IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Detecting ChatGPT-Generated Text using Transformers

[1]Rutam Risaldar
[1]Research student
[1]Department of Computer Science and Data Science,
[1]Keshav Memorial Institute of Technology, Hyderabad, India

Abstract: The detection of ChatGPT-generated text has become a critical task as AI-generated content becomes increasingly prevalent. Traditional statistical methods, such as analyzing burstiness—a measure of the variability in word frequency—provide foundational techniques for identifying patterns characteristic of machine-generated text. Burstiness captures how frequently certain terms appear in quick succession, often revealing unnatural repetition or anomalies typical of AI text (It can be specific to a particular family of LLMs). Other statistical metrics, like TF-IDF, can highlight distinctive features by weighing the importance of words within a corpus. These methods have historically provided a baseline for distinguishing human-written text from its machine-generated counterparts. However, the complexity of language models like ChatGPT necessitates more sophisticated detection approaches. Machine learning-based methods, especially those employing transformer models like RoBERTa, offer a significant advancement in this field. By fine-tuning these models on datasets of human and AI generated text, researchers can train classifiers that capture subtler nuances of generated content. These advanced models analyze context and linguistic patterns far beyond the capabilities of traditional statistical methods. Comparing these two approaches, the machine learning-based methods generally exhibit superior performance in accuracy and robustness, demonstrating their potential for reliably detecting AI-generated text. This paper explores the application of machine learning algorithms and their derivatives to detect text generated by ChatGPT. The study uses a pre-collected dataset of human-written TOEFL essays alongside ChatGPT-generated counterparts to train text classification models. A novel dataset sourced from ChatGPT-3 ensures that the generated texts accurately reflect the language model's style and nuances. The study employs various performance metrics, including accuracy, precision, recall, and F1-score, to assess the models' ability. Experimental results demonstrate the superior performance of the transformer-based approach, highlighting its high accuracy in identifying ChatGPT-generated content. These findings underscore the potential of advanced NLP techniques based on pre-training and how finetuning can help in repurposing to detect AI-generated text, hence aiding in maintaining content authenticity and integrity. Furthermore, this research paves the way for future studies to explore additional features and more approaches to further improve detection capabilities across content modalities.

IndexTerms **– Deep Learning, Transformer finetuning, Text classification, Natural Language Understanding**

INTRODUCTION

Generative models, particularly ChatGPT, have emerged as powerful tools for creating humanlike text among other media modalities. ChatGPT, a variant of the GPT-3 language model developed by OpenAI, is specialized for generating conversational text and can be fine-tuned for various tasks such as question answering, dialogue generation, and other forms of language production [1]. It can produce text remarkably similar to human writing based on the supplied prompts, raising important questions about the ability to distinguish between machine-generated and human-written text. This distinction is crucial for various applications, including academic integrity, content verification in media, and fake news detection. As AI-generated content becomes more prevalent, the need for effective detection mechanisms has become increasingly urgent. Misuse of generative models has led to challenges such as plagiarism in educational settings, the spread of false information, and the erosion of trust in digital communications. Therefore, developing reliable methods to identify AI-generated text is essential for maintaining the integrity of various systems and institutions. This paper addresses this need by exploring and implementing a set of models for text classification to detect text generated by ChatGPT automatically. The training and evaluation of these models are conducted using a dataset comprising essays authored by humans and those generated by ChatGPT. The methodology leverages the advanced capabilities of transformer-based models, specifically focusing on fine-tuning RoBERTa [2], a Robustly Optimized BERT approach known for its effectiveness in NLP tasks. The primary approach adopted in this study with good results is a logical extension of transformer fine-tuning for text classification. By fine-tuning RoBERTa on a specific dataset, the model is adapted to better capture the subtle differences between human and machine-generated text. This process involves several steps, including data pre-processing, model training, and performance evaluation using various metrics. The experimental results are promising, demonstrating that the fine-tuned RoBERTa model achieves a high accuracy of 98% in distinguishing between human-written and ChatGPT-generated text. This high level of accuracy highlights the potential of transformer-based models in addressing the challenges posed by advanced generative models like ChatGPT. The contributions of this work are manifold: 1. Development of a Fine-Tuning Approach for RoBERTa: This study introduces a specialized fine-tuning methodology for the RoBERTa model, tailored to identify text generated by ChatGPT. This involves adjusting the pre-trained model to enhance its ability to classify text based on subtle linguistic features indicative of machine generation. 2. Examination of Generalizability: The study explores the generalizability of the fine-tuned

RoBERTa model by testing it on the text generated by ChatGPT across different, previously unseen topics or domains. This aspect is crucial to ensure that the model can effectively identify AI-generated text in a variety of contexts, beyond the specific dataset used for training. 3. Explainable Visualizations: The research includes visualizations of the model's performance and feature impacts at both the model and instance levels. These serve as an effort towards contributing to explainable AI to ensure reliability..

1. Data Gathering and Preprocessing

The research process for this study involves several key steps, starting with the preparation of the dataset. Once the dataset is prepared and preprocessed using a series of techniques relevant features from the text are selected for analysis. Text normalization used includes tokenization, stemming, and the removal of stop words. In this study, firstly the XGBoost algorithm is used for modeling. XGBoost is a popular library for tasks such as text classification and has been shown to achieve great accuracy and robustness to outliers on a variety of datasets. Post-training the model is evaluated which involves comparing the model's predictions to the true labels of the text data and calculating the required evaluation metrics. Here I outline the data gathering and preprocessing steps taken by the original dataset creators and me.

1.2 Data Acquisition

Plenty of data is available on the web to source human written texts, hence the acquisition of their counterparts required more effort. For authentic human-written text, assessment essays from exams like the TOEFL exam served as a prominent source. A vast repository of TOEFL essays authored by humans exists online, offering valuable insights into writing abilities. The original creators of the dataset [8] took the effort to create two distinct components sourced from different channels. The first segment comprises human-authored essays extracted from a publicly available PDF containing numerous TOEFL preparation essays along with the complementary portion consisting of essays generated by ChatGPT. To ensure ChatGPT's responsiveness across various topics, the authors appended the prompt "Write an essay on the following topic" to each query. This approach enabled them to compile a comprehensive dataset comprising 126 human and 126 nonhuman essays, totalling 252 essays, publicly available on GitHub. The preprocessing phase plays a paramount role in text classification, facilitating the extraction of pertinent features from raw text. Previously stated normalization techniques were employed to refine text data for analysis. For XGBoost classification model training TF-IDF scheme was used which is a widely used method for vectorizing text data was used which quantifies the importance of a word in a document within a corpus by considering both its term frequency (TF) and inverse document frequency (IDF). Table 1. Shows a sample of the dataset for the mentioned question.

Table 1. An example of a topic and two essays written on a topic

Writer	Question: Why do you think people attend college or university
ChatGPT	There are many reasons why people choose to attend college or university. Some of the most common reasons include the desire to gain new knowledge and skills, the opportunity to pursue a career in a particular field, and the chance to earn a higher salary …
Human	College is a place that the students can learn more and new knowledge and experience in it. Of course, different people have different reasons to study in college. For example, some people want to be to go on a further study after they graduate from the college; …

2.2 Preprocessing

Preparing text for analysis is a pivotal stage in the text classification workflow, aimed at extracting salient features from raw textual data. This process, known as text preprocessing, encompasses various techniques including tokenization, lowercasing, stemming, and stop word removal. These methods are instrumental in isolating the most pertinent and informative features from the text corpus, thereby exerting a substantial influence on the efficacy of the subsequent machine-learning model. In my study, text preprocessing is delineated into distinct steps, encompassing tokenization, uncasing, lemmatization, and stop-word removal across implementations. Each step plays a crucial role in refining the textual data and enhancing its suitability for analysis. For text modelling, I opted for the widely employed TF-IDF scheme. TF-IDF, short for term frequency-inverse document frequency, serves as a prevalent methodology for transforming text data into a numerical format conducive to machine learning models. This statistical metric gauges the significance of a word within a document relative to a corpus, thereby encapsulating its contextual importance. Computed by multiplying two key metrics, namely term frequency (TF) and inverse document frequency (IDF), TF-IDF facilitates the creation of feature vectors that capture the essence of the textual data. RoBERTa finetuning requires tokenization based on its pre-trained tokenizer and no other explicit scheme for text modelling was used during the process.

2. METHODOLOGY

The methodology section outline the plan and method that how the study is conducted.

2.1 Word Vector Classifiers

Classical machine-learning techniques, such as logistic regression [3], have been utilized by researchers to create detectors capable of differentiating between machine-generated text and human-written text. For instance, a logistic regression model was developed and trained on a dataset containing web pages and text produced by GPT-2 models. This model relied on term frequency inverse document frequency (TF-IDF) vectors, which highlight the relative importance of individual words and bigrams within the text. Despite its simplicity, this baseline model effectively distinguished between human-written and machine-generated text,

showcasing the potential of traditional machine-learning methods for this task. Researchers have explored various configurations of the GPT-2 language model, with sizes ranging from 117 million to 1542 million parameters, and have used different sampling techniques, including pure sampling and Top-k sampling [4] in their analyses. They observed that larger GPT-2 models tend to produce text that is harder to differentiate from human-written text, while smaller models are more easily detected. Sampling results in text that is more detectable due to the method's tendency to overuse common words, creating statistical anomalies that detection algorithms can identify. Conversely, nucleus sampling produces text that is more challenging to detect.

2.2 Language Models Characteristics

In another compelling study, researchers investigated the influence of modelling choices including model size, and prompt length, on the detection process. They discovered that it is feasible to train a classifier to accurately identify the modelling choices used to generate the text [5]. This suggests that text produced by language models is more sensitive to these choices than previously understood. Another finding stated that it is easier to distinguish between texts generated by different language models compared to distinguishing between machine-generated and human-written text. Traditional machine learning algorithms and basic neural networks proved effective in identifying text generated by language models. These conclusions align with findings from other studies, which noted that text generated by the GPT-3 model was particularly challenging to detect among several other language models.

2.3 XGBoost for Text Classification

XGBoost [6] was one of the implemented algorithms for text classification, leveraging its robustness and high performance across various datasets. The process begins with preprocessing and feature selection to enhance the classification model's accuracy. Feature selection involves identifying a subset of significant words, which simplifies the training data and emphasizes crucial features. The term frequency-inverse document frequency (TF-IDF) method is used for feature selection. TFIDF weights are calculated for each word in the dataset, and a subset of the most important words is chosen based on these values. These selected features are then used to train the XGBoost classification model. After preprocessing and feature selection, the dataset is divided into training and test sets using k-fold cross-validation, a statistical method that assesses model performance by dividing the dataset into kfolds. The model's performance is averaged across the k folds to minimize bias and improve classification accuracy. Results show that the XGBoost model achieves an accuracy rate of 96% when using TF-IDF. SHapley Additive exPlainations (SHAP)[7] values are used to interpret the model, highlighting the importance of input variables.

2.4 RoBERTa Fine tuning

Another approach implemented in this study utilized transformer fine-tuning to detect text generated by ChatGPT. RoBERTa is a transformer-based model known for its exceptional performance in various natural language processing tasks. The methodology involves preprocessing the dataset, feature selection, and training the model for text classification to distinguish between human-written and ChatGPT-generated text. The preprocessing stage includes cleaning the text to remove special characters, extra spaces, and other artefacts that could hinder the tokenization process. The cleaned text is then tokenized using the RoBERTa tokenizer, converting the text into input IDs necessary for model processing. Here feature selection is implicitly handled by the RoBERTa model, which utilizes contextual embeddings to capture the significance of words within the text. Unlike traditional methods, RoBERTa does not rely on manually crafted features like TF-IDF but leverages deep learning to understand the context and importance of words in the dataset. The fine-tuning process involves configuring the RoBERTa model for binary classification, where it learns to differentiate between human-written and ChatGPT-generated text. Training arguments, including batch sizes, number of epochs, and logging configurations, are defined to optimize the fine-tuning process.

2.5 Feature Importance and Interpretation

To interpret the model's decisions, SHAP value plots are used. SHAP values provide insights into the importance of input variables, helping to understand the relationships between input features and the model's output. Visualization through SHAP summary plots orders variables based on their impact on the model's predictions. The results, including comprehensive evaluation metrics and feature importance visualizations, highlight the model's capability to distinguish between human-written and ChatGPT-generated text with high accuracy and reliability.

International Journal of Novel Research and Development
Research Through Innovation

3. Result Analysis

In this section, we evaluate the XGBoost classification and the fine-tuned RoBERTa model's performance. Evaluation results demonstrate an accuracy rate of 98% when using TF-IDF-based features. Detailed results are presented in Figure 1.

	Precision	Recall	F1	Support
Human	1.00	0.92	0.96	29
chatGPT	0.94	1.00	0.99	22
acc			0.96	51
mavg	0.96	0.96	0.96	51
wavg	0.96	0.96	0.96	51

Figure 1. XGBoost model performance metrics

The fine-tuned RoBERTa model's evaluation results are presented in Figure 2.

```
Performance Metrics:
Label          Precision       Recall          F1              Support
Human          1.00            1.00            1.00            29
ChatGPT        1.00            1.00            1.00            22
Accuracy: 1.00
Macro-averaged Precision: 1.00
Macro-averaged Recall: 1.00
Macro-averaged F1 Score: 1.00
Weighted-averaged Precision: 1.00
Weighted-averaged Recall: 1.00
Weighted-averaged F1 Score: 1.00
```

Figure 2. RoBERTa model performance metrics

The RoBERTa model achieves a very high accuracy with due credit to the pre-training process of the transformer model. The scope of this research was also to achieve explainable predictions. To explain the feature's importance in the prediction process SHAP [7] text plot was used which shows the text instance level feature importance along with their dynamic SHAP values. An example is shown in Figure 3 which depicts a text plot showing the instance feature's importance for the true label (ie.ChatGPT)

Figure 3. SHAP Text plot for RoBERTa

A similar approach was adopted using a summary plot for an instance of human text with features that use TF-IDF schema for the XGBoost model. This is depicted in Figure 4 as follows.

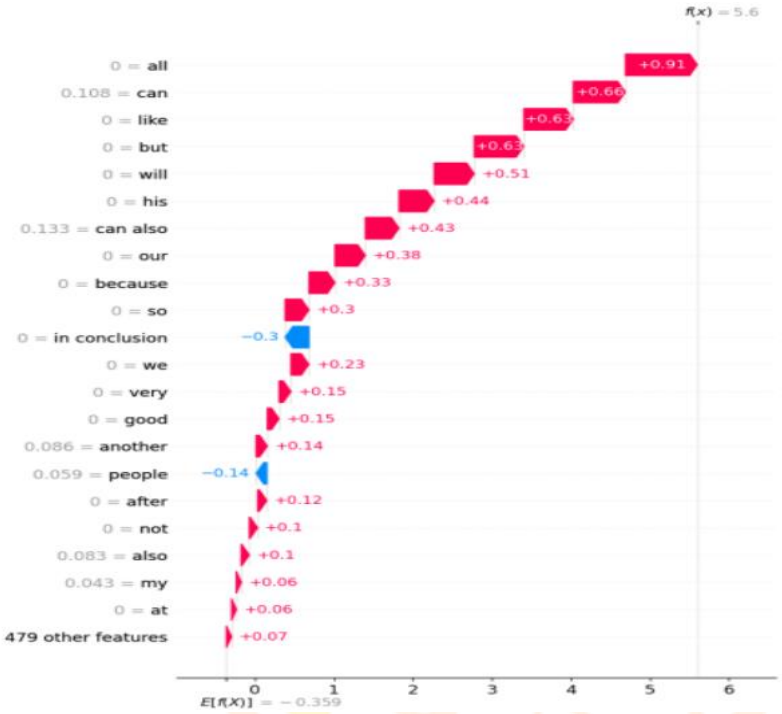

Figure 4. SHAP Summary plot for TF-IDF features

4. Conclusion

In this study, I implemented a proposed machine learning methodology for automatically identifying text generated by ChatGPT. The dataset comprised both human-written and ChatGPT-generated essays and was employed it to train and evaluate the classification models. The findings revealed that my model could distinguish between the two types of text with an accuracy rate of 98%. A significant contribution of my research was the extension of existing approaches and implementation of RoBERTa finetuning. This underscores the crucial role of feature engineering in tasks involving ChatGPT-generated text and implies that similar strategies could be beneficial in other related tasks. Future research could explore the application of LLM in detecting machine-generated text and additional plugins that would focus on the real-time semantic verification of input text. Additionally, investigating the integration of more features or adopting more sophisticated feature engineering methods could further enhance model performance. In summary, my study demonstrates the potential of transformers for the automatic detection of ChatGPT-generated text, providing a valuable tool for researchers to unmask deceptive objectives.

.

REFERENCES

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.
[2]Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[3]W.-K. Chen, Linear Networks and Systems. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.
[4]Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models, 08 2019.
[5] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. pages 8384– 8395, 01 2020.
[6]Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016.
[7] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 12 2017.
[8] Shijaku, Rexhep & Canhasi, Ercan. (2023). ChatGPT Generated Text Detection. 10.13140/RG.2.2.21317.52960.