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

**Muhammad Farid Adilazuarda, Nikolaos Nektarios Arkoulis, Oleksii Chumakov**

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{faridlazuarda, arkoulis.nick, AlexeyChumakov2001}@gmail.com

**Abstract**

Significant progress has been made on text generation by language models, 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 language models, 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). These techniques enable language models to specialize in the detection and categorization of machine-generated texts. They adapt preexisting knowledge to the task at hand, effectively manage language-specific biases, and improve classification performance.

Through a exhaustive examination of a diverse set of machine-generated texts, we deliver insights into the strengths and weaknesses of these methodologies. We illuminate the ongoing necessity for advancement in NLP and the critical importance of developing techniques that can keep pace with the progress of language models research. Our paper offers the following contributions:

1. An exhaustive evaluation of the capabilities of language models 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 language models, 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) 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 (Radford et al., 2019) 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 language models.

**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, underscoring the necessity for continuous progress in alignment with the evolving proficiency of language models.

## 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. Prior to the training process, we implemented a fundamental preprocessing step, which involved the removal of non-ASCII and special characters to refine our dataset and enhance the quality of our results. To enrich the training source of the models as showed in Table 1, 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 provides a measure of 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, which can also be indicative of language proficiency and style.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, the RR is computed as:

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

To illustrate our feature engineering process, Table 1 presents a snapshot of our dataset after the application of our feature calculations. The table showcases a selection of texts and their corresponding labels (0 representing machine-generated text and 1 indicating human-generated text), as well as a range of text features derived from these texts. 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 accurately discerning between human and machine-generated text.

### 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). We found that this setup provides superior performance compared to training separate models for each language.

Table 1: Text feature calculation. Label, AWL: Avg. Word Length, ASL: Avg. Sent. Length, VR: Vocab. Richness, RR: Repetition Rate

| **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 |

During evaluation, we employed the F1 score for each class along with the overall accuracy as our primary metrics, given the F1 score's ability to provide a balanced measure in instances of class imbalance. Further bolstering our evaluation approach, we incorporated a Few-Shot learning evaluation to assess our models' capacity to learn effectively from a limited set of examples. This involved using varying seed quantities, specifically `[200, 400, 600, 800, 1000]` instances, applied across both English and Spanish languages, thus ensuring our models' robustness and their practical applicability in real-world scenarios.

## 4. Experiments

### 4.1. Dataset

Our experiments utilize two multi-class classification datasets, namely Subtask 1 and Subtask 2, as referenced from the Autextification study (Ángel González et al., 2023). Subtask 1 is a document-level dataset composed of **65,907** samples, designed to differentiate between human and machine-generated text. Each sample is assigned one of two class labels: 'generated' or 'human'. Subtask 2, on the other hand, 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 2.

Table 2: Statistics of the datasets.

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

### 4.2. Training and Evaluation Setup

Our approach to fine-tuning language models 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.

[1]https://huggingface.co/

**Multilingual Finetuning**. An integral part of our approach was the independent fine-tuning of each model for two distinct languages - English and Spanish. This strategy was adopted to facilitate the models in effectively capturing the unique linguistic features of each language.

**Few-Shot Learning**. To gauge the performance of our models in few-shot learning scenarios, we systematically increased the count of data points for fine-tuning in increments of 100, ranging from 100 to 500 data points for each language, resulting in a total of 200 to 1000 samples per scenario. The results of the few-shot learning experiments are depicted in Fig. 1. This was computed using the equation:

$$L_{\text{few-shot}} = \frac{1}{n} \sum_{i=1}^{n} L_i$$

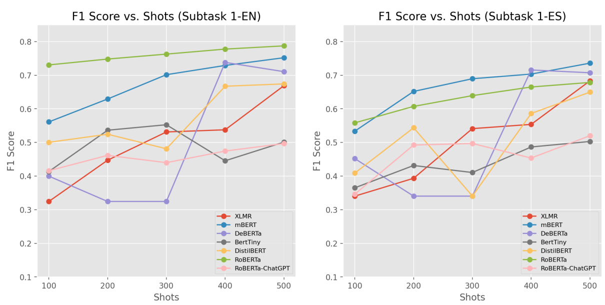
where $n$ is the total number of instances, and $L_i$ is the loss calculated for each individual data point. The loss function ($L$) provides a measure of the model's performance in this few-shot scenario.

# 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.

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.




(a) English (b) Spanish

Figure 1: Subtask 1 Evaluation on Few-Shot Learning

## 5.2. Model Generation Capability

Figure 2 illustrates the error rates of the evaluated models, with ***model E*** exhibiting the highest error rate. This indicates that model E demonstrates superior generation capabilities, enabling it to effectively deceive the detector

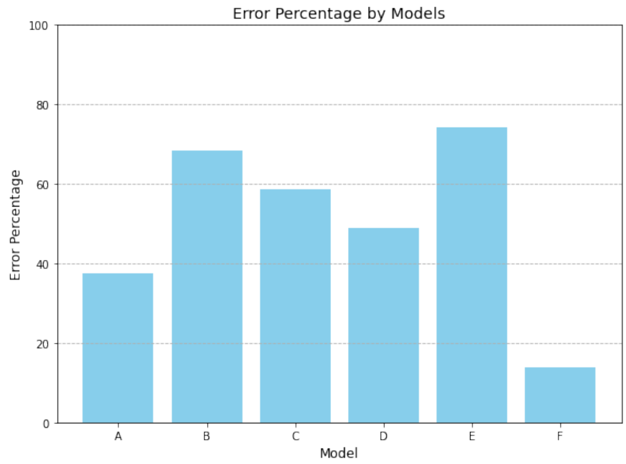



Figure 2: Comparison of Model Error Percentages. The bar chart illustrates the error percentages of multiple models, indicating the proportion of misclassified instances in the dataset. The models, labeled as A, B, C, D, E, and F, were used for prediction. The error rate was computed using mBERT with multilingual fine-tuning.

model. On the other hand, ***model F*** exhibits the lowest error rate, indicating a weaker ability to deceive the detector model. However, it is worth noting that these observations may be influenced by the similarity bias in architecture between the text detector and text generator models employed.

## 5.3. Comparative Analysis of Model Performances

Our analysis from experiments in Table 3 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

Table 3: 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.

| 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% |
| ***Finetuning*** | | | | |
| 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% |

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. The analysis revealed significant variations in performance across these techniques, suggesting the need for continued improvements in this evolving field.

Our findings showed that mBERT emerged as a consistently robust performer across different tasks and experimental settings, making it a potentially reliable choice for such tasks. 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 valuable 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.


## Acknowledgements

We express our profound gratitude to our mentors, Professor Jan Šnajder and Teaching Assistant Josip Jukić, for their invaluable guidance, constructive feedback, and unwavering support throughout the duration of this project. Their expertise and dedication have significantly contributed to the advancement of our research and understanding.


## References


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.

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.

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*.

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

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.

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

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