Understanding and Improving Limitations of Multilingual AI Text Detection

Anonymous EMNLP submission

Abstract

*With the advances in multilingual large language models (LLMs), recent research has embarked on investigating diverse approaches towards multilingual AI-generated text (AI text) detection, including the fine-tuning of monolingual detectors. In this paper, we pinpoint the limitations in the evaluation procedures of current multilingual AI text detection. Our extensive analysis uncovers significant inadequacies in all of the available multilingual datasets, including **(i)** a primary focus on a limited set of languages, **(ii)** imbalanced data distribution between human and AI-generated samples, and **(iii)** a lack of diverse yet rich data collection sources. Amidst these challenges, we propose new methods to **(a)** improve cross-lingual transfer, **(b)** exploit novel fine-tuning strategies, **(c)** analyze the complexities of using neural machine translation (NMT) with monolingual detectors, and **(d)** a detailed analysis on adversarial robustness. Our results facilitate the engineering of a more resilient model for multilingual text detection, demonstrating superior performance and adaptability across a spectrum of languages.*

1 Introduction

Recent advances in natural language processing have led to the creation of powerful large language models (LLMs) like GPT-4 (Achiam et al., 2023), LLaMA-2 (Touvron et al., 2023), etc., enabling the development of technologies such as chatbots and writing assistants. However, the ability of LLMs to imitate human language patterns also presents a risk of misuse, including the generation of deceptive AI-generated text that can undermine trust in information sources and disrupt online discussions (Macko et al., 2023).

Models like T5 (Raffel et al., 2020) and DetectGPT (Mitchell et al., 2023) identify fake news and AI-generated text in English. Yet, the dominance of English in LLMs has evolved with Neural Machine Translation (NMT), now supporting over 200 languages. However, detecting AI-generated text in multilingual contexts poses a significant challenge due to linguistic complexities and a lack of resources in the multilingual domain. Although the success of NMT encourages us to examine whether integrating NMT with English detectors could be deemed effective in handling multilingual text detection, the outcomes were unrewarding (refer to Figure 3). In contrast, researchers aim to fine-tune detectors for only a few languages (Spanish, Russian, & English in MULTITuDE (Macko et al., 2023); Chinese, Urdu, Bulgarian, English, & Indonesian in SemEval (Wang et al., 2024)), hence relying on zero-shot transfer for other languages. However, due to the lack of comprehensive multilingual datasets, initial efforts focused on available datasets and questioned their limitations and inadequacies. Moreover, we observe 4 major flaws that are attributed to the state-of-the-art text detectors:

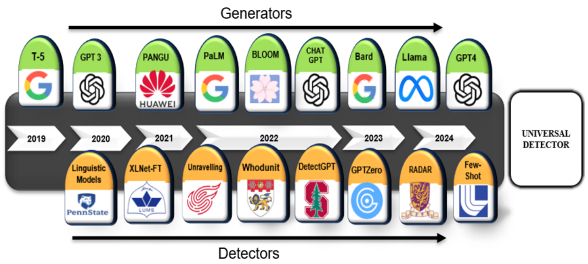

Figure 1: Chronology of AI-text generators and detectors.

(1) ***Sensitive to translations:*** When AI-generated texts in other languages are translated into English using various translators (Tiedemann and Thottingal, 2020; Fan et al., 2021; Zhang et al., 2020), they can evade detection as most of the recent works as translators are trained Neural Networks (NNs) which can eventually be treated as an AI-generated text.

(2) ***Unavailability of cross-linguality:*** Currently available English AI text detectors lack support for detecting languages other than English, resulting in erratic results when applied to non-English texts

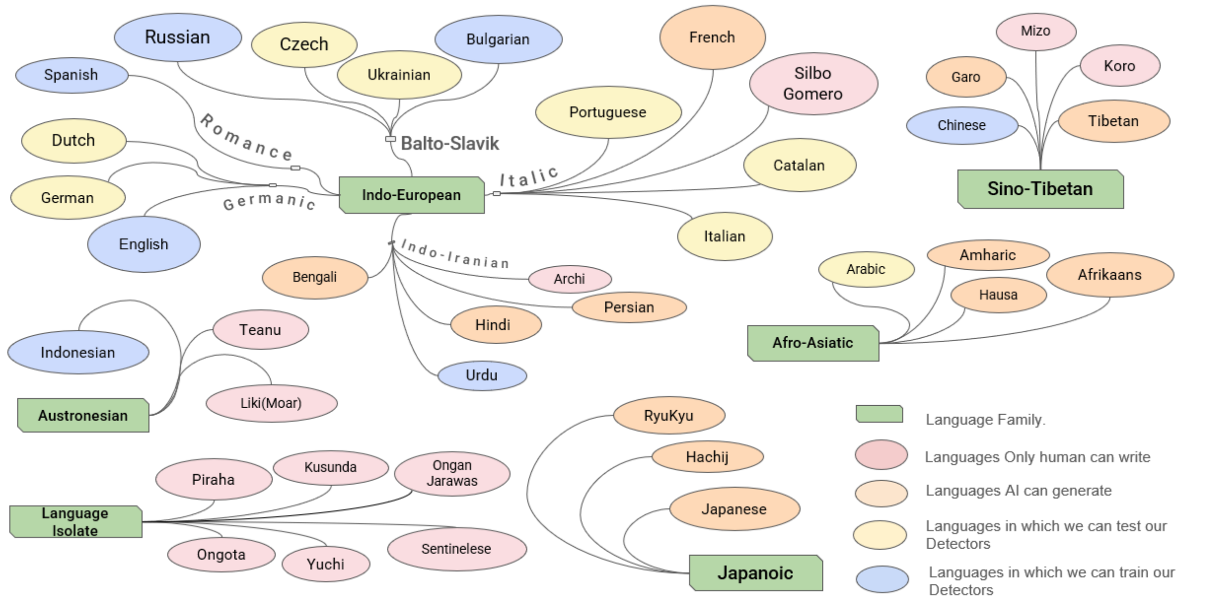

Figure 2: Highlighting the necessity for uniform detectors, reflecting the expanding multilingual capabilities of humans, AI generators, and AI detectors. Advances in society and AI are erasing language barriers, as globalization and urbanization draw people closer.

such as German, Hindi, Russian, etc (Hu et al., 2023; Macko et al., 2023).

(3) ***Sensitive to various writing forms:*** Texts containing poetic elements, personal views, summaries, drama scripts, conversations, and first-person opinions can successfully evade detection (Dugan et al., 2024).

(4) ***Sensitive to dialects:*** Texts written in various English dialects significantly decrease the detector's performance.

Notably, training a detector in an adversarial manner (such as RADAR (Hu et al., 2023)) can enhance models' ability to differentiate between authentic and AI-generated multilingual text, improving detection accuracy, particularly in the realm of paraphrasing, and consequently challenging the generator's capabilities. However, training a model in such a setting (from scratch) requires huge chunks of data (Hu et al., 2023). Researchers have shown that models can be transferred from pre-trained monolingual to multilingual domains through fine-tuning with a much smaller amount of data (Macko et al., 2023). In (Minixhofer et al., 2024), the authors explored the zero-shot transfer capabilities of tokenizers to enable them to process multilingual text. In light of the above facts, we aim to fine-tune RADAR using multi-lingual texts inspired by (Macko et al., 2023) work. The advancement of **mRADAR** (multi-lingual RADAR) is attributed to several improvements against various adversarial robustness analyses (Macko et al., 2024) such as *(i) translation & back-translation*, *(ii) paraphrasing*, and *(iii) back-translation after paraphrasing*. Our key contributions are as follows:

❶ ***Cross-Lingual Transfer Learning:*** We have successfully paved a path to transfer RADAR (Hu et al., 2023) into multilingual settings (*i.e.* **mRADAR**), showcasing its effectiveness and versatility in detecting AI-text across diverse linguistic landscapes. We first conducted extensive analysis on two state-of-the-art multi-lingual datasets.

❷ ***Detailed Analysis on Adversarial Robustness:*** Following the (Macko et al., 2024) work, we introduce two more robustness analyses: (i) translation and (ii) back-translation after paraphrasing. We are the first one to showcase the superiority of models fine-tuned with an adversarial approach across four different robustness aspects compared to state-of-the-art text detectors in multilingual scenarios.

❸ ***Complexities in using NMT with Monolingual Detectors:*** We highlight the limitations of current detection methods and the need to consider translators as a distinct class to reduce detection ambiguities.

2 Related Work

AI-Generated Text Detectors: Prior works in machine-generated text (MGT) detections can be broadly categorized into two sections: *(i) statistical models* and *(ii) fine-tuned models* (Macko et al., 2024). Statistical MGT detection models typically leverage pre-trained LLMs like GPT-2 (Radford et al., 2019) or mGPT (Shliazhko et al., 2024) without further fine-tuning to differentiate AI-generated text by employing metrics such as entropy (Lavergne et al., 2008), rank (Gehrmann et al., 2019), and perplexity. Prominent examples include GLTR (Gehrmann et al., 2019) and DetectGPT (Mitchell et al., 2023).

In contrast, several pre-trained models are available for MGT detection, including RoBERTa-base-OpenAI (Solaiman et al., 2019), RADAR (Hu et al., 2023) which can be used directly in a zero-shot manner, though they are mostly monolingual. Multilingual models like XLM-RoBERTa (Conneau et al., 2019), BERT-base-Multilingual-Cased (Devlin et al., 2019), and mDeBERTa (He et al., 2022) can be fine-tuned on custom datasets for multilingual detection. In recent, authors of (Macko et al., 2023) have beautifully presented a comprehensive multilingual benchmark of a range of detection methods along with a novel multilingual bench-marking dataset, MULTITuDE. Furthermore, SemEval-2024 (Wang et al., 2024) detection competition has made significant strides in multilingual text detection, effectively addressing critical challenges by mitigating class imbalances and dataset biases. Here, our proposed mRADAR facilitates comprehensive evaluation and benchmarking in this field in context of different robustness analysis. These achievements emphasize the importance of continually innovating to keep up with the evolving AI-generated text in different languages and fields.

Robustness Analysis & Authorship Obfuscation To evaluate the adversarial robustness of AI-text detectors, (Macko et al., 2024) work have categorized several existing Authorship Obfuscation (AO) methods into: ***(i) Back-translation:*** It involves translating a text from one language to another and then translating it back to the original (*e.g.*, English → Hindi → English) (Almishari et al., 2014; Altakrori et al., 2022). Here, the resulting backtranslated version will differ subtly from the original, hence making accurate detection more challenging; ***(ii) Paraphrasing:*** It involves rewriting the text in the same language, unlike back-translation that involves translation into another language and back (Lu et al., 2023; Krishna et al., 2024; Sadasivan et al., 2023); and ***(iii) Attacks*** such as an syntactic attack – ALISON (Xing et al., 2024), lexical-based attacks (Pu et al., 2023), and for more information refer to (Macko et al., 2023). In this work, we have instructed two other AOs - (i) translation and (ii) back-translations after paraphrasing. Moreover, we conducted these analyses on two state-of-the-art multi-lingual datasets (*i.e.* SemEval 2024 (Wang et al., 2024) and Multitude (Macko et al., 2023)) in both the scenarios in-order and out-order distribution. Here, beyond analyzing all of these aspects, we have identified that detectors trained in an adversarial manner (with generators) *i.e.* **mRADAR** demonstrate remarkable capabilities in handling these obfuscations. Please refer to Table 3, Section 4.3, Section 4.4, and Figure 3.

3 Methodology

In this section, we discuss the objectives and methods behind our analysis. To begin our analysis, we initially gathered a variety of benchmarking models from MULTITuDE (Macko et al., 2023), RADAR (Hu et al., 2023), and RoBERTa-large (Liu et al., 2019). We have performed assessments on DetectGPT (Mitchell et al., 2023) and other statistical approaches (like rank, as well, but since our paper primarily emphasizes the transfer of monolingual and multilingual LLMs in the field of MGT, we have not included the results in Table 1 for clarity. However, the analysis of the models can be located in the appendix.

3.1 Fine-tuning of detectors

We primarily utilized MULTITuDE's methods and scripts for fine-tuning, but we modified hyperparameters and selected the 3 optimal hyperparameters for RADAR resulting in model versions 1, 2, and 3. Other models were fine-tuned using the same hyperparameters as well. More information can be found in the appendix, where all code for fine-tuning detectors has been provided. Table one presents a comparison between the fine-tuned RADAR versions and the original benchmarks up to our research time.

3.2 Objective of experimentation

We have significant concerns about the ideas that could lead us toward our objective of creating a universal detector, a state-of-the-art model capable of excelling in multilingual settings.

(a) *Will the models, pre-trained for specific detection tasks be able to retain their native properties if we were to finetune them?* This was a noteworthy topic of discussion as it questions even the reason-

Model	Finetuned?	MULTITuDE					SemEval				
		AUROC (↑)	FPR (↑)	TPR (↑)	TNR (↑)	FNR (↑)	AUROC (↑)	FPR (↑)	TPR (↑)	TNR (↑)	FNR (↑)
mDeBERTa*	✔	0.96	0.26	0.98	0.74	0.02	-	-	-	-	-
BERT-base*	✔	0.91	0.47	0.96	0.53	0.04	-	-	-	-	-
OpenAI-RoBERTa*	✔	0.86	0.43	0.94	0.57	0.06	-	-	-	-	-
XLM-RoBERTa*	✔	0.96	0.41	0.99	0.59	0.01	-	-	-	-	-
mDeBERTa	✔	0.83	0.98	0.81	0.014	0.19	0.00	0.50	0.00	0.50	0.00
BERT-base	✔	0.82	0. 97	0.82	0.03	0.11	0.24	0.50	0.40	0.50	0.60
OpenAI-RoBERTa	✔	0.86	0.97	0.84	0.03	0.16	0.91	0.71	0.36	0.29	0.64
XLM-RoBERTa	✔	0.81	0.98	0.82	0.02	0.18	0.56	0.29	0.51	0.71	0.49
RADAR	✗	0.64	0.05	0.17	0.95	0.83	0.39	0.50	0.32	0.50	0.68
RoBERTa-large**	✗	0.74	93.81	99.75	6.18	0.2	0.75	0.65	0.47	0.35	0.53
mRADAR	✔	0.95	0.98	0.86	0.02	0.14	0.91	0.61	0.30	0.39	0.70

Table 1: Performance of detection methods on two benchmark datasets. Here, models are finetuned and tested on same dataset. * Model's performance are taken from MULTITuDE (Macko et al., 2023) paper as it is and fine-tuned on the same script. **RoBERTa (Liu et al., 2019) is ambiguous as the model returns [0,1] for both human and AI *e.g.* (text is human with 0.99 probability with a threshold accuracy of 50%).

Model	Finetuned?	MULTITuDE → SemEval					SemEval → MULTITuDE				
		AUROC	FPR	TPR	TNR	FNR	AUROC	FPR	TPR	TNR	FNR
mDeBERTa	✔	0.94	0.70	0.20	0.30	0.80	0.00	0.89	0.00	0.11	1.00
BERT-base	✔	0.80	0.57	0.32	0.43	0.68	0.60	0.89	0.89	0.11	0.11
OpenAI-RoBERTa	✔	0.97	0.65	0.39	0.35	0.61	0.63	0.90	0.88	0.10	0.12
XLM-RoBERTa	✔	0.83	0.68	0.11	0.32	0.89	0.72	0.92	0.89	0.08	0.11
mRADAR	✔	0.88	0.71	0.37	0.29	0.63	0.56	0.89	0.88	0.11	0.12

Table 2: Performance of detection methods on two benchmark datasets. Here models are finetuned on one trained and tested on another dataset, for *e.g.* MULTITuDE → SemEval signifies that models are finetuned on MULTITuDE but tested on SemEval.

ing for fine-tuning. However, as seen in Table 3 and Table 5, we observe how well the models preserve the native properties.

(b) *Would there be a requirement for making the models multilingual, when we are already witnessing the rise of better translators and a variety of language translation bilingual support?* or whether adding a few layers might help us in handling multilingual texts? To tackle this we used NMT models provided by Helsinki-NLP's Opus-MT (Tiedemann and Thottingal, 2020) and performed the translations twice to check the impacts can be found in Figure 3.

(c) *Do these detectors work well in English (their main language) and in multilingual settings?* To address the absence of a multilingual paraphraser, we incorporated translator layers in both the input and output of the paraphraser. In our experiment in Figure 3 and table 4, we utilized Pegasus (Zhang et al., 2020) for paraphrasing. Given our understanding of how translation layers can distort samples, we stress the importance of further research on multilingual paraphrasers, to accurately assess model performance.

3.3 Evaluation metrics

Evaluating the models is a considerable challenge due to the potential for accuracy and **AUROC** to be deceptive. To address this, we rely heavily on the **confusion matrix** which provides **TPR** *(AI samples are identified as AI samples)* and **TNR** *(Human samples are identified as Human samples)* of the models. In situations where detecting AI and avoiding false accusations of plagiarism by humans *(as the scenario with most of the legal aspects)* is crucial, we consider the absolute variance between TPR and TNR alongside accuracy, and AUROC to select a well-rounded model instead of one that may be biased towards a skewed dataset. moreover, we use **Score -** predefined Scikit-learn accuracy score metric.

3.4 Multilingual Benchmark Dataset

To advance research in multilingual AI-generated text detection, effective multilingual detectors require benchmark datasets for training.

Multilingual datasets play a crucial role in training and evaluating models for detecting AI-generated text across different languages. However, upon closer examination of renowned datasets, we identified several flaws that hinder model generalization and effectiveness:

(a) ***Limited Language Coverage:*** Many datasets lack coverage of widely spoken languages, hindering model generalization. For example, the MULTITuDE dataset primarily focuses on English, Russian, and Spanish, limiting its applicability across diverse linguistic contexts. Similar issues are observed in datasets like SemEval-2024, where English comprises more than 65% of the dataset, thereby questioning its multilingualism.

(b) ***Imbalanced Data Distribution:*** Some datasets exhibit imbalances between human and AI-generated text samples, impacting model measurement and analysis. For instance, the MULTITuDE dataset has significantly more AI samples than human samples, leading to challenges in accurate model evaluation. In contrast, the SemEval dataset maintains a more balanced distribution.

(c) ***Single Source Bias:*** Reliance on a single data collection method, such as web scraping of news articles, introduces biases and limits dataset diversity. For example, the MULTITuDE dataset may suffer from biases inherent to the source platform, affecting model generalization. In contrast, SemEval-2024 Task 8 collects data from various sources like ArXiv and Wikipedia, enhancing dataset diversity. this is also explored by (Dugan et al., 2024)

(d) ***Quality of Data:*** While sample balance is crucial, the quality of text samples also impacts model performance. The MULTITuDE dataset benefits from higher-quality data sourced from news articles, ensuring a more consistent text corpus. However, SemEval's dataset includes noise from sources like Wikipedia, diminishing data quality and suitability for model fine-tuning.

Addressing these challenges is essential to improve the quality and effectiveness of multilingual text detection models. The issues may be linked to the datasets and are likely to continue until we establish a benchmark dataset.

4 Experiments

In our attempts to extend the monolingual model to the multilingual domain, we looked into numerous methodologies, which include fine-tuning as recommended by MULTITuDE, using adversarial training as indicated by RADAR, and using supervised learning akin to prior detectors. Due to the high expense of training multilingual detectors from scratch, our approach has centered on fine-tuning monolingual detectors to be able to cope with multilingual tasks efficiently.

RADAR, which is known for its robustness even after multiple exposures to paraphrasing (n-shots paraphrasing), serves as our foundational model. Hyperparameter tuning has been conducted to identify optimal parameters for RADAR over suggested methods, presented by MULTITuDE . We have fine-tuned models fine-tuned presented in MULTITuDE, OpenAI's RoBERTa, and RADAR itself, yielding conclusive evidence on the conversion of monolingual detectors into the multilingual domain. Currently, our focus has been on datasets like MULTITuDE and SemEval, given the limited availability of resources in this domain.

4.1 Performance of Benchmark Models

We have gathered models presented in MULTITuDE, where authors successfully fine-tuned models for the multilingual domain. Additionally, we included the RADAR Checkpoint and the RoBERTa Checkpoint to investigate their performance. After fine-tuning, we observed a drop in the AUROC score for RoBERTa, suggesting a potential fault in the fine-tuning method. However, when comparing the True Positive Rate (TPR), the RoBERTa model shows an improvement in identifying AI-generated samples, indicating that despite the AUROC drop, the model is becoming more effective in detecting AI content. The findings from our evaluation are as follows: **(a)** The performance of models fine-tuned from the MULTITuDE dataset exhibits a notable decline in accuracy across various datasets. (see Table 14 in Appendix). For instance, MDeBERTa (He et al., 2022) initially demonstrates a high accuracy score of 0.92 when evaluated within the confines of the MULTITuDE dataset. However, when tested on the SemEval dataset, its accuracy significantly drops to 0.52. This substantial decrease of 40 points indicates that MDeBERTa, despite its strong performance on the testing data, loses its performance on other datasets.

(b) Similar trends are observed with other models such as BERT-base (Devlin et al., 2019), which also show a marked decrease in performance when transitioning from MULTITuDE to SemEval. BERT-base's accuracy drops from 0.89 to 0.42, reflecting a reduction of 47 points.

(c) RADAR, in its current version, demonstrates significant difficulties in handling multilingual texts effectively. The AUROC scores for RADAR are notably low, further emphasizing its struggle to distinguish between human-written and AI-generated texts across different languages. RADAR's predictions, referred to as RADAR preds, exhibit discernible limitations.

(d) Three different versions of RADAR, based on the hyperparameters, fine-tuned on the MUL-

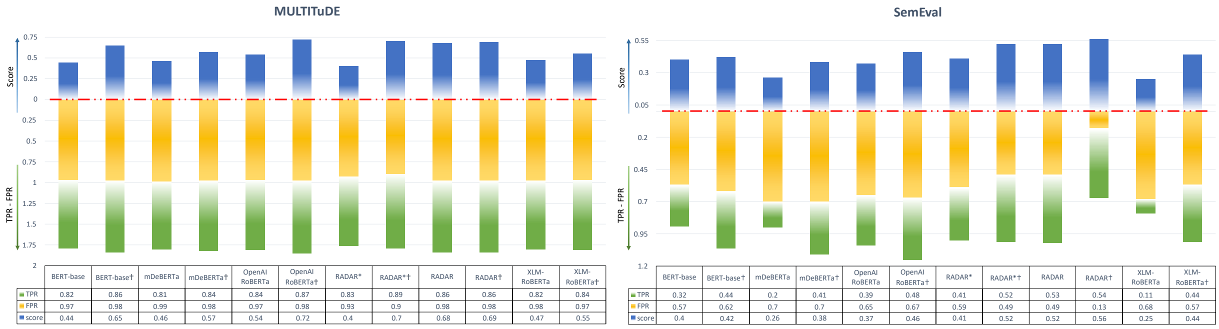

Figure 3: The effects of translation over state-of-the-art detectors on MULTITuDE and SemEval datasets. † Means translated * denotes model trained only on spanish.

TITuDE dataset were analyzed: RADAR-v1, RADAR-v2, and RADAR-v3. These versions consistently show a decrease in accuracy by 10-20 points when tested on external datasets like SemEval.

(e) The significant drops in performance across different models and versions highlight a crucial issue: models trained on the MULTITuDE dataset face substantial challenges in generalizing well to other datasets.

4.2 Model analysis with & w/o translations

Now if we focus on Figure 3, it presents the fine-tuned versions of various models across different datasets, both with and without translations. All models were fine-tuned under the same conditions as The variations in RADAR (v1, v2, v3). However, we have introduced RADAR-v4 and RADAR-Multi both of which are trained on the whole dataset, for more details see our appendix A1. Although, for readability we have reported only the best versions as RADAR- fine tuned. Our observations have the following conclusions:

(a) Models when evaluated on translated datasets exhibit higher accuracies but also demonstrate elevated False Positive Rates (FPR), erroneously labeling human-generated content as AI. This phenomenon may stem from the fact that current translation methods, such as Neural Machine Translation (NMT), also produce AI-generated text which increases the presence of LLM-generated data in a text sample. Consequently, the notion of incorporating a translator as the first layer in a detector, followed by a monolingual detector, is challenged. Although the concept of a bilingual translation approach utilizing over 200 languages seems promising for developing a universal detector, this conclusion underscores the complexities and limitations inherent in current detection methodologies. This is also proven by our table no. 14 in Appendix section A3, which shows 0 TNR and and high FPR and TPR, This result was performed on a non fine tuned monolingual benchmark model released by (Hu et al., 2023)

(b) Despite models showcasing impressive AUROC surpassing 95 within their training and testing environments, their performance significantly declines when evaluated on external datasets, with many models achieving accuracy scores below 40%. Even within the MULTITuDE dataset, the performance of these models remains unsatisfactory. This fragility raises concerns regarding the robustness and generalizability of these models. It's noteworthy to highlight discrepancies between metrics like AUROC and accuracy. While accuracy serves as a standard metric for comparison, AUROC presents a skewed perspective on model performance. These discrepancies may be attributed to dataset nuances. Additionally, providing accuracy scores alongside other metrics facilitates a more comprehensive evaluation of model performance, offering valuable insights for further analysis and comparison.

4.3 Performance after paraphrasing

As the models we have used should be investigated on paraphrasing to comment on their robustness, we generated paraphrased AI Samples but as multilingual paraphrasers are not available for this experiment we translated all the samples to English. Additionally, paraphrasing results for the base RADAR and RoBERTa can be found in the RADAR paper. The findings from our evaluation (presented in Table 3) are as follows:

(a) Many detectors experience a loss exceeding 60%, indicating their unsuitability for paraphrasing tasks. This substantial decrease underscores

Dataset	Model	Score	Acc. Drop over AI
MULTITuDE	BERT-base	0.79	0.21
	mDeBERTa	0.84	0.16
	RADAR-Multi	0.01	0.99
	OpenAI-RoBERTa	0.88	0.12
	RADAR-finetuned	1.00	0.00
	RADAR-es	0.96	0.04
	RADAR-Sem	0.95	0.05
	XLM-RoBERTa	0.83	0.17
SemEval	BERT-base	0.66	0.34
	mDeBERTa	0.82	0.18
	RADAR-Multi	0.00	1.00
	OpenAI-RoBERTa	0.84	0.16
	RADAR-finetuned	1.00	0.00
	RADAR-es	0.87	0.13
	RADAR-Sem	0.01	0.99
	XLM-RoBERTa	0.75	0.25

Table 3: Paraphrased Performance of Benchmark Models (Multi-lingual AI text → translate → English → Paraphrase by Pegasus.)

their inadequacy in accurately identifying and distinguishing between original and paraphrased texts. Such a significant drop in performance highlights the necessity for more robust detectors capable of preserving semantic meaning while detecting paraphrased content effectively. (Refer to Table 3.)

(b) Despite the absence of adversarial fine-tuning, RADAR demonstrates remarkable robustness compared to other models in the study. This resilience suggests that adversarial fine-tuning might not be indispensable for maintaining robustness in detectors. Moreover, it prompts us to ponder whether the properties exhibited by RADAR can be successfully transferred to the multilingual domain. This inquiry not only explores the potential for cross-domain applicability but also raises the overarching question: Can a universal detector, capable of accurately discerning between human-generated and AI-generated texts across various languages and contexts, truly exist?

4.4 Performance of back-translations

Table 4 contains results obtained after back-translation, which involves translating any presented language to English and then back again to the original language. This process was conducted to measure the effect of translation on texts. The observations from this evaluation are:

(a) While versions of RADAR exhibit higher AUROC values in the reported Table, it's prudent to overlook AUROC as it may create an illusion of robust performance in terms of TNR. Instead, a more comprehensive assessment involves comparing scores and both TNR and TPR pairs. Despite our models outperforming others in accuracy, all models here struggle with low TNR, likely influenced by the characteristics of the testing data itself.(refer to Table 4). Also if we have to choose the most optimal model to work upon, we believe we should not go with either accuracies or AUCROC instead a model which have a balanced TPR and TNR should be chosen (in this case RADAR v1, 6 point difference).

(b) This table reveals significantly lower TNR values, primarily attributable to the introduction of two layers of Neural Machine Translation (NMT). This intensified integration of AI translators likely contributes to the diminished TNR observed, especially evident in back-translated texts. This raises a pertinent question: Should we categorize translators as a distinct class? Given the prevalent use of NMT for translation purposes, distinguishing translators as a separate entity could alleviate ambiguity in detection methodologies.

Consider this scenario: a student, proficient only in Chinese, who relies on Neural Machine Translation (NMT) to translate their work into English. If traditional detection methods were used, in academic settings to identify AI-generated content, the student would likely be flagged erroneously. In our society, we acknowledge and credit individuals who translate texts across languages. Therefore, it's essential to consider this situation and ensure that due credit is given to NMT models for their role in enabling communication across linguistic barriers.

4.5 Performance on Back-translation after paraphrasing

As there were no multilingual paraphrasers available at the time of our research, we translated texts from the original language to English, used a paraphraser, and then translated them back to the original language. This method aims to mimic a multilingual paraphraser. However, as previously encountered, this process increases the presence of AI-generated elements, thereby reducing the effectiveness of the paraphrasing. We strongly emphasize the need to develop multilingual paraphrasers to test other benchmark models more accurately. RADAR also opens the way for such advancements.

(a) Despite experiencing a notable drop in accuracy

Model	(a) MULTITuDE				(b) SemEval			
	Acc. (↑)	AUROC (↑)	TPR (↑)	TNR (↑)	Acc. (↑)	AUROC (↑)	TPR (↑)	TNR (↑)
BERT-base	0.53	0.83	0.96	0.84	0.44	0.99	0.56	0.44
mDeBERTa	0.58	0.87	0.98	0.85	0.40	0.83	0.68	0.43
RADAR-Multi	0.11	0.30	0.89	0.92	0.50	0.00	0.50	0.00
OpenAI-RoBERTa	0.63	0.92	0.96	0.86	0.46	0.92	0.56	0.47
RADAR-finetuned	**0.73**	**0.97**	0.96	0.88	0.47	0.89	0.53	0.45
RADAR-es	0.49	0.67	0.92	0.86	0.43	0.99	0.57	0.44
RADAR-Sem	0.27	0.56	0.89	0.89	**0.50**	0.00	0.50	0.00
XLM-RoBERTa	0.61	0.87	0.98	0.85	0.42	0.90	0.64	0.44

Table 4: Analysis on Back-translations (Multi-lingual Human & AI samples → English → back-translate to original language).

Dataset	Model	Score	Para_Drop
MULTITuDE	BERT-base	0.64	0.36
	mDeBERTa	0.86	0.14
	RADAR-Multi	0.00	1.00
	OpenAI-RoBERTa	0.63	0.37
	RADAR-finetuned	**0.93**	0.07
	RADAR-es	0.22	0.78
	RADAR-Sem	0.36	0.64
	XLM-RoBERTa	0.89	0.11
SemEval	BERT-base	0.54	0.46
	mDeBERTa	0.84	0.16
	RADAR-Multi	0.00	1.00
	OpenAI-RoBERTa	0.61	0.39
	RADAR-finetuned	**0.84**	0.16
	RADAR-es	0.25	0.75
	RADAR-Sem	0.00	1.00
	XLM-RoBERTa	**0.87**	0.13

Table 5: Robustness analysis of multi-lingual detectors on back-translation after paraphrasing.

when tested on back-translated paraphrased texts, the RADAR model manages to maintain its ranking. While there is a decrease in accuracy, a modest 7% decline can still be considered a success. (refer to Table 5)

(b) We have successfully transferred the robust properties of the RADAR model without the need for adversarial fine-tuning. This achievement addresses our initial inquiry.

Additionally, we surpass the loss observed in models subjected to adversarial fine-tuning. This leads to a conclusive point regarding the approach to developing a universal detector. We propose training models without adversarial fine-tuning and then transferring them into the multilingual domain. This approach proves to be cost-effective, as it leverages existing models, such as RADAR. However, we encourage further exploration by researchers to investigate models trained multilingually from scratch with adversarial training. Nevertheless, such endeavors are beyond the scope of this paper. Moreover, it's important to note that the current datasets available in this domain may not meet benchmark standards, as previously mentioned. However, the improvement or suggestion of new datasets falls outside the scope of our study.

5 Conclusions

We have presented the following conclusions **(a)** Detectors can be finetuned in multilingual domains and yet can retain their properties as monolingual detectors **(b)** We have demonstrated that existing benchmarks lack robustness in the multilingual domain; however, monolingual models can achieve effectiveness through cross-lingual transfer **(c)** Our research has revealed the flaws in the current benchmark datasets for AI text detection,

6 Limitations

The primary focus of our work is more focused on understanding and experimenting with current benchmarks in the field, we have encountered flaws and reported them, and we have used different ways to evade the impacts of these flaws, however, addressing these issues falls outside the scope of this paper which includes absence of paraphrasers fluent in multiple languages, inadequate multilingual datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mishari Almishari, Ekin Oguz, and Gene Tsudik. 2014. Fighting authorship linkability with crowdsourcing. In *Proceedings of the second ACM conference on Online social networks*, pages 69–82.

Malik Altakrori, Thomas Scialom, Benjamin CM Fung, and Jackie Chi Kit Cheung. 2022. A multifaceted framework to evaluate evasion, content preservation, and misattribution in authorship obfuscation techniques. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2391–2406.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2405.07940*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

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*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ning Lu, Shengcai Liu, Rui He, Qi Wang, Yew-Soon Ong, and Ke Tang. 2023. Large language models can be guided to evade ai-generated text detection. *arXiv preprint arXiv:2305.10847*.

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. *arXiv preprint arXiv:2401.07867*.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. *arXiv preprint arXiv:2405.07883*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1613–1630. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana

Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

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*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. 2024. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*.

Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Alison: Fast and effective stylometric authorship obfuscation. In *AAAI*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

A Appendix

A.1 Dataset Details

For information regarding the Dataset we have used we are referencing the tables mentioned above from their respective authors.

A.2 Training Details

The majority of our experiments were conducted using `GeForce RTX 4090` GPU, totaling approximately 140 GPU hours of computation. The mRADAR (multi-lingual RADAR) are using three sets of hyperparameters, detailed below: Parameter 1: - Gradient size: 6 - Batch size: 32

Parameter 2: - Gradient size: 3 - Batch size: 64

Parameter 3: - Gradient size: 6 - Batch size: 64

The number of epochs for training can be adjusted based on the available GPU capacity, and we have implemented early stopping callbacks in the Macko et al script. Regarding the generator, we experimented with various models including Llama and text-davinci-003. However, our paper only includes details of models fine-tuned on GPT4 text samples. Finetuning on different models has minimal impact on accuracy, typically within the range of ±5 points. For paraphrasing, we used the Pegasus paraphraser, and similar results can be achieved using Dipper. However, we recommend fine-tuning mT5 for paraphrasing purposes to establish benchmarks across multilingual paraphrases. In terms of translation, we primarily utilized the Helsinki Opus MT translators. For languages not supported by Helsinki Opus MT, we employed Facebook m2m 100 base models.

A.3 Model Results

We have conducted several experiments to demonstrate the impact of translators on monolingual base models and statistical models. The table includes experiments on the base version of RADAR and shows a noticeable trend of increasing AUROC and FPR. We did not include statistical detectors in our main paper, and we compared the statistics of RADAR, RoBERTa large open AI detector, and statistical detectors on Wikipedia data. The results above also indicate the diminishing performance of the statistical models.

We have conducted several experiments to demonstrate the impact of translators on monolingual base models and statistical models. The table includes experiments on the base version of RADAR and shows a noticeable trend of increasing AUROC and FPR. We did not include statistical detectors in our main paper, and we compared the statistics of RADAR, Roberta large open AI detector, and statistical detectors on Wikipedia data. The results above also indicate the diminishing performance of the statistical models.

.

RADAR over MULTITuDE without Translation

RADAR					
Language	**AUROC**	TPR	FNR	TNR	FPR
German (de)	**0.67511**	0.207271	0.792729	0.917808	0.082192
English (en)	**0.885298**	0.432701	0.567299	0.949458	0.050542
Spanish (es)	**0.714209**	0.240803	0.759197	0.894366	0.105634
Dutch (nl)	**0.656536**	0.166528	0.833472	0.93311	0.06689
Portuguese (pt)	**0.691891**	0.191534	0.808466	0.898955	0.101045
Russian (ru)	**0.527453**	0.077183	0.922817	0.983333	0.016667
Chinese (zh)	**0.49706**	0.158204	0.841796	0.98	0.02
Arabian (ar)	**0.500888**	0.077085	0.922915	0.973244	0.026756
Ukranian (uk)	**0.541595**	0.064979	0.935021	0.979866	0.020134
Czech (cs)	**0.700578**	0.114274	0.885726	0.983333	0.016667
Catalan (ca)	**0.644315**	0.188624	0.811376	0.956667	0.043333
Average	0.6395	0.17	0.8255	0.95	0.049

Table 6

RADAR over MULTITuDE with Translation

RADAR					
Language	**AUROC**	TPR	FNR	TNR	FPR
German (de)	**77.53**	0.7684914333	0.2315085667	0.6780821918	0.3219178082
English (en)	**88.52**	0.432701	0.567299	0.949458	0.050542
Spanish (es)	**80.39**	0.7566889632	0.2433110368	0.7676056338	0.2323943662
Dutch (nl)	**78.46**	0.8101001669	0.1898998331	0.6588628763	0.3411371237
Portuguese (pt)	**77.05**	0.8675607712	0.1324392288	0.5888501742	0.4111498258
Russian (ru)	**67.19**	0.9578237031	0.04217629692	0.1533333333	0.8466666667
Chinese (zh)	**84.85**	0.9974821653	0.002517834662	0.07333333333	0.9266666667
Arabian (ar)	**76.16**	0.9915754002	0.008424599832	0.03344481605	0.9665551839
Ukranian (uk)	**55.64**	0.9772151899	0.02278481013	0.04026845638	0.9597315436
Czech (cs)	**76.71**	0.8304730013	0.1695269987	0.5133333333	0.4866666667
Catalan (ca)	**61.14**	0.9845253032	0.01547469678	0.03	0.97
Average	74.88	85.22	0.15	0.41	0.59

Table 7

Method :	LOGRANK	Entropy	LogP
Language	**AUROC**	**AUROC**	**AUROC**
German (de)	**18.76**	**27**	**19.34**
English (en)	**17.07**	**50.49**	**18.68**
Spanish (es)	**16.2**	**26.84**	**16.92**
Dutch (nl)	**12.95**	**23.34**	**13.91**
Portuguese (pt)	**20.94**	**30.17**	**21.84**
Russian (ru)	**34.47**	**40.5**	**34.58**
Chinese (zh)	**34.28**	**51.66**	**35.83**
Arabian (ar)	**29.7**	**35.43**	**28.86**
Ukranian (uk)	**32.66**	**36.47**	**31.95**
Czech (cs)	**18.62**	**27.27**	**19.19**
Catalan (ca)	**16.93**	**23.87**	**18.04**
Average	23	34	23.55

Table 8: table contains statistical method performance over MULTITuDE

Pipeline 1

		RADAR	RoBERTa	Logrank *	Logp	Entropy
German	TPR	53.7	98.7	62.7	58.6	49.3
	FPR	9.53	99.4	21.8	26.3	37.3
	FNR	46.3	1.3	31.9	41.4	50.7
	TNR	90.47	0.6	65.4	73.7	62.7
	AUROC	**84.3**	**42.74**	**34.54**	**65.95**	**56.8**
French	TPR	62.7	98.7	64.2	62.7	47.2
	FPR	14	99.7	24.3	28.1	27.5
	FNR	37.3	1.3	30.7	37.3	52.8
	TNR	86	0.3	57.8	71.9	72.5
	AUROC	**80.14**	**40.87**	**36.01**	**67.39**	**61**
Italian	TPR	56.34	98.3	36.16	34.14	22.47
	FPR	20.42	99.8	1.8	1.8	2.9
	FNR	43.65	1.6	57.94	65.83	77.52
	TNR	79.58	0.2	96.9	98.2	97.1
	AUROC	**79.21**	**38.98**	**36.86**	**65.88**	**59.28**

Table 9: Here Pipeline 1 refers to text detection without any translator

Pipeline 2

		RADAR	RoBERTa	Logrank *	Logp	Entropy
German	TPR	94.6	43.8	88.6	86.2	69.6
	FPR	33.68	76.3	56.3	56.9	33.2
	FNR	5.4	56.2	9.8	13.8	30.4
	TNR	66.315	23.7	35.1	43.1	66.8
	AUROC	**91.69**	**25.01**	**23.43**	**64.85**	**51.4**
French	TPR	95.9	46.6	91.2	88.9	75.6
	FPR	58.3	76.7	58.7	59.25	30.73
	FNR	4.1	53.4	7.3	11.1	24.4
	TNR	41.7	23.3	29.8	40.74	69.26
	AUROC	**86.23**	**25.89**	**20.52**	**65.57**	**53.25**
Italian	TPR	95.9	50.94	85.11	81.31	65.53
	FPR	56.41	81.5	48	47.9	54.9
	FNR	4.09	49.05	12.28	18.68	34.46
	TNR	43.58	18.5	43	52.1	45.1
	AUROC	**83.58**	**23.55**	**22.75**	**67.1**	**55.16**

Table 10: Results over Pipeline 2. (pipeline 2 refers to text detection with translators).

The above table shows the imbalance of the testing set in MULTITuDE samples.

Model	MDEBERTA	XLM-Roberta	BERT	Roberta
Metrics				
Accuracy	**92.88**	**93.11**	**82.36**	**89.42**
Total Human	**3,236**	**3,236**	**3,236**	**3,236**
Total AI	**26,059**	**26,059**	**26,059**	**26,059**
Predicted Humans	**1,717**	**1,427**	**194**	**200**
Predicted AI	**25,493**	**25,851**	**23,935**	**25,997**
AUROC	**92.32**	**91.025**	**47.55**	**73.67**
TPR	**97.82**	**99.2**	**91.84**	**99.76**
FPR	**46.94**	**55.9**	**94**	**93.81**
TNR	**53.05**	**44.09**	**5.99**	**6.18**
FNR	**2.17**	**0.79**	**8.1**	**0.2**

Table 11: Multitude model analysis Multitude on Mutitude-test set

The table above shows the data imbalance in training dataset.

Model	MDEBERTA	XLM-Roberta	BERT	Roberta
Metrics				
Accuracy	**93.68**	**95.01**	**83.55**	**89.93**
Total Human	**7,992**	**7,992**	**7,992**	**7,992**
Total AI	**66,089**	**66,089**	**66,089**	**66,089**
Predicted Humans	**5,072**	**4,807**	**383**	**634**
Predicted AI	**64,330**	**65,579**	**61,515**	**65,992**
AUROC	**92.98**	**94.46**	**45.82**	**78.68**
TPR	**97.33**	**99.22**	**93.07**	**99.85**
FPR	**36.53**	**39.85**	**95.2**	**92.06**
TNR	**63.46**	**60.14**	**4.79**	**7.9**
FNR	**2.66**	**0.77**	**6.9**	**0.14**

Table 12: Multitude model analysis Multitude on Mutitude-train set

Language	Model	score	AUROC	FPR	TPR	TNR	FNR
Arabic	radar	**0.61**	**0.89**	**0.97**	**0.85**	**0.03**	**0.15**
Catalan	radar	**0.82**	**0.82**	**1.00**	**0.88**	**0.00**	**0.12**
Czech	radar	**0.81**	**0.85**	**1.00**	**0.88**	**0.00**	**0.12**
German	radar	**0.68**	**0.94**	**0.99**	**0.86**	**0.01**	**0.14**
Spanish	radar	**0.42**	**0.80**	**1.00**	**0.80**	**0.00**	**0.20**
Dutch	radar	**0.68**	**0.95**	**1.00**	**0.86**	**0.00**	**0.14**
Russian	radar	**0.51**	**0.83**	**0.97**	**0.83**	**0.03**	**0.17**
Ukranian	radar	**0.67**	**0.93**	**0.98**	**0.86**	**0.02**	**0.14**
Chinese	radar	**0.87**	**0.60**	**0.98**	**0.89**	**0.02**	**0.11**

Table 13: Detailed analysis language wise of Radar without translation

Model	Train Dataset	Test - Dataset	score	AUROC	FPR	TPR	TNR	FNR
XLM-Roberta	MULTITuDE	MULTITuDE	0.47	0.81	0.98	0.82	0.02	0.18
Openai-Roberta	MULTITuDE	MULTITuDE	0.54	0.86	0.97	0.84	0.03	0.16
RADAR-Multi	MULTITuDE	MULTITuDE	0.11	0.28	0.89	1.00	0.11	NA
RADAR-v2	MULTITuDE	MULTITuDE	0.68	0.95	0.98	0.86	0.02	0.14
RADAR-v1	MULTITuDE	MULTITuDE	0.36	0.52	0.87	0.92	0.13	0.08
RADAR-v3	MULTITuDE	MULTITuDE	0.42	0.73	0.95	0.83	0.05	0.17
RADAR-v4	MULTITuDE	MULTITuDE	0.11	0.00	0.89	NA	0.11	NA
RADAR-v4	SemEval	MULTITuDE	0.28	0.59	0.89	0.87	0.11	0.13
RADAR-es	MULTITuDE(es)	MULTITuDE	0.40	0.62	0.93	0.83	0.07	0.17
Bert-base	MULTITuDE	MULTITuDE-tr	0.65	0.92	0.98	0.86	0.02	0.14
Mdeberta	MULTITuDE	MULTITuDE-tr	0.57	0.86	0.98	0.84	0.02	0.16
XLM-Roberta	MULTITuDE	MULTITuDE-tr	0.55	0.85	0.97	0.84	0.03	0.16
Openai-Roberta	MULTITuDE	MULTITuDE-tr	0.72	0.97	0.98	0.87	0.02	0.13
RADAR-Multi	MULTITuDE	MULTITuDE-tr	0.12	0.43	0.89	0.99	0.11	0.01
RADAR-v2	MULTITuDE	MULTITuDE-tr	0.69	0.95	0.98	0.86	0.02	0.14
RADAR-v1	MULTITuDE	MULTITuDE-tr	0.89	0.42	0.34	0.89	0.66	0.11
RADAR-v3	MULTITuDE	MULTITuDE-tr	0.71	0.96	0.96	0.87	0.04	0.13
RADAR-v4	MULTITuDE	MULTITuDE-tr	0.13	0.43	0.89	0.98	0.11	0.02
RADAR-v4	SemEval	MULTITuDE-tr	0.66	0.88	0.86	0.90	0.14	0.10
RADAR-es	MULTITuDE(es)	MULTITuDE-tr	0.70	0.94	0.90	0.89	0.10	0.11
Bert-base	MULTITuDE	SemEval	0.40	0.80	0.57	0.32	0.43	0.68
Mdeberta	MULTITuDE	SemEval	0.26	0.94	0.70	0.20	0.30	0.80
XLM-Roberta	MULTITuDE	SemEval	0.25	0.83	0.68	0.11	0.32	0.89
Openai-Roberta	MULTITuDE	SemEval	0.37	0.97	0.65	0.39	0.35	0.61
RADAR-Multi	MULTITuDE	SemEval	0.50	NA	0.50	NA	0.50	NA
RADAR-v2	MULTITuDE	SemEval	0.34	0.94	0.71	0.37	0.29	0.63
RADAR-v1	MULTITuDE	SemEval	0.52	0.93	0.49	0.53	0.51	0.47
RADAR-v3	MULTITuDE	SemEval	0.42	0.67	0.55	0.29	0.45	0.71
RADAR-v4	MULTITuDE	SemEval	0.50	NA	0.50	NA	0.50	NA
RADAR-v4	SemEval	SemEval	0.50	NA	0.50	NA	0.50	NA
RADAR-es	MULTITuDE(es)	SemEval	0.41	1.00	0.59	0.41	0.41	0.59
Bert-base	MULTITuDE	SemEval-tr	0.42	0.89	0.62	0.44	0.38	0.56
Mdeberta	MULTITuDE	SemEval-tr	0.38	0.87	0.70	0.41	0.30	0.59
XLM-Roberta	MULTITuDE	SemEval-tr	0.44	0.99	0.57	0.44	0.43	0.56
Openai-Roberta	MULTITuDE	SemEval-tr	0.46	0.67	0.67	0.48	0.33	0.52
RADAR-Multi	MULTITuDE	SemEval-tr	0.50	0.60	0.50	NA	0.50	1.00
RADAR-v2	MULTITuDE	SemEval-tr	0.39	0.76	0.77	0.43	0.23	0.57
RADAR-v1	MULTITuDE	SemEval-tr	0.56	0.63	0.13	0.54	0.87	0.46
RADAR-v3	MULTITuDE	SemEval-tr	0.41	0.80	0.68	0.44	0.32	0.56
RADAR-v4	MULTITuDE	SemEval-tr	0.50	0.58	0.50	NA	0.50	1.00
RADAR-v4	SemEval	SemEval-tr	0.50	0.58	0.50	NA	0.50	1.00
RADAR-es	MULTITuDE(es)	SemEval-tr	0.52	0.98	0.49	0.52	0.51	0.48
BERT-Base	SemEval	MULTITuDE	0.25	0.60	0.89	0.89	0.11	0.11
Mdeberta	SemEval	MULTITuDE	0.11	0.00	0.89	0.00	0.11	0.00
RADAR	SemEval	MULTITuDE	0.18	0.56	0.89	0.88	0.11	0.12
Openai-Roberta	SemEval	MULTITuDE	0.29	0.63	0.90	0.88	0.10	0.12
XLM-Roberta	SemEval	MULTITuDE	0.85	0.72	0.92	0.89	0.08	0.11
BERT-Base	SemEval	SemEval	0.50	0.24	0.50	0.40	0.50	0.60
Mdeberta	SemEval	SemEval	0.50	0.00	0.50	0.00	0.50	0.00
RADAR	SemEval	SemEval	0.36	0.91	0.61	0.30	0.39	0.70
Openai-Roberta	SemEval	SemEval	0.33	0.91	0.71	0.36	0.29	0.64
XLM-Roberta	SemEval	SemEval	0.51	0.56	0.29	0.51	0.71	0.49

Table 14: Detailed data of charts is given below

Language	Model	score	AUROC	FPR	TPR	TNR	FNR
Arabic	radar+tr	**0.73**	**0.96**	**0.96**	**0.87**	**0.04**	**0.13**
Catalan	radar+tr	**0.81**	**0.87**	**1.00**	**0.88**	**0.00**	**0.12**
Czech	radar+tr	**0.72**	**0.97**	**1.00**	**0.87**	**0.00**	**0.13**
German	radar+tr	**0.64**	**0.92**	**1.00**	**0.85**	**0.00**	**0.15**
Spanish	radar+tr	**0.51**	**0.84**	**0.99**	**0.83**	**0.01**	**0.17**
Dutch	radar+tr	**0.57**	**0.88**	**0.98**	**0.84**	**0.02**	**0.16**
Russian	radar+tr	**0.70**	**0.95**	**0.97**	**0.87**	**0.03**	**0.13**
Ukranian	radar+tr	**0.84**	**0.72**	**0.99**	**0.88**	**0.01**	**0.12**
Chinese	radar+tr	**0.66**	**0.93**	**0.99**	**0.86**	**0.01**	**0.14**

Table 15: Detailed analysis language wise of Radar with translation

Language	Model	score	AUROC	FPR	TPR	TNR	FNR
Arabic	Radar+backtranslation	**0.54**	**0.85**	**0.93**	**1.00**	**0.07**	**0.00**
Catalan	Radar+backtranslation	**0.88**	**0.55**	**1.00**	**1.00**	**0.00**	**0.00**
Czech	Radar+backtranslation	**0.88**	**0.55**	**1.00**	**1.00**	**0.00**	**0.00**
German	Radar+backtranslation	**0.80**	**0.93**	**0.99**	**1.00**	**0.01**	**0.00**
Spanish	Radar+backtranslation	**0.49**	**0.83**	**0.99**	**1.00**	**0.01**	**0.00**
Dutch	Radar+backtranslation	**0.84**	**0.73**	**0.99**	**1.00**	**0.01**	**0.00**
Russian	Radar+backtranslation	**0.59**	**0.88**	**0.95**	**1.00**	**0.05**	**0.00**
Ukranian	Radar+backtranslation	**0.82**	**0.82**	**0.93**	**1.00**	**0.07**	**0.00**

Table 16: Detailed analysis language wise of Radar with back-translation