CNLP-NITS-PP at GenAI Detection Task 2: Leveraging DistilBERT and XLM-RoBERTa for Multilingual AI-Generated Text Detection

Annepaka Yadagiri, Reddi Mohana Krishna and Partha Pakray
Department of Computer Science & Engineering
National Institute of Technology Silchar, Assam, India, 788010
{annepaka22_rs, reddi_pg_23, partha}@cse.nits.ac.in

Abstract

In today's digital landscape, distinguishing between human-authored essays and content generated by advanced Large Language Models such as ChatGPT, GPT-4, Gemini, and LLaMa has become increasingly complex. This differentiation is essential across sectors like academia, cybersecurity, social media, and education, where the authenticity of written material is often crucial. Addressing this challenge, the COLING 2025 competition introduced Task 2, a binary classification task to separate AI-generated text from human-authored content. Using a benchmark dataset for English and Arabic, developing a methodology that fine-tuned various transformer-based neural networks, including CNN-LSTM, RNN, Bi-GRU, BERT, DistilBERT, GPT-2, and RoBERTa. Our Team_CNLP-NITS-PP achieved competitive performance through meticulous hyperparameter optimization, reaching a Recall score of 0.825. Specifically, we ranked 18th in the English sub-task A with an accuracy of 0.77 and 20th in the Arabic sub-task B with an accuracy of 0.59. These results underscore the potential of transformer-based models in academic settings to detect AI-generated content effectively, laying a foundation for more advanced methods in essay authenticity verification.

1 Introduction

Large Language Models (*LLMs*), such as ChatGPT [1] have made remarkable advances in generating human-like text responses, leading to diverse and sophisticated outputs tailored to specific prompts. While these advancements open up extensive practical applications, they also present challenges, including concerns over academic integrity and questions surrounding actual authorship. Considerable research has been devoted to distinguishing Machine-Generated Texts (*MGT*) from Human-Written Texts (*HWT*). This has primarily involved model-based techniques (Wang et al., 2023; Bhattacharjee et al., 2023) and statistical analysis methods to examine distinct text characteristics (Hans et al., 2024). Several platforms like GPTZero (Touvron et al., 2023) and Sapling have effectively differentiated MGT from HWT.

Detection of MGT has commonly employed a binary classification approach to distinguish between MGT and HWT. However, advancements in LLMs have blurred these distinctions, challenging the efficacy of straightforward classification techniques. For instance, in statistical detection, the linguistic features of an MGT might closely resemble those typically found in HWT, leading to potential misclassification. Similarly, model-based approaches often need help to generalize effectively; they are typically trained on specific datasets or models and may not perform as accurately as newer models emerge with distinct characteristics. Additionally, many detection systems need more transparency. Although some detection tools attempt to integrate explanatory elements, they often fail to deliver insightful interpretations, as observed in evaluations of models like GPTZero (Touvron et al., 2023).

2 Related Work

Zero-shot detection methods leverage statistical attributes to differentiate MGT from HWT. Research in this area has explored various Language Model (*LM*) driven features such as entropy (He et al., 2023), average log probability scores (Solaiman et al., 2019), and perplexity (Wu et al., 2023) as indicators. As LMs advance, generating increasingly sophisticated text, recent zero-shot detection methods (Mitchell et al., 2023) have evolved to capture high-level characteristics in generated content.

One notable zero-shot detection model, Binoculars (Hans et al., 2024), uses LMs to make next-token predictions across text positions. By analyzing the log perplexity ratio relative to baseline text,

[1] https://chatgpt.com/

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 307–311
January 19, 2025.

Binoculars identifies nuanced discrepancies that help distinguish MGT from HWT effectively. This technique represents an essential advancement in zero-shot detection, adapting to the sophisticated language features characteristic of current LLM outputs.

3 Methodology

3.1 Problem Statement

This research aims to develop a classification system that can identify machine-generated essays, uphold academic integrity, and mitigate the misuse of LLMs in educational contexts. The system receives as input a set of essays authored by both human writers (*including native and non-native speakers*) and by LLMs in both English and Arabic languages.

This task is defined as a binary classification problem, aiming to classify each essay as machine-generated or human-authored. The problem can be formally stated as follows:

- **Input:** A text sample E consisting of n tokens, where $E = \{w_1, w_2, \ldots, w_n\}$.
- **Output:** A binary label $y \in \{0, 1\}$, where:
  - $y = 0$ denotes a human-authored essay,
  - $y = 1$ denotes a machine-generated essay.

To approach this classification task, features $F(E) = \{f_1, f_2, \ldots, f_m\}$ are extracted from each essay E, capturing various linguistic, syntactic, and semantic characteristics. These features may include lexical patterns, syntactic structures, token frequency distributions, and transformer-based embeddings tailored to English and Arabic text properties.

The classification model $f : E \rightarrow y$ seeks to assign a probability $P(y = 1|E)$ that represents the likelihood of the text E being machine-generated. The model's performance is evaluated on a large corpus of annotated text samples, aiming to achieve robust classification across different linguistic profiles and LLM-generated writing styles.

3.2 Dataset Description

The dataset comprises essays authored by humans alongside texts generated by various AI models (Chowdhury et al., 2025). Human-written essays were sourced from the ETS Corpus of Non-Native Written English [2]. For AI-generated content, we utilized outputs from seven distinct open and closed LMs, including GPT-3.5-Turbo, GPT-4o, GPT-4o-mini, Gemini-1.5, Llama-3.1 (*8B*), Phi-3.5-mini, and Claude-3.5. Tables 1,2 present the dataset statistics for English and Arabic. Additionally, Figures 1 and 2 provide a visual comparison between human and AI-generated essays for English and Arabic datasets, respectively.

Label	English	
	Train Count	Dev Count
Human	629	1235
AI	1467	391
Total	2096	1626

Table 1: Dataset Label Counts for English Train and Development

Label	Arabic	
	Train Count	Dev Count
Human	1145	182
AI	925	299
Total	2070	481

Table 2: Dataset Label Counts for Arabic Train and Development

3.3 System Description

This paper presents our approach to the MGT Detection Task 2, aimed at detecting AI-generated content. The task involves classifying whether a given text is machine-generated or human-written, with our solution applied to both Subtask A (*English texts*) and Subtask B (*Arabic academic essays*). For Subtask A, we used the DistilBERT model, while for Subtask B, we employed the XLM-RoBERTa model. In addition, we used a rule-based method to extract semantic features like *average line length, vocabulary richness, word density, POS tags, and stop word frequency* to enhance the model's ability to detect AI-generated text. The DistilBERT model generates contextual embeddings and is followed by a pre-classifier layer to refine the output. We added a fully connected layer to incorporate additional features, using ReLU activation and dropout layers to prevent overfitting. The final output is obtained by concatenating the features and passing them through a classification

[2] https://catalog.ldc.upenn.edu/LDC2014T06

layer with a sigmoid activation to generate probabilities. Our model was trained and used on a system with an Intel Xeon CPU, 64GB RAM, and an NVIDIA Quadro GPU. Finally, we achieved an accuracy of 0.771 in Subtask A and Subtask B with an accuracy of 0.59. This result demonstrates the effectiveness of combining transformer-based models with additional feature-based enhancements in identifying AI-generated content.

Parameter	Value
Activation Function	Sigmoid
Optimizer	AdamW
Loss Function	binary_crossentropy
Learning Rate	2×10^{-5}
Batch Size	16
Number of Epochs	05
Dropout	0.3
ModelCheckpoint	Yes
EarlyStopping	Yes
Patience	2

Table 3: Hyperparameters utilized across all experiments

Model for English Language

To classify human-written versus AI-generated essays, we utilize the DistilBERT transformer model for Sub Task A. DistilBERT is a streamlined version of the original BERT model, designed for computational efficiency without compromising the core interpretative capabilities of BERT. This optimized model is well-suited for tasks requiring nuanced linguistic analysis and rapid processing. DistilBERT's architecture enables the capture of complex linguistic patterns and contextual information, essential for distinguishing subtle differences between human-authored and AI-generated content.

The classifier processes the input text to produce hidden representations h_i for each token. However, in this model, the final prediction is based on the hidden representation of the [CLS] token, designed to capture the aggregated semantic and syntactic information from the entire sequence. The prediction is computed by applying the softmax function to the [CLS] token's hidden state, as shown below:

$$\hat{y}_i = \text{softmax}(W h_{[\text{CLS}]} + b) \quad (1)$$

Here, $h_{[\text{CLS}]}$ represents the hidden representation of the [CLS] token. The parameters W and b are trainable components of the model. The softmax function generates a probability distribution across the two classes: Human-written and AI-generated. The final classification decision is based on the class with the highest probability ($\hat{y}_i$).

Model for Arabic Language

To classify human-written versus AI-generated essays in Arabic (Sub Task B), we adopted the XLM-RoBERTa model. XLM-RoBERTa is chosen for its pre-trained language-specific embeddings, which enhance its performance across multiple languages. This model generates detailed contextual embeddings for each input sequence and passes them through a classification layer for predictions. To improve classification accuracy, we incorporated additional semantic features, such as vocabulary richness, average sentence length, and stop word frequency, which helped capture the distinctions between AI-generated and human-authored essays.

The final prediction is derived from a weighted combination of the model's contextual embeddings and the extracted semantic features. This allows for a robust and accurate classification outcome.

Error Analysis: We utilized the XLM-Roberta model, which was trained on data from 100 languages, including Arabic. However, this model was not explicitly fine-tuned for the Arabic language, which may limit its performance on tasks that require a deep understanding of Arabic syntax and semantics.

3.4 Results Analysis

Among state-of-the-art transformer-based models, DistilBERT demonstrated strong performance on the English dataset, while XLM-RoBERTa proved effective for the Arabic dataset. The DistilBERT model achieved high results on English text classification, with a recall of 0.82, an F1-score of 0.77, and an accuracy of 0.77, highlighting the improved performance achieved through ensemble techniques. In comparison, models for the Arabic dataset showed relatively lower performance, with XLM-RoBERTa emerging as the best performer. XLM-RoBERTa achieved a precision of 0.55, an F1-score of 0.55, and an accuracy of 0.59. These results underscore the challenges in achieving comparable performance with Arabic models and indicate areas for further optimization in multilingual transformer-based text classification.

Model	Acc	Pre	Rec	F1
DistilBERT-En	0.77	0.784	0.82	0.77
XLM-RoBERTa-Ar	0.59	0.55	0.56	0.55

Table 4: Test Results given by Leaderboard

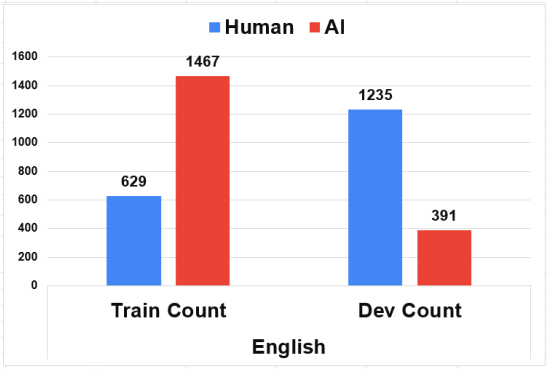

Figure 1: Visual Comparison of English Training and Development Datasets

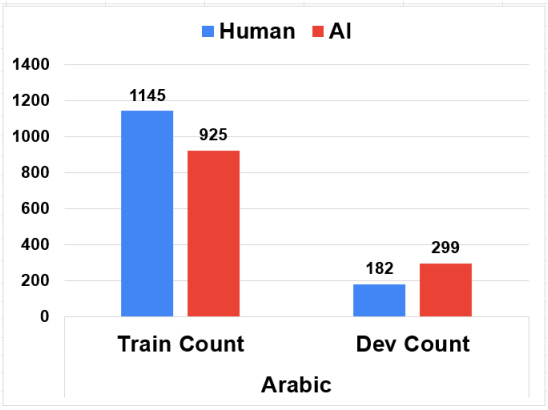

Figure 2: Visual Comparison of Arabic Training and Development Datasets

4 Conclusion

This study demonstrates the effectiveness of transformer-based models, such as DistilBERT and XLM-RoBERTa, distinguishing between human-written and AI-generated essays across English and Arabic. Our Team CNLP-NITS-PP achieved a high detection recall of 0.825 for English and an accuracy of 0.590 for Arabic, indicating these models' adaptability to diverse linguistic contexts. Ensemble methods further improved classification accuracy, underscoring the importance of robust detection systems as AI-generated content continues to proliferate. Future research could investigate additional linguistic features and cross-domain applications to enhance detection performance and address the specific challenges observed with Arabic models.

5 Future Work

We utilized the XLM-Roberta model, which was trained on data from 100 languages, including Arabic. However, this model was not explicitly fine-tuned for the Arabic language, which may limit its performance on tasks that require a deep understanding of Arabic syntax and semantics. We plan to explore models specifically fine-tuned on Arabic datasets for future work. These specialized models are expected better to understand the nuances and complexities of the Arabic language, potentially leading to improved accuracy in detecting AI-generated content in Arabic texts. By focusing on optimized models for Arabic, we aim to enhance the overall performance of our approach in this specific context.

References

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. *arXiv preprint arXiv:2309.03992*.

Shammur Absar Chowdhury, Hind Al-Merekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, Georgios Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

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.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. *arXiv preprint arXiv:2310.08903*.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*.