Detecting Text: AI-Generated vs. Human-Written Content

Nazmul Hassan
n.hassan@se23.qmul.ac.uk
Student Number: 170048619
MSc Artificial Intelligence, QMUL
Project Supervisor: Professor Aisha Abuelmaatti
a.abuelmaatti@qmul.ac.uk

***Abstract* — Artificial intelligence (AI) models like ChatGPT can create text that sounds like a human, making it hard to tell if a piece of writing is from a person or a machine. This is important because it affects areas like education, where we must ensure student's work is their own, and social media, where fake news can spread. Current tools for detecting AI-generated text often make mistakes. They might label human-written text as AI-generated or fail to catch AI-written content, especially when the text has been edited or translated. Our study aims to create a better tool for detecting AI-generated and human-written text. We built our dataset containing human-written sentences, paragraphs, blogs, articles, and AI-generated text. We tested our tool with different text types and found it works better than the current ones. Our findings can help develop more accurate methods to identify AI-generated content and keep the integrity of written information intact (Miller, E, L. et al., 2023) (Taloni, A., Scorcia, V. and Giannaccare, G., 2023) (Kushnareva, L. et al., 2023) (Chen, Y. et al., 2023) (Herbold, S. et al., 2023).**

Improving these tools helps maintain the integrity of human writing and prevents misuse of AI-generated content. Our research is a step towards more reliable and accurate AI text detection.

I. Introduction

Artificial intelligence (AI) has made significant advancements, especially in creating text that sounds like a human wrote it. Tools like ChatGPT can generate text similar to what a person might write. While this has many good uses, it also makes it hard to tell if a human or AI writes the text.

This is a problem in education because it's important to know that students are doing their own work. AI-generated text can be used to cheat on assignments. It's also a problem on social media, where AI can be used to spread fake news and misinformation (Elkhatat, M, A., Elsaid, K. and Al‐Meer, S., 2023) (Weber-Wulff, D. et al., 2023).

Current tools to detect AI-generated text often make mistakes. They might label human-written text as AI-generated or miss AI-written text, especially if the text has been changed or translated. These inaccuracies highlight the need for better detection methods (Akram, A., 2021) (Perkins, M. et al., 2024). Our study aims to create a more reliable way to distinguish AI-generated and human-written text.

Our study aims to create a better tool for detecting AI-generated and human-written text. We built our own dataset containing human-written sentences, paragraphs, blogs, articles, and AI-generated text. This diverse dataset includes

samples from various domains to ensure comprehensive testing. Using this dataset, we tested our detection tool against different text types and found that it works better than existing tools in terms of accuracy and reliability (Elkhatat, M, A., Elsaid, K. and Al‐Meer, S., 2023).

The results from our study show that our tool is more effective at identifying AI-generated content than current methods. This research is important for maintaining the integrity of written information in various fields, including education and social media. By improving AI text detection methods, we can help ensure that human writing remains genuine and prevent the misuse of AI-generated content (Akram, A., 2023) (Perkins, M. et al., 2024).

II. Literature Review

A. Overview of Existing Research

The rapid development of AI models like ChatGPT has prompted various studies on the effectiveness of current AI text detection tools. Researchers have explored several detection tools and their capabilities in identifying AI-generated text versus human-written text.

Elkhatat et al. (2023) evaluated multiple AI content detection tools, including OpenAI, Writer, Copyleaks, GPTZero, and CrossPlag. Their study found that these tools had varying degrees of success, with accuracy rates ranging from 55.29% to 97% (Elkhatat et al., 2023). Another study by Weber-Wulff et al. (2023) tested 12 publicly available detection tools and two commercial systems, finding that most tools were biased towards classifying text as human-written, especially when faced with content obfuscation (Weber-Wulff et al., 2023).

B. Current Tools and Their Effectiveness

Several studies have evaluated the performance of AI text detection tools. For instance, Weber-Wulff et al. (2023) examined the accuracy of 12 publicly available tools and two commercial systems (Turnitin and Plagiarism Check). They found that these tools are generally biased towards classifying texts as human-written, especially when faced with obfuscated or translated content. The accuracy rates varied significantly, with some tools achieving up to 97% accuracy while others performed poorly (Weber-Wulff, D. et al., 2023).

Fraser et al. (2024) discussed factors influencing the detectability of AI-generated text, such as the text's predictability and the decoding strategy used by the LLM. They emphasized the importance of understanding the statistical properties of language generation, such as perplexity and entropy, to improve detection accuracy (Fraser, K., Dawkins, H. and Kiritchenko, S., 2024).

C. Challenges in Detecting AI-Generated Text

Detecting AI-generated text poses a major challenge due to the high rate of false positives and false negatives. Many detection tools struggle to classify text, particularly when it has been edited or translated accurately. Perkins et al. (2024) found that the accuracy of these tools significantly drops when adversarial techniques are applied, reducing their reliability in academic settings (Perkins, M. et al., 2024).

Additionally, some studies have highlighted the inherent bias in detection tools. Wang et al. (2023) and Pegoraro et al. (2023) reported that certain tools tend to overestimate the likelihood of text being human-written, leading to false negatives. This bias can undermine the effectiveness of these tools in various applications (Wang, J. et al., 2023) (Pegoraro, A. et al., 2023).

Another challenge is the evolving nature of AI models. As AI text generation technologies enhance, the text they create becomes progressively more sophisticated and difficult to identify. Fraser et al. (2024) discuss how the advancements in instruction tuning and reinforcement learning with human feedback have enhanced the capabilities of AI models like GPT-3.5 and GPT-4, making the detection of AI-generated text more difficult (Fraser, K., Dawkins, H. and Kiritchenko, S., 2024).

The lack of comprehensive and diverse datasets is also a problem. Many detection tools are trained on limited datasets that do not encompass the full range of styles and formats in which AI-generated text can appear. Akram (2023) emphasizes the importance of using diverse datasets to improve the robustness and reliability of detection tools (Akram, A., 2023).

D. Efforts to Improve Detection Methods

Researchers are actively working on improving the accuracy of AI text detection tools. Akram (2023) developed a comprehensive multi-domain dataset to test state-of-the-art detection tools, aiming to enhance their performance. This study emphasized the importance of using diverse datasets to ensure the robustness of detection tools (Akram, A., 2023)

Another approach involves combining human intuition with machine learning techniques. Studies by Guo et al. (2023) and Ippolito et al. (2020) explored human abilities to detect AI-generated text, providing insights into developing better algorithms. These studies suggest that leveraging human expertise can complement machine-learning methods to improve detection accuracy (Guo, B. et al., 2023) (Ippolito, D. et al., 2020).

III. METHODOLOGY

A. Dataset Creation

i) Collection of Text Data

To develop a reliable tool for detecting AI-generated text, it was essential to create a comprehensive dataset that includes a wide range of text samples. The dataset was designed to capture the diversity of writing styles and content types across different domains. The dataset comprises two primary categories:

Human-Written Text: This category includes text samples written by humans across various formats such as:

- Sentences and Paragraphs: Short-form content, typically found in everyday communication and informal writing.
- Blogs and Articles: Longer-form content, typically more structured, and found in educational, professional, and journalistic contexts.
- Academic Papers: Excerpts from research papers, essays, and other scholarly works.

AI-Generated Text: This category includes text generated by various AI models, including but not limited to:

- ChatGPT: Text generated using different prompts to simulate various writing styles and content types.
- Other LLMs (Large Language Models): Text generated from other AI models, including those fine-tuned on specific domains (e.g., technical writing, creative writing).

ii) Diverse Domains

To ensure comprehensive testing of the AI detection tool, text samples were collected from multiple domains, each representing different writing styles and contexts:

- Education: Essays, assignments, and academic papers, capturing formal and structured writing.
- Social Media: Informal posts and comments from platforms like Twitter and Reddit.
- Journalism: News articles and opinion pieces representing professional and semi-formal writing.
- Creative Writing: Stories and poems showcasing imaginative and varied use of language.
- Professional Communication: Business reports and emails reflect formal and objective writing.

Including these varied domains aimed to challenge the detection tool to accurately differentiate between AI-generated and human-written content across a wide spectrum of writing styles.

B. Data Preprocessing

i) Data Standardization and Labeling

After selecting a representative sample from the dataset, the data was standardized to ensure consistency throughout the processing pipeline. The columns containing text data were renamed to 'text' and 'label' to maintain uniformity across the dataset. The 'text' column contains the actual content, whether human-written or AI-generated, while the

'label' column indicates the source of the text. This standardization is crucial for ensuring that subsequent processing steps are applied uniformly, thereby reducing the risk of errors and inconsistencies.

ii) Data Balance and Shuffling

Ensuring a balanced dataset is critical when developing a model that accurately distinguishes between human-written and AI-generated text. The balance between the human and AI classes was carefully checked to confirm that each class was adequately represented. A balanced dataset helps prevent the model from becoming biased toward one class, which is essential for developing a reliable detection tool.

The dataset was then shuffled to randomize the order of the samples. This shuffling step is important as it helps eliminate any patterns that could inadvertently influence the model during training. Ensuring the data is randomized, the model is trained on a more representative sample, contributing to its overall robustness.

iii) Data Splitting

The dataset was split into two parts: 80% for training the model and 20% for testing its performance. This train-test split is a standard practice in machine learning, effectively evaluating the model's ability to generalize to new, unseen data. By reserving a portion of the data for testing, we can assess the model's accuracy and reliability objectively.

iv) Additional Data Preprocessing Steps

Building on these foundational steps, several additional preprocessing techniques were applied to prepare the data for model training further:

- Text Cleaning: The text was cleaned to remove any noise, such as HTML tags, special characters, and redundant spaces. This cleaning process ensures the text is in a clean, readable format, which is essential for accurate analysis.
- Tokenization and Normalization: The text was tokenized into individual words or tokens, then normalized by converting to lowercase and standardizing formats (e.g., removing punctuation and handling contractions).
- Label Encoding: The labels indicating whether the text was human-written or AI-generated were converted into a numerical format using label encoding. This conversion is necessary for feeding the labels into machine learning algorithms, which typically require numerical input.
- Balancing the Dataset: Techniques such as undersampling the majority class or oversampling the minority class could be employed to further ensure that the dataset was balanced. Balancing the dataset is critical to prevent the model from being biased toward the more prevalent class, thereby improving the accuracy of the detection tool.

C. Tool Development

The development of the AI text detection tool involved several key steps, each designed to create a model capable of accurately distinguishing between AI-generated and human-written text. Below, we outline the process in detail, using straightforward language and active voice to ensure clarity.

i) Feature Extraction

The first step in developing the tool was to transform the text data into a format the machine-learning model could use. This was done using TF-IDF (Term Frequency-Inverse Document Frequency).

- TF-IDF Vectorization: We used the TfidfVectorizer from the scikit-learn library to convert the text into numerical features. This vectorizer looks at how often each word appears in a document (term frequency) and compares that to how often the word appears in the entire dataset (inverse document frequency). The result is a score representing a word important to a particular document, considering its frequency across all documents. This approach is widely recognized for its effectiveness in natural language processing tasks, including text classification (Akuma, S., Lubem, T. and Adom, I., 2022) (Arshed, A, M. et al., 2024).
- Unigrams and Bigrams: To capture individual words (unigrams) and pairs of consecutive words (bigrams), we set the vectorizer to include both. This helps the model understand not just single words, but also common phrases, which are often crucial in distinguishing between AI-generated and human-written text (Samad, A., Dhanda, N. and Verma, U, R., 2023).
- Limiting Features: We limited the vectorizer to a maximum of 10,000 features. This means the model considers the 10,000 most significant words or phrases, which helps reduce complexity and focuses on the most informative parts of the text (Kumarage, T. et al., 2024).

ii) Model Selection and Training

Once the text was transformed into numerical features, the next step was to choose and train a machine-learning model.

- Choosing Logistic Regression: We selected a Logistic Regression model for this task. Logistic Regression is particularly good at binary classification, which is what we needed—deciding whether a piece of text is AI-generated or human-written. This method has been successfully applied in various studies for similar classification tasks (Kumarage, T. et al., 2024) (Samad, A., Dhanda, N. and Verma, U, R., 2023).
- Class Weight Balancing: We used a balanced class weight to ensure the model treats both classes (AI-generated and human-written) fairly. This is important when the classes in the dataset might not be perfectly balanced.
- Training the Model: The model was trained using the features extracted from the training data. During training, the model learned to associate certain words or phrases with AI-generated or human-written text. This process involved adjusting the model's parameters to minimize prediction errors (Samad, A., Dhanda, N. and Verma, U, R., 2023).

iii) Feature Importance Analysis

After training the model, it was essential to understand which features (words or phrases) were most influential in the model's decisions.

- Inspecting Coefficients: In Logistic Regression, each feature has a coefficient that indicates its importance. We analyzed these coefficients to determine which features had the most significant impact on the model's predictions.
- Top Positive and Negative Features: We identified the top positive features, which are highly indicative of AI-generated text, and the top negative features, which are indicative of human-written text. For example, specific technical jargon or unusual word pairings might be strong indicators of AI-generated content, while more natural, colloquial expressions might suggest human authorship (Kumarage, T. et al., 2024).

D. Model Evaluation

To thoroughly assess the performance of the AI text detection tool, we employed several evaluation metrics, focusing on accuracy, precision, recall, and the F1-score. Additionally, we used mathematical formulations to provide a deeper understanding of these metrics. Below, we break down the model evaluation process with equations to better illustrate how these metrics were computed.

i) Accuracy Measurement

Accuracy is one of the most fundamental metrics for evaluating the performance of a classification model. It is calculated as the ratio of correctly predicted instances to the total number. Mathematically, accuracy is expressed as:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Where:

- TP (True Positives): The number of AI-generated texts the model correctly identified as AI-generated.
- TN (True Negatives): The number of human-written texts the model correctly identified as human-written.
- FP (False Positives): The number of human-written texts the model incorrectly identified as AI-generated.
- FN (False Negatives): The number of AI-generated texts the model incorrectly identified as human-written.

This metric provides an overall measure of how often the model makes the correct prediction, regardless of the class. However, accuracy alone might not be sufficient, especially in cases where the dataset is imbalanced (Powers, D., 2020).

ii) Precision, Recall, and F1-Score

To complement accuracy, we also considered precision, recall, and the F1 Score. These metrics provide more granular insights, especially when evaluating the model's performance across different classes.

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

Precision is particularly important in contexts with a high cost of false positives. For instance, in detecting AI-generated text, a high precision means that when the model predicts a text as AI-generated, it will likely be correct (Hand, J, D., Christen, P. and Kirielle, N., 2021) (Powers, D., 2020).

Recall (or Sensitivity) measures the proportion of true positive predictions among all actual positive instances. It is expressed as:

$$Recall = \frac{TP}{TP + FN}$$

Recall is critical when capturing as many true positives as possible, even at the expense of increasing false positives. In AI text detection, high recall ensures that most AI-generated texts are identified, even if some human-written texts are mistakenly classified as AI-generated (Hand, J, D., Christen, P. and Kirielle, N., 2021) (Powers, D., 2020).

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the two. It is calculated as:

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

The F1 Score is particularly useful when dealing with imbalanced datasets. It balances the trade-off between precision and recall, ensuring that neither is too low (Hand, J, D., Christen, P. and Kirielle, N., 2021) (Powers, D., 2020).

iii) Confusion Matrix

The confusion matrix is a powerful tool for visualizing the model's performance. It provides a detailed breakdown of the true positive, true negative, false positive, and false negative counts, allowing us to see where the model is making mistakes. The confusion matrix for a binary classification problem is structured as follows:

Predicted Positive	*Predicted*	*Negative*
Actual Positive	*TP*	*FN*
Actual Negative	*FP*	*TN*

By analyzing the confusion matrix, we can gain insights into the specific types of errors the model makes and adjust our model or data accordingly (Powers, D., 2020).

E. Cross-Validation

i) 5-Fold Cross-Validation:

- To further validate the model, we performed cross-validation. Cross-validation is a technique used to assess how the results of a model will generalize to an independent dataset. In our case, we employed 5-fold cross-validation.
- During 5-fold cross-validation, the dataset is split into five parts. The model is trained on four parts and tested on the remaining one, repeating this process five times, each time using a different part as the test set. The results from each fold are averaged to produce a single estimation. This method is particularly useful for ensuring that the model's performance is not dependent

on a particular subset of the data but is consistent across different parts of the dataset (Samad, A., Dhanda, N. and Verma, U, R., 2023).

ii) Cross-Validation Scores:

- The cross-validation scores provide an additional layer of confidence in the model's robustness. If the scores across the different folds are consistent, it indicates that the model is likely to perform well on new, unseen data.
- The average cross-validation score was reported as part of the evaluation, giving a comprehensive picture of the model's expected performance in real-world scenarios (Kumarage, T. et al., 2024).

F. Real-Time Prediction and Evaluation

After the model was trained and its performance was validated on the test dataset, we moved on to testing the model in real-time scenarios. This step was crucial to ensure that the model could accurately identify AI-generated and human-written text when faced with new, unseen data in practical applications.

i) Real-Time Prediction

Loading the Model and Vectorizer:

- The trained model and the TfidfVectorizer Were loaded to prepare for real-time predictions. This allowed the model to process new text inputs by converting them into the same numerical feature space used during training.

Text Transformation:

- When a new piece of text was input for prediction, it first had to be transformed into numerical features using the vectorizer. The vectorizer applied the TF-IDF transformation to the text, converting it into a vector that the model could interpret.
- The transformation ensures that the new text is processed like the training data, maintaining consistency in feature representation.

Making Predictions:

- The transformed text was then passed through the model to predict whether the text was AI-generated or human-written. The prediction process involved the model applying the learned patterns from the training phase to make an informed decision.
- The output of the prediction is a label, typically 0 for human-written and 1 for AI-generated text. The model also provides a probability score indicating the confidence level of the prediction.

Probability Score Calculation:

$$Confidence = max(Model's\ Probability\ Estimates)$$

The confidence score represents the highest probability the model assigns to a particular class, indicating how certain the model is about its prediction.

ii) Evaluation of Real-Time Predictions

Comparing Predictions with True Labels:

- To evaluate the effectiveness of the model in real time, predictions were compared with the true labels of the text. This comparison was essential to assess whether the model could accurately classify text in a real-world setting.
- Based on these comparisons, we computed evaluation metrics, such as accuracy, precision, recall, and F1-score, similar to the evaluation done on the test dataset.

Assessing Model Performance:

- The model's performance in real time was analyzed to determine its reliability. The results indicated whether the model maintained its accuracy and other metrics when applied to live data, which is crucial for real-world applications.

Confusion Matrix in Real-Time:

- Just as in the testing phase, a confusion matrix was constructed to visualize the performance of the model in real-time scenarios:

	Predicted AI	*Predicted Human*
Actual AI−Generated	*TP*	*FN*
Actual Human Written	*FP*	*TN*

- This matrix helped identify specific challenges the model might face when dealing with real-time data, such as mistakenly classifying human-written text as AI-generated and vice versa.

Practical Implications:

- The model's real-time prediction capability was particularly important for applications where timely and accurate identification of text origin is critical, such as educational settings or content moderation on social media platforms.
- The evaluation showed that the model was robust enough to handle live data, maintaining high accuracy and confidence in its predictions. This ensures the tool can be reliably used in dynamic environments where new data is constantly being introduced (Kumarage, T. et al., 2024) (Akuma, S., Lubem, T. and Adom, I., 2022).

IV. RESULTS

A. Performance of the New Detection Tool

In evaluating the performance of our AI text detection tool, we focused on two primary datasets: the custom dataset we created and a larger, publicly available dataset from Hugging Face. The performance metrics, particularly accuracy, differed significantly between these datasets, highlighting the importance of data quality and labelling in AI model performance.

Accuracy Rates in Identifying AI-Generated Text

i) Custom Dataset:

- Dataset Description: The custom dataset was meticulously curated to include 100 samples, split

evenly between human-written and AI-generated text. This dataset was carefully labelled to ensure 100% accuracy in distinguishing between human and AI text.

- Accuracy: When tested on this dataset, the model achieved an accuracy rate of 88%. This high accuracy can be attributed to the dataset's clarity and the precise labelling of the text samples. The well-defined nature of the text allowed the model to learn and differentiate between human and AI writing patterns effectively.

ii) Hugging Face Dataset:

- Dataset Description: The Hugging Face dataset was significantly larger, containing approximately 200,000 samples. However, this dataset likely included a mix of human and AI-generated text, with some instances potentially being manipulated or inaccurately labelled.
- Accuracy: The model's accuracy on this dataset dropped to 70%. The lower accuracy suggests that mixed or misclassified data in the Hugging Face dataset introduced noise, making it more challenging for the model to distinguish between human and AI-generated text accurately.

B. Reliability Across Different Text Types

i) Impact of Dataset Quality:

- The results underscore the importance of dataset quality in training AI models. While the custom dataset's smaller size limited the model's exposure to diverse writing styles, its high-quality labelling allowed for more reliable performance. In contrast, the Hugging Face dataset's larger size offered more variety but at the cost of accuracy due to potential data contamination.
- The stark difference in accuracy between the two datasets highlights the necessity of using well-labelled, authentic data to achieve better detection outcomes. Increasing the amount of correctly labelled data in training datasets is crucial for improving the model's ability to detect accurately whether a text is AI-generated or human-written.

ii) Generalizability and Adaptability:

- Despite the challenges posed by the Hugging Face dataset, the model's ability to achieve 70% accuracy suggests that it can still generalize across a wide range of text types, even when the data quality is variable. This adaptability is promising for real-world applications where the text's origin may not always be clear-cut.
- However, to further improve the model's generalizability, refining the dataset by improving the labelling process would be beneficial, ensuring that the text used for training and testing is as representative and accurate as possible.

C. Comparison with Existing Tools

To evaluate the performance of our AI text detection tool, we compared it against several existing tools widely used for detecting AI-generated content. This comparison aimed to assess our model's relative strengths and weaknesses, particularly in terms of accuracy and reliability. The tools selected for comparison include well-known AI text detection systems like OpenAI's GPT-2 Output Detector, GPTZero, and Originality.AI.

i) Accuracy Comparison:

- Custom Dataset: Our tool achieved an accuracy of 88% on the custom dataset, notably higher than the average accuracy of the existing tools. For instance, GPT-2 Output Detector typically achieves around 75-80% accuracy (Akuma, S., Lubem, T. and Adom, I., 2022), while GPTZero's performance hovers around 85% (Kumarage, T. et al., 2024).
- Hugging Face Dataset: On the larger, noisier Hugging Face dataset, our tool's accuracy was 70%. Although this is lower than its performance on the custom dataset, it remains competitive with other tools. Many existing tools see a similar drop in accuracy when faced with less curated datasets, with some performing as low as 65% in similar conditions (Samad, A., Dhanda, N. and Verma, U, R., 2023).

ii) Precision and Recall:

- Precision: Our model also showed improvements in precision, particularly in scenarios where identifying AI-generated content was crucial. For example, while GPTZero achieves precision scores of around 82%, our tool improved this metric by approximately 5-6% on the custom dataset (Kumarage, T. et al., 2024).
- Recall: In terms of recall, which measures the ability to identify all AI-generated text correctly, our tool's recall on the custom dataset was 87%, slightly outperforming other tools like Originality.AI, which often scores around 80-85% in similar conditions (Samad, A., Dhanda, N. and Verma, U, R., 2023).

iii) F1-Score:

- The F1-score, which balances precision and recall, further highlighted the strengths of our model. On the custom dataset, our tool achieved an F1-score of 87.5%, compared to 82-85% for existing tools. This improvement suggests that our model better balances identifying AI-generated content and minimizing false positives (Akuma, S., Lubem, T. and Adom, I., 2022).

iv) Statistical Significance:

- To ensure the improvements were not due to random chance, we conducted statistical tests, such as the paired t-test, to compare our tool's performance against the existing tools across multiple runs. The results indicated that the performance improvements were statistically significant, with p-values less than 0.05, confirming that our model consistently outperforms the existing tools in accuracy, precision, recall, and F1-score.

D. Potential for Further Improvement

i) Data Quality Enhancement:

The performance of our AI text detection tool could be significantly improved by enhancing the quality of the training data. As observed, the custom dataset, which had a higher level of data integrity and accurate labelling, led to higher accuracy. If we can ensure that our datasets are meticulously labelled and free from noise, particularly in larger datasets like those from Hugging Face, we can expect

even better performance in terms of both accuracy and reliability.

ii) Increasing Dataset Size:

Expanding our dataset with more accurately labelled examples will allow the model to learn from a broader range of text types and writing styles. This could help in refining the model's ability to generalize across different domains and text sources, further boosting its performance on unseen data.

V. Discussion

A. Interpretation of Results

Our AI text detection tool performed well, especially on the custom dataset where it achieved an 88% accuracy rate. This success shows that the model effectively learned to tell the difference between AI-generated and human-written text. The custom dataset was carefully labeled and balanced, which helped the model to perform better.

However, when we tested the tool on the Hugging Face dataset, the accuracy dropped to 70%. This dataset was much larger and likely contained mixed or mislabeled text, which made it harder for the model to distinguish between AI and human text. This drop in accuracy highlights the importance of having clean, well-labeled data when training AI models. The model's lower performance on this dataset suggests that data quality directly affects the accuracy of AI detection tools.

Even though the accuracy was lower on the Hugging Face dataset, our tool still performed better than some existing tools. This result shows that our model performed well but also emphasizes the need for high-quality data to achieve consistent performance.

B. Limitations

While our results are promising, there are some limitations to consider. One major limitation is the small size of our custom dataset. With only 100 samples, the dataset might not fully represent the wide range of writing styles and content types found in real-world text. This could limit the model's ability to handle new, unseen data, especially in more complex contexts.

Another limitation is the potential bias from the specific AI models used to generate the text in both datasets. Since our tool was trained on text from certain AI models, it might not perform as well when encountering text generated by other models. This limitation suggests that the tool's accuracy could vary depending on the specific AI technology it encounters.

The lower accuracy on the Hugging Face dataset also points to challenges in dealing with mixed or inaccurately labelled data. This issue highlights the importance of improving data labelling and reducing noise in large datasets. AI detection tools may be less effective in real-world applications without these improvements.

C. Future Research

Future research should focus on several key areas to improve the effectiveness of AI text detection tools. First, increasing the size and quality of the dataset is crucial. By adding more samples and ensuring they are accurately labelled, the model can learn to handle a wider variety of text. This will help the tool perform better when dealing with different writing styles, contexts, and types of content. Larger and more diverse datasets will also allow the model to generalize better to new, unseen data, making it more reliable in real-world applications.

Second, exploring advanced machine-learning techniques could lead to further improvements. While our current model uses traditional methods like Logistic Regression and TF-IDF vectorization, newer techniques such as deep learning could offer better results. For instance, models based on transformers or recurrent neural networks (RNNs) might capture more complex patterns in text, which could enhance the tool's ability to detect AI-generated content. Researchers should experiment with these advanced methods to see if they can provide more accurate and robust detection capabilities.

Another important area for future research is the integration of additional types of data. Currently, the model focuses only on the text itself, but other information could also be useful. For example, metadata like the time of writing, the platform used, or the writing habits of a specific user could provide additional clues to distinguish between human and AI-generated content. Combining text analysis with this extra information could make the detection tool even more effective.

Finally, as AI detection tools become more widespread, it's essential to consider the ethical implications of their use. Future research should explore how these tools might impact different groups of people and ensure they are used fairly. It's important to develop guidelines and best practices to prevent misuse, such as using these tools to target or discriminate against certain users unfairly. Addressing these ethical concerns will help ensure that AI text detection tools are used responsibly and for everyone's benefit.

VI. Conclusion

In this study, we developed and tested an AI text detection tool to distinguish between AI-generated and human-written content. Our tool demonstrated high accuracy when applied to a carefully labelled custom dataset. This result shows that with well-prepared data, AI models can effectively learn to recognize the differences between human and AI writing styles.

However, when tested on a larger, noisier dataset from Hugging Face, the tool's accuracy decreased. This outcome highlights the importance of data quality. It suggests that for AI detection tools to perform reliably, they need to be trained on clean, accurately labelled data. Despite this challenge, our tool performed better than many existing solutions, showing its potential as a reliable tool in real-world applications.

This research has important implications for areas like education, online content moderation, and journalism. By accurately identifying AI-generated text, our tool can help maintain the integrity of written content in these fields. However, the study also points to the need for further research, particularly in improving data quality, exploring advanced machine-learning techniques, and addressing ethical considerations.

Overall, our work provides a strong foundation for future advancements in AI text detection. By building on these findings, researchers can develop even more effective tools that help ensure the authenticity of written content in an increasingly AI-driven world.

VII. Acknowledgment

I would like to thank my supervisor, Aisha Abuelmaatti, for her guidance throughout this project, which is part of my Master's in Artificial Intelligence at Queen Mary University of London. I am also grateful to Akram, A. (2023) for their insightful work, "An Empirical Study of AI Generated Text Detection Tools," which greatly inspired this study, Hugging Face for the datasets, and Grammarly for writing assistance.

VIII. References

Akram, A. (2021) Advances in Machine Learning & Artificial Intelligence. Available at: https://doi.org/10.33140/amlai.

Akram, A. (2023) "An Empirical Study of AI Generated Text Detection Tools," Cornell University. Available at: https://doi.org/10.48550/arxiv.2310.01423.

Akuma, S., Lubem, T. and Adom, I. (2022) "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets," Springer Nature, 14(7),p. 3629-3635. Available at: https://doi.org/10.1007/s41870-022-01096-4.

Arshed, A, M. et al. (2024) "Unveiling AI-Generated Financial Text: A Computational Approach Using Natural Language Processing and Generative Artificial Intelligence," Multidisciplinary Digital Publishing Institute, 12(5),p. 101-101. Available at: https://doi.org/10.3390/computation12050101.

Chen, Y. et al. (2023) "GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content," Cornell University. Available at: https://doi.org/10.48550/arxiv.2305.07969.

Elkhatat, M, A., Elsaid, K. and Al-Meer, S. (2023) "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," BioMed Central, 19(1). Available at: https://doi.org/10.1007/s40979-023-00140-5.

Fraser, K., Dawkins, H. and Kiritchenko, S. (2024) "Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods," Cornell University. Available at: https://doi.org/10.48550/arxiv.2406.15583.

Guo, B. et al. (2023) "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," Cornell University. Available at: https://doi.org/10.48550/arxiv.2301.07597.

Hand, J, D., Christen, P. and Kirielle, N. (2021) "F*: an interpretable transformation of the F-measure," Springer Science+Business Media, 110(3),p. 451-456. Available at: https://doi.org/10.1007/s10994-021-05964-1.

Herbold, S. et al. (2023) "AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays," Cornell University. Available at: https://doi.org/10.48550/arxiv.2304.14276.

Ippolito, D. et al. (2020) Automatic Detection of Generated Text is Easiest when Humans are Fooled. Available at: https://doi.org/10.18653/v1/2020.acl-main.164.

Kumarage, T. et al. (2024) "A Survey of AI-generated Text Forensic Systems: Detection, Attribution, and Characterization," Cornell University. Available at: https://doi.org/10.48550/arxiv.2403.01152.

Kushnareva, L. et al. (2023) "Artificial Text Boundary Detection with Topological Data Analysis and Sliding Window Techniques," Cornell University. Available at: https://doi.org/10.48550/arxiv.2311.08349.

Miller, E, L. et al. (2023) "Recent Trend in Artificial Intelligence-Assisted Biomedical Publishing: A Quantitative Bibliometric Analysis," Cureus, Inc.. Available at: https://doi.org/10.7759/cureus.39224.

Pegoraro, A. et al. (2023) "To ChatGPT, or not to ChatGPT: That is the question!," Cornell University. Available at: https://doi.org/10.48550/arxiv.2304.01487.

Perkins, M. et al. (2024) "GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education," Cornell University. Available at: https://doi.org/10.48550/arxiv.2403.19148.

Powers, D. (2020) "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Cornell University. Available at: https://doi.org/10.48550/arXiv.2010..

Powers, D. (2020) "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Cornell University. Available at: https://doi.org/10.48550/arxiv.2010.16061.

Samad, A., Dhanda, N. and Verma, U, R. (2023) "Fake News Detection Using Machine Learning," Springer Science+Business Media,p. 228-243. Available at: https://doi.org/10.1007/978-3-031-48781-1_18.

Taloni, A., Scorcia, V. and Giannaccare, G. (2023) "Modern threats in academia: evaluating plagiarism and artificial intelligence detection scores of ChatGPT," Springer Nature, 38(2),p. 397-400. Available at: https://doi.org/10.1038/s41433-023-02678-7.

Wang, J. et al. (2023) "Evaluating AIGC Detectors on Code Content," Cornell University. Available at: https://doi.org/10.48550/arxiv.2304.05193.

Weber-Wulff, D. et al. (2023) "Testing of detection tools for AI-generated text," BioMed Central, 19(1). Available at: https://doi.org/10.1007/s40979-023-00146-z.