Bridging the Gap: Ensemble Learning-Based NLP Framework for AI-Generated Text Identification in Academia

Layth Hazim
Department of Electrical and Computer Engineering
Altinbas University, Turkey
layth.r.hazim@tu.edu.iq

Oguz Ata
Department of Computer Engineering
Istanbul Atlas University, Turkey
oguz.ata@atlas.edu.tr

Abstract: *Background: the advent of Large Language Models (LLMs), including Chat Generative Pre-trained Transformer (ChatGPT) and Bard, has revolutionised text generation while raising ethical concerns regarding academic integrity. Differentiating Artificial Intelligence-Generated Texts (AIGT) from human-written content is crucial to maintaining transparency and trust in scholarly communication. Objective: this study aims to address the limitations in existing detection methods by introducing a Machine Learning (ML)-based Natural Language Processing (NLP) framework that effectively distinguishes between AI-generated and Human-Written academic texts (HWAI). Methodology: the proposed framework integrates comprehensive preprocessing, Exploratory Data Analysis (EDA), linguistic analysis, and ensemble learning techniques. Text representation was achieved using Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. We employed two diverse datasets, Artificial Intelligence-Generated Academic (AI-GA) and HWAI, to validate the framework's efficacy, ensuring robust classification performance. Results: the ensemble model did better than individual classifiers. On the AI-GA dataset, it achieved state-of-the-art accuracy (98.67%) and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) (99.88%). The HWAI dataset achieved 96.52% accuracy and 99.37% ROC-AUC. These results highlight the framework's capability to identify unique linguistic patterns in AI-generated content. Conclusion: the framework addresses key linguistic and computational challenges and provides a scalable and reliable solution for detecting AI-generated content in academic domains. Future work will explore hybrid human-AI authorship detection and real-time deployment to enhance its practical utility across disciplines.*

Keywords: *Large language models, AI-generated text, human-written text, exploratory data analysis, NLP, ensemble learning.*

Received April 6, 2025; accepted July 10, 2025
https://doi.org/10.34028/iajit/22/6/2

1. Introduction

Large Language Models (LLMs) have transformed Natural Language Processing (NLP) by enabling the generation of high-quality, human-like text. These models, such as OpenAI's GPT-4 and Google's bard, have shown exceptional performance in various domains, including content generation, automated translation, and academic writing. The rapid evolution of these models has brought new opportunities but also raises concerns regarding their impact on authorship and academic integrity [32]. LLMs have improved performance on few-shot and zero-shot challenges lately. The massive language model Chat Generative Pre-trained Transformer (ChatGPT), which OpenAI published on November 30, 2022, [30], has demonstrated previously unheard-of performance in comprehending user inquiries and producing writing that seems human. On March 21, 2023, Google started granting access to Bard; Google developed Bard, a dialogue-based generative Artificial Intelligence (AI) chatbot. Initially constructed on the Language Model for Dialogue Applications (LaMDA) family of LLMs, it was subsequently developed on the Pathways Language Model (PaLM) LLM. In just a few months, Bard garnered much attention and was extensively discussed in the NLP community and other fields [4].

Modern NLP techniques, coupled with Machine Learning (ML) and Deep Learning (DL), have led to significant improvements in Natural Language Generation (NLG), allowing AI to produce coherent and contextually relevant text [26, 41]. AI-powered writing assistants have been widely adopted in academia, journalism, and business communication. However, these advancements come with challenges, including the risk of misinformation, authorship fraud, and the erosion of academic integrity [47]. The ability of AI to generate scholarly articles that closely resemble human-written content has created an urgent need for reliable detection methods [3].

Several studies have attempted to address the issue of AI-generated content detection [37]. However, existing approaches predominantly focus on general text classification rather than distinguishing Artificial Intelligence-Generated Academic (AI-GA) content. This gap in research highlights the pressing need for

robust techniques to differentiate Artificial Intelligence -Generated Text (AIGT) from Human-Written Text (HWT) in scholarly work, ensuring the integrity of academic publications. A number of AI conferences, such as the 61st annual conference of the Association for Computational Linguistics (ACL, 2023) [1] and the 14th International Conference on Machine Learning (ICML, 2023) [18], have revised their authorship criteria in response to this trend.

This work presents a holistic approach fusing Exploratory Data Analysis (EDA), linguistic analysis, and ensemble learning methods to identify AI-GA work. In contrast to previous methods, our proposed method unifies state-of-the-art text representation techniques, Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams, with ensemble classifiers to ensure better accuracy and robustness. With the use of publicly available datasets, this paper guarantees reproducibility and unbiased benchmarking, responding to the increasing demand for transparency in research and academic publications. The theoretical framework of the AIGT system proposed here employed in distinguishing between human-written and AI-written scholarly papers is depicted in Figure 1.

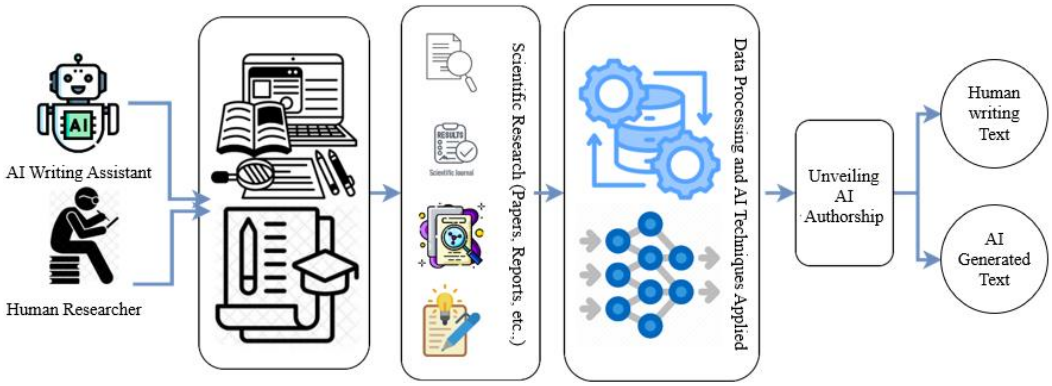


Figure 1. The work overview of AIGT.

2. Literature Review

The literature review defines the boundaries of this study and provides a comprehensive understanding of the existing knowledge in this field. Despite being a relatively new idea, AI text recognition using LLMs has already been the subject of relevant studies. This section reviews the literature on AI-generated material and introduces relevant research on academic AI-generated content identification.

2.1. AI-Generated Content

When technology emerged in the 1950s, computer-generated content primarily focused on music and visual art [6]. The audience was able to easily distinguish early computer-generated material from human-generated content [31]. As AI technology has advanced, visual material produced using methods like Generative Adversarial Networks (GAN) [16] and diffusion models [13] has grown increasingly realistic.

LLMs support the growth of downstream NLP activities while overcoming social and technological obstacles. According to the study [33], Large pre-trained language models may retain and logically deduce the whole of human knowledge obtained from extensive training datasets, alongside acquiring linguistic expertise. The Meta AI team introduced Galactica, an advanced language model capable of storing, integrating, and reasoning with scientific knowledge. as a means of organising scientific knowledge [43]. Galactica beats current models in several scientific NLP tasks, but it has been criticized for perpetuating biases and generating misleading information. For example, studies have shown that Galactica sometimes fabricates scientific references or reinforces existing biases in data-driven content. OpenAI proposes ChatGPT; Bing Chat is powered by the same model as ChatGPT from OpenAI, and Google Bard is an AI language model created by Google. All these tools can generate highly fluent text. Users like ChatGPT over other LLMs because of its accessibility and capacity to provide responses that are both grammatically accurate and understandable for users across a range of areas.

2.2. Potential Risks of AI-Generated Content

AI-generated content has demonstrated an understanding of complex domains such as medicine, necessitating extensive vetting for accuracy and reliability. Additionally, AI-GA articles, such as those produced by SCIgen since 2005, have occasionally passed peer review despite being nonsensical [38]. Even though the context-free grammar-generating approach is relatively simple, such papers have continued to appear in reputable publications over the years [7]. These cases highlight the need for robust detection mechanisms to maintain academic integrity in scholarly publishing.

2.3. AI-Generated Content Detection

The latest developments in models like ChatGPT have

also created questions about the authenticity of AI-generated content, especially in educational settings. While past research has taken various detection methods into account, distinguishing human-written and AI-generated educational content remains a main challenge [10]. The present work aims at two key aspects:

1. How humans perceive the AI-generated content.
2. The effectiveness of the detection models.

2.3.1. Human Behaviour for Content Recognition

As language models like GPT-3 and GPT-4 improve, distinguishing their output from human writing becomes increasingly difficult [8]. For example, research shows that users perceived misinformation generated by Grover as more credible than human-written disinformation [47]. Other studies found that participants failed to reliably distinguish GPT-2-generated poetry from human poetry, and even in contexts like Airbnb profiles or job applications, users achieved poor detection accuracy [22]. Although certain heuristics can help, overall, the evidence suggests humans are becoming less able to detect AI-generated content due to the rising quality of language models [20].

2.3.2. Detection Models for AI-Generated Content

Detecting AIGT is often approached as a binary classification problem [28]. Various models have been trained to distinguish between human-written and Machine-Generated Content (MGC) [10], and some studies also explore attribution to specific generation models [24], as outlined in Table 1.

Recent detectors such as GPTZero-XL (2025) and DetectGPT-v2 (2025) demonstrate improved robustness and generalizability [46]. These systems utilize techniques like perplexity scoring, log-probability curvature, and embedding-based classification to detect subtle inconsistencies in AIGT [42].

Table 1 summarizes major studies across diverse domains, showcasing approaches ranging from Logistic Regression (LR) and LSTM to ensemble methods like eXtreme Gradient Boosting (XGBoost) and robustly optimized BERT approach-Feedforward Neural Network (RoBERTa-FNN) with many achieving accuracy levels above 90%.

Table 1. Overview of key studies on AIGT detection across research domains.

Research Field	Approach used	Dataset	Outcomes
Detecting social media disinformation [40].	Pretrained detectors like RoBERTa-based models.	News and social media texts.	Raised concerns about generalizability of current detection methods.
Investigated scientific content [27].	LR on syntax/semantic features.	Human and AI-generated abstracts.	High F1 score based on structured linguistic features.
Detection scientific paper AI-generated [44].	Word2Vec+LSTM.	COVID-19 Open Research Dataset (CORD-19).	Achieved 98.7% accuracy.
Distinguishing academic Science writing in ChatGPT [12].	Traditional ML on linguistic patterns.	Perspectives articles vs. ChatGPT.	Achieved 99% accuracy.
Distinguish AIGT in the academic field [3].	ML models (LR, RF, SVM, etc.).	500 Q and A responses.	RF performed best across tasks (92.5-93.5% accuracy).
Differentiate the text generated by ChatGPT in two domains (news and social media) [19].	ML classifier comparison.	10,000-word samples.	77% overall accuracy.
Efficiency and authenticity in education and research [11].	Manual review+ChatGPT-assisted generation.	Research paragraph dataset.	Literature more prone to MGC than abstracts.
Created an automated text detection model for humans and ChatGPT [21].	TSA-LSTM-RNN+optimization.	Two benchmark datasets.	Accuracy: 93.17% and 93.83%.
Discriminating the articles from Wikipedia or ChatGPT [39].	DL (CNN), ML (SVM, LR, etc.).	44,162 samples.	LR achieved 97% accuracy.
Detect AIGT writing style in documents [5].	Stylometric+XGBoost/Stacking.	Human vs. ChatGPT documents.	Up to 98% accuracy.
Text authenticity and AI-generated content detection [17].	SBERT, RoBERTa+LR and FNN.	1M balanced samples.	RoBERTa-FNN achieved 99.95% accuracy.
AI-generated abstracts detection [14].	Perplexity score+AI tools.	50 PubMed abstracts.	AUC = 0.7794, up to 95% accuracy.

Despite the tremendous progress in AI-generated content detection, notable gaps remain. Many studies overlook preprocessing steps like EDA and linguistic analysis that could improve model performance. Ensemble learning remains underutilized, particularly in academic text detection. Moreover, datasets often lack diversity in domain representation and class balance.

To address these issues, our study introduces a robust pipeline incorporating text cleaning, EDA, and linguistic features. We apply ensemble learning to two balanced academic datasets AI-generated and Human-Written academic text (HWAI) demonstrating generalizability across contexts. This framework contributes significantly to preserving academic integrity in the face of evolving AI text generation capabilities.

3. System Design and Methodology

This study proposes an effective framework to classify AI-generated versus Human-Written academic text (HWAI). The method integrates advanced preprocessing, EDA, linguistic features, and ML algorithms, along with ensemble learning techniques. The framework achieves research objectives by building an organized workflow for data preprocessing, model training, and evaluation. The proposed methodology integrates comprehensive preprocessing, language analysis, and ensemble learning for optimal

performance. The use of two diverse datasets enhances the generalisability of the findings. Figure 2 below provides an overview of the proposed methodology.

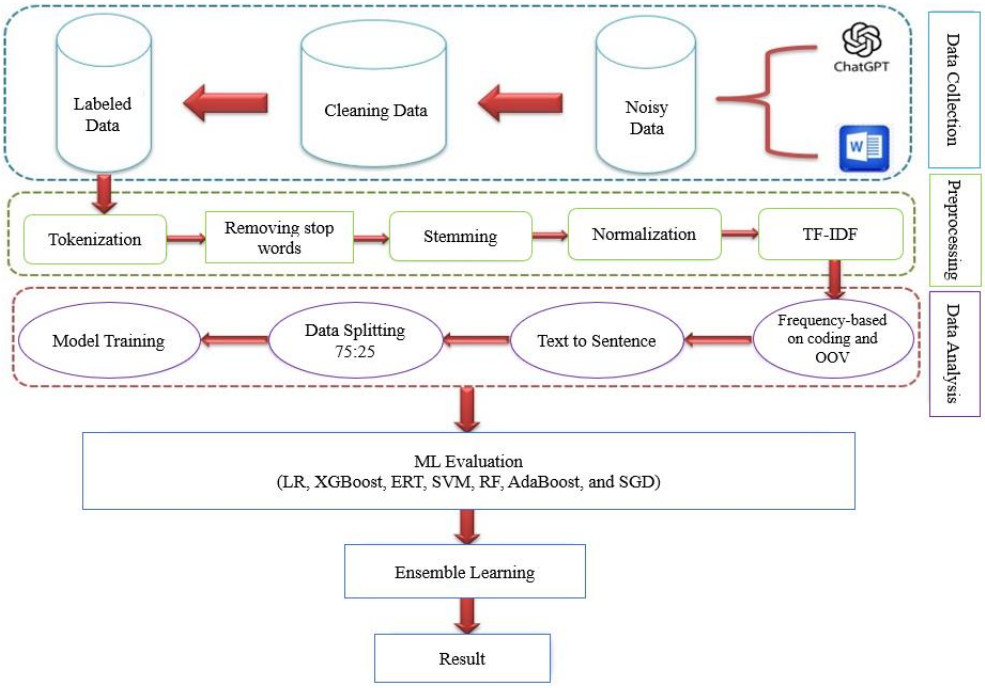


Figure 2. Proposed framework.

3.1. Data Collection

One of the most important phases of academic content processing is collecting data. Given that this is the primary factor influencing our prediction, we have proposed two publicly available datasets for this study.

3.1.1. AI-GA Dataset

- **Source**: the AI-GA AI-generated abstracts dataset consists of texts produced by huge language models such as GPT-3, GPT-4, and other similar AI systems. These texts were collected from open-access repositories, forums, and platforms where AI-generated content is shared [44].
- **Content**: the dataset consists of a list of titles and abstracts, half generated by AI and the other half created by humans. The content covers various topics, including technology, science, literature, and general knowledge.
- **Description**: AI-generates abstractions using modern language generation methods, particularly the GPT-3 model. The original abstracts for the research studies on the COVID-19 pandemic come from a corpus of published papers.
- **Size**: there are 28,662 samples in the AI-GA dataset, and each sample has a title, an abstract, and a label. The dataset is divided into "original abstracts" with 14,331 samples and "AI-generated abstracts" with 14,331 samples, with a total word count of approximately 19,813. The label designates whether the sample is an AI-generated abstract (labelled as 1) or an original abstract (labelled as 0), as seen in Figure 3.

	title	abstract	label
0	Clinical features of culture-proven Mycoplasma...	OBJECTIVE: This retrospective chart review des...	0
1	Nitric oxide: a pro-inflammatory mediator in l...	Inflammatory diseases of the respiratory tract...	0
2	Surfactant protein-D and pulmonary host defense	Surfactant protein-D (SP-D) participates in th...	0
3	Role of endothelin-1 in lung disease	Endothelin-1 (ET-1) is a 21 amino acid peptide...	0
4	Gene expression in epithelial cells in respons...	Respiratory syncytial virus (RSV) and pneumoni...	0
...	...	...	...
28657	Characteristics of the Dynamic of Mobile Networks	\n\nThis empirical study examines the characte...	1
28658	Prediction and Prevention: Interventions to En...	\n\nThis journal articulates the impact of pro...	1
28659	Asthma and Microbes: A New Paradigm	\n\nAsthma is a chronic inflammatory disease o...	1
28660	Diagnostic Knowledge in the Genetic Economy an...	\n\n\nThis paper proposes to investigate the ...	1
28661	High Capacity Vectors	\n\nThis study entitled 'High Capacity Vector...	1

28662 rows × 3 columns

Figure 3. Longitudinal section of the AI-GA dataset.

3.1.2. HWAI Dataset

- **Source**: the HWAI dataset consists of HWT collected from academic papers, online articles, and essays authored by humans from "Wikipedia." The most current iteration of the ChatGPT model produces articles roughly the same size. The sources were chosen to represent various writing styles and domains, modified in March 2023 [39].
- **Content**: this dataset includes academic research papers, opinion articles, and other forms of written content produced by humans. The topics covered are diverse, including but not limited to science, technology, humanities, and social sciences. Wikipedia is a fantastic source for high-quality articles of various lengths and subjects.
- **Size**: the collection of articles consists of four sets. We carefully select and annotate the following sets of text articles: Set 1 (20–100 words), Set 2 (100–200 words), Set 3 (200–300 words), and Set 4 (more than

300 words). The Corpus comprises 44,138 text articles from ChatGPT and Wikipedia, with a total word count of approximately 289,503. There are 22,069 articles in machine-generated and human-generated text amalgamated categories. The class indicates whether the sample text was generated by a computer (designated as 1) or by a person (designated as 0), as seen in Figure 4.

	Text	Class
0	Mississippi Highway 403 and MS 15 in Mathiston...	0
1	State Route 97 northwest of Congress northeast...	0
2	Edwin B. Erickson III was an American politici...	0
3	The Tiger Fire was a wildfire that burned 16,2...	0
4	Fajsz , was Grand Prince of the Hungarians fro...	0
...	...	...
9995	Women in early modern Scotland, between the Re...	0
9996	The Court of Common Pleas was one of the four ...	1
9997	Megara is a fictional character from the Disne...	1
9998	Luan Da is a Chinese literary figure who lived...	1
9999	The New York Herald Tribune was a newspaper pu...	0

44138 rows × 2 columns

Figure 4. Longitudinal section of the HWAI dataset.

3.2. Preprocessing Datasets

To train traditional ML models and build a framework for feature description to distinguish between AIGT and human-written material using language analysis, we will look at well-established NLP techniques for text representation and embedded representation techniques.

3.2.1. Cleaning Datasets

Initially, the dataset underwent preprocessing to display the text articles in an understandable Word format. To be more precise, preprocessing was done on the contents of these articles to minimise their dimensionality and make categorisation easier [45]; six phases make up the cleaning datasets stage: data framing, Tokenisation, normalisation, and feature selection of TF-IDF stop words [35]. The process of cleaning was done by applying Algorithm (1) below:

Algorithm 1: Proposed strategy of data preprocessing and cleaning.

Input: Raw Corpus of text documents.
Output: Processed feature set for text classification.
Begin
Step 1: Data Preprocessing
For each text in the Corpus, do
Remove special characters and HTML tags.
Eliminate line breaks and excessive whitespace.
Normalising text by converting all characters to lowercase for uniformity.
Remove stop words that do not contribute to meaning (e.g., "the", "and", "of").
Delete any numbers or non-alphabetic characters.
Eliminating the frequent terms in the created text, such as "paper," "study,".
End for
End

3.2.2. Exploratory Data Analysis (EDA)

In any ML workflow, EDA is a critical component, and NLP is no exception. The "word cloud," which is more frequently used as a visualisation tool for EDA in the context of text data, will be discussed and applied in this study [29].

The term "wordcloud" refers to a cloud of words that, by presenting the words in a corpus at varying sizes, indicates the frequency of occurrence of each word. A dataset offers a rapid and easy method to determine which terms are most frequently used and visually examine their distribution [9], as stated in Figures 5 and 6, including them in the two suggested datasets.

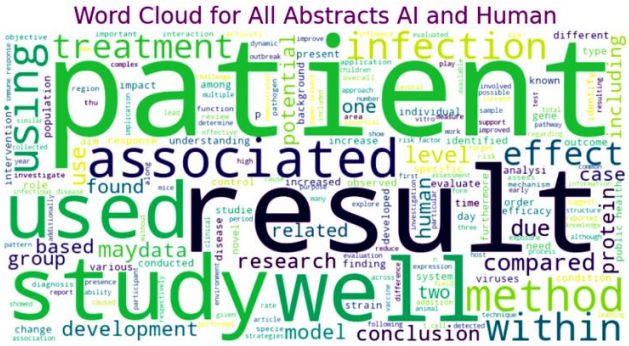


Figure 5. Word cloud of the AI-GA dataset.

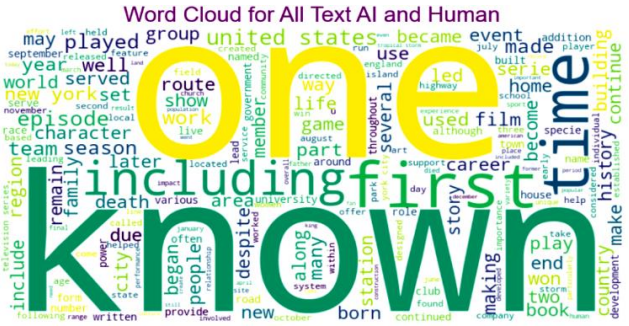


Figure 6. Word cloud of the HWAI dataset.

Word clouds are a straightforward yet efficient method for visualising textual data in a clear and comprehensible format. We inspected to enhance the user experience, facilitating a deeper comprehension of the academic material inside their text.

3.2.3. Linguistic Analysis

We must carry out a text-cleaning process to guarantee that the linguistic analysis of the academic content produced by ChatGPT and humans is founded on high-quality data [21].

In this subsection, we delve further into the datasets supplied by the organisers. Given that the quantity of English-language texts is about equal and the distribution of both labels is balanced (generated and human), We conducted an examination of the texts' lengths from two perspectives (number of words and frequent terms) using n-grams ('1-grams', '2-grams', '3-gram', '4-grams', '5-grams'), which are the outcome

of linguistic analysis, in order to determine which model is better appropriate for this processing [15].

N-grams of text are widely employed in NLP and text mining activities. In essence, they are a set of words that co-occur inside a specific frame, and to compute the n-grams, one typically progresses one word at a time (but in more complex situations, you may advance *X* words) [45]. The number of n-grams for a particular phrase *K* would be as in the Equation (1) if *X*=*Num* of words in sentence *K*:

$$Ngrams_K = X - (N - 1) \quad (1)$$

The purpose of using this method is to visualise what differences occur in the two datasets, such as the length of sentences and the distribution of vocabulary when you choose a combination of datasets, detailed n-gram frequency distributions for both datasets are provided in supplementary Figures 7, 8, 9, 10, 11, and 12.

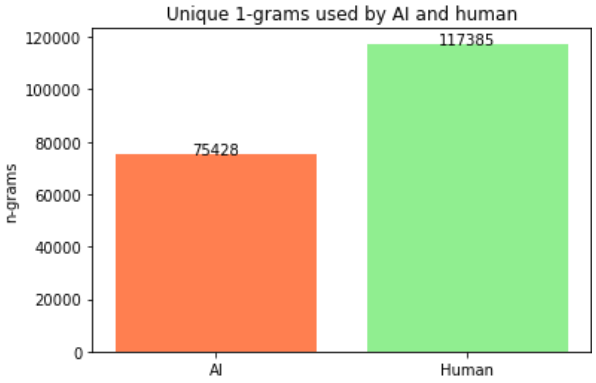


Figure 7. Use 1-grams for AI-GA dataset.

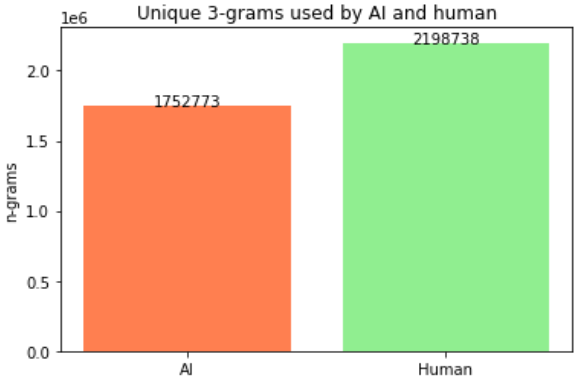


Figure 8. Use 3-grams for AI-GA dataset.

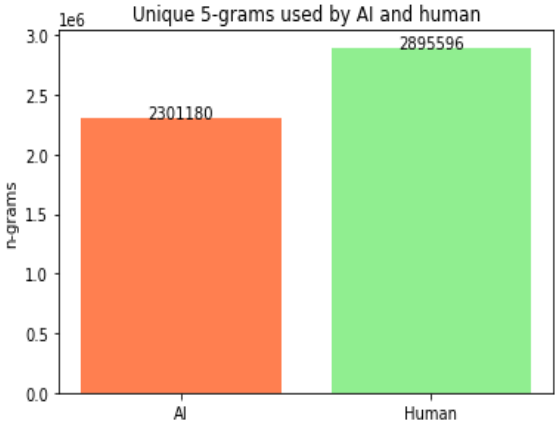


Figure 9. Use 5-grams for AI-GA dataset

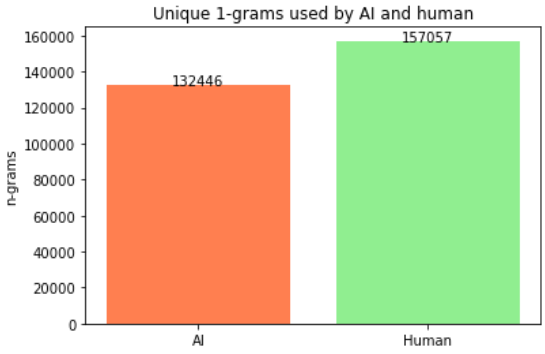


Figure 10. Use 1-grams for the HWAI dataset.

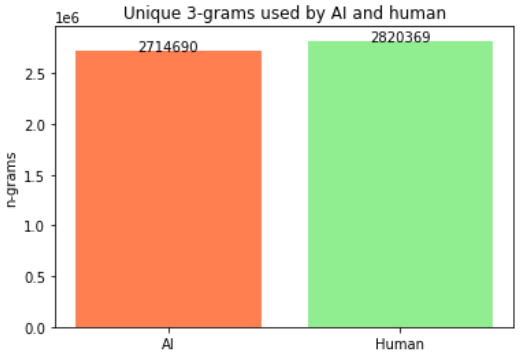


Figure 11. Use 3-grams for the HWAI dataset.

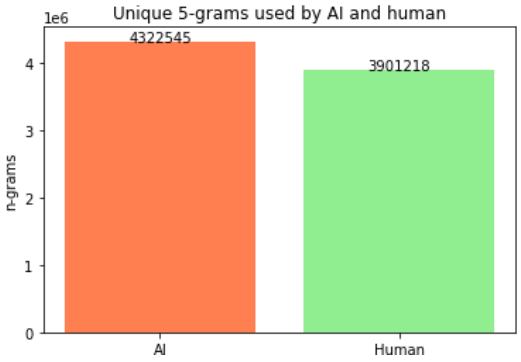


Figure 12. Use 5-grams for the HWAI dataset.

Algorithm (2), which is executed implementation for linguistic analysis as follows:

Algorithm 2: Proposed strategy of linguistic analysis (n-grams).

Input: Raw Corpus of text documents.
Output: Processed feature set for text classification.

Begin
Step 2: Linguistic Analysis
For each preprocessed text in the Corpus, do
Perform Tokenisation by splitting the text into individual words or tokens.
Analyse the text to extract language patterns using N-grams (where N=1 for unigrams, N=2 for bigrams, N=3 for trigrams, 4-gram, 5-gram) and their frequencies using Eq (1).
Identify the most common words and their counts, as well as typos and colloquialisms.
Identify words unique to AI-generated and human-written texts.
Analyse sentence structure, length, and complexity based on linguistic patterns.
Compute statistical linguistic features (part-of-speech distribution).
End for
End

We found in the figures above that the vocabulary of AI-generated sentences was significantly less than that of human-written sentences in the two types of datasets AI-GA and HWAI. This is likely due to the nature of generative AI, which generates word sequences with the highest probability of occurrence.

Additionally, we have included a list of frequently used words that are exclusively used by AIs and those that are only used by humans. Some frequent words used only by humans come from typos such as "because", "because" and "driveless." As other researchers have pointed out, focusing on typos may improve accuracy. Tables 2 and 3 illustrate this.

Table 2. Frequent words in AI and human text of AI-GA dataset.

Top 10 5-grams only used by AI	Freq	Top 10 5-grams only used by human	Freq
presents the findings of a	110	the online version of this	538
the title of this article	96	the online version of this article	538
this article presents a comprehensive	95	available supplementary material	530
presents the results of a	94	material which is available to	530
The title of this article is	94	contains supplementary material, which is	528
presents an analysis of the	91	which is available to authorised	525
presents a novel approach to	84	is available to authorised users	525
this article presents a novel	80	electronic supplementary material, the online	480
this paper presents an analysis	78	supplementary material, the online version	480
this paper presents the findings	77	material, the online version of	480

Table 3. Frequent words in AI and human text of HWAI dataset.

Top 10 5-grams only used by AI	Freq	Top 10 5-grams only used by human	Freq
as a reminder of the	571	it originally aired on the	183
serves as a reminder of	348	originally aired on the fox	180
reminder of the importance of	198	on the Fox Network on	156
is a testament to the	176	series The Simpsons originally	146
a reminder of the importance	165	on NBC in the united	144
served as a reminder of	163	NBC in the United States	143
a popular destination for tourists	148	centres on FBI special agents	143
it is a popular destination for	136	aired on NBC in the	137
that tells the story of	133	race between crews from the	136
home to a variety of	122	show centres on FBI special	127

Depending on the dataset employed, our examination of the text length, word count, and frequency of terms shows a distinct difference between machine-generated and human-generated texts. As we noticed in Table 2, where the AI-GA dataset used is academic, it is characterised by high consistency. Therefore, we observed a significantly lower frequency of frequent words in the AIGT. On the contrary, in the HWAI dataset, which was just scientific data taken from Wikipedia, we noticed in Table 3 that the frequent words in the text generated by the AI are much more than in the text generated by humans.

This conclusion suggests that linguistic analysis, such as word count or word frequency, and EDA, such as 'word cloud', could help distinguish between literature written by AI and human authors. Nevertheless, we conducted a thorough feature engineering process to identify additional text features that would benefit our methodology.

3.2.4. Text Representation

In NLP, text representation is a crucial stage. The primary objective is to convert the raw text input into a numerical format for various machine-learning models. Text representation functions as a mechanism for text algorithms to do various NLP tasks, such as machine translation, sentiment analysis, and text classification. This study will employ the TF-IDF approach for feature extraction. NLP and information retrieval use the TF-IDF method to assess a term's significance within a text or Corpus. Term Frequency (TF) and Inverse Document Frequency (IDF), two separate components, are combined to generate it [34]. Equation (2) delineates the TF component, quantifying the frequency of a term's occurrence in a text. At the same time, Equation (3) elucidates the IDF component, attributing less significance to often-used terms and enhanced significance to infrequently utilised phrases.

$$TF(t,d) = \frac{n_t, d}{\sum_k n_k, d} \tag{2}$$

$$IDF(t,D) = log\frac{N}{df(t)} \tag{3}$$

A word's term frequency increases with its frequency of occurrence in documents, yet its significance (IDF) increases when a word is searched in a particular document and occurs less frequently [2]. The result of multiplying the TF and IDF components is TF-IDF.

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \tag{4}$$

Finally, we multiply the TF and IDF from the previous phases to obtain the IF-IDF, as illustrated in Equation (4). Using the "TfidfVectorizer" on the two datasets, we generate a matrix of TF-IDF features that indicate the relevance and importance of each word inside the text samples. Where the final shape of the AI-GA dataset became (28662, 5000) and the HWAI dataset (44018, 5000). By leveraging the subtle linguistic patterns and contextual variations unique to each source, we can

utilise this numerical representation in combination with a range of ML models to distinguish between text produced by ChatGPT and text written by humans. Algorithm (3) shows our work by extracting features.

Algorithm 3: Proposed strategy of feature extraction.

Input: Raw Corpus of text documents.
Output: TF-IDF feature matrix for text classification.

Begin
Step 3: Feature Extraction using TF-IDF
Define a TfidfVectorizer with specified parameters (max_features=5000, stop_words="English").

Apply the TfidfVectorizer to the preprocessed and linguistically analysed Corpus:
For each preprocessed document in the Corpus, do
Calculate each term's Term Frequency (TF) using Eq (2).
Calculate Inverse Document Frequency (IDF) for each term in the Corpus by using Eq (3).
Compute the TF-IDF score for each term by multiplying TF and IDF using Eq (4).
End for
End

3.3. Detection Models

We will divide the data into two different sections. The first section will be used in data training in our models, which is a more significant part of the work, and then the other part of the data will be used after training the model for the evaluation. We do not recommend using the exact data for both training and evaluation, as the model will acquire new insights from the data. This study will employ a collection of ML algorithms, including LR, XGBoost, highly randomised trees, Support Vector Machine (SVM), Random Forest (RF), AdaBoost, and Stochastic Gradient Descent (SGD), to classify GPT-generated text as a binary classification task [23]. We will use a classification strategy based on the perplexity score and our suggested ML-based technique [36]. Perplexity is typically used to assess how well language models perform in NLP activities like text creation and machine translation. It gauges how effectively a language model can estimate a specific text [25]. Perplexity quantifies the degree of uncertainty in text produced by probabilistic language models. A lower level of confusion suggests that the algorithm can anticipate the text more accurately. Put differently, language models are more likely to create texts with fewer perplexities, whereas humans are more likely to develop texts with greater perplexities.

3.4. Ensemble Learning Building

Our ensemble learning model implements the 'vote classifier', an ensemble technique that integrates the forecasts of many fundamental ML models to enhance overall performance. As seen in Figures 13 and 14, the base models in this study will vary between the two datasets based on the most effective detection.

Figure 13. Ensemble learning of the AI-GA dataset.

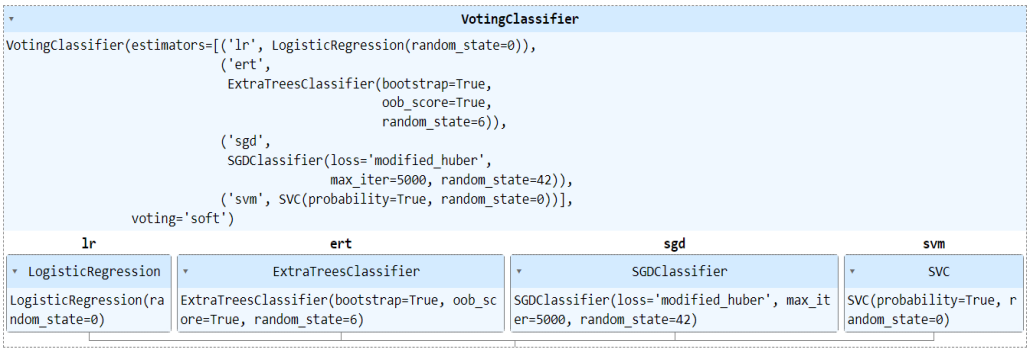


Figure 14. Ensemble learning of the HWAI dataset.

After selecting the techniques, we can gradually train the models using the data. We will assess the model to see if it meets our needs. Now, let us check if our model's assumptions are sound.

4. Results and Discussions

This section presents the experimental results of the proposed framework, detailing model performance, evaluation metrics, and insights derived from the results. The study evaluates seven traditional ML models and an ensemble learning approach using two datasets AI-GA and HWAI.

4.1. Model Training

After feature extraction and preprocessing, the last step is to train several ML models on the two datasets AI-GA and HWAI to tell the difference between text written by humans and text that AI wrote. During the training phase, we fit the models to 75% of the entire data, comprising the TF-IDF feature matrix and related classes (33013, 5000 for the HWAI dataset and 21496, 5000 for the AI-GA dataset). In this work, we examine a variety of models in order to evaluate their respective performances and determine which is best for the given goal. As shown in Table 4 below, this study examined seven ML models in total.

Table 4. Results for the training models of datasets.

Models	AI-GA dataset (ACC.)	HWAI dataset (ACC.)
LR	98.96	97.2
XGBoost	99.92	97.64
ErT	100.00	100.00
SVM	99.94	99.71
RF	100.00	100.00
AdaBoost	95.93	87.71
SGD	99.77	98.23

We design each model with the default configurations of the scikit-learn package and calibrate it to the training dataset. Table 5 illustrates that the fitting phase adjusts the model's parameters to reduce the gaps between the expected and actual classes in the training data. Utilising this strategy, the model proficiently "acquired comprehension" to discern patterns in the "TF-IDF vectorizer" properties that distinguish ChatGPT-generated text from human-authored language.

Table 5. The best parameters utilised for ML models.

Models	Parameters identifier	Parameters description
TF-IDF	TfidfVectorizer (max_features=5000, stop_words="English")	"Max features" is to limit the number of features from the datasets for which we want to calculate the TF-IDF scores. "Stop Words" are common words that frequently appear in text data but carry little meaning, including ("is", "the", "and", "of", etc.).
LR	LogisticRegression (random_state=0)	To ensure consistent outcomes, ML models employ a "Random State" technique to manage any inherent unpredictability.
XGBoost	XGBClassifier (random_state=0)	The random permutation of the features at each split is managed by the "Random State". In order to create a validation set, it also regulates the training data's random splitting.
ErT	ExtraTreesClassifier (random_state=6, bootstrap=True, oob_score=True)	"Random State" presents for decision trees in sci-kit-learn determines which feature to select for a split if there are two equally good splits. "Bootstrap" is a phenomenon of resampling random observations from the datasets to make a new randomised dataset. "Out of Bag (OOB)" score validates the extremely randomised trees model.
SVM	SVC (probability=True, random_state=0)	"Probability" typically refers to the SVM classifiers that will enable probability estimates for class labels. The creation of pseudorandom numbers to shuffle the data for probability estimations is managed by "Random States."
RF	RandomForestClassifier (n_estimators=500, n_jobs=10, bootstrap=True, random_state=42, criterion='entropy')	"N_Estimators" indicates how many decision trees will be employed in the ensemble. The "N_Jobs" parameter determines the number of CPU cores to use for parallelising the training of individual decision trees in the ensemble. "Random State" presents for decision trees in sci-kit-learn determines which feature to select for a split if there are two equally good splits. "Bootstrap" is a phenomenon of resampling random observations from the datasets to make a new randomised dataset. "Criterion" establishes the function that each decision tree in the ensemble uses to gauge the quality of a split.
AdaBoost	AdaBoostClassifier (random_state=1)	"Random State" is used to control the randomness of the algorithm. It is an integer that serves as the seed for the random number generator.
SGD	SGDClassifier (max_iter=5000, loss="modified_huber," random_state=42)	"Max_Iter" establishes the maximum number of epochs or iterations the classifier should go through while being trained. "Loss," the loss function that will be applied during training, is specified. The loss function measures the error between the expected and actual labels. "Random State" is used to control the randomness of the algorithm. It is an integer that serves as the seed for the random number generator.

4.2. Model Evaluation

This categorisation task aimed to discern between these groups according to the text's content. Following model training, we assess the models' performance using standard metrics on the testing data, representing 25% of the entire dataset, a different subset not utilised for training. This section displays the performance of the implemented models on the two datasets. By evaluating the models' ability to generalise to new data, we can determine how well they function in the "real world."
First, we apply typical ML models to two datasets of themes with varying word counts. Table 6 presents the findings. Making predictions involves using the patterns that each trained model learns during training to estimate the classes for the testing data. These classes comprise the matrix of TF-IDF features for the test samples, which indicate whether the text is human or ChatGPT.

We employed criteria like accuracy and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) scores to assess each algorithm's performance. We evaluated the models based on their accuracy and

ROC-AUC. This widely used statistic is unaffected by class imbalance and provides a comprehensive overview of a classifier's performance across potential classification thresholds. Table 6 allows us to evaluate the models' performance and select the most effective model for distinguishing between ChatGPT and human text. Afterwards, we can adjust this model and apply it to various tasks, including content moderation, plagiarism detection, and quality assurance for text-generating systems.

Table 6. Results for the testing models of datasets.

Models	AI-GA dataset (ACC.)	AI-GA dataset (ROC-AUC)	HWAI dataset (ACC.)	HWAI dataset (ROC-AUC)
LR	97.72	98.00	95.79	96.00
XGBoost	97.67	98.00	94.14	94.00
ErT	96.80	97.00	94.55	95.00
SVM	98.20	98.00	96.46	96.00
RF	96.55	97.00	94.23	94.00
AdaBoost	95.59	96.00	87.29	87.00
SGD	98.32	98.00	96.35	96.00
Ensemble learning	98.60	99.88	96.52	99.37

As indicated by Table 6, the ensemble learning by voting classifiers achieved the highest accuracy and ROC-AUC of 98.60% and 99.88% for the AI-GA dataset; also, in the HWAI dataset, ensemble learning by voting classifiers achieved the highest accuracy and ROC-AUC of 96.52% and 99.37, whereas the SVM classifier had a 96.35% accuracy rate in the HWAI dataset and the SGD in the AI-GA dataset achieved an accuracy of 98.32%.

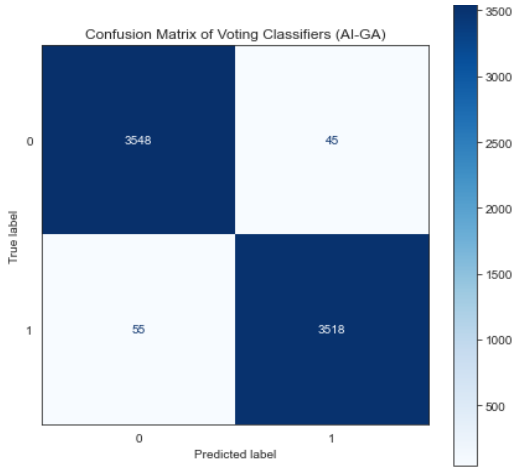


Figure 15. CM for ensemble learning on the AI-GA.

Secondly, the prediction results for the classification problem are summarised in the Confusion Matrix (CM). Count values have been utilised for describing the number of accurate and inaccurate predictions for every one of the classes. The CM yields the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) derivations. The subsequent figures, Figures 15 and 16, show a differentiating AIGT and HWT CM produced by an ensemble learning model employing binary classification on the two datasets AI-GA and HWAI.

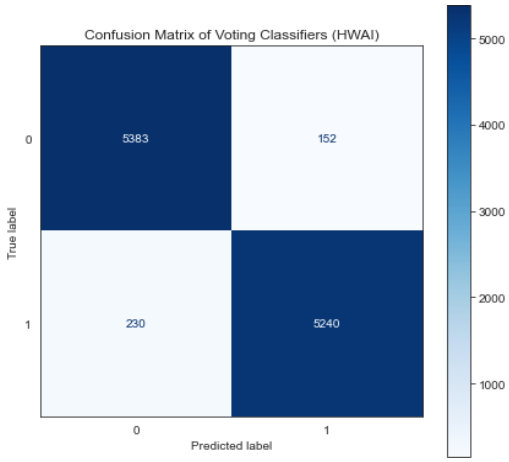


Figure 16. CM for ensemble learning on the HWAI.

We observe that the CM on the AI-GA dataset has the following measurements: TP=3548, FP=45, FN=55, and TN=3518. We also see in Figure 16 that the CM on the HWAI dataset has the following measurements: TP=5383, FP=152, FN=230, and TN=5240. We notice a clear difference in numbers between the two figures since the HWAI dataset is larger than the AI-GA dataset.

Thirdly, we show the outcomes for our models in Figures 17 and 18. However, we employed cross-validation to achieve the conventional classification performance for the two datasets, dividing them into 75% training and 25% testing datasets based on the same recommended partition. We added a metric to these pictures: the Standard Deviation (Std). This is a way to show how different the evaluation metric (like accuracy or ROC-AUC) is across different folds of the data during cross-validation. A higher (Std) suggests that the model's performance is more sensitive to the choice of training and validation data, indicating potential instability or variance in the model's performance.

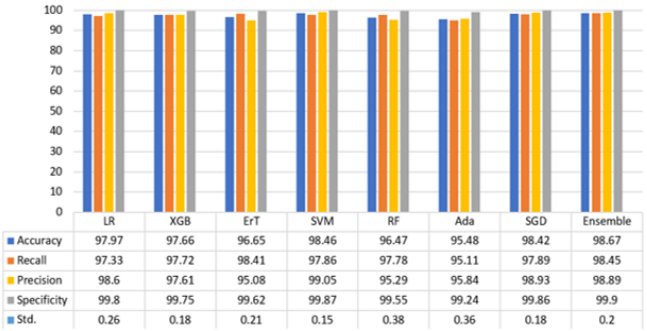
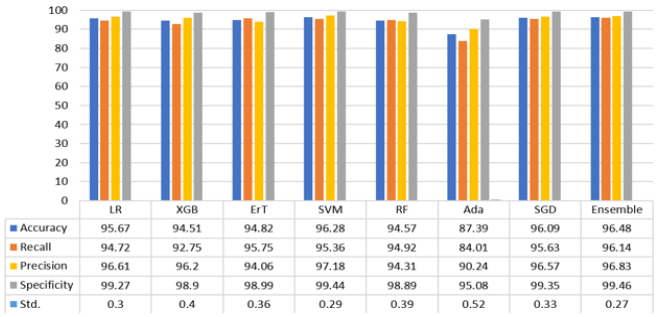


Figure 17. Cross-validation performance measures of our models and Std for AI-GA dataset.

As we observe in Figures 17 and 18, the models maintained almost the same values, while we notice that the values in the specificity metric are all high. This is very useful in our strategy because We employ a metric to evaluate the effectiveness of a binary classification model. It quantifies the ratio of TN accurately recognised by the model relative to the total number of

real negative cases. Conversely, the Std yielded varying percentages for each division.



	LR	XGB	ErT	SVM	RF	Ada	SGD	Ensemble
Accuracy	95.67	94.51	94.82	96.28	94.57	87.39	96.09	96.48
Recall	94.72	92.75	95.75	95.36	94.92	84.01	95.63	96.14
Precision	96.61	96.2	94.06	97.18	94.31	90.24	96.57	96.83
Specificity	99.27	98.9	98.99	99.44	98.89	95.08	99.35	99.46
Std.	0.3	0.4	0.36	0.29	0.39	0.52	0.33	0.27

Figure 18. Cross-validation performance measures of our models and Std for HWAI dataset.

4.3. Benchmarking and Framework Evaluation

We selected the most effective techniques for binary classification of the AI-GA and HWAI datasets. Two further studies employed identical datasets. Consequently, eight models were identified: seven classical strategies (LR, XGBoost, Extra Trees Classifier (ErT), SVM, RF, Ada, and SGD) and one ensemble learning strategy utilising voting classifiers (LR, SGD, ErT, XGBoost, and SVM), as seen in Figures 13 and 14. We evaluated the models' achievements based on their capacity for feature extraction. We compared the current study's results with previous research (Table 7).

Table 7. Baseline comparison of the results of previous studies.

Previous studies	Dataset used	Models	Accuracy
[14]	Radiology abstracts	Perplexity score	95%
[39]	HWAI	LR	97%
[44]	AI-GA	LSTM-w2v	98.7%
Proposed ensemble model	HWAI	Ensemble learning	99.37%
Proposed ensemble model	AI-GA	Ensemble learning	99.88%

The studies enumerated in Table 7 reflect the most recent research and the authors of these datasets, as elaborated in section 3. The analysis of these studies revealed a lack of extensive preprocessing of the dataset. On the other hand, the present study concentrated on addressing the issue with a four-step preprocessing approach, which included cleaning, EDA, linguistic analysis, and text representation, as outlined in section 4. The accuracy achieved in this study surpassed that of prior research, attributable to the hyperparameter optimisation across all features utilising the ensemble learning model, resulting in an accuracy of 99.88%.

The results show that abstracts written by AI show more consistency in linguistic patterns than those written by humans, increasing the classification accuracy in the AI-GA dataset. Such consistency is probably due to optimisation strategies typical for AIGT. In addition, ensemble learning proved to be the most robust method; it outperformed single models by effectively combining predictions and reducing variability across datasets. High accuracy and low variance during cross-validation validate its excellent performance, highlighting its reliability for real-world applications. These results underline the functional relevance of ensemble learning in solving problems related to plagiarism detection and keeping academic integrity in scientific publishing, mainly in situations where a distinction between AI-generated and HWT is essential.

4.4. Key Differences and Advancements in AI-Generated Text Detection

This study advances prior work by introducing ensemble learning, TF-IDF features, EDA, and linguistic analysis, focusing specifically on academic texts. Table 8 compares the previous and current approaches.

- **Methodology**: the earlier study used Sentence-BERT (SBERT) and RoBERTa with LR and FNNs. The new framework applies interpretable ensemble models (voting classifier, SVM, RF, etc.,) with TF-IDF and linguistic features, improving flexibility and transparency.
- **Datasets**: while the previous work used a large single-source dataset (1M samples), this study evaluates across two distinct academic datasets AI-GA and HWAI, enhancing generalization.
- **Performance vs. Interpretability**: although RoBERTa-FNN previously achieved 99.95% accuracy, the new ensemble model (98.60% on AI-GA, 96.52% on HWAI) offers stronger cross-domain adaptability and explainability.
- **Robustness**: improvements include feature-level analysis (n-grams, syntax), statistical validation (Friedman, Wilcoxon tests), and lower performance variance through cross-validation.

Table 8. Comparison of methodologies and contributions of our previous study.

Aspect	Previous study [17]	Current study
Title	Textual authenticity in the AI era.	Bridging the gap: ensemble learning-based NLP framework.
Research focus	Detection using embeddings (SBERT, RoBERTa).	Detection using ensemble learning, TF-IDF, EDA.
Methodology	LR, FNNs, with SBERT and RoBERTa embeddings.	Ensemble learning (voting classifier with LR, SGD, XGBoost, etc.), RF, SVM, and additional ML models.
Dataset	1M samples (balanced).	Two datasets AI-GA, HWAI.
Best accuracy	99.95% (RoBERTa-FNN).	98.60% (AI-GA), 96.52% (HWAI).
Feature extraction	Transformer embeddings.	TF-IDF+linguistic analysis.
Novelty	Focus on transformer-based classification.	Focus on interpretability and real-world utility.
Contribution	Demonstrated transformer power.	Introduced robust, interpretable ensemble framework.
Key findings	FNN>LR with RoBERTa>SBERT.	Ensemble+linguistic analysis improves generalization.

These advancements shift focus from raw performance to real-world usability. While transformer-based methods offer high accuracy, they lack transparency. Our ensemble-based model balances precision, interpretability, and resilience key traits for AI detection tools in academic integrity workflows.

4.5. Discussion

These findings have important theoretical and practical implications for AI in academic content detection. Linguistic analysis showed that AIGTs often use simpler vocabulary and sentence structures. This pattern, especially noticeable in the AI-GA dataset, supports the use of linguistic features as reliable indicators in classification.

Ensemble learning was the most effective approach. It combines the strengths of multiple models to improve accuracy and stability. This makes ensemble strategies well-suited for a large number of text classification tasks.

In practice, the proposed structure can be used for actual applications like plagiarism checking, academic quality assurance, and content management. Because of its high accuracy and low variability, it is a good option to use as an AI written text detection tool.

The study also poses ethical and social concerns. Artificially created content left unaddressed could undermine the credibility of scientific publishing and scholarly work. This research emphasizes that using good detection techniques is essential to maintaining public and institutional confidence and openness.

But detection models also pose ethical problems. False positives are possible, particularly for non-native English authors whose texts might look like AI-produced patterns. To prevent such risks, thresholds for ensemble methods need to be conservatively optimized in favor of precision to prevent misclassifying original scholarly work.

Follow-up studies would involve domain-specific data sets, in-real-time detection software, and the combination of human judgment with AI systems. Such measures would enhance the utility of the framework as well as drive responsible AI use in academia and in the workplace.

5. Conclusions

This paper presented an effective framework for identifying AI-GA content against HWT and confronting a rising concern in academic ethics and fair content generation. The research combined EDA, linguistics analysis, and ML paradigms with a particular focus on ensemble learning models. With two heterogeneous datasets, AI-GA and HWAI, the current study proved an integrative methodology of augmented preprocessing, high-dimensional TF-IDF-based feature extraction, and blending all the linguistic patterns, n-grams, and sentence complexity. In both AI-GA and HWAI datasets, the ensemble learning model surpassed all the separate classifiers with invariably enhanced performance, with accuracy of 98.60% and 96.52% and ROC-AUC of 99.88% and 99.37%, respectively. The results underline the importance of linguistic patterns, such as less diverse vocabulary and simpler sentence structures, in AIGTs. Strong ensemble learning methods combined with such linguistic analysis position the presented framework as an efficient and scalable solution for detecting AI-generated content in academic and professional settings. The work also emphasizes the risks of uncontrolled AIGT, especially in mechanisms of peer review, content moderation, and moral publishing standards. It thus contributes to cutting-edge outcomes and introduces transparency in scholarly communication.

Follow-up papers might expand this foundational framework to more datasets for application in specific areas and introducing semantic and context facets for real-time detection applications and thereby making deployment more pervasive. Furthermore, it would augment usability in functions by further extending the framework towards identifying hybrid text collaboratively produced by humans and AI to fix dataset bias and generalisation across domains.

In addition, upcoming research can be directed toward integrating Hybrid Quantum Machine Learning (HQML) approaches to extend the detection capability. Through the convergence of traditional ML and quantum algorithmic computational power, HQML can likely accelerate learning procedures and unveil deeper patterns in linguistic data. This path holds promise for scalable high-accuracy models that are very much amenable to real-time integration in academic content authentication.

References

[1] ACL, ACL 2023 Policy on AI Writing Assistance, the 61st Annual Meeting of the Association for Computational Linguistics, https://2023.aclweb.org/blog/ACL-2023-policy/, Last Visited, 2025.

[2] Agarwal B. and Mittal N., "Text Classification Using Machine Learning Methods-A Survey," *in Proceedings of the 2nd International Conference on Soft Computing for Problem Solving*, Jaipur, pp. 701-709, 2014. https://doi.org/10.1007/978-81-322-1602-5_75

[3] Alamleh H., Alqahtani A., and Elsaid A., "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning," *in Proceedings of the Systems and Information Engineering Design Symposium*, Charlottesville, pp. 154-158, 2023. https://ieeexplore.ieee.org/document/10137767

[4] Alkaoud M., Alsaqoub M., Aljodhi I., Alqadibi A., and Altammami O., "ACLM: Developing a

Compact Arabic Language Model," *The International Arab Journal of Information Technology*, vol. 22, no. 3, pp. 535-546, 2025. https://doi.org/10.34028/iajit/22/3/9

[5] Berriche L. and Larabi-Marie-Sainte S., "Unveiling ChatGPT Text Using Writing Style," *Heliyon*, vol. 10, no. 12, pp. 1-19, 2024. https://doi.org/10.1016/j.heliyon.2024.e32976

[6] Boden M. and Edmonds E., "What is Generative Art?," *Digital Creativity*, vol. 20, no. 1-2, pp. 21-46, 2009. https://doi.org/10.1080/14626260902867915

[7] Cabanac G. and Labbe C., "Prevalence of Nonsensical Algorithmically Generated Papers in the Scientific Literature," *Journal of the Association for Information Science and Technology*, vol. 72, no. 12, pp. 1461-1476, 2021. https://doi.org/10.1002/asi.24495

[8] Clark E., August T., Serrano S., Haduong N., and et al., "All That's 'Human' is not Gold: Evaluating Human Evaluation of Generated Text," *in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Virtual, pp. 7282-7296, 2021. https://aclanthology.org/2021.acl-long.565/

[9] Coppersmith G. and Kelly E., "Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis," *in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, pp. 22-29, 2014. https://aclanthology.org/W14-3103.pdf

[10] Crothers E., Japkowicz N., and Viktor H., "Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," *arXiv Preprint*, vol. arXiv:2210.07321v4, pp. 1-36, 2023. https://arxiv.org/abs/2210.07321

[11] Dalalah D. and Dalalah O., "The False Positives and False Negatives of Generative AI Detection Tools in Education and Academic Research: The Case of ChatGPT," *The International Journal of Management Education*, vol. 21, no. 2, pp. 100822, 2023. https://doi.org/10.1016/j.ijme.2023.100822

[12] Desaire H., Chua A., Isom M., Jarosova R., and Hua D., "Distinguishing Academic Science Writing from Humans or ChatGPT with over 99% Accuracy Using Off-the-Shelf Machine Learning Tools," *Cell Reports Physical Science*, vol. 4, no. 6, pp. 101426, 2023. https://doi.org/10.1016/j.xcrp.2023.101426

[13] Dhariwal P. and Nichol A., *Advances in Neural Information Processing Systems*, Curran Associates, 2021. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[14] Elek A., Yildiz H., Akca B., Oren N., and Gundogdu B., "Evaluating the Efficacy of Perplexity Scores in Distinguishing AI-Generated and Human-Written Abstracts," *Academic Radiology*, vol. 32, no. 4, pp. 1785-1790, 2025. https://doi.org/10.1016/j.acra.2025.01.017

[15] Fernandez-Hernandez A., Arboledas-Marquez J., Ariza-Merino J., and Jimenez-Zafra S., "Taming the Turing Test: Exploring Machine Learning Approaches to Discriminate Human vs. AI-Generated Texts," *in Proceedings of the Iberian Languages Evaluation Forum*, Jaen, pp. 1-18, 2023. https://ceur-ws.org/Vol-3496/

[16] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., and et al., "Generative Adversarial Networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, 2020. https://doi.org/10.1145/342262

[17] Hazim L. and Ata O., "Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for Text Classification," *in Proceedings of the 16th International Symposium on Electronics and Telecommunications*, Timisoara, pp. 1-6, 2024. https://ieeexplore.ieee.org/document/10797291

[18] ICML, Clarification on Large Language Model Policy LLM, the 14th International Conference on Machine Learning, https://icml.cc/Conferences/2023/llm-policy, Last Visited, 2025.

[19] Islam N., Sutradhar D., Noor H., Raya J., Maisha M., and Farid D., "Distinguishing Human Generated Text from ChatGPT Generated Text Using Machine Learning," *arXiv Preprint*, vol. arXiv:2306.01761v1, pp. 1-6, 2023. http://arxiv.org/abs/2306.01761

[20] Jakesch M., Hancock J., and Naaman M., "Human Heuristics for AI-Generated Language are Flawed," *PNAS*, vol. 120, no. 11, pp. 1-7, 2023. https://doi.org/10.1073/pnas.2208839120

[21] Katib I., Assiri F., Abdushkour H., Hamed D., and Ragab M., "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning," *Mathematics*, vol. 11, no. 15, pp. 1-19, 2023. https://doi.org/10.3390/math11153400

[22] Kobis N. and Mossink L., "Artificial Intelligence Versus Maya Angelou: Experimental Evidence that People Cannot Differentiate AI-Generated from Human-Written Poetry," *Computers in Human Behavior*, vol. 114, pp. 106553, 2021. https://doi.org/10.1016/j.chb.2020.106553

[23] Kowsari K., Meimandi K., Heidarysafa M., Mendu S., and et al., "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, pp. 1-68, 2019. https://doi.org/10.3390/info10040150

[24] Lavoie A. and Krishnamoorthy M., "Algorithmic Detection of Computer Generated Text," *arXiv*

Preprint, vol. arXiv:1008.0706v1, pp. 1-6, 2010. http://arxiv.org/abs/1008.0706

[25] Lewis M., Liu Y., Goyal N., Ghazvininejad M., and et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv Preprint*, vol. arXiv:1910.13461v1, pp. 1-10, 2019. https://doi.org/10.48550/arXiv.1910.13461

[26] Liao W., Liu Z., Dai H., Xu S., and et al., "Differentiate ChatGPT-Generated and Human-Written Medical Texts," *arXiv Preprint*, vol. arXiv:2304.11567v1, pp. 1-15, 2023. http://arxiv.org/abs/2304.11567

[27] Ma Y., Liu J., Yi F., Cheng Q., and et al., "AI vs. Human-Differentiation Analysis of Scientific Content Generation," *arXiv Preprint*, vol. arXiv:2301.10416v2, pp. 1-18, 2023. https://doi.org/10.48550/arXiv.2301.10416

[28] Nguyen-Son H., Tieu N., Nguyen H., Yamagishi J., and Zen I., "Identifying Computer-Generated Text Using Statistical Analysis," *in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kuala Lumpur, pp. 1504-1511, 2017. https://ieeexplore.ieee.org/document/8282270

[29] Oelke D. and Gurevych I., "A Study on Human-Generated Tag Structures to Inform Tag Cloud Layout," *in Proceedings of the International Working Conference on Advanced Visual Interfaces*, Como, pp. 297-304, 2014. https://doi.org/10.1145/2598153.2598155

[30] OpenAI, Chatgpt: Optimizing Language Models for Dialogue (2022), https://openai.com/blog/chatgpt, Last Visited, 2025.

[31] Pataranutaporn P. Danry V., Leong J., Punpongsanon P., and et al., "AI-Generated Characters for Supporting Personalized Learning and Well-Being," *Nature Machine Intelligence*, vol. 3, pp. 1013-1022, 2021. https://www.nature.com/articles/s42256-021-00417-9

[32] Perez-Castro A., Martinez-Torres M., and Toral S., "Efficiency of Automatic Text Generators for Online Review Content Generation," *Technological Forecasting and Social Change*, vol. 189, pp. 122380, 2023. https://doi.org/10.1016/j.techfore.2023.122380

[33] Petroni F., Rocktaschel T., Lewis P., Bakhtin A., and et al., "Language Models as Knowledge Bases?," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing, and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, pp. 2463-2473, 2019. https://aclanthology.org/D19-1250.pdf

[34] Qaiser S., Utara U., Sintok M., Kedah M., and et al., "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25-29, 2018. DOI: 10.5120/ijca2018917395

[35] Rafea L., Ahmed A., and Abdullah W., "Classification of a COVID-19 Dataset by Using Labels Created from Clustering Algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 164-173, 2021. http://doi.org/10.11591/ijeecs.v21.i1.pp164-173

[36] Rosenfeld R., "A Maximum Entropy Approach to Adaptive Statistical Language Modelling," *Computer Speech and Language*, vol. 10, no. 3, pp. 187-228, 1996. https://doi.org/10.1006/csla.1996.0011

[37] Roumeliotis K. and Tselikas N., "ChatGPT and Open-AI Models: A Preliminary Review," *Future Internet*, vol. 15, no. 6, pp. 1-24, 2023. https://doi.org/10.3390/fi15060192

[38] SCIgen, An Automatic CS Paper Generator, https://pdos.csail.mit.edu/archive/scigen/, Last Visited, 2025.

[39] Singh A., Sharma D., Nandy A., and Singh V., "Towards a Large Sized Curated and Annotated Corpus for Discriminating between Human Written and AI Generated Texts: A Case Study of Text Sourced from Wikipedia and ChatGPT," *Natural Language Processing Journal*, vol. 6, pp. 100050, 2024. https://doi.org/10.1016/j.nlp.2023.100050

[40] Stiff H. and Johansson F., "Detecting Computer-Generated Disinformation," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 363-383, 2022. https://doi.org/10.1007/s41060-021-00299-5

[41] Sun S., Zhao W., Manjunatha V., Jain R., and et al., "IGA: An Intent-Guided Authoring Assistant," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Punta Cana, pp. 5972-5985, 2021. https://aclanthology.org/2021.emnlp-main.483/

[42] Sun Z., Zhang Z., Shen X., Zhang Z., and et al., "Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media," *arXiv Preprint*, vol. arXiv:2412.18148v3, pp. 1-29, 2025. https://arxiv.org/abs/2412.18148

[43] Taylor R., Kardas M., Cucurull G., Scialom T., and et al., "Galactica: A Large Language Model for Science," *arXiv Preprint*, vol. arXiv:2211.09085v1, pp. 1-58, 2022. http://arxiv.org/abs/2211.09085

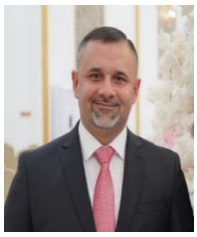
[44] Theocharopoulos P., Anagnostou P., Tsoukala A., Georgakopoulos S., and et al., "Detection of Fake Generated Scientific Abstracts," *in Proceedings of the IEEE 9th International Conference on Big*

Data Computing Service and Applications, Athens, pp. 33-39, 2023. https://ieeexplore.ieee.org/document/10233982

[45] Uysal A. and Gunal S., "The Impact of Preprocessing on Text Classification," *Information Processing and* Management, vol. 50, no. 1, pp. 104-112, 2014. https://doi.org/10.1016/j.ipm.2013.08.006

[46] Wu J., Yang S., Zhan R., Yuan Y., and et al., "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275-338, 2025. https://aclanthology.org/2025.cl-1.8/

[47] Zellers R., Holtzman A., Rashkin H., Bisk Y., and et al., "Defending Against Neural Fake News," *in Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, pp. 9054-9065, 2019. https://dl.acm.org/doi/10.5555/3454287.3455099



Layth Hazim is an academic staff in the Department of Cybersecurity at Tikrit University, Iraq. He earned his B.Sc. degree in Computer Science from Tikrit University in 2007 and his M.Sc. degree from Altinbas University, Turkey, in 2018. He has been serving as the Head of the Computer and Informatics Center (CIC) at Tikrit University since 2020. Currently, he is pursuing his Ph.D. in Electrical and Computer Engineering at Altinbas University. His research interests include Computer Networks, Natural Language Processing, Machine Learning, IoT, Information Security and AI Ethics. Layth has published over 9 peer-reviewed papers.

Oguz Ata is a faculty member in the Department of Computer Engineering at İstanbul Atlas University, Turkey. He received his Ph.D. degree in Computer Engineering from Trakya University in 2012, his M.Sc. degree from Beykent University in 2008, and his B.Sc. degree from Sakarya University in 2004. Dr. Ata has held various academic and administrative positions, including Head of Department at Altınbaş University since 2017 to 2025. His primary research interests encompass Machine Learning, Data Mining, Artificial Intelligence, Cyber Security, and Software Engineering. He has published numerous scholarly articles and supervised multiple graduate theses within these fields.