Telescope: Discovering Multilingual LLM Generated Texts with Small Specialized Language Models

Héctor Cerezo-Costas[1,†], Pedro Alonso-Doval[1,†], Maximiliano Hormazábal-Lagos[1] and Aldan Creo[2,†]

[1] Fundacion Centro Tecnoloxico de Telecomunicacion de Galicia (GRADIANT), Spain
[2]Independent Researcher, Dublin, Ireland

Abstract

This paper introduces Telescope, a Machine-Generated Text (MGT) detection system developed for the IberAuTexTification challenge at IberLEF 2024. Our approach is an adaptation of Binoculars, a technique which involves the change ratio of perplexity and cross-perplexity using two closely related language models to quantify the level of surprise in word selection in the generation of a sentence for identifying MGT. This is supported by an iterative threshold selection process that balances false positives and false negatives. Enhancements include fine-tuning of pretrained linguistic models to improve performance in the minority languages present in the Iberian Peninsula. This approach obtained the best performance measured in a random split of the training used for testing in all the Iberian languages. Results extrapolate to other contexts and generation profiles, with our models finishing third of all participants in the final contest. Telescope demonstrated its robustness and efficacy across diverse linguistic contexts, with significant improvements in MGT detection over Iberian languages. These results highlight the potential of Telescope for enhancing content moderation strategies.

Keywords

Machine-Generated Text (MGT) Detection, Large Language Models, Natural Language Processing, Artificial Intelligence

1. Introduction

Machine-Generated Text (MGT) refers to natural language text produced, extended, or modified by machines, predominantly large language models (LLMs) [1]. Given the continuous advancement in the quality of text that language models are capable of generating and, consequently, the expansion of their application in content generation for various purposes, it is imperative to develop robust methods for detecting MGT across diverse contexts, languages, and domains. These methods are essential for developing effective content moderation strategies and are crucial in mitigating risks associated with MGT, such as disinformation, phishing, and spam [2].

This paper introduces Telescope, a system developed for the IberAuTexTification [3] shared task at IberLEF 2024 [4], which aims to develop models that exploit linguistic form and meaning cues to identify MGT and Human-Written Text (HWT) using a wide variety of models, domains, and languages from the Iberian Peninsula. Our approach primarily focuses on *Subtask_1*, which involves the binary classification of text records as either machine-generated or human-written.

Telescope is an adaptation of Binoculars [5], a technique that utilizes dual metrics: *perplexity* and *cross-perplexity* (a cross-entropy measure of two perplexities) combined in a metric called B, which represents the change ratio of both metrics. This approach quantifies the level of surprise in word selection during sentence generation from the perspective of two closely related language models, with

IberLEF 2024, September 2024, Valladolid, Spain
[†]These authors contributed equally.
hcerezo@gradiant.org (H. Cerezo-Costas); palonso@gradiant.org (P. Alonso-Doval); mhormazabal@gradiant.org (M. Hormazábal-Lagos); aldan.creo@rai.usc.es (A. Creo)
https://github.com/hmightypirate (H. Cerezo-Costas); https://github.com/PedroDoval (P. Alonso-Doval); https://github.com/maxhormazabal (M. Hormazábal-Lagos); https://acmc-website.web.app/intro (A. Creo)
0000-0003-2813-2462 (H. Cerezo-Costas); 0009-0000-8255-3466 (P. Alonso-Doval); 0009-0003-3687-1924 (M. Hormazábal-Lagos); 0000-0002-7401-5198 (A. Creo)

the hypothesis that this ratio will increase in the presence of HWT and decrease in the presence of MGT.

We employ open pretrained language models that achieve state-of-the-art results under optimal conditions; however, their performance degrades when applied to specific domains or languages. Given that this challenge involves languages from the Iberian Peninsula, some of which are considered minority languages, `Telescope` proposes to enhance detection accuracy through a fine-tuning phase to develop specialized models for defined language groups. This process involves adapting models to better comprehend the nuances and syntactic structures of minority languages, which are often underrepresented in mainstream datasets.

As the Binoculars approach provides an indicator of the likelihood of text being HWT, the selection of the decision threshold is critical to the performance of the detection system. To determine this threshold value, we employ an iterative process, experimenting with various threshold values to optimize the trade-off between false positives and false negatives, thereby achieving a higher accuracy rate in MGT identification.

By incorporating these advanced techniques and enhancements, we achieved third place in the *Subtask_1* ranking, demonstrating the effectiveness of our approach across diverse linguistic contexts. The remainder of this paper is organized as follows:

Section 2 provides an overview of current approaches to detect generated text. Section 3 describes the dataset utilized in our approach. Section 4 details the methodology employed to detect MGT texts. Section 5 presents the evaluation and model selection process. Section 6 reports the obtained results. Section 7 discusses the implications of these results. Finally, Section 8 concludes the paper and suggests directions for future research.

2. Background

Current approaches to detect machine-generated text can be categorized into several distinct groups:

- **Machine Learning Classifiers**: These methods employ traditional text classifiers trained on labeled HWT and MGT. This approach can involve fine-tuning pretrained models such as BERT or RoBERTa, or utilizing simpler techniques like logistic regression [6]. However, these methods often encounter challenges when confronted with texts generated by models more sophisticated than those used for detection [7], or when faced with overfitting to the domains and distributions of the training data [8].
- **Statistical Analysis**: These models calculate the probabilities of generated text using the output probability matrix of a model. Such approaches typically account for the common heuristics employed in text generation. Many of these techniques demonstrate superior performance in white-box analysis when the model generating the text to be detected is known a priori. However, their performance deteriorates significantly when the generating model is unknown. Numerous models in this category rely on the perplexity of the generated text [9], incorporating random perturbations with another model to measure probability curvature [10] or analyzing the log rank of tokens with and without perturbations [11].
- **Watermarking**: This category encompasses strategies that require unrestricted access to the text generation process. These methods involve altering output probabilities of the text at each step [12] or directly filtering them [13], thereby creating a hidden pattern in the output text that can be readily detected through statistical analysis. While watermarking is a promising approach, it can be susceptible to evasion through paraphrasing in certain configurations. To enhance the reliability of this scheme, more robust generation heuristics have been proposed [14, 15].
- **Detection via Rewriting**: This innovative method, as employed by some researchers [16], involves tasking an LLM with rewriting the text and subsequently analyzing the result. This approach leverages the tendency of LLMs to perceive MGT as higher-quality text, resulting in fewer modifications. This method offers several advantages, including the ability to utilize any LLM, regardless of its open or proprietary nature, and requires only the final output for analysis.

3. Dataset

The dataset employed in our approach comprises a diverse collection of texts in various languages from the Iberian Peninsula, provided by the IberAuTexTification shared task. The corpus encompasses texts in Spanish (ES), Galician (GL), Basque (EU), Catalan (CA), Portuguese (PT), and English (EN). In total, the dataset consists of $109,663$ texts released by the organization as training data.

We conducted a preliminary analysis of the dataset to ascertain the distribution of texts and the average text length per language. Table 1 presents a comprehensive overview of the text distribution and average text length for each language represented in the corpus. The languages have been detected using *langID* [17].

Table 1
Distribution and Average Text Length of the Training Corpus by Language

Language	Number of Texts	Average Length (characters)
Spanish	22,535	1,121
Catalan	16,372	836
Basque	13,444	874
Galician	11,625	1,006
Portuguese	19,733	1,051
English	25,954	1,196

The corpus was strategically partitioned into three segments. The largest segment, comprising $87,730$ texts, was utilized for model fine-tuning. Two smaller segments, each containing $10,000$ texts (hereafter referred to as $10K$), were employed for evaluation and testing purposes. Throughout this paper, these segments are designated as *eval-10K* and *test-10K*, respectively, to distinguish them from the authentic unlabeled test set released by the organization during the test phase of the IberAuTexTification task.

4. Methodology

4.1. Binoculars Detector

The evaluation of statistical metrics for the detection of MGT is predicated on measuring the presence of a statistical signature characteristic of both HWT and MGT. The underlying rationale is that language models generally adhere to a measurable probability distribution in their word selection during text generation. Specifically, perplexity is employed as a metric to quantify the degree of surprise generated when selecting words that compose a text, given its prior context.

A common assumption has been that when a language model analyzes the perplexity of a text, human-written content will yield higher scores than machine-generated text. Consequently, previous contributions have proposed identifying a threshold to differentiate between human and machine-generated text based on this measure. However, this assumption has been invalidated by the advent of larger, prompt-based models. Their outputs tend to exhibit perplexity metrics similar to human-generated texts due to their enhanced expressiveness. As a result, false positives are not uncommon when analyzing perplexity in isolation.

Binoculars proposes an alternative approach [5], based on a metric B, which represents the rate of the perplexity of an input text s of length L divided by a *cross-perplexity* metric that measures how surprising are the predictions on one model to another (a sort of cross-entropy). Binoculars employs two closely related language models, M_O (observer) and M_P (performer), executing the following steps:

- A *baseline-perplexity* (equation 1) of the input text is calculated. Complex prompts are expected to increase this value. Here $M_P(s_i)$ is the output probability of sentence token i obtained with the model.

$$logPPL_{M_P}(s) = -\frac{1}{L}\sum_{i=1}^{L} logM_P(s_i) \quad (1)$$

- Concurrently, *cross-entropy* is calculated (equation 2), measuring the degree of surprise in the text generation by the performer from the observer's perspective.

$$logxPPL_{M_P,M_O}(s) = -\frac{1}{L}\sum_{i=1}^{L} M_O(s_i) * logM_P(s_i) \quad (2)$$

- The ratio B between these two metrics is calculated (equation 3) and compared to a decision threshold. We adhere to the same order of performer/observer models as in the official Binoculars implementation [18], that differs slightly from the paper notation of the same authors.

$$B = ScoreM_P, M_O(s) = \frac{logPPL_{M_P}(s)}{logxPPL_{M_P,M_O}(s)} \quad (3)$$

The rationale for employing a ratio rather than absolute *baseline-perplexity* and *cross-perplexity* values is to mitigate the impact of complex prompts on perplexity scores. It is anticipated that both metrics will be correlated in MGT, but not in human texts. This implies that human B values are higher than MGT ones, justifying the use of the B ratio with a decision threshold.

In the original work, a specialized model for English was utilized, the FALCON 7B model [19]. Performance in other languages is degraded, although the authors claim the system maintains adequate performance across multiple contexts. They establish the reference threshold using a random split of the reference datasets used to train the system.

We propose two main improvements over the seminal Binoculars work. First, we introduce a fine-tuning step with only generative content to enhance system performance. Second, instead of using a single threshold to discriminate between human and generated content, we obtain multiple thresholds using non-random splits of the data. Consequently, different text sizes, languages, or contexts could potentially have different thresholds.

4.2. Model Fine-tuning

The adjustments implemented by Binoculars in MGT detection are related to the influence of the prompt on MGT perplexity values. Although Binoculars achieves state-of-the-art performance using pretrained models without the need for fine-tuning, this implies that the performance of these systems will be strictly related to the quality of pretraining of the selected models. This is why one of the most relevant drawbacks is the noticeable decrease in performance of this algorithm when dealing with different languages and domains, particularly with minority languages and specific domains.

Our proposed method utilizes the dataset presented in Section 3, separated by languages into groups that allow language models to specialize in minority languages (Basque and Galician) and in another group of more widely spoken languages such as Spanish, English, and Catalan, with a final group for Portuguese. Specifically, we fine-tuned the following models for each language:

- Basque [**`Latxa-7B`**]: A family of LLMs designed for Basque with 7 to 70 billion parameters. The model is based on Llama 2 and was trained on a corpus of 4.3M Basque documents. `Latxa` was released by a team from the HiTZ Center [20].
- Spanish, English, and Catalan [**`FLOR-1B3`**]: BLOOM architecture fine-tuned with 26 billion tokens of Catalan, Spanish, and English text, released by the Barcelona Supercomputing Center [21].
- Galician [**`Carballo-BLOOM-1B3`**]: Model based on `FLOR-1B3` pretrained with CorpusNOS [22], a massive Galician corpus with 2.1B words.

- Portuguese [**BLOOM-1B7**]: Multilingual model composed of over 1.7 billion parameters, supporting 45 natural and 12 programming languages. It was developed by the BigScience initiative using the Jean Zay Public Supercomputer in France [23]. We also experimented with `Aira-2` [24] but did not obtain superior results in our tests.

We utilized the HuggingFace Transformers library [25] to fine-tune the models using the PEFT [26] implementation of LORA, with hyperparameters $\alpha = 16$, dropout $= 0.1$ and rank $= 64$, to reduce training complexity. Each model except the Basque was fine-tuned for 250 steps with a batch size of 16, exclusively feeding randomly shuffled training examples in the relevant language. The model for Basque was fine-tuned for 500 steps with a batch size of 4. In this phase, only the generated data are used. The base model and the fine-tuned models are used as *performer* and *observer*, respectively, using the same strategy as Binoculars. With the exception of `Latxa-7B`, all models have a smaller size than the original `Falcon-7B` to accommodate training in the available hardware. Fine-tuned and base models share the same token dictionary, which is the main requirement for use in Binoculars.

Following the naming convention of Binoculars, we call our models `Telescope`, as they are more focused and powerful in specific scenarios but, in contrast, are less *portable* as they require extra models per language and computational steps for fine-tuning. To differentiate from the baseline that is closer to the Binoculars strategy, we use this name only to refer to the combination that uses the fine-tuned models.

4.3. Selection of a Classification Threshold

The `Telescope` strategy based on Binoculars allows us to obtain a numerical indicator to decide whether a text is HWT or MGT. However, as important as the calculation of the perplexity metrics is the determination of the threshold value to decide the boundary between values that will be classified as HWT or MGT. As mentioned in the methods section, the value of the B-ratio will increase if the text was generated by a human and decrease if it was written by machines.

The optimal threshold value for Binoculars is directly influenced by factors such as text length, communicative context, linguistic environment, and the combination of models selected as *performer* and *observer*, respectively. This is why correctly selecting the threshold value for Binoculars is crucial for accurately determining the classification of a text at its source of generation. As this is a learning process influenced by multiple factors, it is essential to search through a wide set of possible threshold values around their potential combinations.

Given a subset of data from the training corpus, the optimal threshold for this subset is obtained with the AUROC score, using the ground-truth labels. For example, Figure 1 and Figure 2 show the ROC curve for the Spanish texts in the training corpus. In these examples, the best threshold for this split is 0.76 and 0.90, respectively, using this combination of models.

Due to resource constraints, we limit the search for the optimum threshold to a small subset of text size splits (with a maximum of 10 splits), fixing the split by language and the subset of models used. To select the optimum threshold for the training corpus, we employed Optuna [27]. Here, we aim to maximize the Macro-F1 score in the *test-10k*, using the thresholds obtained in the *eval-10k*.

5. Evaluation and model selection

In this comparative analysis, we evaluate four model combinations. Two serve as zero-shot baselines: `FLOR-1B3/FLOR-1B3-instruct` and `BLOOM-1B7/BLOOMZ-1B7`. The third combination, which we designate as `Telescope`, utilizes the fine-tuned and base models, one model pair per language. At test time, language is detected to apply the corresponding model pair and threshold. Finally, we aggregate the scores of all models into a single output, termed `Ensemble`, assuming equal contributions from all combinations.

It is crucial to note that the Binoculars score is not symmetrical, and the designation of a model as *performer* or *observer* significantly impacts the result. Table 2 presents a preliminary study used

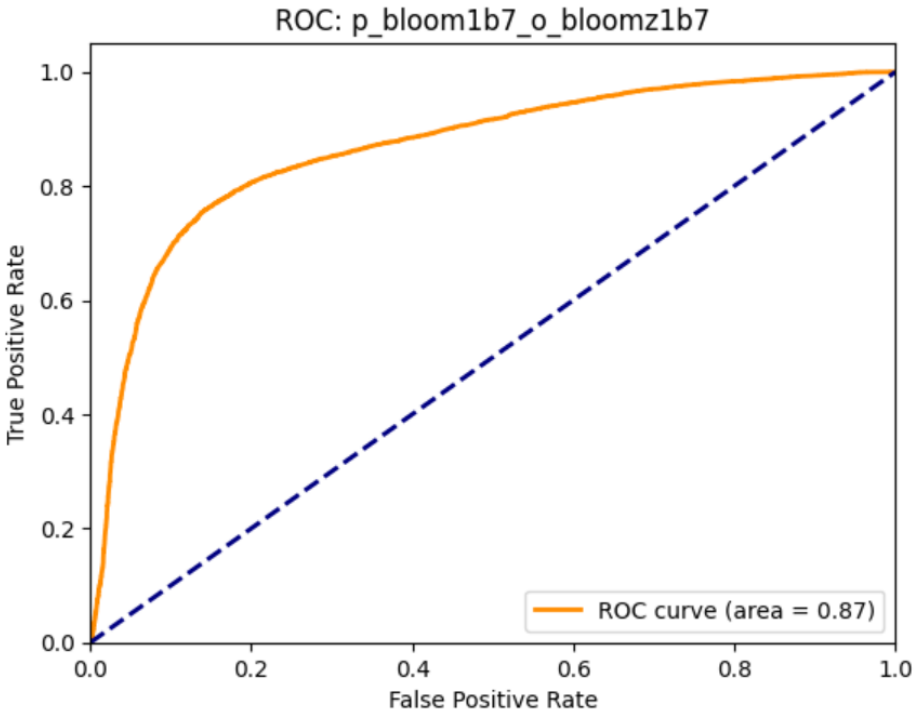

Figure 1: ROC for the Spanish texts of the train corpora with Binoculars using `BLOOM-1B7` as performer and `BLOOMZ-1B7` as observer.

to determine the optimal combination of models for `Telescope`. The results demonstrate that the *fine-tuned* models consistently perform better as *performers* across all language and model combinations for this task.

Table 2
AUROC in the *eval-10K* for different model pairs. Fine-tuned versions have the PEFT suffix. Combinations marked with * were finally selected for `Telescope`.

Performer	Observer	Language	AUROC
`Carballo-1B3`	`Carballo-1B3-PEFT`	GL	0.71
`Carballo-1B3-PEFT`	`Carballo-1B3`	GL	**0.81***
`FLOR-1B3`	`FLOR-1B3-PEFT`	ES	0.84
`FLOR-1B3-PEFT`	`FLOR-1B3`	ES	**0.90***
`FLOR-1B3`	`FLOR-1B3-PEFT`	EN	0.83
`FLOR-1B3-PEFT`	`FLOR-1B3`	EN	**0.89***
`FLOR-1B3`	`FLOR-1B3-PEFT`	CA	0.72
`FLOR-1B3-PEFT`	`FLOR-1B3`	CA	**0.82***
`Aira-2`	`Aira-2-PEFT`	PT	0.72
`Aira-2-PEFT`	`Aira-2`	PT	0.75
`BLOOM-1B7`	`BLOOM-1B7-PEFT`	PT	0.79
`BLOOM-1B7-PEFT`	`BLOOM-1B7`	PT	**0.85***
`BLOOM-1B7`	`BLOOM-1B7-PEFT`	EU	0.61
`BLOOM-1B7-PEFT`	`BLOOM-1B7`	EU	**0.85**
`Latxa-7B`	`Latxa-7B-PEFT`	EU	0.43
`Latxa-7B-PEFT`	`Latxa-7B`	EU	0.79*

For each case (excluding the baselines), thresholds are derived from the *eval-10K* dataset and applied to the *test-10K* dataset to compute the Macro-F1 score. The baselines, being zero-shot, utilize the training split and *eval-10k* to determine the optimal thresholds.

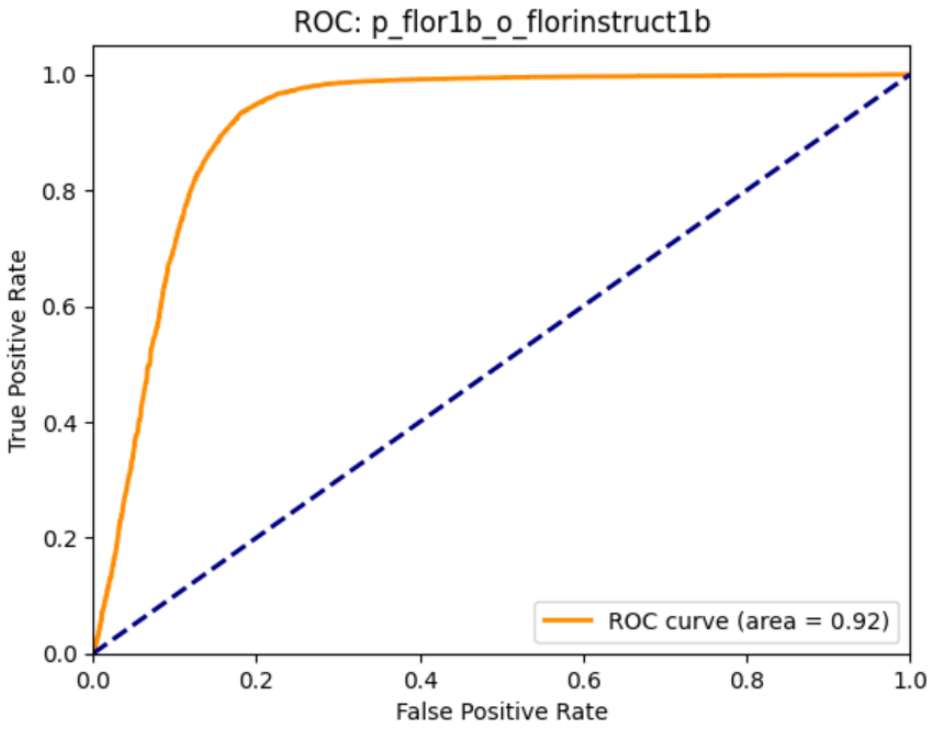

Figure 2: ROC for the Spanish texts of the train corpora with Binoculars using `FLOR-1B3` as performer and `FLOR-1B3-instruct` as observer.

Table 3 illustrates the splits by text size and language, along with the optimal thresholds in the *eval-10k* for the `FLOR-1B3` and `FLOR-1B3-instruct` model combination. Similar analyses are conducted for the other model combinations, as shown in Tables 4 and 5. Notably, the optimal threshold varies considerably from smaller to larger text sizes, although no clear trend is discernible.

Table 3
Optimum threshold for each split using using `FLOR-1B3` as performer and `FLOR-1B3-instruct` as observer.

Text length (in tokens)	GL	ES	EN	CA	PT	EU
$\langle 0, 35 \rangle$	1.14	0.98	1.06	0.97	0.95	1.12
$\langle 35, 166 \rangle$	0.97	0.94	0.92	0.87	0.92	1.01
$\langle 166, 405 \rangle$	0.93	0.87	0.91	0.84	0.89	0.97
$\langle 405, max \rangle$	0.91	0.87	0.91	0.80	0.92	0.99

Table 4
Optimum threshold for each split using using `BLOOM-1B7` as performer and `BLOOMZ-1B7` as observer.

Text length (in tokens)	GL	ES	EN	CA	PT	EU
$\langle 0, 35 \rangle$	0.89	0.79	0.86	0.54	0.47	0.76
$\langle 35, 166 \rangle$	0.80	0.68	0.86	0.54	0.55	0.72
$\langle 166, 405 \rangle$	0.82	0.76	0.86	0.60	0.63	0.50
$\langle 405, max \rangle$	0.81	0.80	0.87	0.66	0.70	0.58

Table 5
Optimum threshold for each split using the Telescope models (fine-tuned/pretrained) per language.

Text length (in tokens)	GL	ES	EN	CA	PT	EU
$\langle 0, 35 \rangle$	0.81	0.91	0.72	0.74	0.69	0.86
$\langle 35, 166 \rangle$	0.74	0.67	0.67	0.66	0.69	0.81
$\langle 166, 405 \rangle$	0.81	0.75	0.75	0.69	0.84	0.76
$\langle 405, max \rangle$	0.80	0.78	0.79	0.71	0.87	0.78

6. Results

To evaluate the performance of the various model combinations and their respective thresholds, we employ the Macro-F1 score on the random split of the training corpus, designated as *test-10K*. Table 6 summarizes these results. Upon initial examination, a significant disparity in performance across different languages is evident. The majority languages (English and Spanish) demonstrate comparatively superior results relative to the minority languages, with Basque exhibiting the lowest score overall. Surprisingly, certain models that purportedly support a given language, such as BLOOM-1B7, yielded lower scores than other model combinations that do not explicitly support it, such as FLOR-1B3.

When a uniform threshold is applied across all languages, as in the original Binoculars paper, the system's performance decreases substantially. Implementing language-specific thresholds and incorporating thresholds based on text length enhances the performance of every model pair configuration. Telescope demonstrates robust performance even with a shared threshold, at least when the training and test data share the same context and model generation techniques.

The utilization of fine-tuned models improves metrics across all languages. However, it is important to note that some biases may be introduced, as we are using generated data from a different split of the training corpus that follows the same distribution as the test data (e.g., context and methods of generation). Nevertheless, the fine-tuning step appears to be particularly beneficial for minority languages that lack substantial support in terms of the number of models and training data used to pretrain the multilingual models.

Table 6
Macro-F1 score per language for different model pairs and threshold selection strategies in the *test-10k*. The configurations marked with an asterisk (*) are those submitted to the IberAutextification 2024 competition. Our model achieved third place in this task [4].

Model Pairs	GL	ES	EN	CA	PT	EU	all
Shared threshold							
FLOR-1B3/FLOR-1B3-instruct	0.77	0.88	0.83	0.69	0.78	0.44	0.76
BLOOM-1B7/BLOOMZ-1B7	0.76	0.77	0.80	0.51	0.67	0.47	0.72
Telescope	0.81	0.90	0.87	0.79	0.80	0.81	0.84
Ensemble	0.77	0.89	0.83	0.72	0.84	0.78	0.82
Thresholds by language							
FLOR-1B3/FLOR-1B3-instruct	0.77	0.88	0.83	0.69	0.78	0.73	0.79
BLOOM-1B7/BLOOMZ-1B7	0.76	0.79	0.85	0.73	0.74	0.62	0.76
Telescope	0.82	0.90	0.87	0.82	0.82	0.82	0.85
Ensemble	0.85	0.89	0.88	0.79	0.84	0.77	0.85
Thresholds by language and size							
FLOR-1B3/FLOR-1B3-instruct *	0.80	0.88	0.85	0.76	0.78	0.72	0.81
BLOOM-1B7/BLOOMZ-1B7	0.79	0.81	0.86	0.74	0.78	0.64	0.78
Telescope *	0.84	**0.91**	**0.90**	**0.83**	**0.86**	**0.81**	**0.87**
Ensemble *	**0.85**	0.90	0.89	0.82	**0.86**	0.78	0.86

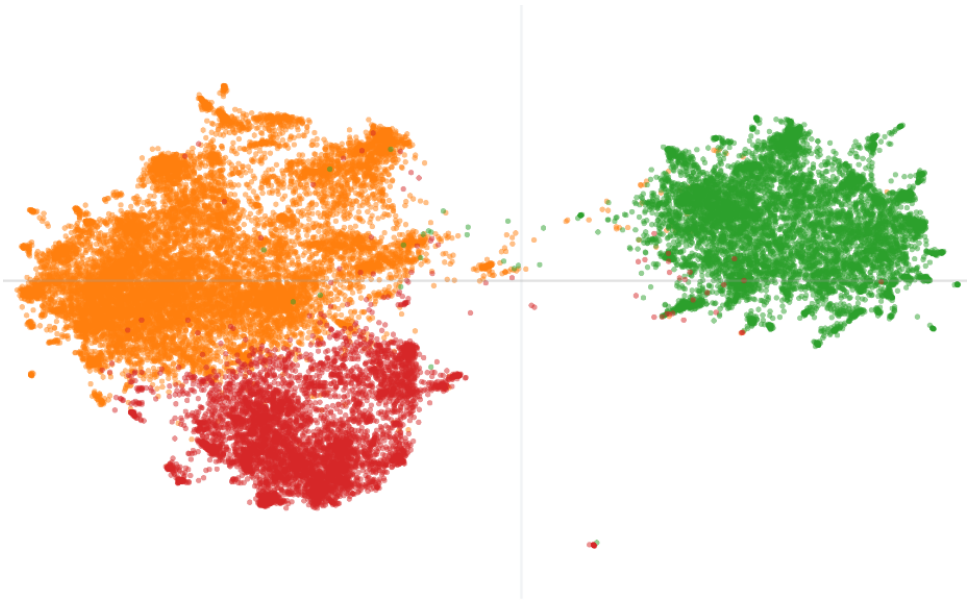

Figure 3: Clusters for a subset of Spanish records from the training data of AuTexTification 2023. The orange cluster primarily comprises tweets, while the red cluster contains *how-to* guides and the green cluster consists of legal and formal texts.

7. Discussion

7.1. Selection of Models

The selection of generative models for certain minority languages is constrained. Recent developments have seen the creation of models for Catalan and Galician based on the FLOR family of models (2023) and the release of Latxa for Basque (2024). However, multilingual models have not demonstrated optimal performance in these languages for text detection, likely due to the limited representation of these languages in the pretraining data. Indeed, models exclusively pretrained in a single language demonstrated superior performance in the majority of cases. Nevertheless, model selection was often driven by availability, as it was frequently the only option for a given language. Our implemented strategy requires two models whose tokenizers share the same vocabulary, further constraining the availability of models for the study.

A secondary limitation arose from the constrained computing resources available. All models (with the exception of Latxa) were fine-tuned on a single NVIDIA GeForce GTX 1080, with a hardware limit of 12GB of RAM. The performance of larger models using this strategy remains a prospective avenue for future research.

7.2. Generalization to Other Contexts

An intriguing question is the extent to which the implemented strategy generalizes to other contexts. To address this, we analyze the system's performance using the training data from Autextification 2023 [28] (utilizing only the Spanish records). In comparison with the 2024 data, the different domains of 2023 are more easily distinguishable.

The 2023 training data comprised three distinct contexts: tweets (orange), *how-to* guides (red), and legal texts (green). These contexts are represented in Figure 3. The clusters were obtained through a three-step pipeline: sentence embeddings vectorization with the `all-MiniLM-L12-v2` model, UMAP for dimensionality reduction, and K-Means with $k = 3$.

We compare the performance of the FLOR-1B3 baseline with that of the fine-tuned models across the three clusters using the area under the receiver operating characteristic curve (AUROC). Table 7

summarizes these results. It is evident that the baseline performs poorly in detecting legal and formal texts, despite these texts being longer in length. This is likely due to the inherent difficulty in detecting these texts using this strategy. The baseline zero-shot model performs no better than random, and the fine-tuned `Telescope` is marginally better but still far from optimal. As observed, fine-tuning the model with generated data appears to be beneficial in all contexts, despite the fact that the training data were obtained from different sources, topics, and generation strategies.

Table 7
AUROC scores of `FLOR-1B3/FLOR-1B3-instruct` and `Fine-tuned/FLOR-1B3` on 2023 training data. The optimal threshold for each context is shown in parentheses.

Cluster	`FLOR-1B3/FLOR-1B3-instruct`	`Telescope`
Tweets	0.82 (0.87)	0.84 (0.86)
How-to guides	0.86 (0.88)	0.89 (0.89)
Legal	0.51 (0.90)	0.66 (0.86)

We conducted a similar analysis with the 2024 data but did not observe substantial gains or drops in performance across the clusters. Consequently, the threshold was calculated only per language and text length.

7.3. Influence of Text Size

Generally, strategies for detecting MGT text tend to exhibit a positive correlation between accuracy and text length, a trend we also observe in our proposal. Figure 4 illustrates the AUROC calculated at text length intervals of 25 tokens for the Spanish and English records using the `FLOR-1B3/FLOR-1B3-instruct` model pairs.

A slight performance drop is observed between 100-300 tokens, which we hypothesize may be attributable to factors other than text size (e.g., the context of the texts or the generation strategy). For texts with fewer than 100 tokens, the proposed system operates far from its optimal working region. This is logical, as this strategy relies on metrics that are averaged over the text length (such as entropy and cross-entropy).

8. Conclusions

This paper presents and evaluates our `Telescope` approach for MGT detection across multiple languages and domains, with a particular focus on languages of the Iberian Peninsula. This work was conducted within the context of the IberAuTexTification task for the IberLEF Campaign, whose primary objective is to develop models that exploit linguistic form and meaning cues to identify automatically generated texts from a wide variety of models, domains, and languages.

Our primary contributions are threefold: *i)* Development of the `Telescope` approach for MGT detection based on the statistical signatures that current language models impart on the texts they generate; *ii)* Evaluation of the system's effectiveness across various contexts; *iii)* Analysis of the influence of factors such as text length and linguistic context on system performance.

These objectives were largely achieved, albeit with certain limitations and constraints detailed below. The `Telescope` system demonstrated efficacy in detecting generated texts across several languages, achieving competitive performance and securing third place in the IberAutextification 2024 competition.

To implement our approach, we fine-tuned a set of pretrained linguistic models for each language in the dataset. We then selected optimal thresholds for the scores based on a split of the language variety and text sizes using the *eval-10K* dataset. Finally, we evaluated the performance of different model combinations and threshold selection strategies on the *test-10K* dataset.

The results revealed notable variations in system performance according to language and text size. While Spanish and English texts exhibited satisfactory performance, Basque texts presented the greatest

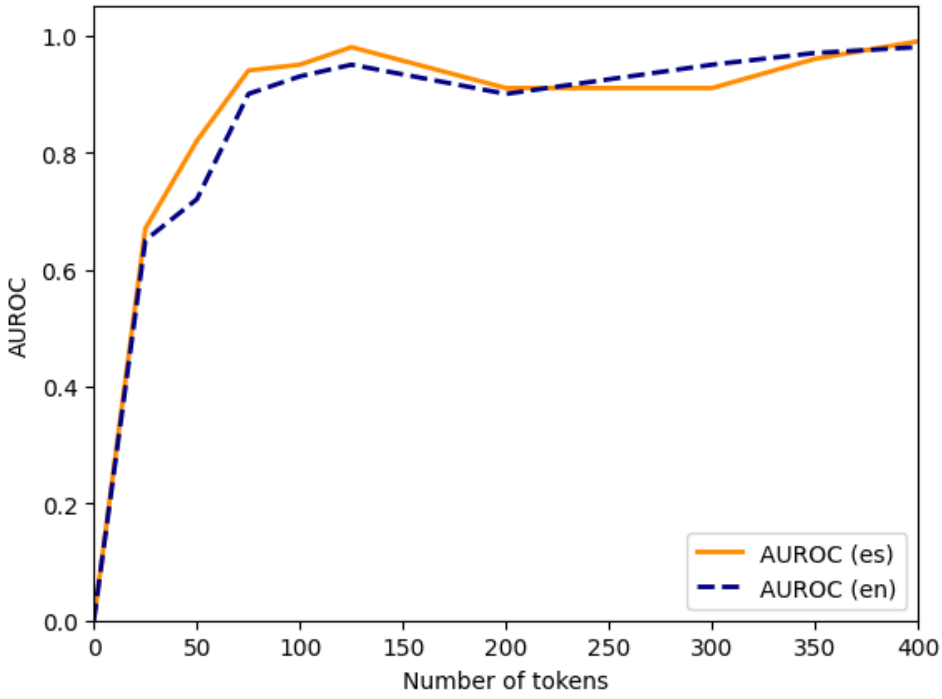

Figure 4: AUROC scores with `FLOR-1B3/FLOR-1B3-instruct` calculated at intervals of token size in the Spanish and English records of the training corpus.

challenges. This was attributed to the limited representation of this language in training corpora, coupled with its linguistic dissimilarity to the other studied languages. These findings underscore the limitations of multilingual models in minority languages that differ significantly in structure from more widely spoken languages. Furthermore, system performance generally improved with longer texts. However, performance drops were identified for text sizes between 100 and 300 tokens, which may be attributed to contextual factors rather than text size itself, affecting measures of surprise such as perplexity.

The most significant limitation was the limited availability of high-quality generative models for minority languages such as Catalan, Galician, and Basque. For these languages, often only one robust option was available, negatively impacting our ability to train and tune models for optimal performance or to conduct comprehensive benchmarking of minority language models. It should be noted that model selection is one of the factors affecting Binoculars performance. Additionally, computational resource constraints prevented us from exploring larger models, more complex architectures, or extending the range of tests for threshold values. This limited our training to a single NVIDIA GeForce GTX 1080 GPU.

We identify three key areas for future research: *i)* Augmentation of robust models for minority languages and expansion of training datasets to improve representation and performance in these languages. *ii)* Exploration of more advanced hardware and optimization techniques to enable the training of larger models. *iii)* Further investigation into how different contexts and textual domains affect system performance to improve generalization results to new areas.

The generalization of the implemented strategy to other contexts is limited and influenced by the aforementioned factors. Therefore, these three areas represent critical avenues for future research. Nevertheless, `Telescope` represents a significant contribution to the field of MGT detection, having yielded successful results within a shared task in the IberLEF campaign.

References

[1] E. Crothers, N. Japkowicz, H. L. Viktor, Machine-generated Text: A Comprehensive Survey of Threat Models and Detection Methods, IEEE Access (2023).

[2] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, S. Parkinson, A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats, IEEE Access 11 (2023) 144497–144529. doi:10.1109/ACCESS.2023.3344653.

[3] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of IberAuTexTification at IberLEF 2024: Detection and Attribution of Machine-Generated Text on Languages of the Iberian Peninsula, Procesamiento del Lenguaje Natural 73 (2024).

[4] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

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

[6] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection, arXiv preprint arXiv:2301.07597 (2023).

[7] G. Jawahar, M. Abdul-Mageed, L. V. Lakshmanan, Automatic Detection of Machine Generated Text: A Critical Survey, arXiv preprint arXiv:2011.01314 (2020).

[8] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, A. Szlam, Real or Fake? Learning to Discriminate Machine from Human Generated Text, arXiv preprint arXiv:1906.03351 (2019).

[9] S. Gehrmann, H. Strobelt, A. M. Rush, GLTR: Statistical Detection and Visualization of Generated Text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 111–116.

[10] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, in: International Conference on Machine Learning, PMLR, 2023, pp. 24950–24962.

[11] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text, arXiv preprint arXiv:2306.05540 (2023).

[12] X. Zhao, Y.-X. Wang, L. Li, Protecting Language Generation Models via Invisible Watermarking, in: International Conference on Machine Learning, PMLR, 2023, pp. 42187–42199.

[13] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A Watermark for Large Language Models, in: International Conference on Machine Learning, PMLR, 2023, pp. 17061–17084.

[14] J. Ren, H. Xu, Y. Liu, Y. Cui, S. Wang, D. Yin, J. Tang, A Robust Semantics-based Watermark for Large Language Model against Paraphrasing, 2024. arXiv:2311.08721.

[15] A. B. Hou, J. Zhang, T. He, Y. Wang, Y.-S. Chuang, H. Wang, L. Shen, B. V. Durme, D. Khashabi, Y. Tsvetkov, SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation, 2024. arXiv:2310.03991.

[16] C. Mao, C. Vondrick, H. Wang, J. Yang, Raidar: Generative AI Detection via Rewriting, arXiv preprint arXiv:2401.12970 (2024).

[17] M. Lui, T. Baldwin, Cross-Domain Feature Selection for Language Identification, in: Proceedings of 5th international joint conference on natural language processing, 2011, pp. 553–561.

[18] ????

[19] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, et al., The Falcon Series of Open Language Models, arXiv preprint arXiv:2311.16867 (2023).

[20] J. Etxaniz, O. Sainz, N. Perez, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An Open Language Model and Evaluation Suite for Basque, arXiv preprint arXiv:2403.20266 (2024).

[21] S. Da Dalt, J. Llop, I. Baucells, M. Pàmies, Y. Xu, A. González-Agirre, M. Villegas, FLOR: On the Effectiveness of Language Adaptation, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 7377–7388.

[22] I. de Dios-Flores, S. Paniagua Suárez, C. Carbajal Pérez, D. Bardanca Outeiriño, M. Garcia, P. Gamallo, CorpusNÓS: A Massive Galician Corpus for Training Large Language Models, 2024. URL: https://doi.org/10.5281/zenodo.11655219. doi:10.5281/zenodo.11655219.

[23] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176B-Parameter Open-Access Multilingual Language Model (2023).

[24] N. Kluge, Nkluge-correa/Aira-EXPERT: release v.01, 2022. URL: https://doi.org/10.5281/zenodo.6989727. doi:10.5281/zenodo.6989727.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[26] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods, https://github.com/huggingface/peft, 2022.

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[28] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of AutexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

Acronyms

CA Catalan. 3, 4, 6–9, 11

EN English. 3, 4, 6–8, 10, 11

ES Spanish. 3–11

EU Basque. 3–11

GL Galician. 3, 4, 6–9, 11

HWT Human-Written Text. 1–3, 5

MGT Machine-Generated Text. 1–5, 10, 11

PT Portuguese. 3–8