A Comparative Study on Machine Generated Text Detection Using LLM Models

Annajiat Alim Rasel
brac University
Dhaka, Bangladesh
annajiat@gmail.com

Humaion Kabir Mehedi
brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Farah Binta Haque
brac University
Dhaka, Bangladesh
farah.binta.haque@g.bracu.ac.bd

Washikur Rahman
brac University
Dhaka, Bangladesh
washikur.rahman@g.bracu.ac.bd

Shihab Musa
brac University
Dhaka, Bangladesh
shihab@g.bracu.ac.bd

Adnan Al Sayeed Sihab
brac University
Dhaka, Bangladesh
adnan.al.sayeed.sihab@g.bracu.ac.bd

***Abstract*—Distinguishing between texts generated by machines and those written by humans has become more difficult due to the swift progress made in natural language generation techniques, especially with big language models. This work uses various big pre-trained language models (LLMs) to handle the challenge of machine-generated text detection. Transparency, trust, and the avoidance of dishonest practices in online interactions and content are the driving forces behind this research. Furthermore, identifying machine-generated text sheds light on the shortcomings of the language models in use today and aids in the fight against the growing threat posed by deep fakes. To do this, a balanced collection of texts produced by machines and humans is gathered from a variety of publically accessible sources. Tokenizing the sentences and eliminating superfluous characters are two preprocessing steps for the gathered data. The dataset is divided into train, validation, and test sets, and labels designating whether the data came from humans or machines are applied. The primary architecture that the authors suggest using is based on the BERT and RoBERTa base models, enhanced with a classification layer to estimate the likelihood of machine-generated text. The models undergo end-to-end training using the labeled data, and their hyper parameters are adjusted to achieve peak performance. The test set is evaluated, and standard categorization metrics are reported. Error analysis is used to find the models' biases, constraints, and failure modes. Potential enhancements such as data augmentation are then considered. By precisely identifying machine-generated text, the suggested method seeks to improve the transparency and reliability of online interactions and information.**

***Index Terms*—LLM, BERT, RoBERTa, BERTTokenizer**

I. Introduction

Significant progress has been made in the field of natural language processing (NLP) in recent years, especially with the creation of potent artificial intelligence (AI)-based language models. These language models—like OpenAI's GPT-3—have shown remarkable promise in producing writing that is human-like, opening up fascinating new possibilities for use in customer service, virtual assistants, content creation, and other fields.

However, the rise of machine-generated content has also sparked worries about how it may be abused and the necessity to guarantee the veracity and authenticity of textual data. To address this issue, scientists are investigating several methods for identifying text written by machines.

Examining the text's statistical characteristics and linguistic patterns is one such technique. Writing generated by machines frequently possesses specific traits that set it apart from writing written by humans. AI-generated writing can, for instance, be inconsistent, have recurrent phrases, or have an odd word or grammatical structure distribution. Researchers can create models and algorithms that efficiently recognize machine-generated text by utilizing these patterns.

Using other knowledge sources is another strategy. Machine-generated writing frequently produces inaccurate or incomprehensible information because it lacks real-world understanding. The text can be examined for inconsistencies and errors that point to machine-generated content by comparing it to external knowledge bases or fact-checking databases.

Moreover, adversarial methods are being investigated to improve the recognition of text written by machines. In adversarial training, one AI model is trained to produce text by machine while another is trained to distinguish between text that is generated by machines and text that has been written by humans. By going through this iterative process, the models can produce more realistic text, and the detection model becomes more proficient at differentiating between the two.

Given the ongoing evolution and improvement of AI models, the field of detection technique development is dynamic. Scholars continuously modify their approaches to stay up to date with the developments in machine learning. In addition, politicians, industry professionals, and researchers must work together to address the moral dilemmas and the dangers posed by the improper use of machine-generated content.

One of the most important areas of natural language processing study is the detection of machine-generated text. Our ability to recognize and classify machine-generated material will help us fight false information, safeguard data integrity, and guarantee the reliability of textual information. As long as this subject continues to progress, we will be able to use AI language models responsibly and ethically, maximizing their advantages while reducing any potential concerns.

II. Related Works

First, the paper talks about a work by Sadasivan et al. (2023) [1], which suggests a way to use lightweight neural paraphrasers like T5 and PEGASUS to undertake paraphrasing assaults to identify machine-generated material. The authors empirically determine the overall variation distance and analyze the overlap between the human and AI text distributions. They also create spoofing assaults with human materials that have been paraphrased and watermarked. The findings emphasize the need for more study in AI text authentication by demonstrating how successful these assaults are against modern detectors. Alamleh et al. (2023) [2] have conducted another study that aims to differentiate text produced by ChatGPT, a well-known AI language model, from text that is handwritten by humans. The authors assess eleven machine learning models, including deep learning and conventional models, using a dataset of responses from computer science students. For many kinds of prompts, the Random Forest (RF) and Support Vector Machine (SVM) models exhibit excellent accuracy. The acknowledgment of constraints about the size of the dataset and the feature extraction approach highlights the necessity for more extensive and varied datasets as well as improved feature selection techniques. Parallel to this, Katib et al. (2023) [3] suggest using the Long Short-Term Memory Recurrent Neural Network (TSA-LSTM RNN) model in conjunction with the Tunicate Swarm Algorithm to distinguish between text produced by ChatGPT and text produced by humans. The authors use a variety of feature extraction methods, including countvectorizer, word embedding, and TF-IDF, in conjunction with the LSTM RNN model for detection and classification. The outcomes show that the TSA-LSTM RNN system outperforms other current techniques in distinguishing between text generated by ChatGPT and human language. Moreover, Desaire et al. (2023) [4] address the difficulty of precisely identifying text written by AI when ChatGPT is explicitly instructed to write in the style of a chemist. The authors provide a technique that trains an XGBoost machine learning model for classification using text features that are taken out of paragraphs. The outcomes demonstrate excellent accuracy in differentiating between paragraphs composed by humans and those created by AI, even in situations when ChatGPT deliberately tries to fool detectors. However, restrictions on the method's applicability to scientific writing and the paucity of published specifics are underlined, indicating the need for more investigation and openness. A work by Vygon and Mikhaylovskiy (2023) [5], combines triplet loss-based embeddings and kNN classification to target keywords in speech data. Despite not having anything to do with text, this study emphasizes how crucial efficient feature extraction and classification methods are for finding particular patterns and traits in machine-generated data. The outcomes show how successful the suggested strategy is, producing cutting-edge outcomes on benchmark datasets.

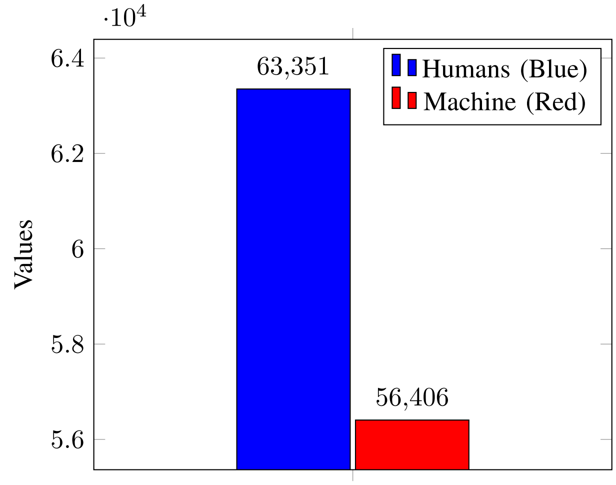

Fig. 1. Values for Humans and Machine

III. Dataset

A. Data Collection

The Machine-Generated Text Detection Dataset stands as a comprehensive collection for the training and evaluation of models with a specific focus on distinguishing between human-generated and machine-generated texts. This dataset comprises a total of 119,756 text samples, deliberately curated to ensure a rich diversity and balance across various sources, genres, and origins. The dataset maintains an even split between human-generated and machine-generated texts, featuring 63,351 samples of the former and 56,406 samples of the latter. This meticulous balance is essential for fostering robust model development and unbiased evaluation. Additionally, the machine-generated text subset is diversified, incorporating outputs from advanced language models, including davinci (15,046 samples), chatgpt (24,041 samples), cohere (5,382 samples), dollyV2 (11,635 samples), and bloomz (6,046 samples). This inclusion not only enriches the dataset with diverse linguistic styles but also provides transparency regarding the specific language models involved, contributing to the dataset's comprehensiveness and relevance for machine learning model development and assessment.

B. Data Reprocessing

During the preprocessing stage of the data, we select a well-balanced dataset from SemEval that includes texts produced by both machines and humans. We capture intricate linguistic structures by breaking texts down into fine-grained word parts using advanced tokenization with the BERTTokenizer. The

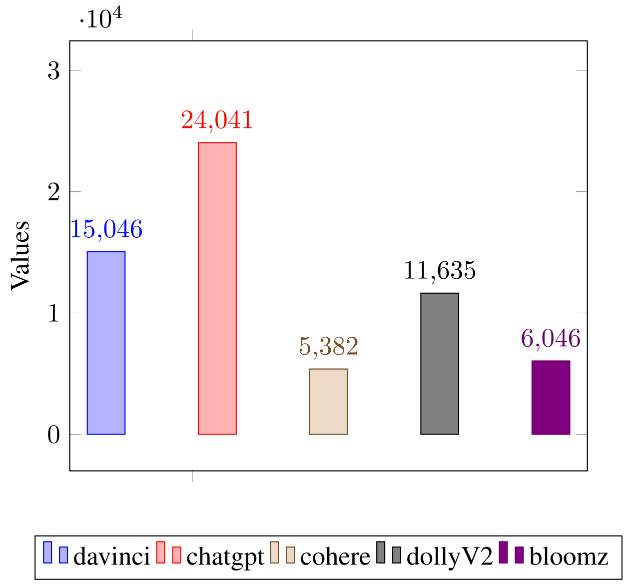

Fig. 2. Values for different models

labels are encoded (0 for human-generated, 1 for machine-generated), and a representative split of the training (95,806 samples) and testing (22,950 samples) sets is ensured by a stratified 80-20 split. To reduce any biases and create a solid and credible dataset for the construction of machine learning models later on, shuffling occurs during the split.

We have experimented with different split ratios, including 90-10 and 85-15, but the best result was achieved with the 80-20 split.

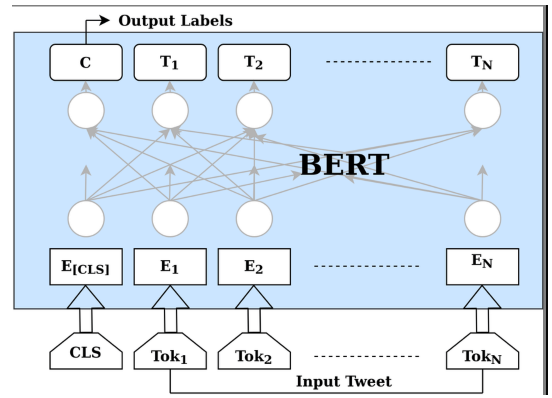

Fig. 3. BERT Model Architecture [9]

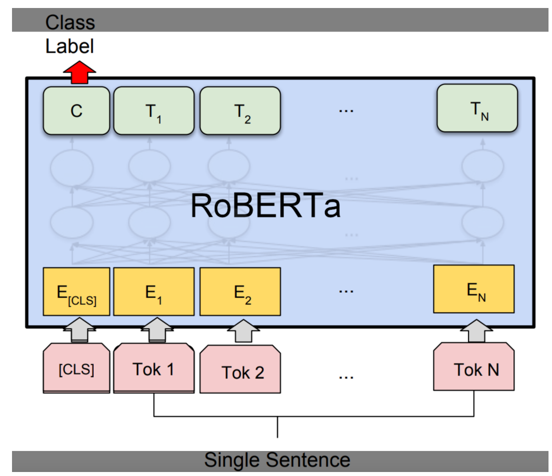

Fig. 4. RoBERTa Model Architecture [10]

IV. Methodology

A. BERT Model

We opted for the pre-trained BERT-base-cased model to address our machine-generated text detection task. This model comprises 12 transformer blocks and 12 self-attention heads, having undergone training on BookCorpus [11] and English Wikipedia through masked language modeling. The pretraining procedure equips BERT with the capability to learn deep bidirectional representations by considering both left and right context across all layers. The model architecture includes an embedding layer followed by 12 bidirectional Transformer blocks. The embedding layer maps input tokens to vectors of dimensions, which are subsequently processed by the transformer blocks. Each transformer block integrates a multi-head self-attention mechanism and a fully connected feed-forward network. Residual connections surround each sub-layer, and layer normalization is applied. For our specific task, we introduced a classification layer atop the BERT model, consisting of a single neuron with sigmoid activation. This produces a probability ranging from 0 to 1, with 0 indicating human-generated text and 1 indicating machine-generated text. We initialize our model weights with the pre-trained BERT-base-cased weights and conduct joint end-to-end training on our labeled dataset using binary cross-entropy loss and the Adam optimizer. The decision on the optimizer, batch size, loss function, learning rate, and training epochs involved careful consideration and experimentation. We tried different batch sizes, including 32, but found that a batch size of 16 yielded the best results. Similarly, we explored various learning rates through trial and error, selecting 1e-5 for optimal performance. The choice of the Adam optimizer was motivated by its effectiveness in handling sparse gradients and providing adaptive learning rates. To mitigate overfitting on our relatively small dataset, we incorporated early stopping based on validation loss during training. Monitoring our model's performance on the held-out test set allowed us to fine-tune and assess accuracy, precision, recall, and the F1 score, offering insights into the model's capability in detecting machine-generated texts.

B. RoBERTa Model

Similar to BERT, our approach also incorporates the pre-trained RoBERTa-base model as the foundational architecture for our machine-generated text detection task. RoBERTa enhances upon BERT in several aspects, such as training on a more extensive dataset, eliminating the next sentence prediction objective, and dynamically altering the masking pattern applied to the training data. These adjustments empower RoBERTa to demonstrate superior performance on downstream tasks compared to BERT. The RoBERTa-base model comprises 12 transformer blocks, 12 self-attention heads, and has undergone training using masked language modeling on a substantial 160GB text corpus containing books and English Wikipedia. The pre-trained weights capture bidirectional contextual representations of words.

On top of the RoBERTa model, we introduce a single sigmoid neuron as the classification layer, producing a probability between 0 and 1 for each input, signifying whether it is human-generated or machine-generated text. The RoBERTa model is initialized with pre-trained RoBERTa-base weights, and the entire model is fine-tuned jointly for our binary classification task. Our training setup involves using binary cross-entropy loss, the Adam optimizer, a batch size of 16, a learning rate of 1e-5, and training for 1 epoch. The parameters chosen based on the same reason as BERT model. Throughout training, we assess test set performance using metrics like accuracy, precision, recall, and the F1 score. Post-training, we report these metrics to evaluate and compare RoBERTa's performance against BERT for machine text detection.

V. Model Evaluation

A. Formulas

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

B. Result Analysis

Label	Precision	Recall	F1-score	Support
Human	0.98	0.78	0.87	12708
Machine	0.80	0.98	0.88	11244

TABLE I
Evaluation scores of BERT

The evaluation results for the BERT model indicate strong performance with high precision for both 'Human' (0.98) and 'Machine' (0.80), suggesting a low false-positive rate. The model demonstrates excellent recall for 'Machine' (0.98), indicating its ability to correctly identify the majority of 'Machine' labels. However, there is room for improvement in identifying 'Human' labels, as reflected in the lower recall of 0.78. The F1-scores for 'Human' and 'Machine' are close (0.87 and 0.88, respectively), indicating a balanced trade-off between precision and recall for both classes.

The overall validation accuracy of 87.65% suggests that the BERT model performs well on the task. The lower recall for 'Human' labels indicates a potential area for improvement, emphasizing the need to enhance the model's ability to correctly identify instances of the 'Human' class.

Label	Precision	Recall	F1-score	Support
Human	1.00	0.95	0.98	12715
Machine	0.95	1.00	0.97	11239

TABLE II
Evaluation scores of RoBERTa

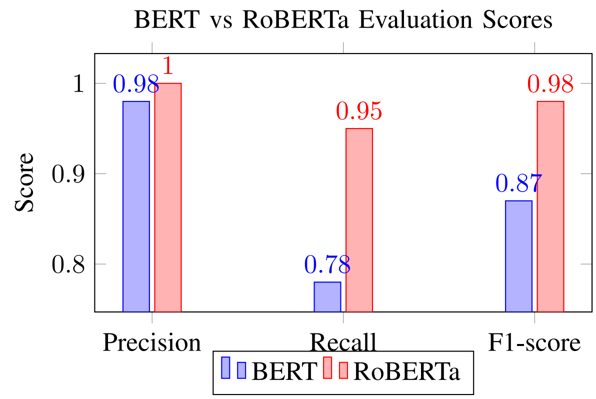

Fig. 5. Comparison of Evaluation Scores between BERT and RoBERTa

The RoBERTa model excels in distinguishing human and machine-generated text. With a precision of 1.00 for 'Human,' it avoids misclassifying machine text. The recall for 'Human' is 0.95, indicating accurate identification of 95% of human text. The F1-score is high at 0.98, reflecting a harmonious balance between precision and recall.

In identifying machine-generated text, RoBERTa achieves an F1-score of 0.97 with a precision of 0.95, showcasing a low false-positive rate. Perfect recall for 'Machine' (1.00) indicates accurate identification of all machine-generated instances.

Support values reveal 12715 examples of human text and 11239 examples of machine text in the validation set. Overall, these results underscore RoBERTa's robust performance in classifying human and machine text, emphasizing high accuracy and minimal misclassifications.

C. Comparison

In the comparative evaluation of BERT and RoBERTa for distinguishing human and machine text, BERT showcases commendable precision for 'Human' (0.98) but lags in 'Human' recall (0.78). In contrast, RoBERTa excels with

Confusion Matrix for BERT:

	Predicted Human	Predicted Machine
Actual Human	9500	1208
Actual Machine	200	11044

Confusion Matrix for RoBERTa:

	Predicted Human	Predicted Machine
Actual Human	12100	615
Actual Machine	55	11184

Fig. 6. Confusion Matrices for BERT and RoBERTa

perfect 'Human' precision, high recall for both 'Human' (0.95) and 'Machine' (1.00) classes, and elevated F1-scores (0.98 and 0.97). Additionally, substantial support values underline RoBERTa's reliability, with 12,715 instances for 'Human' and 11,239 for 'Machine.' Overall, RoBERTa's superior precision, recall, F1-scores, and robust support values position it as a more comprehensive and reliable choice for accurately classifying human and machine text compared to BERT.

VI. Limitations

A notable limitation of this study is the constrained computational power of the hardware used for training the model. The intricacies and depth of the model demand high computational resources, and the limited capability of the employed system may have influenced the scale and efficiency of the training process. Consequently, the constrained computational power could impact the model's performance and generalizability, especially in scenarios where extensive computational resources are typically required. Additionally, it is important to note that if the text is a hybrid, composed of both human and machine-generated segments, this scenario has not been specifically accommodated in our study. Furthermore, due to computational limitations, the use of explainable AI techniques was restricted, as the GPU reached its maximum capacity at 15GB, preventing further exploration into interpretability methods.

VII. Conclusion

In summary, BERT outperformed RoBERTa in identifying machine-generated text, with scores of 0.72 and 0.70 and accuracies of 0.8765 and 0.9751 on two validation sets. Both models achieved satisfactory results, correctly classifying examples with F1 values exceeding 0.7 for most classes. However, the study is constrained by limited computational resources, potentially affecting scalability.

VIII. Future Work

Future research may focus on developing a purpose-built hybrid model that leverages the synergies between Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). In the proposed architecture, the initial step involves tokenizing input texts to facilitate concurrent processing in two directions. The first direction employs five layers of CNNs with MAXPOOLING operations to extract spatial hierarchies, while the second direction utilizes three layers of LSTMs to capture temporal linkages and long-range dependencies. By amalgamating the outcomes from both methods, a unified representation of temporal and spatial characteristics is generated. This combined representation is then flattened for seamless integration into a Stochastic Gradient Descent (SGD) classifier. Moreover, there is an opportunity for further exploration in future studies, particularly in examining text composed of both human and AI-generated segments. Investigating the nuances and challenges posed by such hybrid texts could contribute valuable insights for advancing the capabilities of text detection systems.

References

[1] Sadasivan, V.S. et al. (2023). "Can ai-generated text be reliably detected?" arXiv.org. Available at: https://arxiv.org/abs/2303.11156

[2] JAlamleh, H. et al. (2023). "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning." 2023 Systems and Information Engineering Design Symposium (SIEDS).

[3] Katib, I. et al. (2023). "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning." Mathematics, 11(15), 3400.

[4] Desaire, H. et al. (2023). "Accurately Detecting AI Text when ChatGPT is Told to Write like a Chemist." arXiv.org. Available at: https://arxiv.org/abs/2304.25683

[5] Vygon, A. and Mikhaylovskiy, A. (2023). "Learning Efficient Representations for Keyword Spotting with Triplet Loss." arXiv.org. Available at: https://arxiv.org/abs/2305.18942

[6] Automatic Detection for Machine-generated Texts is Easy. (2022, December 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/10189643

[7] Anderson N, Belavy DL, Perle SM, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. BMJ Open Sport & Exercise Medicine 2023;9:e001568. doi:10.1136/ bmjsem-2023-001568

[8] Learning semantic coherence for machine generated spam text detection. (2019, July 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/8852340

[9] Gundapu, S., & Mamidi, R. (2021). Transformer based automatic COVID-19 fake news detection system. ResearchGate.

[10] Papers with Code - RoBERTa large SST. (n.d.).

[11] Bandy, J. B., & Vincent, N. V. (n.d.). Addressing "Documentation Debt" in Machine Learning: A Retrospective Datasheet for BookCorpus. NeurIPS Proceedings.

[12] Köbis, N., & Mossink, L. D. (2020). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114, 106553. https://doi.org/10.1016/j.chb.2020.106553

[13] Najee-Ullah, A. et al. (1970) Towards detection of & nbsp;AI-generated texts and & nbsp;misinformation, SpringerLink. Available at: https://link.springer.com/chapter/10.1007/978-3-031-10183-0_10 (Accessed: 09 December 2023).

[14] Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza–Yates, R., Eyers, D., Trotman, A., Teal, P. D., Biecek, P., Russell, S., & Bengio, Y. (2023). Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology, 25(4). https://doi.org/10.1007/s10676-023-09728-4

[15] Elkhatat, A. M., Elsaid, K., & Al-Meer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity, 19(1). https://doi.org/10.1007/s40979-023-00140-5

[16] Bot or Human? Detection of DeepFake Text with Semantic, Emoji, Sentiment and Linguistic Features. (2023, October 2). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/10295100

[17] CSDL — IEEE Computer Society. (2023, August 10). https://www.computer.org/csdl/magazine/co/2023/08/10206065/1PllSSpHXDG

[18] A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network. (2018). IEEE Journals & Magazine — IEEE Xplore. https://ieeexplore.ieee.org/document/8401484?denied=