Beyond Turing: A Comparative Analysis of Approaches for Detecting Machine-Generated Text

Muhammad Farid Adilazuarda[*]
MBZUAI   Institut Teknologi Bandung
University of Zagreb, Faculty of Electrical Engineering and Computing
farid.adilazuarda@mbzuai.ac.ae

Abstract

Significant progress has been made on text generation by pre-trained language models (PLMs), yet distinguishing between human and machine-generated text poses an escalating challenge. This paper offers an in-depth evaluation of three distinct methods used to address this task: traditional shallow learning, Language Model (LM) fine-tuning, and Multilingual Model fine-tuning. These approaches are rigorously tested on a wide range of machine-generated texts, providing a benchmark of their competence in distinguishing between human-authored and machine-authored linguistic constructs. The results reveal considerable differences in performance across methods, thus emphasizing the continued need for advancement in this crucial area of NLP. This study offers valuable insights and paves the way for future research aimed at creating robust and highly discriminative models.

1 Introduction

The drive to discern between human and machine-generated text has been a long-standing pursuit, tracing its origins back to Turing's famous 'Turing Test', which explore a machine's ability to imitate human-like intelligence. With the vast and rapid development of advanced PLMs, the capacity to generate increasingly human-like text has grown, blurring the lines of detectability and bringing this research back into sharp focus.

Addressing this complexity, this paper explores two specific tasks: 1) the differentiation between human and machine-generated text, and 2) the identification of the specific language model that generated a given text. Our exploration extends beyond the traditional shallow learning techniques, exploring into the more robust methodologies of Language Model (LM) fine-tuning and Multilingual Model fine-tuning (Winata et al., 2021; Adilazuarda et al., 2023b; Radford et al., 2019). These

---

*Work conducted while visiting University of Zagreb.

techniques enable PLMs to specialize in the detection and categorization of machine-generated texts. They adapt pre-existing knowledge to the task at hand, effectively manage language-specific biases, and improve classification performance. Note that in this experiment, we do not use parameter-efficient strategies even when they have a superior specific-language capabilities. This is due to our constraint to fully fine-tune a language model and given the modular models' limited capabilities in such tasks (Adilazuarda et al., 2023a).

Through an exhaustive examination of a diverse set of machine-generated texts, Our paper offers the following contributions:

1. An exhaustive evaluation of the capabilities of PLMs in categorizing machine-generated texts.

2. An investigation into the effectiveness of employing multilingual techniques to mitigate language-specific biases in the detection of machine-generated text.

3. The application of a few-shot multilingual evaluation strategy to measure the adaptability of models in resource-limited scenarios.

2 Related Works

This study's related work falls into three main categories: machine-generated text detection, identification of specific PLMs, and advancements in language model fine-tuning.

Machine-generated Text Detection: Distinguishing human from machine-generated text has become an intricate challenge with recent advancements in language modeling. Prior research (Schwartz et al., 2018; Ippolito et al., 2020; Jawahar et al., 2020; He et al., 2024; Tian et al., 2023; Bhattacharjee and Liu, 2023; Koike et al., 2023; Yu et al., 2023) has explored nuances separating human and machine compositions. Our work builds

on these explorations by assessing various methodologies for this task.

Language Models Identification: Some studies (Antoun et al., 2023; Guo et al., 2023; Wu et al., 2023; Mitchell et al., 2023; Deng et al., 2023; Su et al., 2023; Li et al., 2023; Liu et al., 2023; Chen et al., 2023) attempt to identify the specific language model generating a text. These efforts, however, are still in growing stages and often rely on model-specific features. Our work evaluates various methods' efficacy for this task, focusing on robustness across a spectrum of PLMs.

Language Model Fine-tuning Advances: Language Model fine-tuning (Howard and Ruder, 2018) and Multilingual Model fine-tuning (Conneau et al., 2020) represent progress in language model customization. They enable model specialization in machine-generated text detection and classification and address language-specific biases, thereby enhancing classification accuracy across diverse languages.

This study intertwines these three research avenues, providing a thorough evaluation of the mentioned methodologies in machine-generated text detection and classification.

2.1 Dataset

Our experiments utilize two multi-class classification datasets, namely Subtask 1 and Subtask 2, as referenced from the publicly available Autextification dataset (Ángel González et al., 2023). Subtask 1 is a document-level dataset composed of **65,907** samples. Each sample is assigned one of two class labels: 'generated' or 'human'. Subtask 2, serves as a Model Attribution dataset consisting of **44,351** samples. This dataset includes six different labels - A, B, C, D, E, and F - representing distinct models of text generation. A detailed overview of the statistics related to both Subtask 1 and Subtask 2 datasets is provided in Table 1.

Language	Subtask	|Train|	|Valid|	|Test|	#Class
English	Subtask 1	27,414	3,046	3,385	2
	Subtask 2	18,156	2,018	2,242	6
Spanish	Subtask 1	25,969	2,886	3,207	2
	Subtask 2	17,766	1,975	2,194	6

Table 1: Statistics of the datasets.

3 Methods

3.1 Shallow Learning

We conducted an evaluation of two distinct shallow learning models, specifically Logistic Regression and XGBoost, utilizing Fasttext word embeddings that were trained on our preprocessed training set. FastText's subword representation captures fine morphological details. This is useful in detecting differences between the often overly formal structured machine-generated text and the morphologically rich human-generated text.

Prior to the training process, we implemented a fundamental preprocessing step involving non-ASCII and special characters removal. As showed in Table 2, we propose embedding on four lexical complexity measures aimed at quantifying different aspects of a text:

Average Word Length (AWL): This metric reflects the lexical sophistication of a text, with longer average word lengths potentially suggesting more complex language use. Let $W = \{w_1, w_2, ..., w_n\}$ represent the set of word tokens in the text. The AWL is given by:

$$AWL = \frac{1}{n} \sum_{i=1}^{n} |w_i|$$

Average Sentence Length (ASL): This measures syntactic complexity, with longer sentences often requiring more complex syntactic structures. Let $S = \{s_1, s_2, ..., s_m\}$ represent the set of sentence tokens in the text. The ASL is defined as:

$$ASL = \frac{1}{m} \sum_{j=1}^{m} |s_j|$$

Vocabulary Richness (VR): This ratio of unique words to the total number of words is a measure of lexical diversity. If UW represents the set of unique words in the text, the VR is calculated as:

$$VR = \frac{|UW|}{n}$$

Repetition Rate (RR): The ratio of words occurring more than once to the total number of words, indicative of the redundancy of a text. If RW represents the set of words that occur more than once, RR is computed as:

$$RR = \frac{|RW|}{n}$$

Table 2 presents a snapshot of our dataset after the application of our feature calculations. These include Average Word Length (**AWL**), Average Sentence Length (**ASL**), Vocabulary Richness (**VR**), and Repetition Rate (**RR**). By computing these features, we aimed to capture distinct textual characteristics that could aid our models in discriminating human and machine-generated text.

Text	Label	AWL	ASL	VR	RR
you need to.	generated	3.12	49.50	0.96	0.04
The Comm..	generated	4.92	62.56	0.69	0.09
I pass my...	human	3.55	90.00	0.90	0.10

Table 2: Text feature calculation. Label, AWL: Avg. Word Length, ASL: Avg. Sent. Length, VR: Vocab. Richness, RR: Repetition Rate

3.2 Language Model Finetuning

In this study, we employed multiple models: XLM-RoBERTa, mBERT, DeBERTa-v3, BERT-tiny, DistilBERT, RoBERTa-*Detector*, and ChatGPT-*Detector*. The models were fine-tuned on single and both languages simultaneously using multilingual training (Bai et al., 2021).

During evaluation, we employed the F1 score for our primary metrics. Furthermore, we incorporated a Few-Shot learning evaluation to assess our models' capacity to learn effectively from a limited set of examples for their practical applicability in real-world scenarios. This involved using varying seed quantities of `[200, 400, 600, 800, 1000]` instances, applied across both English and Spanish languages.

4 Experiments

Our approach to fine-tuning PLMs remained consistent across all models under consideration. We utilized HuggingFace's Transformers library[1], which provides both pre-trained models and scripts for fine-tuning. Utilizing a multi-GPU setup, we employed the AdamW optimizer (Loshchilov and Hutter, 2019), configured with a learning rate of `1e-6` and a batch size of 64. To prevent overfitting, we implemented early stopping within 3 epochs patience. The models were trained across a total of 10 epochs.

Multilingual Finetuning. An integral part of our approach was the models fine-tuning using En-

[1]https://huggingface.co/

glish and Spanish data to capture the unique linguistic features of each language.

Few-Shot Learning. To see the performance of the models in few-shot learning scenarios, employ few-shot learning experiments ranging from 200 to 1000 samples combination from the English and Spanish training data. The results of the few-shot learning experiments are depicted in Fig. 1.

5 Results and Discussion

5.1 Distinguishing Capability

From the few-shot learning experiments, the models' performance varied significantly in distinguishing between human and machine-generated text. In the default evaluation, multilingually-finetuned mBERT outperformed the other models in English, and single-language finetuned mBERT exhibited the highest score in Spanish. However, In the few-shot experiment setting, the RoBERTa-*Detector* demonstrated the most robust distinguishing capability, scoring up to 0.787 with 1000 samples.

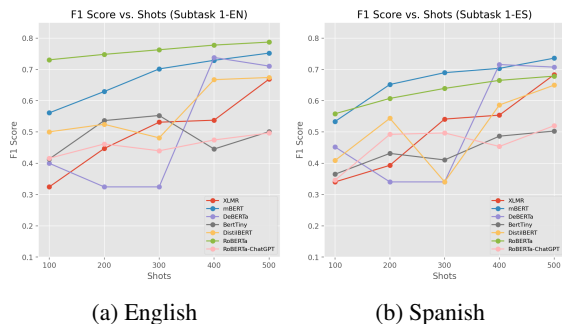|	
(a) English	(b) Spanish

Figure 1: Subtask 1 Evaluation on Few-Shot Learning

When comparing these results, we can observe that mBERT maintains strong performance in both the few-shot learning experiments and the single language experiments. It suggests that mBERT could provide a reliable choice across different tasks and experimental settings in both Subtasks.

5.2 Model Generation Capability

Model	A	B	C	D	E	F
Error(%)	37.62	68.43	58.55	48.89	**74.24**	13.81

Table 3: Comparison of Model Error Percentages. The models, labeled as A, B, C, D, E, and F, were used for prediction. The error rate was computed using mBERT with multilingual fine-tuning.

Figure 3 illustrates the error rates of the evaluated models, with **Model E** has the highest error

Model	Subtask 1		Subtask 2	
	English-F1	Spanish-F1	English-F1	Spanish-F1
Shallow Learning + Feat. Engineering				
Logistic Regression	65.67%	63.87%	38.39%	42.99%
XGBoost	71.52%	71.53%	38.47%	41.08%
Fine-tuning				
XLM-RoBERTa	78.80%	76.56%	27.14%	30.66%
mBERT	<u>85.18%</u>	**83.25%**	<u>44.82%</u>	<u>45.16%</u>
DeBERTa-V3	81.52%	72.58%	43.93%	28.28%
TinyBERT	63.75%	57.83%	15.38%	13.02%
DistilBERT	84.97%	78.77%	41.53%	35.61%
RoBERTa-*Detector*	84.01%	75.18%	34.13%	22.10%
ChatGPT-*Detector*	68.33%	64.64%	23.84%	25.45%
Multilingual Finetuning				
mBERT	84.80%	<u>82.99%</u>	**49.24%**	**47.28%**
DistilBERT	**85.22%**	80.49%	41.64%	35.59%

Table 4: F1 Score for Various Models in English and Spanish for Subtask 1 and 2. **Bold** and <u>underline</u> denote first and second best, respectively.

rate at 74.24%. In this context, a higher error rate is interpreted positively, indicating that Model E has the strongest capability to generate deceptive text. This could mean that Model E is best at creating text that is complex or nuanced enough to trick the detector into making incorrect judgments. **Model F**, conversely, shows the lowest error rate at 13.81%. This suggests that it is the least capable at generating deceptive text compared to the other models. It might produce more predictable or simpler text that the detector can easily identify as generated, hence fewer errors in detection.

However, it's worth noting that the performance might be influenced by "similarity bias in architecture" between the detector and generator models. This means if the generator and detector models are structurally similar, they might share certain biases or weaknesses, which could skew the error rates. For instance, if both models are based on a similar underlying technology (like a specific version of BERT adapted for multilingual contexts, mentioned as mBERT with multilingual fine-tuning), they might inherently perform similarly in certain tasks or languages, affecting the observed error rates.

5.3 Comparative Analysis of Model Performances

Our analysis from experiments in Table 4 reveals variations in the performance of the models for both tasks: differentiating human and machine-generated text, and identifying the specific language model that generated the given text. For

the first task, mBERT emerges as the top performer with English and Spanish F1 scores of 85.18% and 83.25% respectively, in the fine-tuning setup. This performance is closely followed by DistilBERT's English F1 score of 84.97% and Spanish score of 78.77%. In the multilingual fine-tuning configuration, DistilBERT edges out with an English F1 score of 85.22%, but mBERT retains its high Spanish performance with an F1 score of 82.99%.

In the second task, mBERT continues to excel, achieving F1 scores of 44.82% and 45.16% for English and Spanish respectively in the fine-tuning setup. It improves further in the multilingual fine-tuning setup with English and Spanish scores of 49.24% and 47.28%. However, models such as XLM-RoBERTa and TinyBERT show substantial performance gaps between the tasks. For example, XLM-RoBERTa excels in the first task with English and Spanish F1 scores of 78.8% and 76.56%, but struggles with the second task, with F1 scores dropping to 27.14% and 30.66%. Similarly, TinyBERT shows a notable performance drop in the second task.

The performance disparity suggests that the two tasks require distinct skills: the first relies on detecting patterns unique to machine-generated text, while the second demands recognition of nuanced characteristics of specific models. In conclusion, mBERT demonstrates a consistent and robust performance across both tasks. However, the findings also underscore a need for specialized models or strategies for each task, paving the way for future work in the design and fine-tuning of models for

these tasks.

6 Conclusion

This study performed an exhaustive investigation into three distinct methodologies: traditional shallow learning, Language Model fine-tuning, and Multilingual Model fine-tuning, for detecting machine-generated text and identifying the specific language model that generated the text. Our findings showed that mBERT is a robust discriminator model across different tasks and settings. However, other models like XLM-RoBERTa and TinyBERT showed a noticeable performance gap between the tasks, indicating that these two tasks might require different skillsets. This research provides insights into the performance of these methodologies on a diverse set of machine-generated texts. It also highlights the critical importance of developing specialized models or strategies for each task.

Limitations

This study provides a comprehensive comparison and analysis of models' abilities to distinguish between human and machine-generated texts. However, it relies on datasets from the Autextification competition, which withholds the specific models used for text generation in Subtask 1. As a result, in Subtask 2, our classification is based on anonymous labels (A, B, C, D, E, F), without insight into the actual models. This lack of transparency limits our assessment of potential data biases or architectural effects on the classification results. Future work that overcomes these limitations could enhance the depth and accuracy of the analysis.

Acknowledgements

References

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, and Ayu Purwarianti. 2023a. The obscure limitation of modular multilingual language models.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwari-

anti. 2023b. Indorobusta: Towards robustness against diverse code-mixed indonesian local languages.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model generated text: Is chatgpt that easy to detect?

Junwen Bai, Bo Li, Yu Zhang, Ankur Bapna, Nikhil Siddhartha, Khe Chai Sim, and Tara N. Sainath. 2021. Joint unsupervised and supervised training for multilingual asr.

Amrita Bhattacharjee and Huan Liu. 2023. Fighting fire with fire: Can chatgpt detect ai-generated text?

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Gpt-sentinel: Distinguishing human and chatgpt generated content.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 8:264–282.

Zhijie Deng, Hongcheng Gao, Yibo Miao, and Hao Zhang. 2023. Efficient detection of llm-generated texts with a bayesian surrogate model.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Discriminating between human-produced and machine-generated text: A survey. *arXiv preprint arXiv:2012.03358*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples.

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

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

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

Roy Schwartz, Oren Tsur, Ari Rappoport, and Eyal Shnarch. 2018. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1806–1815.

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

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A third party large language models generated text detection tool.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Nenghai Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance.

José Ángel González, Areg Sarvazyan, Marc Franco, Francisco Manuel Rangel, María Alberta Chulvi, and Paolo Rosso. 2023. Autextification.

A Dataset Statistics

Figure 2 presents a comparative visualization of feature-engineered dataset statistics for Subtask 1, encompassing both English and Spanish languages. The distribution patterns across the datasets for each language are delineated by average word and sentence length, alongside vocabulary richness and repetition rate. Notably, the visualizations elucidate the differences between human-generated and machine-generated text, with the human-generated text typically showcasing greater variability in sentence length and vocabulary richness.

Figure 3 offers a detailed feature comparison for Subtask 2, showcasing statistical analyses of engineered datasets in both English and Spanish. This figure provides insights into the average word and sentence length distributions, as well as vocabulary richness and repetition rate across different labels, significantly expanding upon the foundational comparisons of Subtask 1.
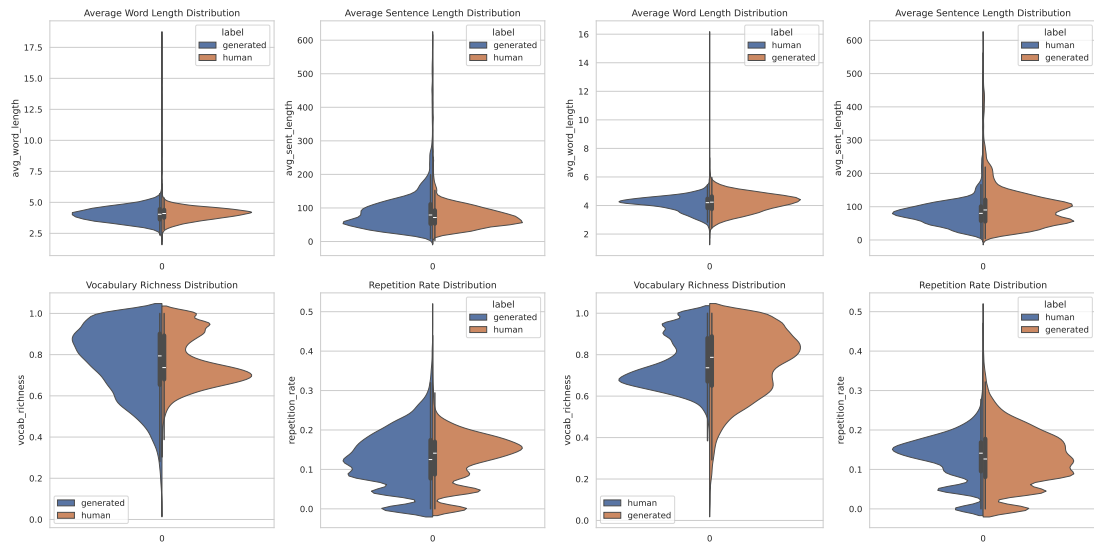
(a) English   (b) Spanish

Figure 2: Subtask 1 feature engineered dataset statistics.
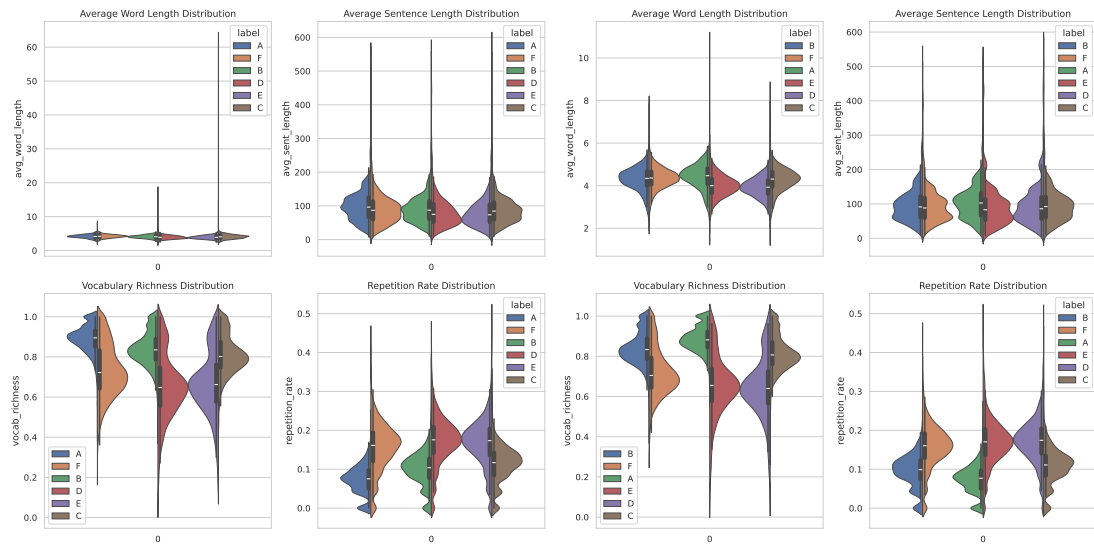
(a) English   (b) Spanish

Figure 3: Subtask 2 feature engineered dataset statistics.

B   Feature Engineered Dataset Samples

We present samples from our feature-engineered dataset, which has been specifically curated to facilitate the analysis of textual features that may distinguish between human-generated and machine-generated text. The dataset consists of text snippets, each labeled as either 'human' or 'generated', representing the origin of the text. The features engineered for this analysis include Average Word Length (AWL), Average Sentence Length (ASL), Vocabulary Richness (VR), and Repetition Rate (RR).

Tables 5 and 6 display subsets of our dataset, illustrating the distribution of these features across texts labeled as 'human' or 'generated'. These samples exhibit the variability within and between categories, forming the basis for subsequent analysis aiming to identify patterns and markers indicative of the text's origin. The engineered features are expected to contribute to the development of models capable of differentiating between human and machine-generated text.

| Text	Label	AWL	ASL	VR	RR
you need to stop the engine and wait until it stops. This is how I would do it: // Check if its safe	generated	3.120	49.500	0.960	0.040
I have not been tweeting a lot lately, but I did in November, and it was a really good month. I also	generated	3.160	49.500	0.840	0.120
I pass my exam and really thankgod for that but idk where will I go for shsmy result is ah	human	3.550	90.000	0.900	0.100
@PierreJoye i have a server already, thanks for the offer the problem is time, as always :p (ill be done	human	3.400	104.000	0.920	0.080
Crying because I have to cry for you?. No. No, no, no. Itll be all right. I	generated	2.458	14.200	0.708	0.208

Table 5: English feature engineered dataset on Subtask 1.

| Text	Label	AWL	ASL	VR	RR
Mam, por qu no me despertaste? Te hable 5 veces, te grite, te prend la luz y te abr	human	2.827	41.000	0.826	0.087
. Artculo 2. Los Estados miembros aplicarn las medidas necesarias para cumplir la presente Directiva a ms tardar el 31 de diciembre de 1981. Artculo 3. Los destinatarios de la presente Directiva sern los Estados miembros. Hecho en Luxemburgo, el 30 de junio de 1981.	human	4.353	43.500	0.647	0.216
Mi memoria es: 5% de los mdicos tienen una alta vocacin y por lo tanto son buenos profesionales, el resto es prescind	generated	3.840	118.000	0.960	0.040
APROBAR el proyecto de resolucin que se adjunta como Anexo I, por la cual se aprueba la solicitud presentada por el seor Csar Enrique Vega Arvalo (CP N PI:KEY), con domicilio en calle 7 N 3080 Quilicura, comuna de Santiago. Artculo 2. Notifquese y publquese. Dado en La Moneda, a los veintisiete das del mes de diciembre de dos mil diecinueve. Curso de Photoshop CS6 Bsico para	generated	3.937	74.600	0.797	0.114
De pequeo Dios me dio a elegir entre tener una memoria increble o un pito gigante y no me acuerdo lo que eleg	human	3.784	109.000	0.957	0.043

Table 6: Spanish feature engineered dataset on Subtask 1.

C   Evaluation on Subtask 2

In Figure 4, we observe the evaluation of few-shot learning performance across various models for Subtask 1 in both English and Spanish, denoted as Subtask 2-EN and Subtask 2-ES respectively. The F1 Score versus the number of shots (examples) is plotted, providing a clear illustration of how model performance scales with the amount of provided training data. Notable trends include the progressive improvement of models like RoBERTa and its variant RoBERTa-ChatGPT with increasing data, as well as the comparatively high performance of XLM-R in both languages.
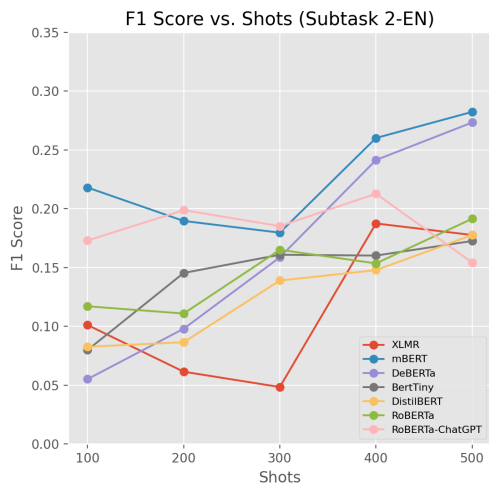
(a) English
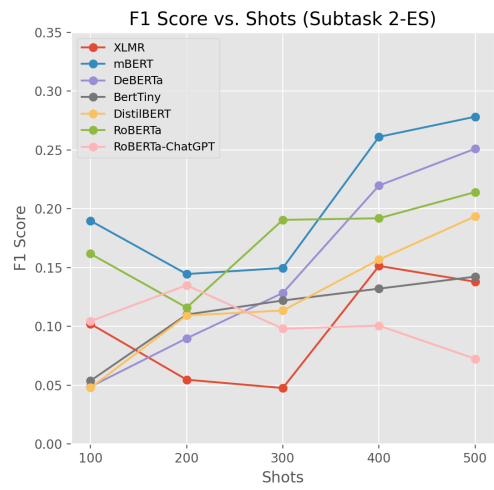
(b) Spanish

Figure 4: Subtask 1 Evaluation on Few-Shot Learning