NootNoot at SemEval-2024 Task 8: Fine-tuning Language Models for AI vs Human Generated Text detection

Sankalp Bahad[1]
IIIT Hyderabad
sankalp.bahad@research.iiit.ac.in

Yash Bhaskar[1]
IIIT Hyderabad
yash.bhaskar@research.iiit.ac.in

Parameswari Krishnamurthy[2]
IIIT Hyderabad
param.krishna@iiit.ac.in

Abstract

In this paper, we introduce a machine-generated text detection system designed to tackle the challenges posed by the proliferation of large language models (LLMs). With the rise of LLMs such as ChatGPT and GPT-4, there is a growing concern regarding the potential misuse of machine-generated content, including misinformation dissemination. Our system addresses this issue by automating the identification of machine-generated text across multiple subtasks: binary human-written vs. machine-generated text classification, multi-way machine-generated text classification, and human-machine mixed text detection. We employ the RoBERTa Base model and fine-tune it on a diverse dataset encompassing various domains, languages, and sources. Through rigorous evaluation, we demonstrate the effectiveness of our system in accurately detecting machine-generated text, contributing to efforts aimed at mitigating its potential misuse.

1 Introduction

Large language models (LLMs) are becoming mainstream and easily accessible, bringing in an explosion of machine-generated content over various channels, such as news, social media, question-answering forums, educational, and even academic contexts. Recent LLMs, such as ChatGPT and GPT-4, generate remarkably fluent responses to a wide variety of user queries. The articulate nature of such generated texts makes LLMs attractive for replacing human labor in many scenarios. However, this has also resulted in concerns regarding their potential misuse, such as spreading misinformation and causing disruptions in the education system. Since humans perform only slightly better than chance when classifying machine-generated vs. human-written text, there is a need to develop automatic systems to identify machine-generated text with the goal of mitigating its potential misuse.

The advent of sophisticated large language models (LLMs), including ChatGPT and GPT-4, has catalyzed a surge in artificially generated text across myriad domains, from news media to social platforms, educational resources, and scholarly publications. These neural network models exhibit an unprecedented capacity to produce natural language, enabling the automation of written content creation. However, the human-like fluency of LLM outputs has concurrently raised serious concerns surrounding potential misuse.

With syntactically coherent and topically relevant text, LLMs could plausibly disseminate misinformation, plagiarize or falsify documents, and automate persuasion-based attacks on a massive scale. The integration of models like ChatGPT into education has additionally ignited fierce debate; while proponents highlight opportunities for personalized instruction, critics argue LLMs enable academic dishonesty and undermine human knowledge acquisition. Amidst this controversy, institutions urgently seek policies to uphold academic integrity.

Alarmingly, humans perform only marginally better than random chance at distinguishing machine-generated versus human-authored text. Developing reliable technical systems to automatically detect AI content has therefore become a research priority. The goal is to provide educators, moderators, and end users tools to identify LLM outputs, thereby mitigating potential dangers from increasingly accessible, human-like models.

Constructing robust LLM detectors demands interdisciplinary collaboration, combining machine learning advances with insights from fields like ethics, media studies, and education. With judicious coordination across stakeholders, experts aim to actualize benefits of LLMs for automation while curtailing risks of misinformation, deception, and cheating.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 918–921
June 20-21, 2024 ©2024 Association for Computational Linguistics

2 Dataset

The dataset provided by the organizers of this shared task (Wang et al., 2024a) comprises a diverse collection of texts encompassing various domains, languages, and sources. The dataset is structured to address the three subtasks outlined: binary human-written vs. machine-generated text classification (Subtask A), multi-way machine-generated text classification (Subtask B), and human-machine mixed text detection (Subtask C).

For Subtask A (Binary Classification), the dataset consists of a balanced corpus of human-written and machine-generated texts. The human-written texts are sourced from various publications, academic papers, forums, and social media platforms. The machine-generated texts are generated by state-of-the-art language models such as ChatGPT, GPT-4, cohere, davinci, bloomz, and Dolly. These texts cover a wide range of topics to ensure diversity and representativeness.

For Subtask B (Multi-Way Classification), the dataset includes texts generated by each of the six specified language models: ChatGPT, cohere, davinci, bloomz, Dolly, and human-written texts. The texts are annotated to indicate their respective sources, enabling the classification task to determine the origin of each text accurately.

For Subtask C (Human-Machine Mixed Text Detection), the dataset contains texts where the first part is human-written, and the subsequent part is machine-generated. Annotations demarcate the boundary where the transition from human to machine-generated text occurs. This allows for training and evaluating models on detecting the boundary between human and machine-generated segments within a single text.

The dataset is preprocessed to remove noise, standardize formatting, and ensure consistency across texts.

Subtask	Train	Dev
A (Monolingual)	119,757	5,000
A (Multilingual)	172,417	4,000
B	71,027	3,000
C	3,649	505

Table 1: Dataset Statistics for Each Subtask

3 Methods

We employed a fine-tuning approach using the XLM-RoBERTa Base model (Conneau et al., 2020) for the task of machine-generated text detection.

We chose Roberta-base as our base model for fine tuning as XLM-RoBERTa is a multilingual language model optimized for classification tasks. It is pretrained on massive multilingual data, and has a robust architecture and performance enable efficient fine-tuning across diverse text classification problems with state-of-the-art accuracy

The fine-tuning process involves initializing the RoBERTa Base model with pre-trained weights and then fine-tuning it on our specific dataset for the tasks of binary human-written vs. machine-generated text classification (Subtask A), multi-way machine-generated text classification (Subtask B), and human-machine mixed text detection (Subtask C).

During fine-tuning, we optimize the model's parameters using stochastic gradient descent (SGD) with backpropagation. We employ task-specific loss functions, such as cross-entropy loss for classification tasks and mean squared error (MSE) for mixed text detection. Additionally, we utilize techniques such as dropout regularization to prevent overfitting and gradient clipping to stabilize training.

The RoBERTa Base model is fine-tuned separately for each subtask, with hyperparameters tuned using grid search or random search techniques. We split the dataset into training, validation, and test sets to facilitate model training and evaluation, ensuring that the model generalizes well to unseen data.

4 Results

We present the performance metrics achieved by our machine-generated text detection system on each of the subtasks: binary human-written vs. machine-generated text classification (Subtask A), multi-way machine-generated text classification (Subtask B), and human-machine mixed text detection (Subtask C). The evaluation metrics include F1 score (macro and micro) and accuracy.

4.1 Subtask A: Binary Classification

Epoch	F1 Macro	F1 Micro	Accuracy
1	0.85431	0.85463	0.85463
2	0.81726	0.81918	0.81918
3	0.80595	0.80859	0.80859

Table 2: Performance Metrics for Subtask A (Monolingual)

Epoch	F1 Macro	F1 Micro	Accuracy
1	0.65693	0.69128	0.69128
2	0.71308	0.72564	0.72564
3	0.64664	0.68958	0.68958

Table 3: Performance Metrics for Subtask A (Multilingual)

From the tables of results of Subtask A, we can observe that in Monolingual case we get a better accuracy and F1-Score. The scores are in the range of 0.8 to 0.85, which decrease with the increase in number of epochs. Hence, the best score is observed in the model trained only for 1 epoch. This pattern indicates that the model can possibly be overfitting on the data.

In case of Multilingual, we observe the best scores in the model trained for 2 epochs.

4.2 Subtask B: Multi-Way Classification

Epoch	F1 Macro	F1 Micro	Accuracy
1	0.80686	0.8065	0.8065
2	0.85083	0.851	0.851
3	0.83146	0.83117	0.83117
4	0.84295	0.84328	0.84328
5	0.86936	0.86794	0.86794

Table 4: Performance Metrics for Subtask B

From the results of Subtask B, we can observe that the model trained for epoch 5 performs the best. Based on the Micro and Macro F1 scores in the table, we can observe that since the Macro F1 increasing over epochs indicates the model is improving at predicting each individual class correctly. The Micro F1 is also increasing which suggests that overall predictive capability on the aggregate data is improving. However, Micro F1 can be influenced by performance on majority classes.

4.3 Subtask C: Human-Machine Mixed Text Detection

Epoch	MSE
1	63.13998
2	33.09197
3	28.01411
4	30.04774
5	27.12254

Table 5: Performance Metrics for Subtask C

From the results of Subtask C, we can observe that the provided mean squared error (MSE) values indicate the model loss decreased with each epoch of training from an initial value of 63.13998 at epoch 1 to 27.12254 at epoch 5. The difference in MSE between epoch 1 and 2 implies that the model is overfitting on the training data. The lowest MSE score is observed in Epoch 5.

5 Conclusion

In this study, we proposed a machine-generated text detection system capable of addressing three subtasks: binary human-written vs. machine-generated text classification (Subtask A), multi-way machine-generated text classification (Subtask B), and human-machine mixed text detection (Subtask C).

Our system leverages the RoBERTa Base model, fine-tuned on a diverse dataset comprising texts from various domains, languages, and sources. Through extensive experimentation and evaluation, we achieved promising results (Wang et al., 2024b) across all subtasks.

For Subtask A, our system demonstrated robust performance in distinguishing between human-written and machine-generated texts, achieving high F1 scores and accuracy across multiple epochs. Similarly, in Subtask B, where the classification involves identifying the source language model among multiple candidates, our system achieved competitive performance, indicating its effectiveness in multi-way classification scenarios.

In Subtask C, where the objective is to detect boundaries between human-written and machine-generated segments within a single text, our system showed reasonable performance, albeit with some room for improvement. Future work could focus on refining the model architecture and exploring additional features to enhance the system's performance in this challenging task.

Overall, our study highlights the importance and feasibility of developing automatic systems for detecting machine-generated text, contributing to efforts aimed at mitigating the potential misuse of large language models in various contexts.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico, Mexico.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta.