TechExperts(IPN) at GenAI Detection Task 1: Detecting AI-Generated Text in English and Multilingual Contexts

Gull Mehak
Amna Qasim
Abdul Gafar Manuel Meque
Nisar Hussain
Grigori Sidorov
Alexander Gelbukh
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Mexico City, Mexico

Abstract

The ever-increasing spread of AI-generated text, driven by the considerable progress in large language models, entails a real problem for all digital platforms: how to ensure content authenticity. The team TechExperts(IPN) presents a method for detecting AI-generated content in English and multilingual contexts, using the google/gemma-2b model fine-tuned for COLING 2025 shared task 1 for English and multilingual. Training results show peak F1 scores of 97.63% for English and 97.87% for multilingual detection, highlighting the model's effectiveness in supporting content integrity across platforms.

1 Introduction

The rise of large language models (LLMs), such as GPT-4, has significantly increased the volume of AI-generated content across various digital platforms. These models can generate coherent and contextually relevant text, making it much more difficult for users to distinguish between human-authored and machine-generated content. The recent rise in AI-generated content is making many question the credibility and reliability of information, especially regarding journalism, academia, and social media, where the integrity of the content is critical. This has brought the need to develop effective methods to detect AI-generated content to an all-time high (Fraser et al., 2024).

Recent gains in the capabilities of LLMs have brought new challenges to their detection. Approaches such as reinforcement learning with human feedback and instruction tuning have given these models more versatility to follow even complex prompts and thus develop plausible responses that further complicate the detection problem (Abdali et al., 2024). Traditional detection methods, which rely on identifying patterns of word choice, sentence structure, or perplexity, are often insufficient as these models improve in mimicking human writing styles (Goddard et al.). One avenue of research lies in resorting to transformer-based models in the detection effort, whereby such models make possible fine-grained differentiation of human-generated from AI-generated texts through fine-tuning curated datasets (Zhao et al., 2024). These models have demonstrated high accuracy in identifying AI-generated content, even when the text is short or resembles typical human writing (Mao et al., 2024).

The proposed research introduces an approach that utilizes the google/gemma-2b model, an advanced Large Language Model (LLM), to identify AI-generated content across English and multilingual contexts, using both the COLING_2025_MGT_en and COLING_2025_MGT_multilingual datasets. By leveraging google/gemma-2b, a powerful multilingual LLM, our approach aims to enhance detection precision through sophisticated machine learning techniques. This method is expected to contribute significantly to maintaining content integrity and mitigating risks associated with the improper use of AI-generated textual content across diverse linguistic landscapes.

This paper is structured as follows: Section 2 reviews related work in AI-generated content detection, Section 3 describes the methodology and dataset used, Section 4 presents the experimental results, and Section 5 discusses the findings and their implications. Finally, we conclude with potential future directions for research in this field.

2 Related Work

Detecting AI-generated content has become a critical research area due to advancements in large language models (LLMs) like GPT-4. These models can produce content that closely mimics human writing, raising concerns about authenticity across academia, journalism, and social media. Early

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 161–165

methods relied on lexical, syntactic, and stylistic features, but these often fell short as modern LLMs became more sophisticated. Detecting machine-generated text is a complex task[1].

Recent advancements involve machine learning, particularly transformer models like RoBERTa and BERT, which show high accuracy when fine-tuned on human and machine-generated datasets. Studies have demonstrated significant improvements using these models on specialized datasets (Zeng et al.). Advanced approaches leverage token-level analysis, focusing on log probabilities and entropy to detect patterns typical of AI-generated text. This strategy exploits the probabilistic nature of LLMs, identifying subtle deviations from human writing [4]. Ensemble methods have also effectively combined models like RoBERTa with domain-specific classifiers. Techniques such as paraphrasing and back-translation further enhance robustness, allowing better generalization across different text sources (Wang et al., 2024).

Emerging trends focus on hybrid approaches, blending linguistic features with machine learning models to capture nuances that traditional statistical methods miss. Zero-shot learning methods are also being explored, enabling detection without explicit examples, though with mixed success (Mitchell et al., 2023). Ethical considerations are increasingly important, particularly avoiding biases that might misclassify content from non-native English writers. Future research aims to develop inclusive systems that ensure high detection accuracy across diverse user demographics (Fraser et al., 2024).

In summary, while advancements in machine learning and transformer-based models have strengthened AI-generated text detection, challenges remain, particularly in addressing diverse linguistic contexts. Motivated by these backgrounds, we employ the google/gemma-2b model (a multilingual LLM) to localize AI-generated text in English and multilingual with COLING_2025_MGT_en and COLING_2025_MGT_multilingual datasets, respectively. This approach aims to enhance detection accuracy, supporting efforts to uphold content integrity and responsible AI use across varied languages and settings.

[1]University of Pennsylvania School of Engineering and Applied Science. (2024, August 16). *Detecting machine-generated text: An arms race with the advancements of large language models*. ScienceDaily. Retrieved December 12, 2024, from https://www.sciencedaily.com/releases/2024/08/240816121550.htm

3 Methodology

This section outlines the datasets employed and the proposed google/gemma-2b model used for both English and Multilingual settings.

3.1 Dataset

In this work, we use two primary datasets to identify AI-generated (AG) texts in English and multilingual texts: COLING_2025_MGT_en, COLING_2025_MGT_multilingual. The datasets (see Table 1) are rich and diverse, including human- and machine-generated examples across finance, medicine, social media feeds, and scientific literature.

3.1.1 English Dataset

(COLING_2025_MGT_en): This dataset includes 610,767 samples in total, with 228,922 human-written and 381,845 machine-generated texts. The development set contains 261,758 samples (98,328 human and 163,430 machine).

3.1.2 Multilingual Dataset

(COLING_2025_MGT_multilingual): Spanning languages such as Chinese, Italian, Arabic, Russian, Bulgarian, and Urdu, this dataset comprises 629,384 training samples, split into 253,625 human-written and 375,759 machine-generated texts. The development set includes 271,215 samples (107,467 human and 163,748 machine).

3.2 Proposed Model

This study utilizes the google/gemma-2b model, a transformer-based architecture for detecting AI-generated content across English and multilingual contexts. As illustrated in Figure 1, the process begins with tokenizing input text, where each token is converted into a vector representation. Positional embeddings are added to these token vectors to preserve sequence information, allowing the model to recognize word order and contextual relationships—an approach common in transformer architectures (Vaswani, 2017). This step is essential for distinguishing nuanced linguistic patterns that differentiate human-generated content from AI-generated text.

The model's core lies in the Decoder Block, where multiple layers process these embeddings to refine the token representations further. Each layer employs multi-head self-attention to capture diverse contextual relationships across tokens, enabling the model to focus on various aspects of the

Dataset	Training Set	Development Set	Grand Total
English	Human: 228,922	Human: 98,328	872,525
	Machine: 381,845	Machine: 163,430	
Multilingual	Human: 253,625	Human: 107,467	900,599
	Machine: 375,759	Machine: 163,748	

Table 1: Datasets (English + Multilingual) Details

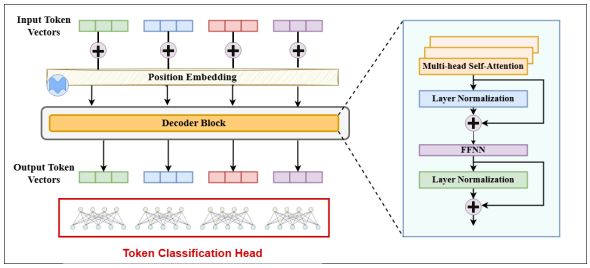

Figure 1: Architectural view of proposed gemma model

text, a technique effective in capturing semantic relationships (Devlin, 2018). Following this, each token embedding passes through a feed-forward neural network (FFNN), which enhances representation depth, allowing the model to interpret complex language structures (Radford et al., 2019). Additionally, layer normalization and residual connections stabilize the outputs and ensure critical information flows through the layers without degradation, as demonstrated in various transformer-based models (He et al., 2016). Finally, the output vectors are passed to a token classification head, which labels each token, distinguishing AI-generated content from human-written text with high precision (Liu, 2019).

4 Results and Analysis

This section presented the experimental setup, evaluation metrics, and training results, demonstrating the proposed model's high accuracy and robust multilingual detection performance.

4.1 Experimental Setup

The proposed gemma model was implemented in the Python-based PyTorch framework. High resources were used for model training on Google Colab Pro Plus. Due to the size of COLING_2025_MGT_en and COLING_2025_MGT_multilingual datasets (Wang et al., 2025), we used only a subset of all these data for experiment capabilities. For each data set, 60,000 examples were sampled from the training set and 10,000 from the development set. We used these stratified samples to train the final model, which confirms a balanced representation across classes for both English and multilingual datasets. Using this method, we could efficiently train the model while keeping the detector's performance strong. Table 2 gives details on the hyperparameter settings.

4.2 Evaluation Measures

We measured the model's accuracy, precision, recall, and F1 score (Mehak et al., 2023). So, accuracy shows us correctness in general; precision is the ratio of correctly identified AI detections to all detected cases by AI, and recall shows how well your model can detect AI instances out of everything. It incorporates false positives and negatives, i.e., identifying a balance between precision and recall (F1 score). Combining these metrics gives a good assessment of how well the model performs in classifying AI-generated versus human-created text.

4.3 Training Results on COLING_2025_MGT_en Dataset

The model obtained high accuracy and F1 across three training epochs for the English dataset. Training loss reduced over epochs, showcasing stable learning, while validation loss fluctuated minimally. As shown in Table 3, the third epoch reached the highest scores for accuracy and F1, indicating excellent detection of AI-generated text in English.

4.4 Training Results on COLING_2025_MGT_multilingual Dataset

The results of the multilingual data set highlighted high precision and F1 scores across three epochs. The training loss was minimized regularly, while the validation was slightly varied (Table 4). These metrics within and across the five languages achieved optimal scores at epoch 3, indicating effective AI-content detection and efficiency in training and testing data separation (accuracy, F1, precision, recall).

Hyperparameter	Value
Model	google/gemma-2b
Epochs	3
Learning Rate	0.0001
Train Batch Size	64, and 56
Eval Batch Size	64, and 56
Seed	42
Optimizer	Adam (betas = (0.9, 0.999), epsilon=1e-08)
Learning Rate Scheduler Type	Linear
Mixed Precision Training	Native AMP

Table 2: Hyperparameter settings

Epoch	Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	938	0.2079	0.0868	0.9662	0.9730	0.9702	0.9758
2	1876	0.0424	0.0938	0.9688	0.9748	0.9829	0.9668
3	2814	0.0089	0.1577	0.9704	0.9763	0.9763	0.9763

Table 3: Results of proposed gemma model on COLING_2025_MGT_en dataset

Epoch	Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	1072	0.1048	0.0781	0.9691	0.9751	0.9734	0.9767
2	2144	0.0373	0.0925	0.9701	0.9757	0.9817	0.9698
3	3216	0.0073	0.1267	0.9737	0.9787	0.9802	0.9772

Table 4: Results of proposed gemma model on COLING_2025_MGT_multilingual dataset

4.5 Test Results

The final model evaluation was conducted through blind submissions on the Codabench platform. Our model achieved competitive results, securing 5th place for Subtask A (English) with an F1 score of 0.8153 and 6th place for Subtask B (Multilingual) with an F1 score of 0.74.

4.6 Results Discussion

The results demonstrate google/gemma-2b's strengths in accurately detecting AI-generated content across English and multilingual datasets. Its advanced multilingual capabilities and high precision and recall scores underscore its effectiveness in capturing subtle linguistic patterns across varied languages. This robust performance reflects gemma-2b's adaptability and precision, making it a reliable multilingual AI content detection tool.

5 Conclusion and Future Directions

The google/gemma-2b model effortlessly detects AI-generated content and performs well even in multilingual contexts, as the context has not only been in English. This study shows that it could be used as a content authenticity tool. In the Future, we are planning to expand detection not just with new models but also by increasing languages to help more communities (Arabic, Urdu, Persian, Chinese), adapting better from one AI model to another, and finally enhancing how real-time we could be so that it won't only be used for moderation content but even fact-checking.

6 Acknowledgment

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, and grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. 2024. Decoding the ai pen: Techniques and challenges in detecting ai-generated text. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6428–6436.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2024. Detecting ai-generated text: Factors influencing detectability with current methods. *arXiv preprint arXiv:2406.15583*.

Jamal Goddard, Yuksel Celik, and Sanjay Goel. Beyond the human eye: Comprehensive approaches to ai text detection.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: generative ai detection via rewriting. *arXiv preprint arXiv:2401.12970*.

Gull Mehak, Iqra Muneer, and Rao Muhammad Adeel Nawab. 2023. Urdu text reuse detection at phrasal level using sentence transformer-based approach. *Expert Systems with Applications*, 234:121063.

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.

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gaševic, and Guanliang Chen. Detecting ai-generated sentences in human-ai collaborative hybrid texts: Challenges, strategies, and insights.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.