Identifyng Human-Written, BARD and ChatGPT Generated Text using Machine and Deep Learning

Manu Dev
Cluster Innovation Centre
University of Delhi
Delhi, India
manudev0004@gmail.com

Abhishek Kumar
Cluster Innovation Centre
University of Delhi
Delhi, India
abhishek15902@gmail.com

Madugula Jayaram
Cluster Innovation Centre
University of Delhi
Delhi, India
jayaram.kumar23@gmail.com

***Abstract*—The use of an Artificial Intelligence (AI) language models, like GPT 3.5/4 and Bard has raised an issue in academic and scholar writing works, distinguishing between the human written text and text generated by AI is difficult. This research aims to explore the effectiveness and accuracy of machine learning algorithms in identifying human written content versus content generated by GPT and Bard. To achieve this we collected responses from GPT, Bard and Human using various prompts such as essay, poem, stories. Subsequently we extensively evaluated machine and deep learning models, including Decision Trees (DT), K-Nearest Neighbors Classifier (KNN), Random Forest Classifier (RF), Support Vector Machine (SVC), Generalized Linear Model (GLM) and Bidirectional Encoder Representations from Transformers (BERT). Throughout this investigation, we gained insights into the difficulties in differentiating between humans vs AI written content. This research is a step, in safeguarding honesty especially, in an era where most of the people rely on AI generated content. As both humans and AI (represented by GPT and Bard) coexist within scholarly settings our findings hold importance in upholding the integrity of intellectual conversations. The aim of our study, we want to figure out how good machine and deep learning models can be to tell if something was written by a human, BARD and ChatGPT. Overall, our goal is to use different models and evaluate them and try to improve accuracy and efficiency further by combining different models together, thus safeguarding the human authored work.**

***Index Terms*—DT — KNN — RF — SVM — GLM — BERT —AI**

I. Introduction

This document dives into the challenges and concerns that arise from the use of AI language models regarding their potential impact, on academic and scholarly integrity. As AI is becoming more and more powerful, its generated content becomes more and more difficult to differentiate from human authored works. The purpose of this study is to assess the effectiveness of machine learning algorithms in discerning between human written and AI generated text. The study evaluates machine and deep learning algorithms using a data sets of 500 samples each, Total of 1500 consisting of essay, poem and stories prompts. After that increasing the only human written text just for checking the impact of biased datasets. By comparing the performance of models the study pinpoints the suitable machine and deep learning model for this task. It provides insights into how machine learning algorithms can detect AI generated text while emphasizing the importance of further research, in this field to uphold academic and scholarly integrity. We aim to make these computer programs better at noticing the small differences in how humans and machines write. We're also looking at the challenges when the programs need to understand the changing abilities of AI and how the context of what's written can affect their accuracy.

II. Related Work

ChatGPT and Bard are an AI model currently being used by millions of users. Rapid Improvement in AI has raised concerned on the use of AI in academics. Students are relying on AIs for their works. While these AI are improving day by day it is hard to detected them as well [1]. The related work done in base paper examine the ability to distinguish between human-written and AI-generated text. Experiments were conducted to study the accuracy of detecting AI-generated text and Human written text by using different models. The datasets include Text and Programming code generated by ChatGPT 2 and Human written [2].

III. Methodology

In this section we will present the datasets used in this study, followed by a discussion of the feature extraction process.

A. Data Collocation Phase

Generating the AI-text is quite easier we had to simply write prompts. One can used API for this purpose too but in our case we found out that ChatGPT API uses older version and replies were not not reliable so we have copy paste the results from its web version. For human it was quite hard to give prompts and get back the result so instead we used old books of poem,essay and stories and wrote prompts for similar generation of text with ChatGPT 3.5 and bard. A datasets of 1500 data points, 500 each was gathered from the replies from ChatGPT, Bard using various prompts. We collected all these data and add it in spreadsheet [2]. In spreadsheet we marked : ChatGPT - 0, Human -1 , Bard - 2.

B. Preprocessing

The input text is preprocessed to ensure consistency and compatibility with the model's architecture. This may involve tasks like:

- Manual Filtering This involves manually filtering out duplicate data points. Duplicate data can introduce biases and skew the learning process. By removing duplicates, you ensure that the model learns from a diverse and representative dataset. We also removed some unreadable signs and text copied from the web.
- Averaging: Averaging likely refers to incorporate averaged features into the model. For example, if BERT generates multiple contextualized embedding for a sentence, averaging them can provide a more stable and representative input for the BERT model.
- Matrix Input: Converting data into a matrix format suitable for both Random Forest and BERT. This could involve transforming textual data into numerical representations that Random Forest can handle, while also preparing the input in a format suitable for BERT embedding.
- Train-Test Split: The datasets was split in to 80/20 of training and testing. This way we used randomly 80% of our total datasets for training our and remaining datasets for testing our accuracy.

IV. Experiment

In this section, we will present the different models used for the purpose of this project.

A. Model Evaluation

The models assessed are as follows:
1) Decision Tree Classifier (DT)
2) K-Nearest Neighbors Classifier (KNN)
3) Random Forest Classifier (RF)
4) Support Vector Machine (SVM)
5) Generalized Linear Model (GLM)
6) Bidirectional Encoder Representations from Transformers (BERT)

Decision Tree Classifier (DT) is a machine learning algorithm that organizes data into a tree-like structure of decisions, where each internal node represents a feature-based decision, each branch indicates the outcome of that decision, and each leaf node signifies the final predicted class or label. It learns from labeled datasets during training, adjusting decision rules to minimize errors, and is widely used for classification tasks due to its simplicity, interpretability, and effectiveness in handling both numerical and categorical data.
K-Nearest Neighbors Classifier (KNN) used machine learning algorithm that is primarily used for its simplicity and ease of implementation. Because of its simplicity and convenience of use, the machine learning algorithm K-Nearest Neighbors Classifier (KNN) is utilized. No assumptions are necessary regarding the distribution of the fundamental data. In addition, it is adaptable to categorical and numeric data, rendering it a versatile option for a wide range of datasets utilized in classification and regression endeavors. This non-parametric technique bases its predictions on how similar the data points in a particular datasets are to one another. By comparison, K-NN is less susceptible to outliers than other algorithms. The K-NN algorithm locates the K nearest neighbors to a given data point using a distance metric, such as Euclidean distance, to determine its neighbors. By calculating the majority vote or average of the K neighbors, the class or value of the data point is subsequently ascertained. Using this method enables the algorithm to anticipate outcomes based on the local structure of the data and adjust to various patterns.
Random Forest works by creating a bunch of decision trees, each trained on a different random subset of the data. For each tree, only a random set of features is considered when making decisions. When you want a prediction for a new data point, each tree gives its own prediction, and the final result is determined by a majority vote (for classification) or averaging (for regression) of all the individual tree predictions. This ensemble approach makes Random Forest more accurate and less prone to over-fitting compared to a single decision tree [4].
Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression tasks. It works by mapping data points into a multi-dimensional space and finding the optimal hyperplane (separator) that maximizes the margin between different categories. Support vectors, the data points closest to the hyperplane, play a key role in determining its position. SVM is versatile and can handle non-linear patterns through the kernel trick, which transforms the input space. It excels in finding a robust and efficient decision boundary, making it widely applicable in various domains such as image recognition, text classification, and bio informatics.
GLM models enable the construction of a linear correlation between predictors and response variables, despite the fact that their underlying relationship is nonlinear. The link function establishes a connection between the response variable and a linear model, thereby enabling this.An extension of the conventional linear model, this statistical model accommodates a wider variety of data types and distributions. Briefly, it serves as a structure to facilitate comprehension of the interconnections among variables. GLM supports non-normally distributed data and many response types, including binary outcomes, counts, and proportions, in contrast to conventional linear models. The framework comprises three primary elements—a probability distribution, a linear predictor, and a link function.
Bidirectional Encoder Representations from Transformers (BERT) is a NLP technique language model that excels in understanding context. It is a deep learning model which recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions. BERT is a neural-network-based technique that uses interconnected nodes or neurons in a layered structure like resembles the

human brain. Bert is pretrained model so it understands the language very well, so it can be a good classifier. [5]

RESULTS

In this section, we evaluate and analyze the performance of various machine learning models. Accuracy was drastically increased after preprocessing the data. But to improve the accuracy further any model alone can't be sufficient so we tried to combine different models. In long run it is better to combine a machine learning model and deep learning model together for reliable and more accurate result. So we choose the models from both the machine and deep learning with highest accuracy scores i.e Random Forest (RF) and BERT. For the Hybrid model, the training and testing were performed on a higher performance computing device specifically designed for machine learning tasks.

We compare the accuracy of all models from base Paper. As you can see in Table I, the base paper experiment, the RF classifier achieved the highest accuracy of 93 %, while the SVM classifier achieved an accuracy of 91.50 %. And the BERT classifier achieved the worst accuracy of 73.46 %.

From Table I, in our paper , the RF classifier achieved the highest accuracy of 98.33 %, while the SVM classifier achieved an accuracy of 95 %. And the BERT classifier achieved the worst accuracy of 65 %.

Below are the accuracy of our models:

TABLE I: Accuracy

Model	Base Paper	Our paper
DT	87.00%	86.66%
KNN	65.50%	65.00%
RF	93.00%	93.33%
SVM	91.50%	91.00%
GLM	91.50%	92.50%
BERT	73.46%	65.00%

B. Hybrid Model Results

After improving the accuracy of model, as can be seen from Table II, when we took 500 each data points of Human written, ChatGPT and BARD generated text, the RF classifier achieved the accuracy of 88.36 %, while the BERT classifier achieved an accuracy of 86.04 %.And the Hybrid Models classifier achieved the accuracy of 96.34 % .

Similarly, we changes the dataset to 2500 data points of human text, 700 each data points of ChatGPT and BARD generated text, the RF classifier achieved the accuracy of 92.27 %, while the BERT classifier achieved an accuracy of 91.86 %. And the Hybrid Models classifier achieved the accuracy of 98.43 % .

Again, we changes the datasets 15000 data points of human text, 700 each data points of ChatGPT, BARD generated text, the RF classifier achieved the accuracy of 94.54 %, while the BERT classifier achieved an accuracy of 95.52 %.And the Hybrid Models classifier achieved the accuracy of 97.06 % . These additional Human data was added just to test the Accuracy in biased datasets. Although this was just an experiment and result were improved, but it was only due to increased no Human written text in of training and test datasets.

TABLE II: Hybrid Model Result

Human+GPT+BARD	RF Model	BERT Model	Hybrid Model
500+500+500	88.36%	86.04%	96.34%
2500+700+700	92.72%	91.86%	98.43%
15000+700+700	94.54%	95.52%	97.06%

C. Figures

On the figures below: ChatGPT - 0, Human - 1, Bard -2. Below results are generated on 1500 datasets (500 each of ChatGpt, Human and Bard) at random_state=42.

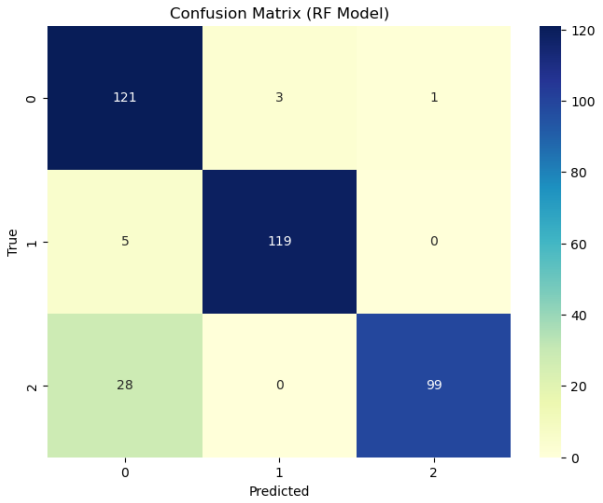

Fig. 1. Confusion Matrix of RF Model

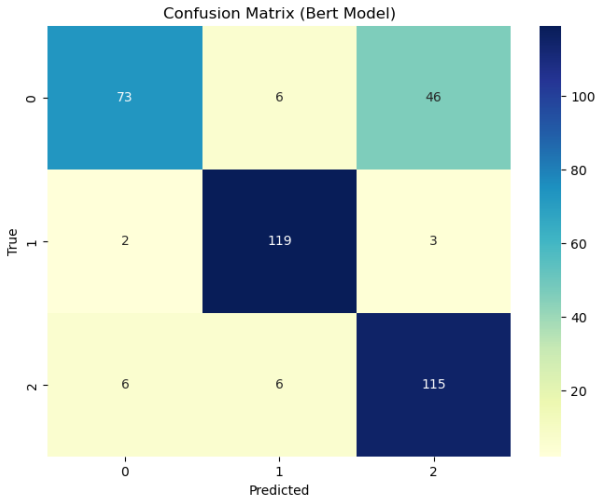

Fig. 2. Confusion Matrix of BERT Model

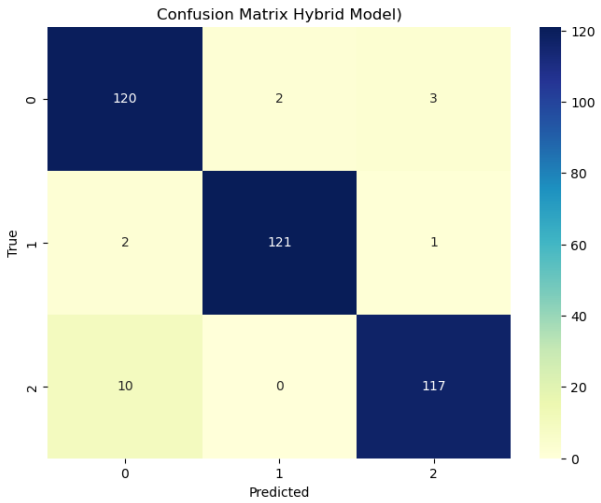

Fig. 3. Confusion Matrix of Hybrid Model

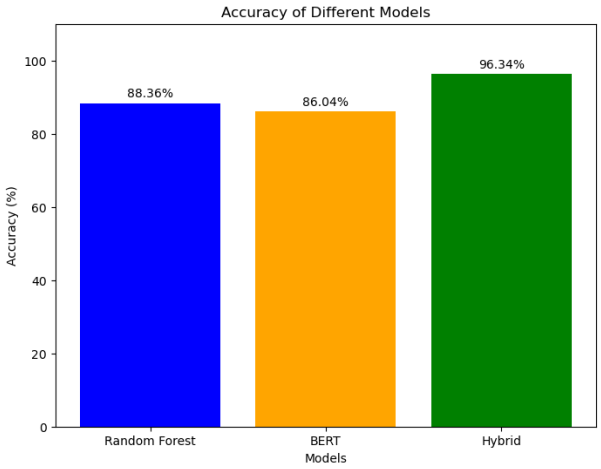

Fig. 4. Accuracy of different Models

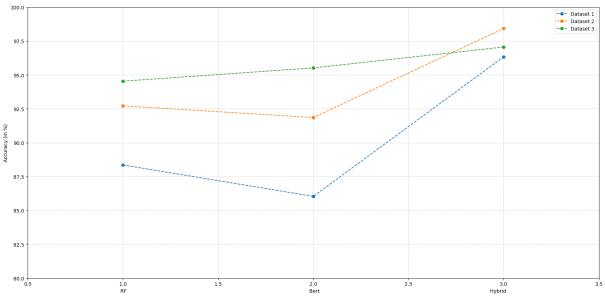

Fig. 5. Accuracy at different Datasets

V. Conclusions

In this paper, we estimated the efficiency and the accuracy of Machine and Deep Learning algorithm in distinguishing between Bard, ChatGPT-generated text and Human-written text. We found that alone a model cannot be accurate and reliable so we merged two good performing models in order to achieve higher results. From Fig 1 we can see that RF model was bad at predicting the Bard Generated Text and from Fig 2, we can see BERT is under-performing in detecting the ChatGPT generated text so by merging both of them Hybrid Model performed too well as seen in Fig 3 and 4. Combining two models imcreased the results drastically. Further datasets should be unbiased for better accuracy, increasing only one data can cause biases in datasets, although it will show increased accuracy because of method we used. But if we check its accuracy on other datasets which wasn't used for training the accuracy will increase as the model will predict more Human because it was trained on more Human data. Since we had only 500 data of ChatGPT and Bard we were not able to test this.

Finally if we increase our datasets it will increase the reliability of the models. Further it can be used to detect AI generated text in any work. For future we can make a website where user can upload their documents such as PDFs, Word File etc and then our model will detect the percentage of ChatGpt, Human and Bard Generated text. There are lot to do in this area, improvements can be done further.

References

[1] Islam, I., & Islam, M. N. (2023). Opportunities and challenges of chatgpt in academia: A conceptual analysis. Authorea Preprints.

[2] Alamleh, H., AlQahtani, A. A. S., ElSaid, A. (2023, April). Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. In 2023 Systems and Information Engineering Design Symposium (SIEDS) (pp. 154-158). IEEE.

[3] Abhishek Kumar, Manu Dev, Madugula Jayaram. Dataset of Human-text, BARD and ChatGPT generated text. (https://github.com/Abhishek-kumar0503/Dataset/tree/main), 2023

[4] Xu, B., Guo, X., Ye, Y., Cheng, J. (2012). An improved random forest classifier for text categorization. J. Comput., 7(12), 2913-2920.

[5] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.