	0.85	0.81
ChatGPT-Detector-RoBERTa-Chinese	Aya-101	0.75	0.79	0.73	0.63	0.75	0.70	0.82	0.75	0.78	0.61	0.52	0.61	0.72	0.67	0.61	0.61	0.67	0.73	0.60	0.60	0.71	0.76	0.67
	GPT-3.5-Turbo-0125	0.59	0.63	0.64	0.59	0.70	0.65	0.86	0.73	0.76	0.65	0.52	0.57	0.68	0.63	0.57	0.69	0.68	0.62	0.56	0.60	0.63	0.88	0.66
	Gemini	0.74	0.84	0.80	0.78	0.88	0.72	0.97	0.90	0.94	0.95	N/A	0.76	0.89	0.78	0.71	0.86	0.90	0.81	0.73	0.78	0.87	0.78	0.80
	Mistral-7B-Instruct-v0.2	0.73	0.77	0.79	0.69	0.82	0.78	0.92	0.84	0.84	0.87	0.66	0.73	0.82	0.73	0.63	0.84	0.84	0.76	0.66	0.67	0.76	0.75	0.76
	OPT-IML-Max-30b	0.79	0.84	0.74	0.60	0.70	0.76	0.82	0.74	0.73	0.65	0.58	0.52	0.68	0.63	0.60	0.62	0.59	0.74	0.62	0.60	0.70	0.63	0.65
	Vicuna-13b	0.74	0.86	0.81	0.63	0.82	0.80	0.93	0.82	0.74	0.73	0.73	0.60	0.81	0.70	0.62	0.73	0.76	0.82	0.62	0.68	0.72	0.88	0.72
	v5-Eagle-7B-HF	0.74	0.87	0.81	0.71	0.88	0.81	0.95	0.89	0.88	0.83	0.76	0.59	0.83	0.72	0.63	0.77	0.78	0.83	0.66	0.65	0.80	0.95	0.76
Longformer Detector	Aya-101	0.22	0.37	0.25	0.36	0.32	0.41	0.53	0.34	0.35	0.38	0.36	0.39	0.35	0.29	0.35	0.31	0.31	0.36	0.35	0.37	0.43	0.27	0.36
	GPT-3.5-Turbo-0125	0.39	0.50	0.37	0.42	0.44	0.50	0.59	0.48	0.40	0.39	0.43	0.48	0.46	0.43	0.44	0.40	0.42	0.45	0.43	0.43	0.55	0.54	0.45
	Gemini	0.67	0.61	0.47	0.42	0.46	0.82	0.80	0.48	0.50	0.92	N/A	0.52	0.49	0.44	0.46	0.42	0.36	0.64	0.34	0.33	0.80	0.76	0.53
	Mistral-7B-Instruct-v0.2	0.33	0.48	0.37	0.58	0.51	0.54	0.63	0.50	0.61	0.64	0.53	0.61	0.62	0.49	0.56	0.51	0.49	0.47	0.59	0.55	0.66	0.50	0.52
	OPT-IML-Max-30b	0.17	0.42	0.28	0.33	0.32	0.39	0.50	0.33	0.41	0.44	0.38	0.42	0.43	0.31	0.35	0.35	0.35	0.29	0.33	0.44	0.51	0.26	0.36
	Vicuna-13b	0.28	0.47	0.24	0.55	0.42	0.53	0.66	0.42	0.48	0.58	0.50	0.54	0.50	0.40	0.49	0.41	0.45	0.49	0.51	0.50	0.60	0.42	0.46
	v5-Eagle-7B-HF	0.31	0.49	0.30	0.61	0.49	0.58	0.81	0.50	0.60	0.67	0.68	0.63	0.61	0.47	0.56	0.51	0.48	0.52	0.59	0.56	0.68	0.54	0.53
RoBERTa-large-OpenAI-Detector	Aya-101	0.76	0.46	0.46	0.19	0.40	0.72	0.73	0.43	0.32	0.37	0.46	0.31	0.20	0.39	0.31	0.35	0.36	0.57	0.35	0.45	0.43	0.69	0.43
	GPT-3.5-Turbo-0125	0.60	0.41	0.45	0.16	0.34	0.59	0.48	0.34	0.25	0.28	0.51	0.20	0.17	0.31	0.23	0.26	0.28	0.48	0.27	0.38	0.30	0.61	0.35
	Gemini	0.65	0.30	0.51	0.05	0.14	0.76	0.14	0.11	0.05	0.41	N/A	0.06	0.04	0.13	0.10	0.08	0.05	0.47	0.08	0.08	0.29	0.43	0.21
	Mistral-7B-Instruct-v0.2	0.69	0.39	0.38	0.10	0.30	0.72	0.40	0.25	0.22	0.14	0.30	0.16	0.11	0.30	0.19	0.17	0.21	0.49	0.26	0.34	0.30	0.55	0.31
	OPT-IML-Max-30b	0.86	0.60	0.54	0.30	0.52	0.79	0.71	0.49	0.32	0.51	0.67	0.34	0.24	0.51	0.36	0.40	0.40	0.70	0.48	0.54	0.60	0.83	0.49
	Vicuna-13b	0.80	0.45	0.40	0.10	0.29	0.79	0.57	0.30	0.17	0.21	0.47	0.18	0.12	0.26	0.17	0.20	0.23	0.54	0.21	0.31	0.36	0.60	0.33
	v5-Eagle-7B-HF	0.78	0.40	0.34	0.08	0.28	0.83	0.62	0.24	0.11	0.33	0.48	0.13	0.07	0.20	0.15	0.16	0.21	0.52	0.18	0.26	0.26	0.52	0.31
ruRoBERTa-ruatd-binary	Aya-101	0.50	0.61	0.57	0.47	0.48	0.45	0.60	0.54	0.48	0.47	0.49	0.53	0.56	0.57	0.48	0.50	0.43	0.64	0.52	0.57	0.50	0.45	0.53
	GPT-3.5-Turbo-0125	0.46	0.66	0.57	0.45	0.46	0.39	0.58	0.50	0.45	0.46	0.45	0.51	0.50	0.52	0.45	0.45	0.41	0.69	0.48	0.48	0.52	0.22	0.50
	Gemini	0.17	0.25	0.53	0.14	0.19	0.11	0.19	0.17	0.11	0.30	N/A	0.10	0.12	0.24	0.17	0.15	0.08	0.30	0.12	0.11	0.23	0.26	0.18
	Mistral-7B-Instruct-v0.2	0.42	0.69	0.53	0.50	0.49	0.35	0.58	0.52	0.49	0.45	0.38	0.57	0.53	0.56	0.48	0.49	0.41	0.79	0.52	0.52	0.69	0.52	0.53
	OPT-IML-Max-30b	0.64	0.82	0.63	0.54	0.52	0.50	0.66	0.55	0.54	0.50	0.44	0.57	0.60	0.59	0.52	0.53	0.43	0.92	0.61	0.65	0.91	0.58	0.59
	Vicuna-13b	0.31	0.69	0.54	0.45	0.43	0.36	0.66	0.49	0.43	0.51	0.57	0.53	0.49	0.51	0.44	0.45	0.34	0.82	0.45	0.50	0.62	0.57	0.51
	v5-Eagle-7B-HF	0.35	0.72	0.53	0.50	0.44	0.29	0.66	0.51	0.51	0.59	0.50	0.56	0.50	0.51	0.51	0.46	0.33	0.78	0.45	0.50	0.66	0.45	0.52

Table 21: Per-LLM AUC ROC performance of pre-trained MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Generator	Test Language [AUC ROC]																						
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
Aya-101-MultiSocial	Aya-101	0.92	0.98	0.97	0.98	0.97	0.96	0.96	0.96	0.98	0.92	0.91	0.98	0.99	0.97	0.97	0.96	0.96	0.93	0.97	0.92	0.90	0.93	0.96
	GPT-3.5-Turbo-0125	0.99	1.00	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.97	0.96	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.98	0.99	0.99
	Gemini	0.90	0.96	0.88	0.95	0.92	0.92	0.96	0.96	0.99	0.85	N/A	0.96	0.99	0.92	0.94	0.94	0.96	0.83	0.92	0.89	0.87	0.94	0.93
	Mistral-7B-Instruct-v0.2	0.99	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.98	0.99	0.99	1.00	0.99	1.00	0.98	0.98	0.98	0.99
	OPT-IML-Max-30b	0.98	0.99	0.96	0.97	0.94	0.96	0.92	0.95	0.95	0.84	0.77	0.95	0.98	0.95	0.97	0.95	0.96	0.98	0.96	0.91	0.95	0.99	0.95
	Vicuna-13b	1.00	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.98	0.95	0.99	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.97	0.99	0.99
	v5-Eagle-7B-HF	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.98	0.94	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	0.98	0.99	0.99	0.99
BLOOMZ-3b-MultiSocial	Aya-101	0.92	0.97	0.96	0.96	0.95	0.94	0.96	0.96	0.96	0.86	0.71	0.94	0.99	0.94	0.94	0.95	0.92	0.90	0.94	0.84	0.85	0.90	0.94
	GPT-3.5-Turbo-0125	0.99	0.99	0.99	0.98	0.98	0.97	0.99	0.99	0.98	0.90	0.84	0.97	0.99	0.97	0.97	0.99	0.96	0.97	0.97	0.91	0.93	0.99	0.98
	Gemini	0.91	0.93	0.88	0.96	0.91	0.93	0.97	0.97	0.99	0.91	N/A	0.95	0.99	0.90	0.92	0.96	0.93	0.82	0.89	0.75	0.77	0.96	0.93
	Mistral-7B-Instruct-v0.2	0.96	0.99	0.98	0.98	0.97	0.97	1.00	0.98	0.99	0.95	0.81	0.98	1.00	0.93	0.96	0.99	0.97	0.96	0.98	0.91	0.94	0.98	0.97
	OPT-IML-Max-30b	0.95	0.98	0.93	0.94	0.89	0.93	0.93	0.94	0.94	0.84	0.76	0.90	0.97	0.89	0.92	0.93	0.90	0.96	0.94	0.86	0.92	0.97	0.93
	Vicuna-13b	0.99	0.99	0.98	0.98	0.98	0.98	1.00	0.99	0.98	0.93	0.88	0.97	0.99	0.97	0.98	0.99	0.96	0.97	0.98	0.90	0.91	0.99	0.98
	v5-Eagle-7B-HF	0.98	0.99	0.99	0.99	0.98	0.98	1.00	1.00	0.99	0.94	0.86	0.99	1.00	0.98	0.98	1.00	0.99	0.98	0.98	0.95	0.95	1.00	0.99
Falcon-rw-1b-MultiSocial	Aya-101	0.93	0.97	0.97	0.97	0.95	0.94	0.96	0.96	0.97	0.88	0.82	0.95	0.99	0.95	0.95	0.95	0.93	0.91	0.95	0.85	0.87	0.91	0.94
	GPT-3.5-Turbo-0125	0.98	0.99	0.99	0.98	0.98	0.98	0.99	0.98	0.98	0.93	0.89	0.98	0.99	0.98	0.98	0.98	0.97	0.97	0.97	0.91	0.94	0.99	0.98
	Gemini	0.86	0.96	0.89	0.95	0.92	0.93	0.99	0.94	0.98	0.92	N/A	0.95	0.99	0.91	0.92	0.92	0.92	0.87	0.88	0.72	0.87	0.90	0.93
	Mistral-7B-Instruct-v0.2	0.95	0.99	0.98	0.98	0.97	0.97	1.00	0.97	0.99	0.96	0.93	0.98	0.99	0.95	0.97	0.98	0.98	0.94	0.97	0.89	0.92	0.95	0.97
	OPT-IML-Max-30b	0.96	0.98	0.94	0.95	0.90	0.95	0.93	0.93	0.94	0.84	0.72	0.92	0.98	0.90	0.93	0.92	0.92	0.95	0.95	0.88	0.91	0.97	0.93
	Vicuna-13b	0.98	0.98	0.98	0.99	0.98	0.98	1.00	0.98	0.99	0.95	0.93	0.97	0.99	0.97	0.98	0.98	0.97	0.96	0.98	0.89	0.88	0.98	0.98
	v5-Eagle-7B-HF	0.98	0.99	0.98	0.99	0.98	0.98	1.00	0.99	0.99	0.96	0.92	0.99	1.00	0.98	0.98	0.99	0.99	0.97	0.98	0.94	0.95	0.99	0.98
Llama-3-8b-MultiSocial	Aya-101	0.94	0.98	0.98	0.99	0.97	0.96	0.97	0.97	0.98	0.90	0.81	0.98	0.99	0.97	0.97	0.96	0.95	0.94	0.97	0.91	0.92	0.95	0.96
	GPT-3.5-Turbo-0125	0.99	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	0.95	0.93	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.96	0.98	1.00	0.99
	Gemini	0.89	0.96	0.91	0.96	0.95	0.89	0.99	0.97	0.98	0.91	N/A	0.98	0.99	0.94	0.96	0.96	0.96	0.87	0.94	0.91	0.88	0.96	0.95
	Mistral-7B-Instruct-v0.2	0.99	1.00	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.98	0.99	0.99	0.99	0.99	0.99	0.97	0.98	0.99	0.99
	OPT-IML-Max-30b	0.98	0.99	0.97	0.98	0.94	0.96	0.94	0.96	0.96	0.85	0.72	0.96	0.99	0.95	0.97	0.96	0.95	0.98	0.97	0.92	0.96	0.99	0.96
	Vicuna-13b	1.00	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.98	0.92	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.96	0.97	1.00	0.99
	v5-Eagle-7B-HF	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.98	0.98	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	0.98	0.99	1.00	0.99
Mistral-7b-v0.1-MultiSocial	Aya-101	0.94	0.98	0.97	0.98	0.97	0.95	0.97	0.97	0.98	0.90	0.89	0.98	0.99	0.97	0.97	0.96	0.95	0.94	0.96	0.90	0.91	0.94	0.96
	GPT-3.5-Turbo-0125	0.99	1.00	1.00	1.00	0.99	0.98	1.00	0.99	0.99	0.95	0.96	0.99	1.00	0.99	0.99	0.99	0.99	0.98	0.96	0.98	1.00	0.99	0.99
	Gemini	0.91	0.98	0.91	0.97	0.96	0.93	0.99	0.99	0.98	0.95	N/A	0.99	0.99	0.95	0.96	0.97	0.96	0.90	0.93	0.90	0.92	0.95	0.96
	Mistral-7B-Instruct-v0.2	0.98	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.96	0.97	1.00	1.00	0.98	0.99	0.99	0.99	0.99	0.99	0.97	0.98	0.99	0.99
	OPT-IML-Max-30b	0.98	0.99	0.96	0.98	0.94	0.97	0.95	0.96	0.96	0.86	0.79	0.96	0.99	0.95	0.97	0.95	0.95	0.98	0.97	0.91	0.96	0.99	0.96
	Vicuna-13b	0.99	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.95	0.95	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.97	0.99	0.99
	v5-Eagle-7B-HF	0.98	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.96	0.97	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.99
XLM-RoBERTa-large-MultiSocial	Aya-101	0.89	0.97	0.94	0.97	0.95	0.94	0.93	0.94	0.95	0.81	0.73	0.94	0.98	0.95	0.95	0.91	0.93	0.90	0.96	0.87	0.86	0.85	0.93
	GPT-3.5-Turbo-0125	0.98	0.99	0.98	0.99	0.98	0.97	0.99	0.98	0.97	0.92	0.88	0.98	0.99	0.98	0.98	0.98	0.98	0.97	0.98	0.95	0.96	0.98	0.98
	Gemini	0.89	0.96	0.78	0.96	0.93	0.94	0.98	0.96	0.99	0.84	N/A	0.98	0.99	0.91	0.95	0.96	0.97	0.89	0.95	0.88	0.91	0.93	0.94
	Mistral-7B-Instruct-v0.2	0.98	0.99	0.96	0.99	0.97	0.97	0.99	0.96	0.97	0.96	0.91	0.98	0.99	0.93	0.98	0.98	0.98	0.97	0.97	0.94	0.96	0.95	0.97
	OPT-IML-Max-30b	0.94	0.97	0.88	0.94	0.88	0.90	0.88	0.92	0.91	0.69	0.47	0.90	0.95	0.88	0.93	0.88	0.91	0.93	0.90	0.82	0.86	0.89	0.90
	Vicuna-13b	0.99	0.99	0.96	0.99	0.98	0.97	0.99	0.98	0.97	0.91	0.80	0.98	0.99	0.97	0.99	0.98	0.98	0.98	0.97	0.94	0.94	0.96	0.98
	v5-Eagle-7B-HF	0.98	0.99	0.98	1.00	0.99	0.98	0.99	0.99	0.99	0.95	0.83	0.99	1.00	0.98	0.99	0.99	0.99	0.98	0.99	0.97	0.98	0.98	0.99
mDeBERTa-v3-base-MultiSocial	Aya-101	0.88	0.97	0.92	0.97	0.95	0.93	0.94	0.94	0.97	0.86	0.74	0.96	0.99	0.95	0.95	0.93	0.93	0.89	0.95	0.89	0.87	0.88	0.93
	GPT-3.5-Turbo-0125	0.97	0.99	0.98	0.99	0.98	0.97	0.99	0.98	0.98	0.91	0.82	0.98	0.99	0.98	0.98	0.97	0.97	0.96	0.98	0.95	0.96	0.99	0.98
	Gemini	0.83	0.95	0.81	0.92	0.88	0.87	0.94	0.93	0.97	0.78	N/A	0.92	0.99	0.89	0.91	0.91	0.91	0.80	0.88	0.81	0.85	0.83	0.90
	Mistral-7B-Instruct-v0.2	0.98	0.99	0.96	0.99	0.97	0.97	0.99	0.97	0.98	0.96	0.86	0.99	1.00	0.95	0.98	0.98	0.99	0.97	0.98	0.96	0.96	0.97	0.97
	OPT-IML-Max-30b	0.96	0.99	0.91	0.95	0.91	0.92	0.90	0.93	0.93	0.75	0.54	0.93	0.97	0.91	0.95	0.92	0.93	0.95	0.94	0.89	0.92	0.97	0.93
	Vicuna-13b	0.99	0.99	0.96	0.99	0.98	0.98	0.99	0.98	0.98	0.95	0.86	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.98	0.96	0.95	0.98	0.98
	v5-Eagle-7B-HF	0.97	0.99	0.98	0.99	0.99	0.98	0.99	0.99	0.99	0.95	0.87	0.99	1.00	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.98	0.98	0.99

Table 22: Per-LLM AUC ROC performance of fine-tuned MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Platform	Test Language [AUC ROC]																					
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
Binoculars	Discord	N/A	N/A	0.78	0.76	0.80	N/A	0.86	0.81	0.75	0.67	0.70	0.83	0.82	0.80	0.84	0.81	0.77	N/A	0.72	N/A	N/A	N/A	0.79
	Gab	0.61	0.62	0.47	0.58	0.69	0.71	0.78	0.72	0.70	0.72	N/A	0.72	0.68	0.71	0.67	0.72	0.67	0.69	0.62	0.70	0.63	0.75	0.68
	Telegram	0.71	0.69	0.61	0.77	0.74	0.83	0.81	0.77	0.75	0.72	N/A	0.79	0.82	0.72	0.78	0.76	0.79	0.65	0.72	0.77	0.63	0.73	0.73
	Twitter	0.67	0.69	0.64	0.76	0.82	0.92	0.81	0.67	0.84	N/A	N/A	0.85	0.79	0.72	0.82	0.80	0.77	0.85	0.87	N/A	N/A	0.80	0.74
	WhatsApp	0.82	N/A	0.73	0.66	0.79	N/A	0.75	0.81	0.52	N/A	N/A	N/A	N/A	0.74	N/A	0.70	0.85	0.66	N/A	N/A	N/A	N/A	0.73
DetectLLM-LRR	Discord	N/A	N/A	0.95	0.98	0.91	N/A	0.92	0.93	0.94	0.83	0.75	0.94	0.98	0.94	0.98	0.95	0.96	N/A	0.90	N/A	N/A	N/A	0.94
	Gab	0.74	0.80	0.58	0.80	0.74	0.79	0.78	0.75	0.71	0.75	N/A	0.77	0.81	0.72	0.82	0.75	0.77	0.72	0.71	0.74	0.77	0.78	0.69
	Telegram	0.76	0.86	0.63	0.94	0.69	0.94	0.81	0.79	0.92	0.77	N/A	0.94	0.97	0.74	0.88	0.86	0.92	0.79	0.89	0.96	0.74	0.76	0.75
	Twitter	0.81	0.87	0.75	0.91	0.91	0.96	0.85	0.78	0.87	N/A	N/A	0.94	0.94	0.85	0.95	0.91	0.90	0.93	0.95	N/A	N/A	0.92	0.75
	WhatsApp	0.87	N/A	0.87	0.96	0.80	N/A	0.66	0.89	0.91	N/A	N/A	N/A	N/A	0.86	N/A	0.80	0.95	0.69	N/A	N/A	N/A	N/A	0.70
Fast-Detect-GPT	Discord	N/A	N/A	0.64	0.81	0.81	N/A	0.86	0.82	0.70	0.76	0.76	0.81	0.79	0.81	0.83	0.84	0.80	N/A	0.65	N/A	N/A	N/A	0.79
	Gab	0.70	0.68	0.57	0.69	0.72	0.60	0.76	0.71	0.62	0.64	N/A	0.72	0.65	0.74	0.73	0.72	0.71	0.69	0.64	0.69	0.72	0.78	0.70
	Telegram	0.76	0.65	0.62	0.83	0.73	0.69	0.81	0.74	0.70	0.70	N/A	0.83	0.81	0.72	0.78	0.76	0.80	0.69	0.72	0.83	0.69	0.72	0.74
	Twitter	0.71	0.65	0.62	0.82	0.82	0.78	0.81	0.65	0.67	N/A	N/A	0.83	0.78	0.72	0.77	0.79	0.78	0.87	0.77	N/A	N/A	0.85	0.75
	WhatsApp	0.85	N/A	0.69	0.75	0.79	N/A	0.77	0.80	0.70	N/A	N/A	N/A	N/A	0.72	N/A	0.77	0.80	0.64	N/A	N/A	N/A	N/A	0.77
LLM-Deviation	Discord	N/A	N/A	0.95	0.98	0.92	N/A	0.93	0.95	0.95	0.86	0.81	0.94	0.99	0.94	0.98	0.96	0.97	N/A	0.91	N/A	N/A	N/A	0.94
	Gab	0.75	0.80	0.58	0.82	0.76	0.82	0.76	0.77	0.75	0.78	N/A	0.78	0.82	0.73	0.83	0.75	0.78	0.73	0.70	0.76	0.81	0.81	0.70
	Telegram	0.79	0.85	0.62	0.94	0.72	0.93	0.81	0.79	0.92	0.80	N/A	0.95	0.98	0.76	0.89	0.86	0.93	0.79	0.90	0.97	0.74	0.77	0.74
	Twitter	0.83	0.88	0.74	0.92	0.92	0.97	0.85	0.75	0.88	N/A	N/A	0.94	0.95	0.82	0.96	0.92	0.91	0.93	0.96	N/A	N/A	0.92	0.74
	WhatsApp	0.90	N/A	0.88	0.97	0.82	N/A	0.64	0.89	0.91	N/A	N/A	N/A	N/A	0.87	N/A	0.82	0.96	0.70	N/A	N/A	N/A	N/A	0.70
S5	Discord	N/A	N/A	0.94	0.97	0.91	N/A	0.91	0.93	0.93	0.86	0.79	0.93	0.98	0.93	0.97	0.95	0.96	N/A	0.89	N/A	N/A	N/A	0.93
	Gab	0.74	0.79	0.59	0.81	0.74	0.82	0.73	0.76	0.76	0.77	N/A	0.77	0.80	0.72	0.81	0.74	0.77	0.72	0.67	0.74	0.79	0.81	0.69
	Telegram	0.78	0.84	0.62	0.93	0.70	0.93	0.79	0.77	0.91	0.77	N/A	0.94	0.97	0.75	0.88	0.85	0.92	0.79	0.89	0.95	0.74	0.76	0.73
	Twitter	0.80	0.87	0.71	0.91	0.90	0.97	0.83	0.73	0.86	N/A	N/A	0.94	0.93	0.78	0.95	0.91	0.90	0.93	0.94	N/A	N/A	0.90	0.74
	WhatsApp	0.90	N/A	0.86	0.97	0.83	N/A	0.62	0.87	0.90	N/A	N/A	N/A	N/A	0.86	N/A	0.81	0.96	0.70	N/A	N/A	N/A	N/A	0.70

Table 23: Per-platform AUC ROC performance of statistical MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Platform	Test Language [AUC ROC]																					
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
BLOOMZ-3b-mixed-Detector	Discord	N/A	N/A	0.96	0.89	0.87	N/A	0.87	0.90	0.87	0.80	0.66	0.84	0.86	0.91	0.91	0.89	0.78	N/A	0.71	N/A	N/A	N/A	0.87
	Gab	0.69	0.71	0.77	0.69	0.67	0.62	0.77	0.69	0.73	0.75	N/A	0.61	0.75	0.69	0.70	0.69	0.53	0.59	0.70	0.54	0.58	0.64	0.66
	Telegram	0.81	0.79	0.73	0.81	0.76	0.85	0.83	0.80	0.84	0.78	N/A	0.80	0.88	0.74	0.77	0.80	0.75	0.75	0.76	0.84	0.71	0.71	0.78
	Twitter	0.82	0.68	0.70	0.71	0.78	0.68	0.81	0.68	0.77	N/A	N/A	0.81	0.71	0.70	0.70	0.80	0.59	0.73	0.90	N/A	N/A	0.68	0.72
	WhatsApp	0.82	N/A	0.88	0.46	0.78	N/A	0.76	0.87	0.81	N/A	N/A	N/A	N/A	0.67	N/A	0.81	0.78	0.80	N/A	N/A	N/A	N/A	0.79
ChatGPT-Detector-RoBERTa-Chinese	Discord	N/A	N/A	0.88	0.72	0.86	N/A	0.95	0.88	0.85	0.73	0.62	0.66	0.82	0.71	0.70	0.75	0.83	N/A	0.68	N/A	N/A	N/A	0.77
	Gab	0.57	0.81	0.65	0.55	0.74	0.68	0.88	0.74	0.74	0.72	N/A	0.60	0.66	0.72	0.58	0.75	0.68	0.65	0.75	0.61	0.74	0.78	0.67
	Telegram	0.65	0.75	0.73	0.66	0.82	0.78	0.90	0.80	0.83	0.75	N/A	0.61	0.83	0.63	0.62	0.75	0.80	0.78	0.60	0.74	0.74	0.81	0.70
	Twitter	0.86	0.90	0.73	0.67	0.77	0.78	0.90	0.79	0.72	N/A	N/A	0.61	0.68	0.76	0.58	0.71	0.70	0.88	0.77	N/A	N/A	0.83	0.74
	WhatsApp	0.82	N/A	0.77	0.36	0.91	N/A	0.89	0.85	0.83	N/A	N/A	N/A	N/A	0.58	N/A	0.71	0.83	0.81	N/A	N/A	N/A	N/A	0.79
Longformer Detector	Discord	N/A	N/A	0.25	0.41	0.47	N/A	0.68	0.44	0.40	0.50	0.46	0.50	0.40	0.42	0.44	0.36	0.33	N/A	0.36	N/A	N/A	N/A	0.44
	Gab	0.56	0.39	0.32	0.44	0.42	0.49	0.71	0.45	0.48	0.56	N/A	0.49	0.45	0.41	0.39	0.48	0.50	0.54	0.34	0.46	0.48	0.49	0.48
	Telegram	0.37	0.78	0.37	0.55	0.36	0.56	0.67	0.43	0.55	0.51	N/A	0.55	0.61	0.47	0.54	0.44	0.41	0.58	0.49	0.45	0.62	0.46	0.51
	Twitter	0.24	0.16	0.30	0.38	0.49	0.67	0.65	0.31	0.43	N/A	N/A	0.38	0.44	0.33	0.47	0.38	0.40	0.31	0.16	N/A	N/A	0.49	0.38
	WhatsApp	0.22	N/A	0.41	0.63	0.30	N/A	0.52	0.53	0.42	N/A	N/A	N/A	N/A	0.44	N/A	0.42	0.38	0.31	N/A	N/A	N/A	N/A	0.44
RoBERTa-large-OpenAI-Detector	Discord	N/A	N/A	0.25	0.05	0.12	N/A	0.44	0.10	0.18	0.32	0.52	0.12	0.09	0.09	0.07	0.09	0.09	N/A	0.22	N/A	N/A	N/A	0.16
	Gab	0.58	0.56	0.30	0.25	0.38	0.75	0.51	0.33	0.29	0.30	N/A	0.38	0.20	0.35	0.31	0.33	0.37	0.50	0.45	0.37	0.43	0.56	0.40
	Telegram	0.72	0.21	0.50	0.11	0.42	0.74	0.50	0.32	0.17	0.27	N/A	0.10	0.09	0.27	0.17	0.20	0.10	0.36	0.22	0.18	0.35	0.62	0.33
	Twitter	0.80	0.60	0.44	0.17	0.19	0.75	0.55	0.42	0.21	N/A	N/A	0.20	0.16	0.31	0.17	0.15	0.28	0.66	0.12	N/A	N/A	0.59	0.42
	WhatsApp	0.82	N/A	0.26	0.20	0.32	N/A	0.55	0.17	0.21	N/A	N/A	N/A	N/A	0.18	N/A	0.27	0.07	0.71	N/A	N/A	N/A	N/A	0.40
ruRoBERTa-ruatd-binary	Discord	N/A	N/A	0.51	0.41	0.25	N/A	0.53	0.42	0.35	0.41	0.36	0.45	0.43	0.43	0.40	0.35	0.25	N/A	0.32	N/A	N/A	N/A	0.40
	Gab	0.38	0.59	0.62	0.46	0.46	0.40	0.61	0.47	0.40	0.53	N/A	0.48	0.46	0.55	0.47	0.44	0.42	0.71	0.50	0.45	0.47	0.45	0.49
	Telegram	0.39	0.63	0.56	0.47	0.50	0.32	0.55	0.44	0.46	0.54	N/A	0.50	0.49	0.45	0.40	0.43	0.32	0.74	0.45	0.51	0.60	0.45	0.48
	Twitter	0.39	0.67	0.55	0.31	0.38	0.23	0.58	0.49	0.45	N/A	N/A	0.37	0.39	0.50	0.41	0.42	0.33	0.69	0.23	N/A	N/A	0.25	0.49
	WhatsApp	0.41	N/A	0.58	0.57	0.37	N/A	0.54	0.49	0.43	N/A	N/A	N/A	N/A	0.43	N/A	0.48	0.34	0.70	N/A	N/A	N/A	N/A	0.50

Table 24: Per-platform AUC ROC performance of pre-trained MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Platform	Test Language [AUC ROC]																					
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
Aya-101-MultiSocial	Discord	N/A	N/A	0.99	1.00	0.97	N/A	0.99	0.99	1.00	0.96	0.90	0.99	1.00	0.99	0.99	0.99	0.99	N/A	0.99	N/A	N/A	N/A	0.99
	Gab	0.91	0.96	0.97	0.96	0.96	0.95	0.96	0.96	0.94	0.94	N/A	0.94	0.98	0.97	0.97	0.96	0.95	0.92	0.96	0.92	0.89	0.94	0.94
	Telegram	0.97	0.99	0.97	0.98	0.98	0.98	0.98	0.97	0.99	0.94	N/A	0.99	1.00	0.96	0.98	0.97	0.99	0.97	0.98	1.00	0.96	0.99	0.98
	Twitter	0.99	0.99	0.98	0.97	0.97	0.99	0.97	0.99	0.98	N/A	N/A	1.00	0.98	0.97	0.99	0.98	0.98	0.98	1.00	N/A	N/A	0.97	0.98
	WhatsApp	0.99	N/A	0.97	0.98	0.95	N/A	0.98	0.98	1.00	N/A	N/A	N/A	N/A	0.97	N/A	0.97	0.99	0.95	N/A	N/A	N/A	N/A	0.97
BLOOMZ-3b-MultiSocial	Discord	N/A	N/A	0.99	1.00	0.97	N/A	0.99	0.99	0.99	0.97	0.84	0.98	1.00	0.98	0.99	0.99	0.99	N/A	0.97	N/A	N/A	N/A	0.99
	Gab	0.87	0.96	0.96	0.93	0.92	0.93	0.96	0.95	0.94	0.88	N/A	0.90	0.97	0.92	0.94	0.95	0.89	0.88	0.95	0.81	0.81	0.93	0.91
	Telegram	0.95	0.98	0.94	0.97	0.96	0.98	0.98	0.97	0.98	0.86	N/A	0.97	1.00	0.92	0.94	0.97	0.97	0.94	0.95	0.98	0.90	0.98	0.96
	Twitter	0.99	0.98	0.98	0.94	0.96	0.99	0.98	0.99	0.92	N/A	N/A	0.99	0.97	0.94	0.97	0.97	0.96	0.99	0.98	N/A	N/A	0.99	0.97
	WhatsApp	0.99	N/A	0.97	0.98	0.97	N/A	0.98	0.98	0.99	N/A	N/A	N/A	N/A	0.95	N/A	0.98	0.99	0.97	N/A	N/A	N/A	N/A	0.98
Falcon-rw-1b-MultiSocial	Discord	N/A	N/A	0.99	1.00	0.97	N/A	0.99	0.98	0.99	0.96	0.87	0.98	1.00	0.99	0.99	0.99	0.99	N/A	0.97	N/A	N/A	N/A	0.99
	Gab	0.88	0.94	0.97	0.94	0.93	0.94	0.97	0.93	0.94	0.92	N/A	0.91	0.97	0.93	0.95	0.92	0.89	0.88	0.93	0.80	0.87	0.92	0.92
	Telegram	0.94	0.98	0.93	0.97	0.96	0.98	0.98	0.96	0.99	0.89	N/A	0.98	1.00	0.94	0.95	0.95	0.98	0.94	0.96	0.98	0.91	0.97	0.96
	Twitter	0.98	0.98	0.98	0.95	0.96	0.99	0.98	0.98	0.95	N/A	N/A	0.97	0.96	0.95	0.96	0.97	0.96	0.99	0.99	N/A	N/A	0.99	0.97
	WhatsApp	0.99	N/A	0.95	0.97	0.96	N/A	0.99	0.97	0.99	N/A	N/A	N/A	N/A	0.95	N/A	0.96	0.99	0.96	N/A	N/A	N/A	N/A	0.97
Llama-3-8b-MultiSocial	Discord	N/A	N/A	0.99	1.00	0.97	N/A	0.99	0.99	1.00	0.96	0.88	0.99	1.00	0.99	1.00	0.99	0.99	N/A	0.98	N/A	N/A	N/A	0.99
	Gab	0.90	0.95	0.98	0.97	0.97	0.95	0.98	0.97	0.95	0.94	N/A	0.96	0.99	0.97	0.97	0.96	0.95	0.93	0.97	0.92	0.88	0.96	0.95
	Telegram	0.97	0.99	0.97	0.99	0.98	0.98	0.98	0.98	0.99	0.94	N/A	0.99	1.00	0.97	0.98	0.97	0.98	0.97	0.98	1.00	0.96	0.99	0.98
	Twitter	0.99	0.99	0.99	0.97	0.98	0.98	0.99	0.99	0.97	N/A	N/A	0.99	0.98	0.98	0.99	0.98	0.98	0.98	1.00	N/A	N/A	0.98	0.98
	WhatsApp	0.99	N/A	0.98	0.98	0.96	N/A	0.99	0.98	1.00	N/A	N/A	N/A	N/A	0.98	N/A	0.98	0.99	0.98	N/A	N/A	N/A	N/A	0.98
Mistral-7b-v0.1-MultiSocial	Discord	N/A	N/A	0.99	1.00	0.97	N/A	0.99	0.99	1.00	0.96	0.91	0.99	1.00	0.99	1.00	0.99	0.99	N/A	0.98	N/A	N/A	N/A	0.99
	Gab	0.90	0.97	0.97	0.98	0.97	0.96	0.98	0.97	0.94	0.92	N/A	0.96	0.99	0.97	0.98	0.96	0.94	0.94	0.95	0.91	0.89	0.96	0.95
	Telegram	0.97	0.99	0.97	0.99	0.98	0.98	0.98	0.98	0.99	0.91	N/A	0.99	1.00	0.97	0.98	0.97	0.99	0.97	0.97	1.00	0.96	0.99	0.98
	Twitter	0.99	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.96	N/A	N/A	0.99	0.98	0.97	0.98	0.98	0.98	0.99	0.99	N/A	N/A	0.98	0.98
	WhatsApp	0.99	N/A	0.97	0.98	0.96	N/A	0.99	0.98	0.99	N/A	N/A	N/A	N/A	0.97	N/A	0.98	0.99	0.98	N/A	N/A	N/A	N/A	0.98
XLM-RoBERTa-large-MultiSocial	Discord	N/A	N/A	0.96	0.99	0.96	N/A	0.98	0.97	0.99	0.89	0.72	0.98	1.00	0.98	0.99	0.97	0.98	N/A	0.99	N/A	N/A	N/A	0.97
	Gab	0.88	0.94	0.95	0.93	0.94	0.94	0.95	0.95	0.90	0.87	N/A	0.94	0.96	0.95	0.95	0.93	0.94	0.90	0.95	0.88	0.84	0.92	0.93
	Telegram	0.94	0.98	0.93	0.98	0.97	0.97	0.96	0.96	0.98	0.87	N/A	0.98	0.99	0.94	0.97	0.96	0.97	0.96	0.96	0.98	0.93	0.94	0.96
	Twitter	0.98	0.99	0.94	0.96	0.96	0.97	0.97	0.97	0.97	N/A	N/A	0.95	0.97	0.94	0.98	0.96	0.97	0.98	0.98	N/A	N/A	0.96	0.96
	WhatsApp	0.98	N/A	0.90	0.96	0.97	N/A	0.97	0.97	0.97	N/A	N/A	N/A	N/A	0.93	N/A	0.95	0.97	0.95	N/A	N/A	N/A	N/A	0.96
mDeBERTa-v3-base-MultiSocial	Discord	N/A	N/A	0.98	1.00	0.97	N/A	0.98	0.98	0.99	0.91	0.75	0.98	1.00	0.98	0.99	0.99	0.99	N/A	0.97	N/A	N/A	N/A	0.98
	Gab	0.85	0.96	0.92	0.94	0.93	0.90	0.95	0.93	0.92	0.88	N/A	0.92	0.97	0.94	0.94	0.92	0.92	0.89	0.92	0.88	0.79	0.91	0.91
	Telegram	0.94	0.99	0.93	0.98	0.95	0.97	0.97	0.95	0.98	0.90	N/A	0.98	1.00	0.93	0.96	0.96	0.97	0.96	0.96	0.99	0.94	0.96	0.96
	Twitter	0.98	0.98	0.96	0.93	0.95	0.97	0.96	0.97	0.97	N/A	N/A	0.96	0.96	0.95	0.97	0.95	0.96	0.96	0.99	N/A	N/A	0.95	0.96
	WhatsApp	0.97	N/A	0.94	0.96	0.96	N/A	0.96	0.97	1.00	N/A	N/A	N/A	N/A	0.96	N/A	0.95	0.99	0.95	N/A	N/A	N/A	N/A	0.96

Table 25: Per-platform AUC ROC performance of fine-tuned MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Generator	Platform [AUC ROC]					
		Discord	Gab	Telegram	Twitter	WhatsApp	all
Binoculars	Mistral-7B-Instruct-v0.2	0.77	0.62	0.68	0.68	0.67	0.68
	Aya-101	0.76	0.65	0.68	0.70	0.72	0.69
	Gemini	0.88	0.79	0.84	0.83	0.84	0.83
	GPT-3.5-Turbo-0125	0.75	0.64	0.71	0.68	0.69	0.68
	OPT-IML-Max-30b	0.71	0.60	0.64	0.66	0.65	0.64
	v5-Eagle-7B-HF	0.85	0.76	0.80	0.81	0.79	0.79
	Vicuna-13b	0.82	0.73	0.76	0.80	0.77	0.76
DetectLLM-LRR	Mistral-7B-Instruct-v0.2	0.94	0.63	0.71	0.71	0.68	0.71
	Aya-101	0.91	0.64	0.69	0.70	0.66	0.70
	Gemini	0.96	0.80	0.84	0.82	0.78	0.83
	GPT-3.5-Turbo-0125	0.93	0.65	0.72	0.71	0.67	0.71
	OPT-IML-Max-30b	0.87	0.60	0.68	0.67	0.63	0.67
	v5-Eagle-7B-HF	0.98	0.78	0.81	0.83	0.77	0.82
	Vicuna-13b	0.96	0.73	0.77	0.80	0.73	0.78
Fast-Detect-GPT	Mistral-7B-Instruct-v0.2	0.65	0.51	0.58	0.59	0.62	0.58
	Aya-101	0.80	0.71	0.74	0.75	0.77	0.75
	Gemini	0.89	0.84	0.88	0.86	0.88	0.87
	GPT-3.5-Turbo-0125	0.75	0.66	0.73	0.68	0.74	0.71
	OPT-IML-Max-30b	0.75	0.61	0.65	0.66	0.71	0.67
	v5-Eagle-7B-HF	0.88	0.81	0.84	0.86	0.86	0.85
	Vicuna-13b	0.81	0.74	0.75	0.82	0.82	0.78
LLM-Deviation	Mistral-7B-Instruct-v0.2	0.95	0.63	0.69	0.70	0.67	0.70
	Aya-101	0.92	0.65	0.70	0.70	0.66	0.70
	Gemini	0.97	0.80	0.82	0.82	0.78	0.82
	GPT-3.5-Turbo-0125	0.93	0.65	0.71	0.69	0.67	0.71
	OPT-IML-Max-30b	0.89	0.62	0.69	0.67	0.64	0.68
	v5-Eagle-7B-HF	0.98	0.78	0.80	0.83	0.77	0.82
	Vicuna-13b	0.96	0.73	0.76	0.80	0.73	0.78
S5	Mistral-7B-Instruct-v0.2	0.93	0.63	0.69	0.69	0.67	0.69
	Aya-101	0.91	0.65	0.70	0.70	0.66	0.71
	Gemini	0.96	0.78	0.81	0.81	0.77	0.81
	GPT-3.5-Turbo-0125	0.91	0.64	0.70	0.68	0.66	0.70
	OPT-IML-Max-30b	0.87	0.62	0.69	0.67	0.64	0.69
	v5-Eagle-7B-HF	0.98	0.78	0.80	0.83	0.77	0.81
	Vicuna-13b	0.96	0.73	0.75	0.80	0.73	0.78

Table 26: Per-platform per-LLM AUC ROC performance of statistical MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Generator	Platform [AUC ROC] Discord	Gab	Telegram	Twitter	WhatsApp	all
BLOOMZ-3b-mixed-Detector	Mistral-7B-Instruct-v0.2	0.87	0.62	0.76	0.70	0.78	0.74
	Aya-101	0.92	0.75	0.85	0.81	0.85	0.83
	Gemini	0.76	0.47	0.63	0.52	0.58	0.59
	GPT-3.5-Turbo-0125	0.91	0.70	0.80	0.75	0.86	0.80
	OPT-IML-Max-30b	0.81	0.67	0.76	0.69	0.79	0.73
	v5-Eagle-7B-HF	0.92	0.70	0.83	0.78	0.86	0.81
	Vicuna-13b	0.89	0.68	0.81	0.78	0.83	0.79
ChatGPT-Detector-RoBERTa-Chinese	Mistral-7B-Instruct-v0.2	0.85	0.70	0.75	0.74	0.85	0.76
	Aya-101	0.70	0.64	0.65	0.71	0.73	0.67
	Gemini	0.88	0.73	0.78	0.81	0.86	0.80
	GPT-3.5-Turbo-0125	0.72	0.59	0.66	0.65	0.73	0.66
	OPT-IML-Max-30b	0.65	0.63	0.64	0.70	0.72	0.65
	v5-Eagle-7B-HF	0.80	0.71	0.74	0.80	0.82	0.76
	Vicuna-13b	0.76	0.68	0.71	0.77	0.81	0.72
Longformer Detector	Mistral-7B-Instruct-v0.2	0.51	0.55	0.59	0.44	0.47	0.52
	Aya-101	0.32	0.38	0.40	0.28	0.37	0.36
	Gemini	0.50	0.54	0.58	0.47	0.52	0.53
	GPT-3.5-Turbo-0125	0.37	0.50	0.49	0.44	0.44	0.45
	OPT-IML-Max-30b	0.36	0.38	0.40	0.27	0.34	0.36
	v5-Eagle-7B-HF	0.53	0.55	0.60	0.41	0.52	0.53
	Vicuna-13b	0.43	0.48	0.53	0.38	0.44	0.46
RoBERTa-large-OpenAI-Detector	Mistral-7B-Instruct-v0.2	0.11	0.38	0.30	0.38	0.33	0.31
	Aya-101	0.25	0.48	0.40	0.49	0.52	0.43
	Gemini	0.07	0.25	0.22	0.29	0.21	0.21
	GPT-3.5-Turbo-0125	0.17	0.40	0.33	0.42	0.38	0.35
	OPT-IML-Max-30b	0.29	0.56	0.47	0.57	0.57	0.49
	v5-Eagle-7B-HF	0.11	0.36	0.30	0.38	0.40	0.31
	Vicuna-13b	0.13	0.39	0.33	0.40	0.41	0.33
ruRoBERTa-ruatd-binary	Mistral-7B-Instruct-v0.2	0.47	0.52	0.54	0.53	0.53	0.53
	Aya-101	0.42	0.57	0.52	0.53	0.57	0.53
	Gemini	0.16	0.17	0.16	0.20	0.19	0.18
	GPT-3.5-Turbo-0125	0.40	0.52	0.49	0.53	0.56	0.50
	OPT-IML-Max-30b	0.45	0.63	0.60	0.62	0.64	0.59
	v5-Eagle-7B-HF	0.46	0.50	0.53	0.49	0.53	0.52
	Vicuna-13b	0.41	0.51	0.53	0.51	0.53	0.51

Table 27: Per-platform per-LLM AUC ROC performance of pre-trained MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Generator	Platform [AUC ROC] Discord	Gab	Telegram	Twitter	WhatsApp	all
Aya-101-MultiSocial	Mistral-7B-Instruct-v0.2	1.00	0.98	0.99	0.99	1.00	0.99
	Aya-101	0.99	0.91	0.97	0.96	0.95	0.96
	Gemini	0.98	0.86	0.95	0.93	0.93	0.93
	GPT-3.5-Turbo-0125	1.00	0.98	0.99	1.00	1.00	0.99
	OPT-IML-Max-30b	0.97	0.92	0.97	0.97	0.96	0.95
	v5-Eagle-7B-HF	1.00	0.98	1.00	1.00	0.99	0.99
	Vicuna-13b	1.00	0.98	0.99	1.00	1.00	0.99
BLOOMZ-3b-MultiSocial	Mistral-7B-Instruct-v0.2	1.00	0.94	0.97	0.98	0.99	0.97
	Aya-101	0.98	0.88	0.94	0.95	0.96	0.94
	Gemini	0.99	0.85	0.94	0.93	0.96	0.93
	GPT-3.5-Turbo-0125	0.99	0.95	0.98	0.99	0.99	0.98
	OPT-IML-Max-30b	0.96	0.87	0.94	0.95	0.95	0.93
	v5-Eagle-7B-HF	1.00	0.96	0.99	0.99	0.99	0.99
	Vicuna-13b	0.99	0.95	0.98	0.99	0.99	0.98
Falcon-rw-1b-MultiSocial	Mistral-7B-Instruct-v0.2	1.00	0.93	0.97	0.98	0.98	0.97
	Aya-101	0.98	0.89	0.95	0.96	0.96	0.94
	Gemini	0.98	0.84	0.94	0.94	0.94	0.93
	GPT-3.5-Turbo-0125	0.99	0.96	0.98	0.99	0.99	0.98
	OPT-IML-Max-30b	0.96	0.88	0.94	0.95	0.95	0.93
	v5-Eagle-7B-HF	1.00	0.96	0.99	0.99	0.99	0.98
	Vicuna-13b	0.99	0.95	0.97	0.99	0.99	0.98
Llama-3-8b-MultiSocial	Mistral-7B-Instruct-v0.2	1.00	0.98	0.99	0.99	1.00	0.99
	Aya-101	0.99	0.92	0.97	0.97	0.97	0.96
	Gemini	0.99	0.89	0.95	0.94	0.95	0.95
	GPT-3.5-Turbo-0125	1.00	0.98	0.99	1.00	1.00	0.99
	OPT-IML-Max-30b	0.97	0.93	0.97	0.97	0.97	0.96
	v5-Eagle-7B-HF	1.00	0.99	1.00	1.00	1.00	0.99
	Vicuna-13b	1.00	0.98	0.99	1.00	1.00	0.99
Mistral-7b-v0.1-MultiSocial	Mistral-7B-Instruct-v0.2	1.00	0.97	0.99	0.99	1.00	0.99
	Aya-101	0.99	0.92	0.97	0.96	0.97	0.96
	Gemini	0.99	0.91	0.96	0.95	0.97	0.96
	GPT-3.5-Turbo-0125	1.00	0.98	0.99	1.00	1.00	0.99
	OPT-IML-Max-30b	0.97	0.92	0.97	0.97	0.97	0.96
	v5-Eagle-7B-HF	1.00	0.98	1.00	1.00	1.00	0.99
	Vicuna-13b	1.00	0.98	0.99	0.99	1.00	0.99
XLM-RoBERTa-large-MultiSocial	Mistral-7B-Instruct-v0.2	0.99	0.95	0.97	0.97	0.99	0.97
	Aya-101	0.96	0.89	0.93	0.95	0.92	0.93
	Gemini	0.97	0.89	0.96	0.94	0.96	0.94
	GPT-3.5-Turbo-0125	0.98	0.96	0.98	0.99	0.99	0.98
	OPT-IML-Max-30b	0.91	0.87	0.91	0.93	0.91	0.90
	v5-Eagle-7B-HF	1.00	0.97	0.99	0.99	0.99	0.99
	Vicuna-13b	0.99	0.96	0.97	0.99	0.99	0.98
mDeBERTa-v3-base-MultiSocial	Mistral-7B-Instruct-v0.2	0.99	0.94	0.98	0.98	0.99	0.97
	Aya-101	0.98	0.88	0.94	0.94	0.93	0.93
	Gemini	0.97	0.80	0.92	0.88	0.90	0.90
	GPT-3.5-Turbo-0125	0.99	0.95	0.98	0.98	0.99	0.98
	OPT-IML-Max-30b	0.95	0.89	0.94	0.94	0.94	0.93
	v5-Eagle-7B-HF	1.00	0.97	0.99	0.99	0.99	0.99
	Vicuna-13b	0.99	0.96	0.98	0.99	0.99	0.98

Table 28: Per-platform per-LLM AUC ROC performance of fine-tuned MGT detectors category. N/A refers to not enough samples per each class (at least 10) making AUC ROC value irrelevant.

Detector	Train Language	Test Language [AUC ROC]																					
		ar	bg	ca	cs	de	el	en	es	et	ga	gd	hr	hu	nl	pl	pt	ro	ru	sk	sl	uk	zh	all
Aya-101	en	0.89	0.97	0.84	0.97	0.92	0.95	0.97	0.93	0.96	N/A	N/A	0.96	0.99	0.87	0.95	0.95	0.97	0.93	N/A	N/A	0.90	0.91	0.93
	es	0.91	0.97	0.90	0.96	0.95	0.93	0.95	0.96	0.96	N/A	N/A	0.96	0.99	0.90	0.95	0.95	0.97	0.95	N/A	N/A	0.90	0.91	0.94
	ru	0.92	0.98	0.89	0.95	0.92	0.94	0.95	0.93	0.96	N/A	N/A	0.96	0.98	0.89	0.95	0.93	0.97	0.97	N/A	N/A	0.94	0.91	0.94
	en-es-ru	0.93	0.98	0.90	0.96	0.94	0.94	0.96	0.95	0.96	N/A	N/A	0.96	0.99	0.89	0.95	0.95	0.97	0.96	N/A	N/A	0.91	0.92	0.94
BLOOMZ-3b	en	0.75	0.91	0.73	0.89	0.87	0.90	0.95	0.74	0.88	N/A	N/A	0.90	0.95	0.84	0.87	0.90	0.92	0.87	N/A	N/A	0.82	0.61	0.82
	es	0.78	0.85	0.84	0.84	0.85	0.82	0.90	0.93	0.83	N/A	N/A	0.82	0.88	0.80	0.80	0.90	0.88	0.85	N/A	N/A	0.79	0.61	0.81
	ru	0.69	0.86	0.57	0.80	0.80	0.76	0.81	0.56	0.80	N/A	N/A	0.78	0.83	0.73	0.75	0.74	0.74	0.90	N/A	N/A	0.84	0.64	0.72
	en-es-ru	0.87	0.90	0.83	0.87	0.89	0.92	0.95	0.92	0.86	N/A	N/A	0.85	0.92	0.79	0.85	0.92	0.90	0.90	N/A	N/A	0.86	0.88	0.86
Falcon-rw-1b	en	0.74	0.74	0.78	0.85	0.86	0.87	0.95	0.87	0.88	N/A	N/A	0.91	0.95	0.79	0.85	0.89	0.92	0.81	N/A	N/A	0.74	0.80	0.83
	es	0.78	0.73	0.79	0.85	0.89	0.91	0.86	0.93	0.86	N/A	N/A	0.90	0.92	0.82	0.87	0.92	0.93	0.85	N/A	N/A	0.75	0.80	0.83
	ru	0.68	0.85	0.71	0.76	0.78	0.86	0.82	0.83	0.82	N/A	N/A	0.84	0.80	0.71	0.82	0.80	0.88	0.90	N/A	N/A	0.84	0.75	0.78
	en-es-ru	0.82	0.87	0.83	0.91	0.90	0.93	0.96	0.92	0.91	N/A	N/A	0.93	0.97	0.84	0.89	0.92	0.94	0.89	N/A	N/A	0.84	0.87	0.89
Llama-3-8b	en	0.85	0.96	0.76	0.89	0.89	0.90	0.97	0.91	0.93	N/A	N/A	0.96	0.98	0.84	0.92	0.94	0.93	0.91	N/A	N/A	0.85	0.73	0.87
	es	0.80	0.94	0.85	0.79	0.90	0.88	0.91	0.95	0.86	N/A	N/A	0.90	0.95	0.84	0.92	0.92	0.92	0.90	N/A	N/A	0.86	0.64	0.83
	ru	0.76	0.95	0.67	0.79	0.84	0.88	0.86	0.81	0.76	N/A	N/A	0.83	0.89	0.80	0.90	0.83	0.82	0.95	N/A	N/A	0.90	0.59	0.78
	en-es-ru	0.92	0.98	0.89	0.93	0.95	0.95	0.97	0.96	0.95	N/A	N/A	0.97	0.99	0.89	0.95	0.95	0.95	0.96	N/A	N/A	0.93	0.93	0.93
Mistral-7b-v0.1	en	0.77	0.83	0.81	0.86	0.89	0.85	0.96	0.86	0.89	N/A	N/A	0.92	0.92	0.82	0.81	0.88	0.92	0.80	N/A	N/A	0.75	0.46	0.82
	es	0.77	0.82	0.86	0.85	0.89	0.75	0.83	0.93	0.89	N/A	N/A	0.91	0.94	0.80	0.78	0.91	0.91	0.82	N/A	N/A	0.76	0.51	0.82
	ru	0.78	0.94	0.76	0.89	0.82	0.85	0.82	0.84	0.89	N/A	N/A	0.92	0.90	0.74	0.80	0.87	0.88	0.95	N/A	N/A	0.90	0.44	0.82
	en-es-ru	0.90	0.94	0.87	0.93	0.91	0.92	0.95	0.93	0.93	N/A	N/A	0.95	0.97	0.85	0.86	0.93	0.94	0.96	N/A	N/A	0.91	0.68	0.90
XLM-RoBERTa-large	en	0.82	0.92	0.77	0.94	0.88	0.92	0.96	0.91	0.95	N/A	N/A	0.96	0.99	0.87	0.91	0.94	0.96	0.88	N/A	N/A	0.81	0.83	0.90
	es	0.87	0.95	0.86	0.95	0.91	0.94	0.92	0.94	0.95	N/A	N/A	0.97	0.99	0.86	0.91	0.94	0.97	0.90	N/A	N/A	0.83	0.85	0.91
	ru	0.92	0.98	0.87	0.96	0.89	0.95	0.92	0.91	0.96	N/A	N/A	0.97	0.98	0.86	0.94	0.93	0.96	0.96	N/A	N/A	0.92	0.91	0.93
	en-es-ru	0.91	0.97	0.88	0.97	0.92	0.96	0.96	0.93	0.97	N/A	N/A	0.98	0.99	0.90	0.94	0.95	0.97	0.95	N/A	N/A	0.90	0.90	0.94
mDeBERTa-v3-base	en	0.83	0.96	0.76	0.97	0.84	0.93	0.96	0.88	0.94	N/A	N/A	0.96	0.99	0.84	0.92	0.94	0.96	0.90	N/A	N/A	0.84	0.74	0.90
	es	0.86	0.95	0.82	0.96	0.87	0.87	0.94	0.92	0.94	N/A	N/A	0.95	0.98	0.85	0.93	0.94	0.95	0.91	N/A	N/A	0.87	0.82	0.90
	ru	0.89	0.96	0.82	0.96	0.85	0.92	0.94	0.88	0.95	N/A	N/A	0.94	0.99	0.84	0.95	0.93	0.94	0.95	N/A	N/A	0.91	0.87	0.91
	en-es-ru	0.88	0.96	0.82	0.96	0.86	0.90	0.95	0.90	0.95	N/A	N/A	0.95	0.99	0.86	0.93	0.94	0.95	0.93	N/A	N/A	0.89	0.85	0.91

Table 29: Cross-lingual AUC ROC performance of the selected MGT detectors fine-tuned monolingually (*en*, *es* and *ru*) and multilingually (*en-es-ru*), evaluated based on Telegram data (for training as well as for testing). N/A refers to not enough samples (at least 2000) in MultiSocial Telegram data.

Detector	Train Platform	Test Platform [AUC ROC]					
		Discord	Gab	Telegram	Twitter	WhatsApp	all
Aya-101	Discord	0.99	0.82	0.89	0.81	0.92	0.89
	Gab	0.97	0.96	0.95	0.96	0.94	0.96
	Telegram	0.98	0.92	0.97	0.95	0.96	0.95
	Twitter	0.98	0.92	0.94	0.98	0.94	0.95
	WhatsApp	0.97	0.88	0.94	0.93	0.98	0.94
	all	0.98	0.95	0.97	0.97	0.97	0.97
BLOOMZ-3b	Discord	0.98	0.85	0.87	0.84	0.88	0.88
	Gab	0.93	0.92	0.89	0.91	0.89	0.91
	Telegram	0.97	0.90	0.94	0.92	0.93	0.93
	Twitter	0.95	0.89	0.88	0.98	0.87	0.91
	WhatsApp	0.96	0.90	0.91	0.92	0.97	0.93
	all	0.98	0.94	0.95	0.97	0.96	0.96
Falcon-rw-1b	Discord	0.98	0.80	0.87	0.82	0.85	0.86
	Gab	0.95	0.93	0.92	0.95	0.91	0.93
	Telegram	0.98	0.90	0.95	0.95	0.93	0.94
	Twitter	0.97	0.91	0.93	0.98	0.92	0.94
	WhatsApp	0.97	0.91	0.93	0.95	0.98	0.94
	all	0.97	0.91	0.94	0.96	0.94	0.94
Llama-3-8b	Discord	0.99	0.83	0.82	0.72	0.86	0.83
	Gab	0.95	0.96	0.93	0.97	0.90	0.94
	Telegram	0.98	0.94	0.97	0.97	0.96	0.97
	Twitter	0.96	0.90	0.92	0.98	0.91	0.93
	WhatsApp	0.97	0.89	0.93	0.93	0.98	0.94
	all	0.98	0.94	0.95	0.97	0.95	0.96
Mistral-7b-v0.1	Discord	0.98	0.82	0.88	0.86	0.88	0.88
	Gab	0.95	0.95	0.93	0.96	0.92	0.94
	Telegram	0.98	0.94	0.97	0.97	0.95	0.96
	Twitter	0.96	0.90	0.90	0.98	0.91	0.93
	WhatsApp	0.96	0.90	0.92	0.94	0.97	0.93
	all	0.96	0.92	0.93	0.96	0.94	0.94
XLM-RoBERTa-large	Discord	0.99	0.87	0.90	0.83	0.94	0.90
	Gab	0.97	0.93	0.92	0.90	0.90	0.92
	Telegram	0.98	0.92	0.96	0.93	0.95	0.95
	Twitter	0.98	0.93	0.94	0.97	0.94	0.95
	WhatsApp	0.97	0.88	0.92	0.88	0.96	0.92
	all	0.98	0.94	0.96	0.96	0.96	0.96
mDeBERTa-v3-base	Discord	0.99	0.86	0.89	0.84	0.93	0.89
	Gab	0.97	0.93	0.93	0.92	0.94	0.94
	Telegram	0.98	0.90	0.95	0.93	0.95	0.94
	Twitter	0.97	0.90	0.92	0.97	0.94	0.94
	WhatsApp	0.98	0.91	0.93	0.90	0.97	0.93
	all	0.98	0.91	0.94	0.95	0.95	0.95

Table 30: Cross-platform evaluation of the selected fine-tuned MGT detectors.