Unmasking AI-Generated Texts Using Linguistic and Stylistic Features

Muhammad Irfaan Hossen Rujeedawa[1], Sameerchand Pudaruth[2], Vusumuzi Malele[3]
Department of Information and Communication Technologies, FoICDT, University of Mauritius, Reduit, Mauritius[1, 2]
School of Computer Science and Information Systems, Vaal Campus, North-West University, Vanderbijlpark, South Africa[3]

***Abstract*—As Artificial Intelligence (AI) generated texts become increasingly sophisticated, distinguishing between human-written and AI-generated content presents a growing challenge. Reliably detecting AI-generated texts is of primary importance in fields that involve a lot of text such as journalism, education and law. In this study, several methods for detecting AI-generated texts by analysing a range of linguistic and stylistic features were investigated. It incorporated features such as text length, punctuation count, vocabulary richness, readability indices and sentiment polarity, to identify patterns in AI-generated content. Out of the six machine learning classifiers which were tested, the Random Forest classifier achieved the highest accuracy of 82.6%. A dataset of 483,360 essays was used in this study. Thus, the findings of this study provide a framework for the development of more sophisticated detection tools that can be applied to various real-world scenarios.**

***Keywords*—AI-generated texts; human-written texts; machine learning; linguistic features; stylistic features**

I. Introduction

In today's world, with Artificial Intelligence (AI) being widely used, it is crucial to ensure that the information encountered is authentic. The technology behind AI has improved to the point that computers are now capable of generating texts that closely mimics human writing. For example, imagine reading an article or news report, only to discover that the writing does not come from a human being but from an AI. This can significantly impact the readers' trust and confidence in digital media. This worrying situation not only highlights the need for us to dig deep into AI-generated texts, but also to detect the texts generated by these wordsmiths. Unmasking AI-generated text involves distinguishing between texts created by an AI and those authored by humans. This has become very important as AI continues to penetrate deeper into many aspects of our daily lives. It is significant in the academic field in maintaining academic integrity by preserving the authenticity of academic works, by ensuring that they are human-authored and not AI-generated. The latest advancements in AI, especially in natural language processing (NLP), have contributed to challenges such as the dissemination of false information and cases of identity fraud [1]. The use of AI technology has led to more artificial texts in different areas. While this has its benefits, it also brings challenges regarding how trustworthy and reliable the information could be.

As AI technologies advance and are being used in more areas, especially in the educational field, the challenge associated with detecting AI-generated content is becoming more complex. Many detection models face challenges in effectively distinguishing between human and AI-generated texts due to difficulties in finding differences in linguistic patterns and stylistic details. The increasing sophistication of AI models will make this task even more challenging. This research aims to address this gap by investigating advanced linguistic and stylistic features, including readability scores such as the Flesch Reading Ease and the Gunning Fog Index, vocabulary richness and sentiment polarity. By exploring these features deeper, this research aims to enhance detection accuracy and provide a better understanding of the differences between human and AI-generated content. Additionally, understanding these nuances can help educational institutions develop better policies regarding the use of AI-generated content in academic settings. The potential benefits of this work can help maintain the integrity of digital content, ensure authenticity of content and prevent the potential misuse of AI technologies. Furthermore, it could also be applied in various other fields such as journalism, and legal documentation, where the authenticity of text is of primary importance.

AI-generated texts can mean many things, like chatbot conversations, content creation, and automated translation from one language to another. While the research aims to develop a reliable method for detecting texts generated by an AI, it is important to note that the field of AI is rapidly evolving. Continuous research will be required to ensure detection methods remain effective as AI technologies advance.

This paper is organised as follows. Section II presents the literature review on techniques to recognise AI-generated texts. Section III presents the methodologies used in creating the dataset and developing the system. The results are presented and discussed in Section IV and Section V concludes the paper.

II. Literature Review

This section looks at the research and methods that have been developed to spot AI-generated text. It explores what has been achieved so far and what challenges still exist. By reviewing the work done in this field, this section aims to give a clear understanding of the current techniques and how effective they are, helping to guide the development of new approaches.

Shah et al. [1] explored different methods for identifying AI-generated texts and discussed various ways for detecting such texts, including syllable count, the length of words and length of sentences. The study employed different machine learning algorithms, Explainable AI (xAI) libraries (LIME and SHAP), and stylistic features (readability, lexical features and

variety and depth of vocabulary). A dataset using Wikipedia articles and two large language models (LLMs) to generate 10,000 articles from each were utilized. The LLMs were combined and shuffled to create two final datasets for the experiments. XAI analysis was performed to determine which features had the highest impact on determining the classification of an article. The xAI analysis revealed that Herdan's C had the highest impact on classification, with a metric of 0.92 for AI texts and 0.89 for human texts. Their ensemble model showed impressive effectiveness, achieving a precision reaching 93% in distinguishing between AI-authored and human-written text.

Elkhatat et al. [2] assessed the efficiency of several AI-generated content detection systems in differentiating between human-made and AI-generated content. The researchers generated 15 paragraphs each from ChatGPT 3.5 and 4, discussing cooling towers in the engineering process, along with five human-generated control responses, for assessment purposes. They used tools for detecting AI-generated content developed by Copyleaks, GPTZero, OpenAI, Writer and Crossplag to classify these paragraphs. The findings for the contents produced by GPT 3.5 indicated a strong level of consistency. However, GPTZero and WRITER classified some AI-generated content as "very unlikely AI generated" and "unclear if AI generated," respectively. However, the result of the detectors on GPT-4 content was not as reliable. Some GPT-4 content got "very unlikely AI generated" results from Crossplag, Writer and GPTZero. When looking at the control responses, it was clear that the effectiveness of the detectors was not completely trustworthy, as many of the human-generated texts resulted in "likely AI generated" by Writer and GPTZero. When examining the outcome of the result of GPT 3.5, the OpenAI Classifier was best at spotting AI-generated content, getting a perfect score of 100%. However, it struggled more with recognizing human-generated content, scoring 0% in this area. GPTZero did well overall, with a 93% score for spotting AI content and 80% for human content. GPT 4 had lower scores overall, with Copyleaks being the best at spotting AI content with 93%, and Crossplag being the best at recognizing human content with 100% accuracy.

Ma et al. [3] investigated the distinction between AI-generated and human-generated scientific content, focusing on scenarios where scientific AI writing assistants are extensively used in scientific writing. They assembled a dataset comprising human-written abstracts and AI-generated abstracts, created from LLMs using optimised prompts containing scientific information. The researchers conducted a human evaluation to detect AI-generated texts. Evaluators were presented with 20 scientific paper abstracts and 20 Wikipedia item descriptions, some of which were human-written and some generated by ChatGPT. The human evaluators achieved a 66% F1 score. Based on the results of human evaluations, the authors created a framework to describe features that can distinguish text authored by AI from text produced by humans. This framework is based on syntax, semantics and pragmatics. The framework categorised features into four dimensions: writing style, coherence, consistency and argument logistics. To statistically analyse the differences from human-generated and AI-written texts, the researchers built separate logistic models for syntax, semantics, and pragmatics. Subsequently, they applied the RoBERTa large OpenAI Detector to the test dataset, achieving an F1 score of 88.3%.

Crothers et al. [4] conducted an extensive survey on the threat models posed by modern text generation systems, as well as the existing ways for detecting machine-generated texts. The survey categorised natural language generation (NLG) approaches into both neural and non-neural methods. Recent non-neural methods have employed reinforcement learning, particularly hierarchical reinforcement learning which uses Markov Decision Process (MDP) agents to develop ideal policies for generating texts. Deep reinforcement learning employing neural networks has been applied to understand policy gradient methods. The analysis of threat models identified four major attack categories: facilitating malware and social engineering, online influence campaigns, exploiting AI authorship, and spam and harassment. Statistical techniques are used to differentiate between text generated by machines and text generated by humans. They also discussed NLM-based approaches, including zero-shot classification as well as fine-tuning pre-trained language models like BERT. Additionally, human-aided methods were explored, which combined statistical and neural approaches with human analyst review for text detection. The paper offers a summary of threat models and methods for the identification of AI-generated texts, highlighting the advancements and challenges in this field.

Wang et al. [5] introduced SeqXGPT, a novel system for spotting AI-generated texts (AIGT) on a sentence-by-sentence basis, as opposed to classifying entire documents. The authors proposed different setups for AIGT detection tasks: (1) Particular-Model Binary AIGT Detection, which distinguishes text written by a specific known AI system from human-written text; (2) Mixed-Model Binary AIGT Detection, which identifies AI-generated content without identifying the exact model of origin; and (3) Mixed-Model Multiclass AIGT Detection, where the objective is to both identify the model who generated the text and detect AIGT. The datasets used were obtained from documents in SnifferBench, which includes human-written and AI-authored sentences. SeqXGPT looks at each sentence in a document one by one and decides if it was created by an AI or not. SeqXGPT consists of three main parts: (1) Perplexity Extraction and Alignment involves extracting lists of token-wise log probabilities from various public open-source models, which serve as the original features. (2) The Feature Encoder processes list of word-wise log probabilities as features that represent how well a model understands semantic and syntactic structures. It employs convolutional networks to extract local features from the input, converting them into a hidden feature space. These resulting features are then passed to a context network based on self-attention layers, enabling the model to capture long-range dependencies and generate contextualized features; and (3) Linear Classification Layer, a straightforward linear classifier is trained to assign each word's features to various labels, ultimately selecting the most common label as the sentence's final category. SeqXGPT achieved a 97.6% F1 score on Particular-Model Binary AIGT Detection, a 95.7% F1 score on

Mixed-Model Multiclass AIGC Detection, and a 95.3% F1 score on Mixed-Model Binary AIGT Detection.

Tulchinskii et al. [6] demonstrated that the intrinsic dimension of a text can be a valuable metric for distinguishing between natural and generated texts. They used the Wiki40b dataset for human text samples. For multilingual text detection experiments, they created a dataset called WikiM in 10 languages produced by GPT3.5-turbo. In experiments assessing cross-domain and paraphrase robustness, they used datasets from Wikipedia and Reddit. They used two consecutive sentences from Wikipedia or a question from Reddit as prompts to produce the texts using OPT13b, GPT2-XL and GPT3.5 and they produced a StackExchange dataset using GPT3.5. They estimated the dimension of each text sample by obtaining embeddings that are specific to each token in the text using a pre-trained transformer encoder. For English, they used RoBERTa-base, and for other languages, they used XLM-R. Each embedding was viewed as a location (point) in Euclidean space. They created a basic classifier for identifying artificial text which uses PDH (Persistence Homology Dimension) as the single feature and trained a logistic regression model using a dataset containing both human-written and AI-generated texts. The results revealed that the intrinsic dimension of human-generated texts typically ranges from 9 to 10, whereas for generated texts, it is around 8, regardless of the specific text generator used.

Mindner et al. [7] aimed to understand the distinctions between natural language and artificially written content. They used Wikipedia articles to create an English text corpus covering 10 different topics. They utilised five text corpora: basic AI-generated, basic AI-rephrased, advanced AI-generated, advanced AI-rephrased, and basic human texts. They incorporated various feature categories for classification and they created systems for detecting text generation, which were trained, fine-tuned, and evaluated using both basic AI-authored texts and human written texts. They developed detection systems for basic text rephrasing, which were trained, fine-tuned, and evaluated using human-generated and AI-altered texts. They developed sophisticated detection systems for text generation and systems to detect rephrased texts, which were trained, fine-tuned, and evaluated using both human-written and advanced artificially authored texts. They achieved an F1-score of 98.0% for distinguishing between basic man-made and artificially made texts, an F1-score of 78.9% for distinguishing between basic human-made and artificially rephrased texts. For advanced human-made and AI-made texts, they achieved an F1-score of 96.9%, and for advanced human-made and AI-rephrased texts, they achieved an F1-score of 81.7%.

Kumari et al. [8] presented a novel detector called DEMASQ. Its purpose is to reliably identify ChatGPT-generated information by addressing the differences in text composition biases between man and machine-made content, as well as human modifications used to circumvent earlier detection techniques. DEMASQ is a model based on energy that uses new components including a Doppler-effect-inspired optimization and explainable AI methods to produce a variety of perturbations. Three main elements make up DEMASQ's approach: a model based on energy that captures the behaviour of both human and ChatGPT activity, an adapted Doppler effect, and the Integrated Gradient (IG) method for assessing hybrid texts that combine information generated by ChatGPT and humans. The Doppler effect is used in the energy-based model to quantify energy, with waves standing in for texts and drumhead vibrations for source frequencies. Wave frequencies are added to the cost function of the model to improve the training process. DEMASQ was evaluated using a benchmark dataset that included human and ChatGPT questions from a variety of domains. DEMASQ significantly outperformed previous detection techniques, attaining an accuracy of up to 74.5%. The academic abstract datasets for Task 1, Task 2, and Task 3 were used to train their model [9]. These datasets categorised the CheckGPT analysis into three tasks: 1) Full abstracts authored by the GPT model, 2) Partially completed abstracts by the GPT model, and 3) Enhanced abstracts by the GPT model. After being retrained on the Task 1, Task 2, and Task 3 datasets, DEMASQ showed accuracy rates of 96.4%, 88.7%, and 82.5% for the respective tasks. Additionally, DEMASQ was assessed in comparison to Task 1, 2, and Task 3's paraphrased texts. For each task (1, 2 and 3), the corresponding accuracy was 76.9%, 68.7%, and 58.3%, after rephrasing.

Bao et al. [10] introduced Fast-DetectGPT, which assumes that humans and machines choose different words when generating text. The method suggests that machine-generated text shows a specific pattern in the way word probabilities change. If the pattern has a certain shape (positive curvature), the text is flagged as AI-generated. Otherwise, if the pattern is more flat (close to zero), the text is likely to be human-written. Fast-DetectGPT uses a novel three step procedure: 1) Sample – generates alternative text samples, 2) Conditional Score – calculates the word probability pattern using a scoring model, and 3) Compare – compares the word probability patterns of the text and samples to determine the curvature. Six datasets were used to cover several topics and languages: XSum used for news articles, SQuAD for Wikipedia contexts, WritingPrompts for story writing, WMT16 English and German for different languages, and PubMedQA for biomedical research question answering. They randomly selected between 150 to 500 human-written examples from each dataset as negative samples and generated an equal number of positive samples. The authors compared their new system with DetectGPT, which employs a perturbation step alongside a more efficient sampling approach. The findings indicate that Fast-DetectGPT outperforms DetectGPT by approximately 75%.

Mitrovic et al. [11] focused on detecting ChatGPT-generated text in restaurant reviews. Their approach involved two main parts. They utilised a model trained to distinguish between human-written text and ChatGPT-generated texts. They started with a Transformer-based model that was pre-trained for classifying sequences. Next, they fine-tuned this model to identify whether a text sample was ChatGPT-generated or human-generated. Finally, they evaluated the model's performance by comparing its classification scores to the ground truth. The authors used three datasets, a publicly available dataset containing human made reviews about a restaurant and two datasets generated by ChatGPT consisting

of restaurant reviews. One of the generated datasets has been obtained by rephrasing the reviews from the human dataset. The authors conducted two experiments. In the first experiment, the human and the ChatGPT generated datasets were used while the human dataset and the ChatGPT rephrased dataset were used in the second experiment. In addition to their machine learning (ML) based approach, they created a way to classify text based on its perplexity score. First, they split the dataset containing human and ChatGPT-generated text into two sets. Then, they used GPT-2 to calculate the perplexity score for each text in the training group. Finally, they used the perplexity score from the training group to classify the text in the test group. The result showed that the ML-based approach outperformed the Perplexity-based approach in both experiments, achieving an accuracy of 98% compared to 84% in Experiment 1 and 79% compared to 69% in Experiment 2.

Yan et al. [12] compared essays written by humans with those written by an AI. They first did a detailed study with a small number of essays to look at different aspects. Then, they conducted a bigger research in which they developed and tested two detectors: one that utilised e-rater features and another that utilised a modified version of the RoBERTa language model. They used OpenAI's GPT-3 to generate AI essays and trained the RoBERTa model with a dataset of 8,000 essays, including 4,000 with added spelling mistakes. The fine-tuned RoBERTa model got a precision of 99.75%, outperforming the support vector classifier, which had a precision of 96%. They found that the AI essays had no grammatical errors compared to the human essays.

Current research on detecting AI-generated texts highlights several challenges. Many tools struggle with accuracy as AI models for text generation are becoming more advanced, making it harder to distinguish between human and AIGT. Most of the existing works have focused on using words, n-gram frequencies, and part-of-speech tags to build their detector. However, there is a lack of studies which uses readability scores and sentiment polarity within their set of features. Many of the existing studies also rely on a black-box approach to make their classification. Moreover, the datasets used are often very small. This study seeks to improve detection accuracy by developing a more transparent machine learning model using linguistic and stylistic features and sentiment polarity by using a much larger dataset of human-written and AI-generated essays.

III. Methodology

This study employed a quantitative research design to investigate the effectiveness of using linguistic and stylistic features for detecting AI-generated texts. The research process was divided into three main phases: dataset preparation, feature engineering, and development of a web application.

The choice of dataset plays an important role in ensuring the accuracy and reliability of our model. The dataset used ns this study was downloaded from Kaggle [13]. The dataset consists of 487,235 essays, which comprise 305,797 human-written and 181,438 AI-generated essays. The dataset consists of texts from various topics and is in a comma-separated value (CSV) file and was built by gathering data from multiple sources, adding them together and removing duplicates. Feature engineering was then carried out on the dataset. Text length, punctuation count, vocabulary richness, readability scores (Gunning Fog Index and Flesch Reading Ease) and sentiment polarity were calculated and added to the dataset to provide the model with more features for training.

After feature engineering, the records which contained missing values, were erroneous or were flagged as outliers (records with unusual Gunning Fog Index or Flesch Reading Ease) were removed from the dataset. A total of 3,875 records were deleted which resulted in a dataset with 483,360 valid records, out of which 180,311 were AI-generated and 303,049 were human-written. The whole dataset consists of 190,383,692 words. Table I provides a statistical analysis of the dataset. Table II and Table III provide the statistical analysis of the dataset for the AI-generated and human-written essays respectively. Table IV describes the features that have been used to differentiate between human-written and AI-generated texts.

TABLE I Statistical Analysis of the Dataset

Features	Mean Value	Minimum Value	Maximum Value
Text Length	393	75	1668
Punctuation Count	48	1	388
Gunning Fog Index	10.73	5	35
Flesch Reading Ease	63.7	0.16	99.97
Vocabulary Richness	0.43	0.05	0.86
Sentiment Polarity	0.16	-0.625	0.82

TABLE II Statistical Analysis of AI-Generated Texts

Features	Mean Value	Minimum Value	Maximum Value
Text Length	345	75	1238
Punctuation Count	46	4	258
Gunning Fog Index	11.54	5	28.75
Flesch Reading Ease	53.6	0.35	99.97
Vocabulary Richness	0.45	0.11	0.86
Sentiment Polarity	0.17	-0.376	0.70

TABLE III Statistical Analysis of Human-Written Texts

Features	Mean Value	Minimum Value	Maximum Value
Text Length	422	75	1668
Punctuation Count	49	1	388
Gunning Fog Index	10.24	5	35
Flesch Reading Ease	69.69	0.16	99.87
Vocabulary Richness	0.43	0.05	0.74
Sentiment Polarity	0.15	-0.625	0.817

TABLE IV LIST OF FEATURES

Features	Description
Sentiment Polarity	Sentiment analysis can be used to tell if a text is human written or AI generated. It involves categorising text as positive, negative or neutral. A negative score signifies negative sentiment, while a positive score represents positive sentiment. Gillham (2024) conducted an analysis against 100 articles generated by three LLMs for their sentiment and concluded that texts generated by LLMs are closer to the neutral part of the sentimental scale [14]. This difference in sentiment analysis between humans and LLMs can be a useful way to classify the texts as AI-generated or human-written.
Gunning Fog Index	Gunning Fog Index is a readability metric that estimates the number of years of education needed to understand a piece of text [15]. It is calculated based on the average sentence length and the percentage of complex words (defined as words with three or more syllables).
Flesch Reading Ease	The Flesch Reading Ease is a readability metric for a piece of text [16]. Kincaid et al. (1975) states that the Flesch Reading Ease formula is the most widely recognised and validated score among all readability metrics. This metric analyses average sentence length (ASL) and the average syllables per word (ASW) to assess the readability of a piece of text [16].
Vocabulary Richness	Vocabulary richness refers to the diversity of words used in a text and can be used to flag texts as AI generated or human authored since AI have the tendency to have a more diverse vocabulary set than humans [17].
Word Count and Punctuation Count	Calculates the number of words and punctuation present in the text. The characters classified as punctuations include these 32 characters: !, “, #, $, %, &, ‘, (,), *, +, ,, -, ., /, :, ;, <, =, >, ?, @, [, \,], ^, _, `, {, |, } and ~. Humans often use punctuation to convey emotion, emphasise points, or structure their writing while AI models use punctuation by following a pattern. This difference can help us in classifying the texts as AI-generated or human-written.

Fig. 1 presents the website architecture design and demonstrates how data is being passed from the client’s web browser to the Flask application and then predicted using a machine learning (ML) model. The text entered by the user is sent to the server, which is then sent to the Flask application. At the Flask application, the text entered by the user is validated and if everything is fine, the data undergoes feature engineering and text preprocessing. The last step involves the ML model to predict whether the preprocessed data is either human-written or AI-generated. The result goes from the ML model to the Flask application, to the server and is then rendered on the user’s browser.

Two activity diagrams are provided, one for the client side (Fig. 2) and one for the server side (Fig. 3) to clearly distinguish between the interactions that occur on each side of the application. The flowchart in Fig. 2 demonstrates the operation that happens at the client side of the application. The user enters the website and the interface of the application is displayed in his/her browser. The user will be presented with a textarea and a file upload button which he/she can use to enter text in the application. The user will then either paste texts manually in the textarea or upload a document for processing using the upload button. Once the user presses on the “detect button”, the text he/she has entered will be sent to the server-side for prediction. Lastly, the server side returns the result, which is displayed on the client’s browser. The flowchart in Fig. 3 demonstrates the operations that happen on the server side.

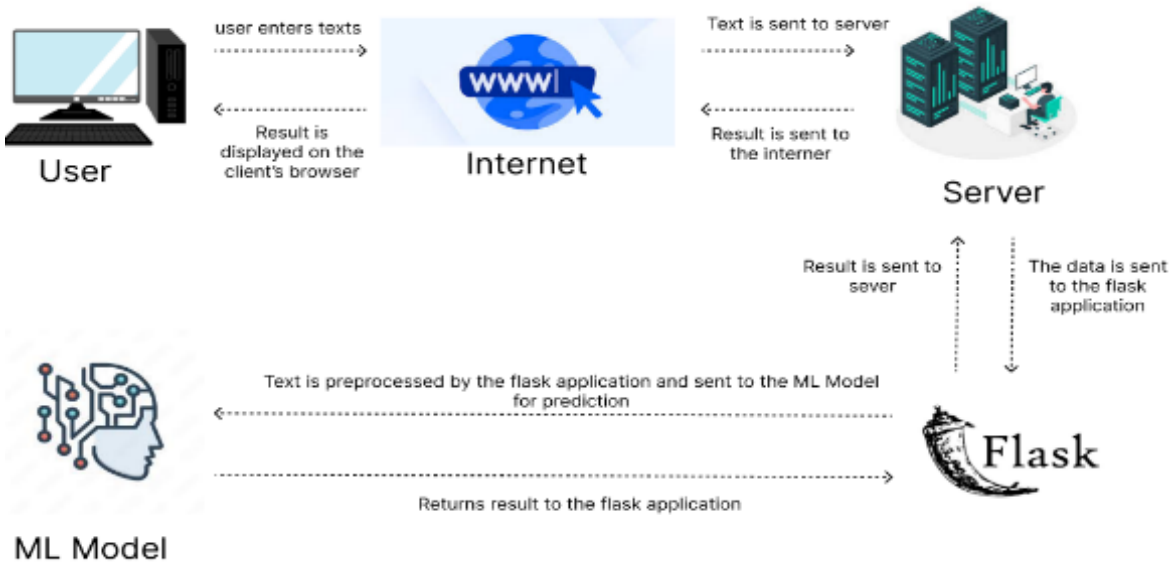

Fig. 1. Website architectural design.

Upon submitting the form by clicking the "detect text" button, the client application sends the data to the server. The server first checks for empty submissions. If the form is empty, it returns an error message to the client. If valid data is provided, the application retrieves the text from the textarea or uploaded document. After preprocessing and feature engineering, the model and vectorizer are loaded to prepare the data for prediction. Finally, the AI model determines if the input text is AI-generated or not and sends the result back to the client for display. However, the AI-model is applied on each batch of 200 words. The results can be different for each text segment. This strategy allows for a more nuanced evaluation of each section independently, rather than providing a single verdict for an entire document. A snapshot of the application is shown in Fig. 4.

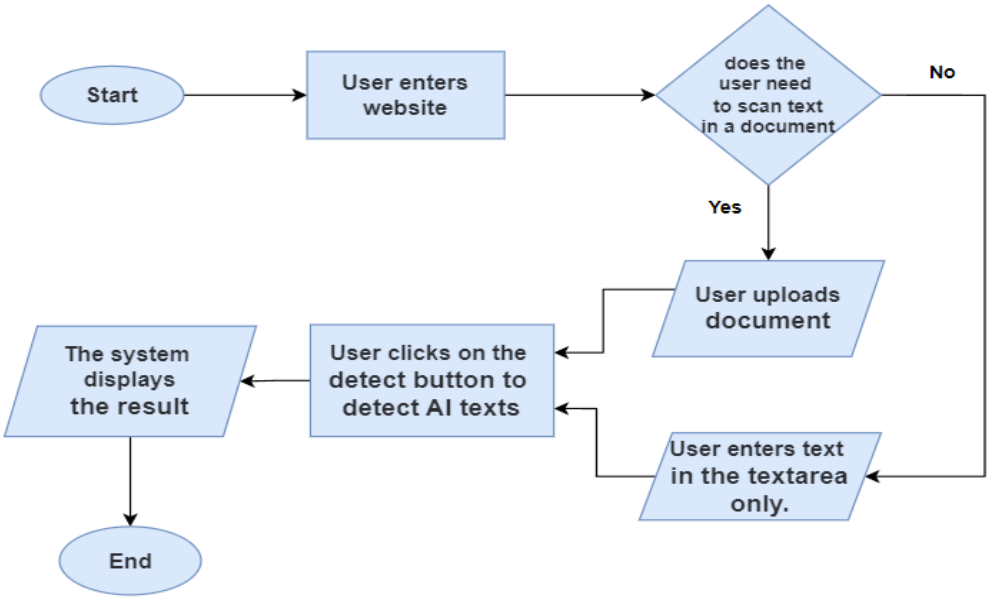

Fig. 2. Operations at the client side.

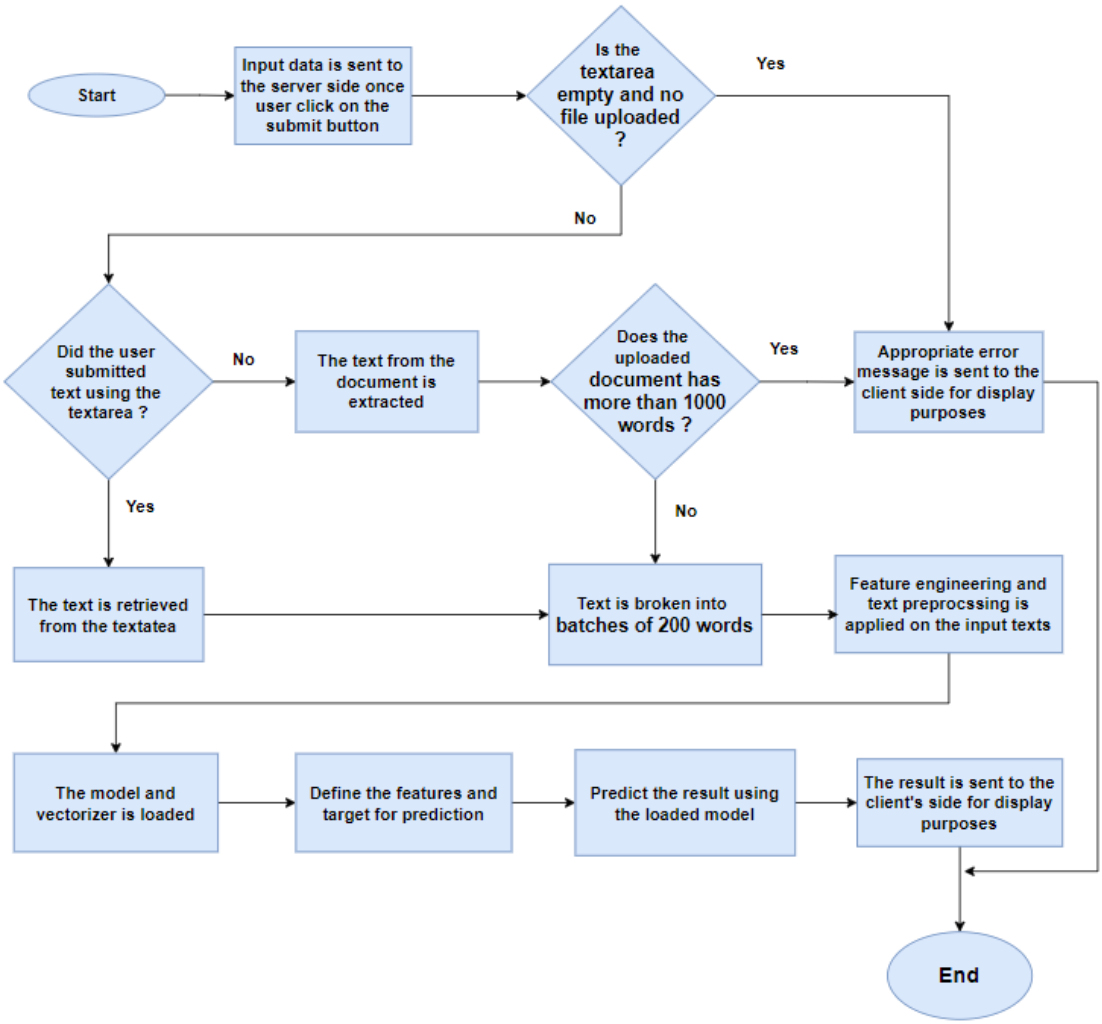

Fig. 3. Operations on the server side.

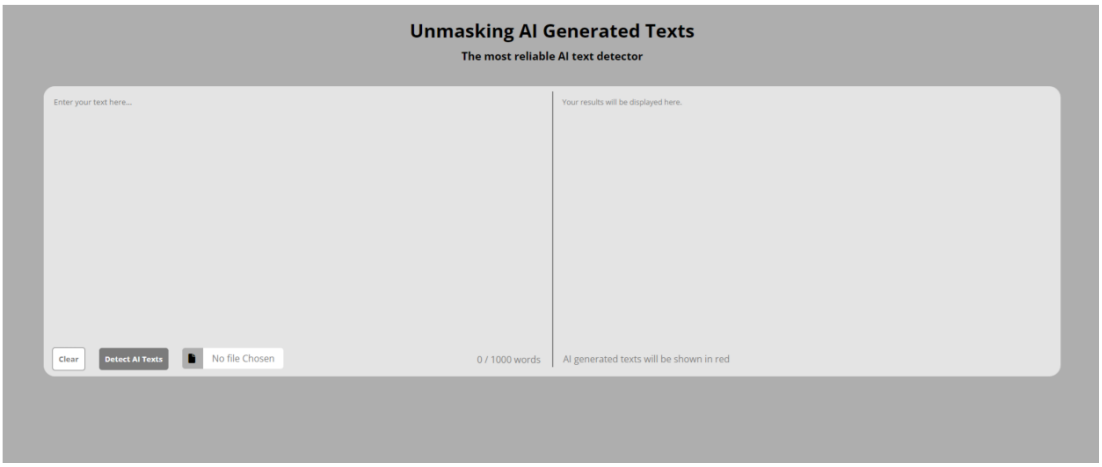

Fig. 4. GUI of the application.

IV. Experiments and Results

This section details the process of building the application, including the design and integration of key components. By outlining the implementation details, this section aims to demonstrate how the theoretical concepts are translated into a functional application, highlighting the practical aspects of the application. In evaluating the performance of our machine learning models, we utilise several key metrics: accuracy, recall, precision and the F1 Score. Table V shows the result of all the trained machine learning (ML) models.

TABLE V Result for All the Trained Models

Machine Learning Model	Metrics			
	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.94	0.93	0.94	0.94
SVM	0.8	0.73	0.88	0.75
Random Forest	**1.00**	**1.00**	**1.00**	**1.00**
Decision Tree	0.99	0.99	0.99	0.99
Gradient Boosting	0.96	0.95	0.97	0.96
XGBoost	**1.00**	**1.00**	**1.00**	**1.00**

Both Random Forest and XGBoost achieved an accuracy of 100% during training. An accuracy of 99% was achieved with Decision Tree. Gradient Boosting and Logic Regression scored above 90%. Only the scores for SVM were low.

The evaluation process is driven by a custom dataset that has been created to match the need of this study. The dataset comprises a balanced collection of human-generated and AI-generated texts. The dataset consists of 20 records, 10 AIGT and 10 human written texts. The AIGT records were obtained from ChatGPT and the human written records were obtained from Wikipedia articles [18-21], The Guardian [22-23] and from existing research papers [24-27]. These papers were selected because they were written well before AI-generated texts became available. The texts cover various topics such as sports, health, arts and science. The average text length of the articles is 368 words. The dataset consists of only two columns: text and label. The text represents the actual text that needs to be predicted as human or AI and the label column can have two values: human or AI to indicate who wrote the text. Table VI shows the result of the model evaluation process.

TABLE VI Evaluation Result of the Trained ML Models

Classifier	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.609	0.55	0.8	0.46
SVM	0.609	0.55	0.8	0.46
Random Forest	0.826	0.8	0.88	0.81
Decision Tree	0.565	0.5	0.28	0.36
Gradient Boosting	0.739	0.7	0.84	0.69
XGBoost	0.652	0.81	0.6	0.55

The Random Forest model scored an accuracy of 82.6% in predicting the nature of the texts found in the custom dataset. Table VII illustrates Random Forest's performance when applied to the custom dataset.

TABLE VII Prediction of Model using Custom Dataset

Record Number	Text Nature	Source	Prediction
1	AI	ChatGPT	AI
2	AI	ChatGPT	AI
3	AI	ChatGPT	AI
4	AI	ChatGPT	Human
5	AI	ChatGPT	AI
6	AI	ChatGPT	AI
7	AI	ChatGPT	Human
8	AI	ChatGPT	AI
9	AI	ChatGPT	AI
10	AI	ChatGPT	AI
11	Human	[26]	Human
12	Human	[27]	Human
13	Human	[24]	Human
14	Human	[18]	Human
15	Human	[19]	AI
16	Human	[20]	Human
17	Human	[22]	Human
18	Human	[25]	Human
19	Human	[23]	AI
20	Human	[21]	Human

During the evaluation, the model misclassified two AI-generated texts as human (record number 4 and 7) and two human-written texts as AI (record number 15 and 19). This may be due to the overlapping linguistic and stylistic features between the two categories. Additionally, biases present during the model's training process could have influenced the results. Thus, the accuracy of the model in this scenario is 80%. In conclusion, the evaluation of the proposed model demonstrates its effectiveness in distinguishing AI-generated texts from human-written ones.

V. Conclusion

This research explored the detection of AI-generated texts through linguistic features, stylistic features and sentiment polarity. Six different machine learning models were trained, with the Random Forest model emerging as the most accurate, achieving an accuracy of 82.6%. As a result, the Random Forest model was selected for its performance in identifying AI-generated content. However, this study also has its limitations. The model struggles with shorter texts, as these often lack the necessary linguistic and stylistic features that longer texts provide, making accurate detection more challenging. Additionally, the rapid advancement of AI text generation technologies makes it difficult to continuously adapt detection methods. Lastly, the Random Forest model was trained and tested on AIGT from only one LLM, specifically ChatGPT 3.5, meaning its ability to detect content produced by other large language models (LLMs) remains unverified. As future works, we intend to extend the systems so that it can also detect AIGT in languages other than English. Moreover, it would also be interesting to investigate whether there is an impact on the detection accuracy for more advanced GPT models such as GPT-4 and GPT 4.5.

References

[1] A. Shah, P. Ranka, U. Dedhia, S. Prasad, S. Munni, and K. Bhowmick, "Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence Using Stylistic Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 10, pp. 1043–1053, 2023. Available: https://dx.doi.org/10.14569/IJACSA.2023.01410110.

[2] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *Int. J. Educ. Integr.*, vol. 19, Art. 17, 2023. Available: https://doi.org/10.1007/s40979-023-00140-5.

[3] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, "AI vs. Human - Differentiation Analysis of Scientific Content Generation," *arXiv*, vol. 2301.10416v2, 2023. Available: https://doi.org/10.48550/arXiv.2301.10416.

[4] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023. Available: https://doi.org/10.1109/ACCESS.2023.3294090.

[5] P. Wang, L. Li, K. Ren, B. Jiang, D. Zhang, and X. Qiu, "SeqXGPT: Sentence-Level AI-Generated Text Detection," *arXiv*, vol. 2310.08903v2, 2023. Available: https://doi.org/10.48550/arXiv.2310.08903.

[6] E. Tulchinskii, K. Kuznetsov, L. Kushnareva, D. Cherniavskii, S. Barannikov, I. Piontkovskaya, S. Nikolenko, and E. Burnaev, "Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts," *arXiv*, vol. 2306.04723v2, 2023. Available: https://doi.org/10.48550/arXiv.2306.04723.

[7] L. Mindner, T. Schlippe, and K. Schaaf, "Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT," *arXiv*, vol. 2308.05341v1, 2023. Available: https://doi.org/10.1007/978-981-99-7947-9_12.

[8] K. Kumari, A. Pegoraro, H. Fereidooni, and A. Sadeghi, "DEMASQ: Unmasking the ChatGPT Wordsmith," *arXiv*, vol. 2311.05019v1, 2023. Available: https://dx.doi.org/10.14722/ndss.2024.231190.

[9] Z. Liu, Z. Yao, F. Li, and B. Luo. Check me if you can: Detecting chatgpt-generated academic writing using checkgpt. arXiv:2306.05524, 2023.

[10] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, "Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature," *arXiv*, vol. 2310.05130, 2024. Available: https://doi.org/10.48550/arXiv.2310.05130.

[11] S. Mitrovic, D. Andreoletti, and O. Ayoub, "ChatGPT or human? Detect and explain," *arXiv*, vol. 2301.13852v1, 2023. Available: https://doi.org/10.48550/arXiv.2301.13852.

[12] D. Yan, M. Fauss, J. Hao, and W. Cui, "Detection of AI-generated Essays in Writing Assessments," *Psychological Test and Assessment Modeling*, vol. 65, no. 1, pp. 125–144, 2023.

[13] S. Gerami, "AI vs Human Text," *Kaggle*, 2023. Available: https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text/data.

[14] J. Gillham, "Study finds popular LLMs make content more neutral in sentiment," *Originality.ai*, August 8, 2024. Available: https://originality.ai/blog/study-popular-llms-make-content-neutral-sentiment?utm_source=chatgpt.com.

[15] S. Zhou, H. Jeong, and P. Green, "How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards," *IEEE Transactions on Professional Communication*, 60(1), 97–111, 2017. https://doi.org/10.1109/tpc.2016.2635720.

[16] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Research Branch Report 8-75*, Institute for Simulation and Training, 1975. Available: https://stars.library.ucf.edu/istlibrary/56.

[17] K. Kettunen, "Can type-token ratio be used to show morphological complexity of languages?" *J. Quant. Linguist.*, vol. 21, no. 3, pp. 223–245, 2014. Available: https://doi.org/10.1080/09296174.2014.911506.

[18] "Natural environment," *Wikipedia, The Free Encyclopedia*, September 24, 2024. Available: https://en.wikipedia.org/w/index.php?title=Natural_environment&oldid=1247530947.

[19] "The Arts," *Wikipedia, The Free Encyclopedia*, September 25, 2024. Available: https://en.wikipedia.org/w/index.php?title=The_arts&oldid=1247701579.

[20] "Governance," *Wikipedia, The Free Encyclopedia*, September 21, 2024. Available: https://en.wikipedia.org/w/index.php?title=Governance&oldid=1246804909.

[21] "Transport," *Wikipedia, The Free Encyclopedia*, September 14, 2024. Available: https://en.wikipedia.org/w/index.php?title=Transport&oldid=1245664293.

[22] K. Riddle, "Is it right to force someone into rehab? The man whose life inspired a landmark law," *The Guardian*, May 13, 2024. Available: https://www.theguardian.com/society/article/2024/may/13/rehab-forced-addiction-treatment.

[23] J. Hinchliffe, "Australia's first genetically modified fruit is ripe for a taste test. Could it avert a global banana apocalypse?" *The Guardian*, September 6, 2024. Available: https://www.theguardian.com/australia-news/article/2024/sep/07/cavendish-banana-genetically-modified-qcav-4.

[24] C. Mayer, "Financial Systems, Corporate Finance, and Economic Development," *Asymmetric Information, Corporate Finance, and Investment*, pp. 307–332, 1990. Available: https://www.nber.org/system/files/chapters/c11477/c11477.pdf.

[25] H. Liu, “In-flight Entertainment System: State of the Art and Research Directions,” *Second Int. Workshop Semantic Media Adapt. Pers.*, 2007. Available: https://doi.org/10.1109/SMAP.2007.37.

[26] M. . Prince, V. Patel, S. Saxena, M. Maf, J. Maselko, M. R. Philips, and A. Rahman, “No health without mental health,” *The Lancet*, vol. 370, no. 9590, pp. 859–877, 2007. Available: https://doi.org/10.1016/S0140-6736(07)61238-0.

[27] S. Hooper and L. P. Rieber, “Teaching with technology,” *Teaching: Theory into practice*, pp. 154–170, 1995.