Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Review of Detecting Text generated by ChatGPT Using Machine and Deep-Learning Models: A Tools and Methods Analysis

Shaymaa Dhyaa Aldeen Ahmed[1,*], Thekra Abbas[1] and Ayad Rodhan Abbas[2]

[1]Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq.
[2]Department of Computer Sciences, University of Technology, Baghdad, Iraq.

ARTICLE INFO

Article history:
Received November 6, 2024
Revised February 15, 2025
Accepted February 23, 2025
Available online March 1, 2025

Keywords:
Text Detection
Machine learning
Deep Learning
Natural Language Processing
Transform Coding
ChatGPT

ABSTRACT

Recently, generative models, such as ChatGPT, have gained considerable attention because of their capacity to generate text almost identical to that produced by humans. However, ChatGPT raises several concerns, particularly regarding the integrity of academic work, the protection of personal information and security, the reliance on artificial intelligence (AI), the evaluation of learning, and the precision of information. Distinguishing between writing generated by machines and text that humans wrote is one of the most critical issues at present. The purpose of this literature review is to provide a comprehensive, up-to-date analysis of the most recent methods for identifying text that ChatGPT created. It examines more than 60 academic papers, especially research articles published after the model's release in 2022, and analyzes state-of-the-art machine learning, deep learning, and hybrid approaches for detecting AI-generated text. The review categorizes detection methods into statistical models, transformer-based architectures, perplexity-based techniques, and human-assisted evaluation. The findings indicate that deep learning models, particularly the Robustly Optimized BERT Pretraining Approach (RoBERTa) and Cross-lingual Language Model with RoBERTa Architecture, have high detection accuracy (up to 99%), whereas traditional statistical methods exhibit limitations in distinguishing complex AI-generated content. This work recommends the use of machine and deep learning techniques and human reviewers in ongoing efforts to distinguish between AI-generated and human-written text. However, given the increasing sophistication and complexity of models, such as ChatGPT, detection techniques have to be continuously improved and innovated to ensure reliability and maintain the integrity of content across various sectors.

1. Introduction

Over the years, natural language processing (NLP) has developed into an important field of research that focuses on improving the capability of computer systems to understand and generate human language for communication purposes. Recent progress in this field has produced language models that apply machine learning (ML) techniques to create text resembling human language and learn from extensive textual datasets [1]. The ability of models to generate text similar to human language has considerably affected communication, language learning, and education. The ChatGPT system is one of the most well-known large language models (LLMs). It was launched by OpenAI and made available to the public in November of 2022 [2, 3].

Generative Pretrained Transformer (GPT) has gained substantial attention in NLP. The development of ChatGPT is a remarkable achievement in NLP and signifies a major

* Corresponding author.
E-mail address: shaymaadhyaa@uomustansiriyah.edu.iq
DOI: 10.24237/djes.2025.18102

progression toward the creation of advanced computer systems proficient in understanding and creating natural language [2]. ChatGPT is trained using a vast amount of textual material, and it provides prompt responses that are relevant and coherent to the context in which they are being used.

As an LLM, ChatGPT can translate languages, generate text, create multiple types of content, and respond to questions informally. Although it is still evolving, it can already perform numerous tasks [3]. However, despite the many benefits that AI-powered text-generating tools offer, they might be misused in various ways, including facilitating scams and phishing attacks, spreading misinformation, and fabricating academic solutions. The quality and reliability of the content produced by these technologies have also elicited much concern [4-6].

Differentiating between manually written and ChatGPT-generated text is crucial to prevent the misuse of AI-generated content, especially in education and content creation [7]. Therefore, this work provides a systematic review and comparative analysis of the feature extraction techniques, ML models, deep learning methods (RoBERTa and BERT), and hybrid approaches used to detect text generated by artificial intelligence (AI). The key contributions of this work are explained as follows:

- This work provides a comprehensive review of detection methods, such as feature extraction, ML models, deep learning (RoBERTa and BERT), and human reviewers. Other related studies generally focused on specific detection techniques or tools only (e.g., RoBERTa and GPT-2) and did not always provide a comprehensive review.
- The role of human reviewers is examined together with the roles of ML and deep learning models, whereas some related studies did not mention the role of human reviewers and focused on the technology only.
- Multiple methods are compared in this review, and the need for continuous innovation due to evolving AI models, such as ChatGPT, is highlighted. Meanwhile, many related works focused on a single method or model only.
- This work discusses the role of explainable AI, particularly Shapley additive explanation (SHAP) values, in enhancing the understanding of detection models. By contrast, related studies rarely mentioned explainable AI methods.

Overall, this work recommends a combination of automated tools and human reviewers for the effective detection of AI-generated text.

The structure of the paper is as described as follows. Section 2 briefly describes the evolution of language models. Section 3 covers related studies and categorizes them into four areas (roles of ChatGPT, key differences between human- and AI-written contents, online AI content detection tools and methods for detecting ChatGPT-generated text). Section 4 outlines the features used for text detection, and Section 5 explains the role of explainable AI in text detection. Section 6 shows the evaluation metrics, Section 7 presents a comparative study and evaluation of relevant articles, and Section 8 discusses the challenges in text detection. Section 9 highlights the general findings and potential future directions, and Section 10 provides the concluding remarks.

2. Evolution of Language Models Leading Up to ChatGPT

In the last few decades, remarkable advancements have been made in the process of developing language models, leading to the sophisticated models we see today, such as ChatGPT. The evolution of language models is briefly described in the following subsections.

2.1 Transformer Models

Transformer models, initially introduced by Vaswani et al. [3], have transformed NLP by enabling the effective handling of long-range relationships without relying on recurrent architectures, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. In AI detection tasks, Transformer-based architectures can be employed to determine whether a piece of text is created by

humans or AI on the basis of patterns, such as syntactic structure, word choice, and sentence length [4]. The primary innovation involved in these architectures is self-attention, which provides a model with the ability to focus dynamically on contextually important terms, hence improving speed and comprehension in tasks that require understanding language dependencies over long distances within a text. This mechanism enables Transformer-based models to concurrently handle complete sequences, in contrast to RNN-based models that sequentially process words.

The performance of sequence-to-sequence models in tasks, such as machine translation, has been greatly improved by the introduction of the attention mechanism by Bahdanau et al. [5]. These models can use this mechanism for relevant regions of the input sequence. In ML, attention can be modeled as assigning weights to data, with useful or essential pieces of data being given large weights to draw increased focus. Several deep learning and advanced NLP methodologies employ attention mechanisms [6].

Transformers are essential tools in NLP and computer vision. Vaswani et al. [3] presented a Transformer model that relies on attention mechanisms to handle long-range dependencies. Transformers eschew recurrence in favor of self-attention, enabling parallelization and efficient training.

2.2 Pre-trained language models

A. BERT

The groundbreaking model known as Bidirectional Encoder Representations from Transformers (BERT) was developed by Jacob et al. [7]. This model was pretrained on huge text corpora by employing masked language modeling, and it was fine-tuned on specific tasks. BERT attains state-of-the-art scores across various NLP benchmarks.

B. GPT

GPT is another state-of-the-art method for NLP. It is a stack of Transformer decoders built on the Transformer architecture. It was initially trained on a large corpus of text, followed by a focus on specific tasks for further refinement. GPT-2 and GPT-3 followed GPT, with GPT-3 being particularly notable for its 175 billion parameters, enabling it to generate text that is extremely cohesive and contextually relevant [8].

2.3 ChatGPT

ChatGPT and Reinforcement Learning from Human Feedback (RLHF): The framework of OpenAI's ChatGPT supports the GPT-3.5 series, which is reinforced with RLHF. The model is then trained and fine-tuned based on feedback from users to improve its ability to generate useful and safe responses. It has been proven that ChatGPT is superior in generating conversational responses, translating language to code, and other complex NLP tasks.

Transformer coding has certain issues despite its benefits. First, Transformer models, especially large ones such as GPT-3 and BERT, demand considerable computational power for training and fine-tuning. Building these models from the ground up or fine-tuning pretrained versions requires powerful GPUs/TPUs, plenty of memory, and substantial processing capabilities, making these models costly and beyond the reach of researchers or small organizations [9]. Second, when trained on large, diverse datasets, Transformer models perform exceptionally well. However, they frequently need to improve in languages or areas with limited data. Fine-tuning on specific domains requires domain-specific training data, but such data are not always available [10, 11].

3. Related works

The emergence of ChatGPT and LLMs has led to concerns regarding their misuse in various contexts, such as spreading misinformation, committing plagiarism, influencing public opinion, and enabling cheating and fraud. Consequently, the differentiation between AI- and human-generated content has become a critical area of research [12, 13]. This literature review explored AI-generated content detection tools and methods. Online sources and academic databases such as Google Scholar and Web of Science were used for the review. Only multilingual studies within the last four years were searched for these studies were assessed in terms of their relevance to the research topics

and inclusion criteria. AI-generated content detection and its effects on numerous fields were examined. Studies that concentrated on the technical aspects of AI or content generated by AI were excluded.

The related studies were categorized as follows: the roles of ChatGPT, the key differences between human- and AI-written text, online AI content detection tools, and methods for detecting ChatGPT-generated text.

3.1 The Roles of ChatGPT

The proliferation of LLMs, such as OpenAI's ChatGPT, has considerably affected multiple domains, including education and healthcare. Li et al. [10] presented a comprehensive analysis of available literature concerning ChatGPT application in the medical field. Imran and Almusharraf [11] conducted a systematic literature evaluation of the 30 most relevant studies to investigate and discuss ChatGPT's usage as an academic writing assistant. Their analysis highlighted the diverse perspectives and contexts associated with utilizing ChatGPT as a writing assistant and the strategies for effective engagement with the tool.

3.2 Key Differences between Human- and AI-written Text

Although ChatGPT is a good resource for scientific writing, technology can never replace human expertise. AI tools, such as ChatGPT, cannot understand the whole writing process. Although they can produce grammatically relevant and accurate content, they may require assistance in capturing the entire meaning or the target audience's specific needs. Human intervention is required to ensure that the generated text is suitable for its intended purpose. AI tools can produce text that may sometimes be inaccurate or unsuitable (may contain factual errors or unsupported assumptions). AI-generated literature typically requires further editing and formatting to match the specific needs of the intended audience [14]. Ma et al. [15] indicated that AI-generated text differs substantially from human-generated text. Furthermore, AI-generated scientific abstracts lack external consistency with real scientific knowledge and necessitate additional insights. Table 1 outlines the differences between text written by humans and text produced by AI across various aspects.

3.3 Online AI Content Detection Tools (Off-the-shelf Detectors)

Off-the-shelf detectors include OpenAI's detector. GPTZero and ZeroGPT are examples under the framework of machine-generated text detection. They were developed by different organizations and can be used by researchers, educators, or other users to identify text that might have been generated by AI models. These tools leverage various algorithms and techniques to analyze text and determine its likely origin (whether human or machine generated). Pegoraro et al.'s [16] "To ChatGPT or not to ChatGPT: That is the question!" thoroughly assessed the latest techniques in ChatGPT detection. The authors validated and tested online detection algorithms and tools with a benchmark dataset that included responses from popular social media platforms and covered a range of subjects, such as finance and medicine. They found that existing detection tools cannot accurately identify ChatGPT-generated information. Many of the tools have accuracy rates below 50%.

Chaka [17] assessed the accuracy and reliability of various online AI content identification systems to distinguish between AI-generated and human-created materials. Their assessment is essential because it enables academics and educational institutions to identify between human-written and AI-generated content.

Uzuna [12] investigated the methods employed to identify AI-generated content and their consequences across different domains. The methods they analyzed included metadata analysis, stylometry, and online tools, and they placed particular emphasis on ethical and legal considerations related to privacy, intellectual property, and content ownership. The objective of this work was to analyze the potential effects that widespread use of AI-generated content could exert on the labor market for human content providers [18].

Table 1: Differences between text written by human and AI generated text across various aspects

Aspect	Human-Written Text	AI-Generated Text
Creativity and Originality [19]	Exhibits unique ideas and perspectives, often reflecting personal experiences and emotions.	Tends to produce content that is formulaic and lacks true innovation, relying on learned patterns.
Coherence and Flow[20]	Demonstrates a natural progression of ideas with logical transitions, even if minor inconsistencies are present.	May have sudden topic transitions or extremely constant patterns, making it mechanical.
Emotional Depth and Tone [8, 21]	Conveys genuine emotions and a distinct personal voice, creating a connection with the reader.	Often lacks emotional depth, resulting in text that feels impersonal or detached.
Grammar and Consistency [21]	Contains occasional grammatical errors and stylistic variations, reflecting human imperfection.	Typically maintains perfect grammar and consistency, sometimes leading to unnaturally flawless text.
Repetition and Redundancy [21]	Varies word choice and sentence structure, minimizing redundancy.	May repeat phrases or use redundant expressions due to pattern-based generation.
Factual Accuracy [19]	Capable of critical analysis and real-time fact-checking, ensuring information accuracy.	May produce outdated or incorrect information, lacking real-time verification capabilities.
Bias and Subjectivity [22]	Reflects personal opinions and cultural context, providing subjective insights.	Aims for neutrality but may unintentionally show biases in training data)..
Adaptability and Personal Voice [19]	Develops a distinct voice, adapting style and tone to various audiences and contexts.	Lacks a personal voice, often producing generic content without adaptability.
Efficiency and Speed [19]	Requires time for research, drafting, and editing, especially for complex topics.	Generates content rapidly, suitable for tasks needing quick turnaround.

3.4 Methods for Detecting ChatGPT-generated Text

Black-box and white-box techniques are the two primary classifications that are commonly used for detection approaches. Access to language models at the API level is the foundation of black-box detection, which is used for data gathering, feature extraction, and classifier creation. However, it suffers from the lack of influence over the model's inner workings. Meanwhile, white-box detection can control and trace its results effectively because it has complete access to the language model [16].

Jawahar et al.'s research in 2020 [23] examined machine-generated text detection. It highlighted three main approaches to detection, as follows:

- Starting from scratch to train a classifier;
- Using a language model for zero-shot detection;
- Using a language model that has been refined to function as a classifier.

Crothers et al. [24] conducted a survey and presented a comprehensive overview of the risks, methods, and detection approaches connected with generated text. The important topics covered by the authors included the effects of multiple domains on detection tasks, the difficulties presented by adversarial attacks, and the social implications of machine-generated text. However, this study did not include assessing or analyzing the outcomes of different approaches. This limitation was primarily due to the need for a standard experimental setup and the use of various datasets and measures.

A. Black-box Techniques

Detection approaches with black-box techniques that can be used to detect human-written and ChatGPT-generated text can be categorized as follows:

a) Perplexity-based Detection Methods

One of the well-known approaches involves the use of perplexity scores. Perplexity measures the uncertainty of a language model in predicting a sequence of words. Low perplexity scores typically indicate text generated by models because these models are well-optimized to produce coherent sequences. Vasilatos et al. [25] developed HowkGPT, a tool that utilizes context-aware perplexity analysis to distinguish between human-written and AI-generated academic assignments. By leveraging a pretrained GPT-2 model and metadata for context-specific thresholds, HowkGPT demonstrates improved accuracy in detecting AI-generated text in educational settings.

b) ML Models

Classic supervised ML models, such as multinomial naïve Bayes, random forest, support vector machine, and K-nearest neighbor, have been trained and assessed using a wide variety of parameters and training approaches [1, 26, 27]. Shijaku and Canhasi [28] developed an ML model that can differentiate between human-written articles and those generated by ChatGPT. They trained and evaluated their model by using a dataset of essays produced by human writers and ChatGPT. Their model was built on the XGBoost classification model [29], and they described their experimentation as using two different feature extraction schemas, namely, term frequency–inverse document frequency (TF-IDF) and a set of hand-crafted features.

c) Deep Learning Models

Diverse forms of deep learning models (e.g., based on Transformer and sequential) have been employed to detect AI generated text. These models can extract linguistic features from text to identify patterns typical of AI-generated or human-written content [30]. Some of these models are described below.

Transformer-based Models

i. **BERT and Variants:** BERT's bidirectional architecture enables the comprehension of context from both sides, resulting in enhanced efficacy in classification problems. According to Jacob et al. [7], fine-tuning BERT on datasets containing human and AI-generated text shows promise in detection tasks.

ii. **RoBERTa:** As an optimized version of BERT, RoBERTa enhances performance through robust training strategies [31]. Its ability to handle nuanced differences in text makes it suitable for distinguishing human- and AI-generated content [2, 28, 32].

Recent research has also explored Transformer-based models with enhanced detection capabilities. Alshammari et al. [33] fine-tuned Transformer models, such as AraELECTRA and the Cross-lingual Language Model with RoBERTa Architecture, to specifically detect AI-generated Arabic text. Their approach, which incorporates a dediacritization layer, considerably improves detectionand achieves results with as high as 100% accuracy in certain configurations.

The study of Oghaz et al. [34] specifically addressed the classification and detection of content generated by OpenAI's ChatGPT by using deep Transformer models. They compiled a comprehensive dataset of human-written and AI-generated content and employed several ML and deep learning models; the efficacy of these models in distinguishing between the two types of content was assessed. The experimental results highlighted the superior performance of the Transformer-based models. Particularly, a custom RoBERTa model obtained an accuracy and F1-score of 0.991 and 0.992, respectively.

Previous studies have explored the use of Transformers for similar tasks. Tien and Labbe [35] utilized grammatical structure similarity to detect AI-generated sentences, and Labbe et al. [36] focused on vocabulary richness and sentence structure to identify computer-generated scientific text. Although effective, these approaches do not leverage the full potential of Transformer-based models.

Recent studies have revealed the effectiveness of Transformers in detecting AI-generated content across various domains. Adelani et al. [37] employed GPT-2 to generate fake reviews and BERT-based classifiers for detection and achieved high accuracy. Similarly, Stiff et al. [38] evaluated Transformer-based detection algorithms for

identifying disinformation in social media posts; they emphasized the need for robust models to withstand adversarial attacks. Beresneva [39] conducted a systematic review of ML techniques for detecting computer-generated text and highlighted the advantages of Transformer based models in capturing linguistic features and phrase frequency. These models are more effective in handling the complexity and diversity of AI-generated content compared with traditional ML approaches. Antoun et al. [40] developed and evaluated ChatGPT detectors for English and French text and investigated their robustness against popular attack techniques and on out-of-domain data. RoBERTa and ELECTRA Transformer models were used for English text [32] [40]. Two pretrained transformer models, namely, CamemBERT and CamemBERTa.6, were used for French text [41].

Sequence Classification Models

1. **LSTMs**

LSTMs can capture sequential dependencies, making them useful for text classification. However, their effectiveness is limited compared with that of Transformer-based models.

2. **Bidirectional LSTMs (BiLSTMs)**

By considering forward and backward contexts, BiLSTMs demonstrate improved performance compared with traditional LSTMs but still fall short of Transformers in handling complex language patterns. Katib et al. [32] developed the TSA-LSTMRNN model to distinguish between human- and ChatGPT-generated text automatically. The TSA-LSTMRNN model aims to analyze the decision-making process and identify any detectable patterns. Feature extraction, classification using LSTMRNN, and parameter adjustment by using TSA are the three stages that compose the model. The performance of their TSA-LSTMRNN approach was evaluated using benchmark databases.

3. **Hybrid Methods: Combinations of the Approaches Above**

Statistical methods have been combined with deep learning approaches, and the Giant Language Model Test Room (GLTR) is a product of such combination. For example, Gehrmann et al. [42] used statistical features alongside deep learning models. Notably, GLTR utilizes the probability distribution of predicted tokens from a language model to identify irregularities in the text generation process. Guo et al. [43] deployed a RoBERTa-based classifier by applying three sample techniques from deep learning and common ML: a deep classifier for QA detection, a deep classifier for single-text recognition, and the use of GLTR Test-2 characteristics to train a logistic regression model. The basis of the two deep classifiers (one for QA and one for single text) is RoBERTa, an effective pretrained Transformer model.

Nguyen et al. [44] applied statistical analysis and logistic regression to identify computer-generated text and achieved notable improvements in accuracy. Transformer-based models, such as RoBERTa and DistilBERT, exhibit high performance as a result of their capacity to retrieve contextual information and dependencies included within text [44, 45]. The DistilBERT-based detector outperforms perplexity-based classification. However, its performance suffers when the modified dataset generated by ChatGPT is considered [5, 45]. Figure 1 illustrates the frequency of related work on the basis of the methods employed.

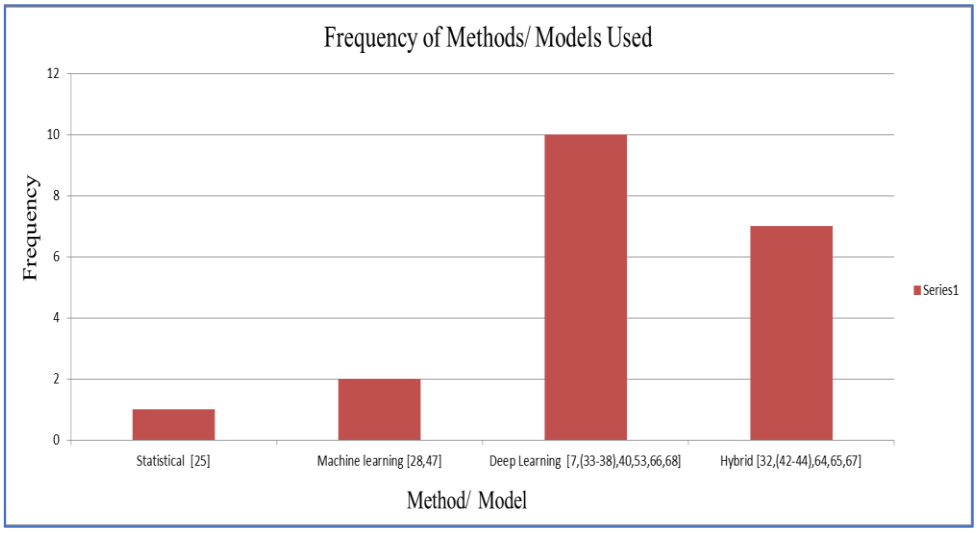

Figure 1. Frequency of various statistical, machine & deep learning methods and hybrid models used in research

B. White-box Techniques

Approaches in white-box detection include zero-shot detection, which leverages pretrained generative models, such as GPT-2 or Grover, and fine-tuning-based detection, which involves adopting pretrained models for the detection task [16].

- Zero-shot detection techniques, such as DetectGPT, use log probabilities to identify AI-generated text. While effective for certain models, these techniques may not generalize well across different AI systems. Grover, another notable model, can generate and detect fake news, outperforming other models such as BERT and FastText in specific scenarios.
- Fine-tuning-based detection methods have shown promising results. Studies have found that fine-tuning-based RoBERTa models consistently outperform equivalent-capacity GPT-2 models. However, these approaches still face challenges in detecting text generated by advanced models, such as ChatGPT.
- Other approaches have also been explored. For instance, Gehrmann et al. [42] developed GLTR, a tool that uses statistical detection and visualization techniques to highlight potential AI-generated text. This method was tested on GPT-2 prompts and human-generated articles from social media. Online tools, such as ZeroGPT and OpenAI's Text Classifier, were adopted as additional methods for detecting AI-generated content.

In summary, the landscape of AI-generated content detection is rapidly evolving, with Transformer-based models (e.g., RoBERTa) and advanced techniques (e.g., zero-shot and fine-tuning-based detection) showing substantial promise [16]. Table 2 shows the type, domain, advantages, disadvantages, and attained performance of some related studies.

Table 2: Type, domain, advantages, disadvantages and attained performance of some related studies

Ref &year	Type	Domain	Advantages	Disadvantages	Attained Performance
Uzun[12] (2023)	Research article	Academic	Address concerns related to academic integrity, analysing the benefits and issues of AI-generated content and the philosophical frameworks in which we could deal with the challenges posed by resulting and potentially disruptive technologies in the future.	The constraints of the study include the use of secondary sources and the possibility of bias in selecting and interpreting research.	Not directly mentioned in the research article. However, it highlights several tools and techniques that have been used to detect AI-generated content
Dhaini, et al. [46] (2023)	Survey	Academic	They find a survey of this kind desirable by considering the substantial opportunities and threats presented by ChatGPT.	Recent methods are primarily pre-prints published on "arXiv", reflecting the rapid nature of research in this field. Additionally, they restrict their focus to scholarly articles and omit non-academic internet tools due to unknown training and internal workings.	Not directly mentioned in the research article, the evaluation of performance is discussed in relation to factors like model robustness, text length, and detection accuracy
Safi and Naini [47] (2023)	Research article	Education	Their machine learning model perfectly discriminated between student and AI responses, and their text characterization approach was able to identify human-detectable-authorship stylistic cues in text by students versus AI.	Firstly, using other groups of students and types of tasks can need to improve the model to the generalizability of our findings. Secondly, students using ChatGPT are likely to alter their responses to some degree.	Accuracy rate : 93.5%
Vasilatos, et al [25] (2023)	Research article	Education	"HowkGPT" improves its analysis by setting category-specific thresholds according to metadata, improving detection accuracy.	A restricted dataset of human and ChatGPT-generated responses to the questions in academic topic across several areas.	Evaluated using AUC (Area Under Curve), recall, and F1 score to determine classification performance.
Imran and Almusharraf [11] (2023)	Review article	Higher Education	It is essential to understand its role as a tool that aids and supports both instructors and learners, as they are advantageous tools that enhance, and simplify the academic process	Firstly, this study focuses solely based on the knowledge that is now available concerning the usage of ChatGPT as a writing assistant in higher education, excluding other functions and uses. Secondly, the depth and quality of the reviewed research may vary, which could impact the generalizability and robustness of the study's findings. Additionally, Due to the rapid evolution of NLP technologies like ChatGPT, the literature that was available at the time of the study might not accurately reflect the most recent developments. This study's focus on higher education may limit its applicability to other educational environments.	Not explicitly quantified in terms of traditional performance metrics.

Ref &year	Type	Domain	Advantages	Disadvantages	Attained Performance
Katib, et al. [32] (2023)	Research Article	General	The results showed that the "TSA-LSTMRNN" system outperformed other recent methods, achieving maximum accuracies of 93.17% and 93.83% on human- and ChatGPT-generated datasets, respectively.	Firstly, Computational Complexity, especially when processing large amounts of text. This could limit its practical application in real-time scenarios, such as content moderation or plagiarismdetection. Secondly, evolving ChatGPT Capabilities: As ChatGPT continues to develop and improve its text generation capabilities, the machine learning effectiveness could decline	Accuracy: 93.17% for human text and 93.83% for ChatGPT text.
Antoun, et al. [40] (2023)	Research Article	General	Identify ChatGPT-generated text, effectively, revealing resilience and a certain level of robustness against primary attack methods in in-domain contexts.	Vulnerabilities are visible in out of domain scenarios.	F1-scores : above 99% , 94% for English and French languages respectively
Chaka [48] (2024)	Review Article	General	It is recommended that a combination of traditional anti-plagiarism tools and modern AI detectors, along with human raters and reviewers, be used in a continuous effort to distinguish between writings that humans write and those that AI generates.	The variation in the efficacy of all the analyzed anti-plagiarism detection systems and all the AI detectors. Both sets of tools demonstrate a lack of reliability in accurately detecting content.	None of the AI detection tools met the 100% accuracy frequently exhibited both false positives and false negatives
Li, et al. [10] (2024)	Review article	medical	Give an overview of the various ways in which ChatGPT can be utilized in the medical field.	The majority of reviewed publications being either reviews or editorial comments may constrain or introduce bias into our understanding of the actual clinical efficacy and applicability of the subject matter.	Accuracy levels around 71.7%

4. Features used for text detection

Feature extraction is an essential process in transforming raw data into a collection of properties or features that can be utilized in ML models or other analysis techniques. The ultimate goal of feature extraction is to reduce data dimensionality and improve computational efficiency, in addition to retaining the most essential information for the specific task. Examining the differences between text composed by humans and text generated by AI involves analyzing various linguistic, semantic, and stylistic features. A brief overview of each is provided below.

A. Linguistic Features

These features pertain to the structural aspects of language, including syntax, grammar, and vocabulary usage. AI-generated materials often exhibit certain linguistic patterns that differ from those in human writing. Georgiou [49] analyzed human-authored and AI-generated essays and revealed considerable differences in components, such as consonants, nouns, word stress, verbs, pronouns, and the use of complex words.

B. Semantic Features

Semantic features relate to the meaning of words and sentences and their interpretation. AI-generated text may sometimes produce

content that lacks deep understanding or context, leading to subtle semantic inconsistencies. Ma et al. [15] indicated that although AI can produce scientifically valid information, it's result may still be lacking in depth and overall quality compared with human-written content.

C. Stylistic Features

Stylistic features relate to the unique manner in which language is used, including tone, voice, and writing style. AI-generated text may lack the nuanced style characteristic of human authors. Opara [50] introduced StyloAI, a model that uses 31 stylometric features to distinguish AI-generated text, and achieved accuracy rates of 81% and 98% on different datasets. Ma et al. [15] employed feature-based style to construct a GPT generated text detection model. An overview of different feature extraction techniques across the text domain is given below.

- **Word count**: number of words in the provided text
- **Word density:** text-provided average word length; it is determined by dividing the number of characters by the number of words
- **Punctuation count**: total number of punctuations in the given text
- **Title word count:** number of title words (each initial letter of the text is capitalized) in the given text
- **Uppercase word count:** number of words beginning with an uppercase letter
- **Noun count**: total number of noun lexicons
- **Verb count:** number of verb lexicons
- **Adjective count**: number of adjective lexicons
- **Adverb count**: number of adverb lexicons
- **Pronoun count**: number of pronoun lexicons

Term frequencies and N-gram features:

- **Vector count**: frequency of terms within the vocabulary
- **Bigram words**: bigram model used to examine the TF-IDF features at the word level, constrained to no more than 5,000 characteristics
- **Trigram words:** trigram model used to examine the TF-IDF characteristics at the word level, constrained to a maximum of 5,000 features
- **BiTrigram characters**: bigram–trigram model used to examine the TF-IDF characteristics at the character level, restricted to a maximum of 5,000 features [51].

Word embedding, count vectorizers, and TF-IDF are the primary foci of the TSA-LSTMRNN approach, which was utilized for the feature extraction process in [32] and [47]. The features are classified into four dimensions, namely, style of writing, coherence, consistency, and argument logistics.

5. Role of Explainable AI in Text Detection

Explainability-techniques are methods used to understand how machine learning models make predictions, particularly when dealing with tasks like detecting whether a text was generated by AI or written by a human [52]. These techniques aim to provide insights into the key attributes and terms that affect classification, allowing for a better understanding of human and ChatGPT writing styles. This is particularly useful in debugging detectors, as they may identify the major phrases that cause misclassification, hence facilitating more effective analysis of such models [46], key concepts explain as follow:

A. SHAP (SHapley Additive exPlanations) values were used to interpret the importance of each feature in the model, providing insights into the detection model [28].

B. GLTR (Giant Language Model Test Room): is an explainability technique focuses on analyzing the likelihood of words in a sequence being generated by a language model. By examining the statistical distribution of word choices, GLTR helps identify if a text exhibits patterns typical of AI-generated content. GLTR operates by examining the probability of each word in a sequence based on its position in the sentence and its likelihood of occurring in that context, according to a large pre-trained language model like GPT-2. The technique highlights which words are more likely or less likely to appear, providing insights into

whether the text follows the common statistical patterns found in AI-generated text or the more varied, unpredictable nature of human writing [42].

C. **Polish Ratio (PR):** The Polish Ratio is a readability and complexity metric that compares the average word length and the average sentence length in a text. It provides an indication of the structural characteristics of a piece of writing, reflecting its complexity and predictability. The Polish Ratio can be used as one of several metrics to evaluate whether a text is AI-generated or human-written. By comparing the Polish Ratio of a text against known patterns from human-written and AI-generated datasets, it becomes possible to identify text that exhibits statistical regularities typical of AI writing. However, the Polish Ratio is generally used in conjunction with other techniques (such as SHAP, and GLTR) to get a more comprehensive assessment of the text [42]. Yang et al. [53] suggest using the Polish Ratio (PR) to assist in the explanation of the detection model that indicates the degree of modification of the text by ChatGPT by using two separate explanation methods, GLTR and PR, assist in supporting the final decision and providing more rational explanations during the decision-making process..

6. Evaluation metrics

Various evaluation metrics are commonly employed to assess detection models that classify text as either AI-generated or human-written. These metrics are essential for evaluating the reliability and performance of models.

A. Accuracy

Accuracy is a simple metric that provides an overall view of model performance [54]. It refers to the percentage of correctly classified instances (human-written and AI-generated) out of the total instances [55]. Safi and Naini [47] used feature extraction and deep learning (TSA-LSTMRNN) and achieved an accuracy of 93.5% in distinguishing student responses from AI-generated ones. The combination of statistical features (TF-IDF) with deep learning models (LSTM and RNN) improves the ability to differentiate AI-generated text. Katib et al. [32] achieved 93.17% accuracy for human text and 93.83% accuracy for ChatGPT text by using TSA-LSTMRNN. Li et al. [10] reported that ChatGPT detection achieves 71.7% accuracy in the medical domain but exhibits reduced effectiveness in specialized areas. This finding suggests that specialized domains (e.g., medicine, law, and scientific writing) present substantial challenges to AI text detection models possibly because of domain-specific terminology and nuances.

B. Precision

Precision refers to the proportion of correctly identified AI-generated texts (true positives) to the total number of texts predicted as AI-generated (true positives plus false positives). It helps assess how many of the predicted AI texts are actually AI-generated [56]. Antoun et al. [40] demonstrated that Transformer-based models, such as RoBERTa, achieve high precision in detecting AI-generated text. By contrast, Shijaku and Canhasi [28] employed XGBoost with TF-IDF features and achieved enhanced precision in identifying AI-generated content.

C. Recall

Recall refers to the ratio of correctly identified AI-generated texts (true positives) to the total number of actual AI-generated texts (true positives plus false negatives); it measures how many of the actual AI-generated texts the model successfully identified. Antoun et al. [40] achieved high recall rates, which means few false negatives. Vasilatos et al. [25] improved recall in distinguishing AI-generated academic writing by using perplexity-based detection methods.

D. F1 Score

The F1 score is the harmonic mean of the precision and recall metrics, and it provides a balance between the two measures. It is especially useful when the data are imbalanced (more human-written texts than AI-generated ones) [3]. Antoun et al. [40] reported remarkable classification performance by achieving an F1 score of over 99% for English and 94% for

French. This result highlights the efficiency of Transformer models in identifying text generated by AI. Similarly, Oghaz et al. [34] used a customized RoBERTa model and achieved an F1 score of 0.991, indicating that the model is highly reliable.

E. Confusion Matrix

The confusion matrix shows the counts of true positives, false positives, true negatives, and false negatives. It provides a detailed breakdown of classification results and helps understand the types of errors made by the model [57].

F. Log Loss (Cross-entropy Loss)

Log loss uses probabilities instead of hard class labels to determine how accurate the model's results are. A small log loss means that the predictions are likely to be correct and have high confidence. Log loss is helpful when the model provides probability scores instead of binary classifications [58].

These metrics, particularly when they are used together, help evaluate the effectiveness of a detection model in distinguishing between human-written and AI-generated texts. Figure 2 compares studies that used accuracy, precision, recall, and F1 score metrics.

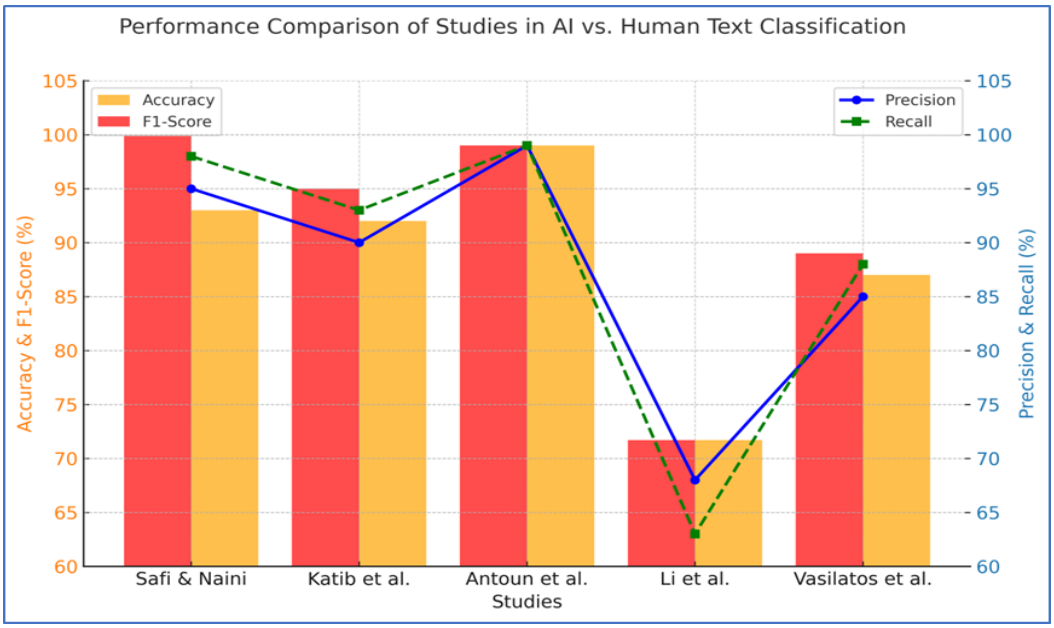

Figure 2. Comparison of the accuracy, precision, recall and F1-Score of different studies

Figures 2 illustrate how different studies achieved varying levels of success in distinguishing human-written text from AI-generated content. The accuracy is generally high. For instance, some studies, such as that of Antoun et al., show near-perfect accuracy and an F1 score of around 99%, indicating strong model performance. The studies of (Safi and Naini) and Katib et al. have high accuracy (around 93%) with high F1-scores,, indicating good performance in both metrics. The study of Li et al. has a noticeable gap between accuracy (71.7%) and F1 score (71.7%), suggesting that although the model identifies AI-generated content correctly to some extent, the balance between false positives and false negatives still needs improvement.

The overall trend where in the Transformer-based models, such as those used by Antoun et al., achieve high precision and recall, resulting in strong balance, as evidenced by the high F1 score. The variability across the studies indicates that although some models (e.g., those of (Safi and Naini) and Katib et al.) perform well, potential for improvement still exists, especially in specialized domains or highly specific datasets. Traditional models, such as those used in the study of Li et al., struggle when they are applied to specialized fields, such as medical text, because these areas often involve complex terminology, jargon, and specific nuances that the models may not be well-equipped to handle. As a result, in certain fields, these models might not be successful in

accurately distinguishing AI-generated text from human-written content.

7. Comparative Study and Evaluation

Several studies have compared the performance of different detection methods.

A. RoBERTa vs. BERT

RoBERTa generally outperforms BERT because of its robust training methodology, training objective, data size and diversity, and dynamic masking [31].

B. Statistical vs. Deep Learning Approaches

Although statistical methods, such as GLTR, provide quick insights, deep learning models, particularly Transformer-based ones, offer superior accuracy and robustness. In their work titled MUGC: Machine-generated versus User-generated Content Detection, Xie et al. [59] evaluated eight conventional ML algorithms for distinguishing between human- and machine-generated content across three datasets: abstracts, essays, and poems. The authors also performed a comprehensive comparative analysis of linguistic characteristics and revealed differences in readability, bias, moral, and affective assessments between machine- and human-generated content. Meanwhile, Dhaini et al. [46] conducted a survey and provided a comprehensive overview of methods and techniques for detecting ChatGPT-created text. They reported and discussed interesting aspects, such as comparative analysis of text produced by humans and ChatGPT and general insights into the characteristics of generated text. Dugan et al. [60] explored human detection capabilities in simple binary classification tasks and evaluated language models by comparing different generative systems and examining how specific model attributes affect human performance.

In addition, Safi and Naini [47] conducted an experiment to compare responses from students and ChatGPT for a typical course assignment. They perfectly discriminated between student and AI responses, and their text characterization approach identified human-detectable authorship stylistic cues in text written by students and AI. Islam et al. [61] presented a model that can recognize manually written and ChatGPT-generated text. They conducted a comparative analysis of 11 ML and deep learning algorithms during the process of classification. The proposed model was tested using a Kaggle dataset with 10,000 texts, 5,204 of which were human-written content collected from news outlets and social media platforms. The proposed method achieves 77% accuracy when the GPT-3.5 corpus is used. However, the study has a limitation, that is, the dataset that was utilized in the process of model training may not be representative of all human-written text or text generated by ChatGPT. This limitation could lead to the model misclassifying some ChatGPT-generated text as human-written text or vice versa. Pegoraro et. al [16] reviewed several simple classifiers, such as logistic regression models, trained on features (e.g., unigram, bigram, and TF-IDF) and achieved up to 97% accuracy in some cases.

However, these models show low effectiveness in the presence of short text. By contrast, advanced approaches, such as those that use RoBERTa-based models, demonstrate enhanced effectiveness in differentiating AI-generated text from human-written content.

8. Challenges in Text detection

The key challenges in detecting ChatGPT-generated text include the following:

1. The sophisticated nature of ChatGPT's text, which closely mimics human writing;
2. The lack of transparency in ChatGPT's training setup and model architecture that makes the development of robust detection methods difficult;
3. The need for feature engineering and model training that can accurately capture the small differences that exist between text generated by machines and manually written text;
4. Models trained on specific datasets may struggle with out-of-distribution data, which affect their robustness.

These challenges pose serious issues in addressing ChatGPT in general. Determining whether ChatGPT's results are reproducible or not is difficult because of ChatGPT's closed-

source nature and the lack of comprehensive information regarding its training and dataset. Changes to models, decommissioning of models, or substantial changes to the price of access can occur at any time. These concerns are balanced by the substantial opportunities and risks associated with ChatGPT [53]. Wang et al. [62] reported that detectors cannot be easily and effectively generalized to cases from unfamiliar domains or LLMs.

9. General findings

The studied articles indicate that baseline models that are trained exclusively on ChatGPT-generated text struggle to identify instances of a polishing attack. This is evident from the tendency of these models to label nearly all polished samples as being written by humans. Errors in detection models can be caused by several factors, such as human-written text being enhanced by AI-driven editing tools or the writing style closely resembling that of ChatGPT [53]. In light of these challenges, this article recommends a combined approach that uses both automated tools (ML and deep learning) and human supervision. This recommendation highlights a critical point of distinction from other studies, which relied solely on either automated methods or human judgment. Such a hybrid approach aims to address biases, enhance detection accuracy, and develop robust deep learning models with improved performance. Future studies should focus on the following:

- The detection of paraphrased AI-generated content, which remains to be a major challenge because of the adaptive nature of advanced language models, should be given attention.
- The robustness of detection tools should be evaluated in specialized domains, including medical and legal settings, where domain-specific nuances complicate detection accuracy.

Table 3 provides a detailed overview of related studies and includes information on the datasets used, languages analyzed, detection methods employed, and SHAP analysis performed. The same studies are arranged based on the language used in Figure 3.

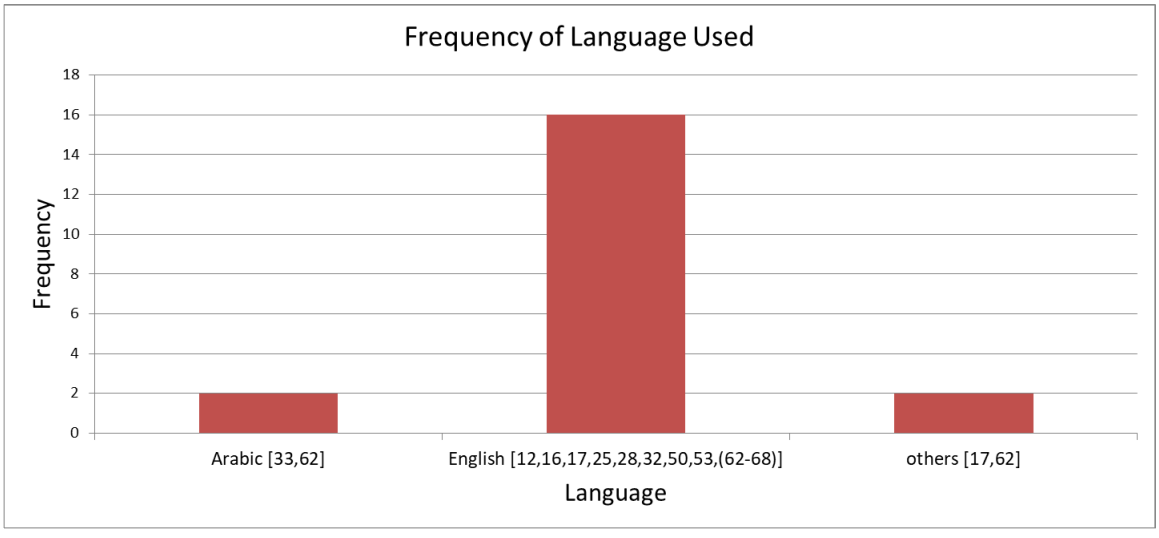

Figure 3. Frequency of related works according to the used language

Table 3: Overview of related studies

Ref &year	Dataset	Language	Detection Methods	SHAP
Shijaku and Canhasi [28] (2023)	252 essays, evenly split between human written and ChatGPT created texts (126 each)	English	Text preprocessing: tokenization, lowercasing, and stemming, Feature Extraction :TF-IDF (Term-Frequency-Inverse-Document-Frequency) and hand crafted features Model Training :XGBoost for training detection model	Yes
Wang et al. [62] (2023)	English: Wikipedia, WikiHow , Reddit (ELI5), arXiv, and PeerRead . Arabic: Wikipedia Chinese: Baike/Web QA, question answering (QA), Russian: RuATD for Urdu, Indonesian and Bulgarian: news	Arabic, Chinese, Bulgarian, English, Russian, Urdu and Indonesian.	Deep Neural Networks: RoBERTa for classification. XLM-R classifier Logistic Regression with GLTR Fearures	No
Gao et al. [63] (2023)	Collected five research abstracts from five medical publications with high-impact, and requested ChatGPT to produce abstracts of the researches based on titles of the journals'.	English	GPT-2 Output Detector plagiarism detector website and -iThenticate	No
Pegoraro, et al. [16] (2023)	The benchmark dataset comprises human and ChatGPT-generated questions and answers across the medical, open Q&A, and financial fields, along with user-generated responses from prominent social networking platforms.	English	AI-generated text detection Tools	No
Chaka [17] (2023)	The data was generated utilizing "ChatGPT, Chatsonic, and YouChat". Three distinct sets of English prompts were entered into three different AI chatbots; each assigned a specific chatbot to create this content. The prompts were presented to the three AI chatbots on two different dates.	English , German, French and Spanish	GPTZero, Writer.com's AI Content Detector, OpenAI Text Classifier, Copyleaks -AI Content -Detector, and GLMTR are five AI content tools.	No
Uzun [12] (2023)	Diverse dataset	English	Copyleaks, Turnitin, metadata analysis, and stylometric analysis.	No
Safi and Naini [47] (2023)	AI-generated responses Unrestricted and restricted	English	NBC algorithm and TF-IDF weight matrix	No
Yang, et al. [53] (2023)	The HPPT dataset includes polished and original matched abstracts, and similarity metrics used to indicate the degree of polishing.	English	RoBERTa	PR
Vasilatos, et al. [25] (2023)	Academic assignments dataset	English	Determine perplexity scores for answers that were created by ChatGPT and those that students wrote.	No
Katib, et al [32] (2023)	Two benchmark databases	English	Word embedding, count vectorizers and TF-IDF (Term- Frequency-Inverse-Document-Frequency)for extract feature. The "LSTMRNN" model is employed for the detecting procedure. The TSA is used to select the parameters of the "LSTMRNN" method.	No

Ref &year	Dataset	Language	Detection Methods	SHAP
Opara, [50] (2024)	Three datasets: poems, abstracts, and essays.	English	RoBERTa	No
Alshammari, et al. [33] (2024)	The dataset includes 43,958 examples of HWT and AIGTs from multiple sources, including "ChatGPT-4, ChatGPT 3.5, and BARD". Additionally, there is a customized dataset with 3078 samples.	Arabic	AraELECTRA and XLM-R	No
Zhang et al. [64] (2024)	Pile and slimpajama datasets, these datasets undergo rigorous filter diverse datasets based on various criteria, including text length and presence of code or mathematical symbols.	English	Mixed methodology that integrates conventional TF-IDF strategies with sophisticated machine learning algorithms, including "Bayesian classifiers, Stochastic Gradient Descent (SGD), Categorical Gradient Boosting (CatBoost), and 12-instances of Deberta-v3-large models".	No
Prova [65] (2024)	The dataset is categorized into two distinct classes: content produced by AI, labeled as 1, and content generated by human, labeled as 0. A total of 3,000 data points were collected, comprising 1,500 samples produced by humans and 1,500 samples generated by artificial intelligence.	English	XGB Classifier, SVM, and BERT	No
Alhijawi, et al. [66] (2024)	AIGTxt dataset contains 3000 records collected from published academic articles across ten domains and categorized into three classes: Human-written, ChatGPT-generated, and Mixed text.	English	AI-Catcher model integrates two deep-learning models, multilayer perceptron (MLP) and convolutional neural networks (CNN).	No
Hui [67] (2024)	This study uses a dataset containing 1,737,000 text messages that cover a variety of topics, including politics, current events, health, science and technology, sports news, personal communications, and others	English	GAN-Based Feature Extraction.and RF-Based Detection.	No
Wang, et al. [68] (2024)	Private dataset, which contains 708 texts labelled with 0 (AI-generated texts)and 670 labelled with 1 (non-AI-generated)	English	BERT	No

10. Conclusion

This study reviewed AI-generated text detection methods and revealed the advantages of Transformer-based models (e.g., RoBERTa and BERT) over traditional approaches. Although deep learning methods achieve high accuracy, challenges remain in multilingual detection, domain-specific identification, and robustness against adversarial attacks. Data from the selected studies were analyzed to assess the effectiveness of current detection tools and methods and identify the challenges faced in various domains. One limitation of this review is its reliance on secondary sources, which may introduce bias in study selection and interpretation. Nevertheless, this literature review provides valuable insights for guiding future research. Researchers should focus on

enhancing explainability, establishing standardized benchmarks, improving real-time detection, and enhancing the robustness of detection models in handling domain-specific text, which presents challenges to AI text detection because of the terminologies and nuances involved. Meanwhile, practitioners should integrate hybrid detection systems and regularly update tools. Future studies should also explore the detection of paraphrased AI content and strengthen adversarial defense with multilingual datasets. As models, such as ChatGPT, become increasingly complex, human supervision remains essential to ensure content accuracy, relevance, and integrity across sectors and languages, highlighting the need for the continuous innovation of detection methods.

References

[1] A. Aliwy, A. Abbas, and A. Alkhayyat, "NERWS: Towards improving information retrieval of digital library management system using named entity recognition and word sense," Big Data and Cognitive Computing, vol. 5, no. 4, p. 59, 2021, doi: 10.3390/bdcc5040059.

[2] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in *2023 IEEE Global Engineering Education Conference (EDUCON)*, 2023: IEEE, pp. 1-9, doi: 10.1109/educon54358.2023.10125121.

[3] A. Vaswani, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA., 2017.

[4] A. J. Yousif and M. H. Al-Jammas, "Real-time Arabic Video Captioning Using CNN and Transformer Networks Based on Parallel Implementation," *Diyala Journal of Engineering Sciences* vol. 17, no. 1, pp. 84-93, 2024, doi: 10.24237/djes.xxxx.13301

[5] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *Preprint* 2014, doi: 10.48550/arXiv.1409.0473.

[6] M. H. Al-Tai, B. M. Nema, and A. Al-Sherbaz, "Deep learning for fake news detection: Literature review," *Al-Mustansiriyah Journal of Science*, vol. 34, no. 2, pp. 70-81, 2023, doi: http://doi.org/10.23851/mjs.v34i2.1292.

[7] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, vol. 1: Minneapolis, Minnesota, p. 2.

[8] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

[9] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, no. 09, pp. 13693-13696.

[10] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, "ChatGPT in healthcare: a taxonomy and systematic review," Computer Methods and Programs in Biomedicine, vol. 245, p. 108013, 2024.

[11] M. Imran and N. Almusharraf, "Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature," *Contemporary Educational Technology*, vol. 15, no. 4, p. ep464, 2023.

[12] L. Uzun, "ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content," *Language Education and Technology*, vol. 3, no. 1, 2023.

[13] M. T. Younis, N. M. Hussien, Y. M. Mohialden, K. Raisian, P. Singh, and K. Joshi, "Enhancement of ChatGPT using API wrappers techniques," *Al-Mustansiriyah Journal of Science*, vol. 34, no. 2, pp. 82-86, 2023.

[14] J. Huang and M. Tan, "The role of ChatGPT in scientific communication: writing better scientific review articles," *American journal of cancer research*, vol. 13, no. 4, p. 1148, 2023.

[15] Y. Ma *et al.*, "AI vs. Human--Differentiation Analysis of Scientific Content Generation," *arXiv preprint arXiv:2301.10416*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.10416.

[16] A. Pegoraro, K. Kumari, H. Fereidooni, and A.-R. Sadeghi, "To ChatGPT, or not to ChatGPT: That is the question!," *arXiv preprint arXiv:2304.01487*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.01487.

[17] C. Chaka, "Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools," *Journal of Applied Learning and Teaching*, vol. 6, no. 2, 2023, doi: https://doi.org/10.37074/jalt.2023.6.2.12.

[18] C. Chen, J. Fu, and L. Lyu, "A pathway towards responsible ai generated content," *arXiv preprint arXiv:2303.01325*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.01325.

[19] Hackernoon, "AI-Generated vs. Human-Written Text: A Technical Analysis," 2023. [Online]. Available: https://hackernoon.com/ai-generated-vs-human-written-text-technical-analysis.

[20] P. S. University. "Q&A: Increasing Difficulty Detecting AI vs Human Writing." https://www.psu.edu/news/information-sciences-and-technology/story/qa-increasing-difficulty-detecting-ai-versus-human (accessed 10-Feb-2025.

[21] Originality.ai. "How to Identify AI-Generated Text." https://originality.ai/blog/identify-ai-generated-text (accessed 10-Feb-2025.

[22] Reddit. "How Do You Distinguish Whether an Article Is AI-Generated or Human-Written?" https://www.reddit.com/r/Futurology/comments/zn7qb9/how_do_you_distinguish_whether_an_article_is (accessed 10-Feb-2025.

[23] G. Jawahar, M. Abdul-Mageed, and L. V. Lakshmanan, "Automatic detection of machine generated text: A critical survey," *arXiv preprint arXiv:2011.01314*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2011.01314.

[24] 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, doi: 10.1109/ACCESS.2023.3294090.

[25] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, and M. Maniatakos, "Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis," *arXiv preprint arXiv:2305.18226*, 2023.

[26] B. K. Al-Windi, A. H. Abbas, and M. S. Mahmood, "Using Texture Analyses and Statistical Classification for Detection Plant Leaf Diseases," *Al-Mustansiriyah Journal of Science*, vol. 32, no. 5, pp. 1-4, 2021, doi: http://doi.org/10.23851/mjs.v32i5.1115.

[27] H. A. Alatabi and A. R. Abbas, "Sentiment analysis in social media using machine learning techniques," *Iraqi Journal of Science*, vol. 61, no. 1, pp. 193-201, 2020, doi: 10.24996/ijs.2020.61.1.22.

[28] R. Shijaku and E. Canhasi, "ChatGPT generated text detection," *Publisher: Unpublished*, 2023, doi: 10.13140/RG.2.2.21317.52960.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794, doi: http://dx.doi.org/10.1145/2939672.2939785.

[30] A. J. Yousif and M. H. Al-Jammas, "A Lightweight Visual Understanding System for Enhanced Assistance to the Visually Impaired Using an Embedded Platform," *Diyala Journal of Engineering Sciences*, pp. 146-162, 2024, doi: https://djes.info/index.php/djes/article/view/1377.

[31] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[32] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, "Differentiating chat generative pretrained transformer from humans: detecting ChatGPT-generated text and human text using machine learning," *Mathematics*, vol. 11, no. 15, p. 3400, 2023, doi: https://doi.org/10.3390/math11153400.

[33] H. Alshammari, A. El-Sayed, and K. Elleithy, "Ai-generated text detector for arabic language using encoder-based transformer architecture," *Big Data and Cognitive Computing*, vol. 8, no. 3, p. 32, 2024, doi: https://doi.org/10.3390/bdcc8030032.

[34] M. M. D. Oghaz, K. Dhame, G. Singaram, and L. B. Saheer, "Detection and Classification of ChatGPT Generated Contents Using Deep Transformer Models," *Authorea Preprints*, vol. 11, 2023, doi: 10.22541/au.167702907.35890747.

[35] N. M. Tien and C. Labbé, "Detecting automatically generated sentences with grammatical structure similarity," *Scientometrics*, vol. 116, no. 2, pp. 1247-1271, 2018.

[36] C. Labbé, D. Labbé, and F. Portet, "Detection of computer-generated papers in scientific literature," *Creativity and universality in language*, pp. 123-141, 2016.

[37] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection," in *Advanced information networking and applications: Proceedings of the 34th international conference on advanced information networking and applications (AINA-2020)*, 2020: Springer, pp. 1341-1354.

[38] H. Stiff and F. Johansson, "Detecting computer-generated disinformation," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 363-383, 2022.

[39] D. Beresneva, "Computer-generated text detection using machine learning: A systematic review," in *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, 2016: Springer, pp. 421-426.

[40] W. Antoun, V. Mouilleron, B. Sagot, and D. Seddah, "Towards a Robust Detection of Language Model Generated Text: Is ChatGPT that Easy to Detect?," *arXiv preprint arXiv:2306.05871*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.05871.

[41] K. Clark, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[42] S. Gehrmann, H. Strobelt, and A. M. Rush, "Gltr: Statistical detection and visualization of generated text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy, 2019, vol. , pp. 111–116.

[43] B. Guo *et al.*, "How close is chatgpt to human experts? comparison corpus, evaluation, and detection," *arXiv preprint arXiv:2301.07597*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.07597.

[44] H.-Q. Nguyen-Son and I. Echizen, "Detecting computer-generated text using fluency and noise features," in *Computational Linguistics: 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16–18, 2017, Revised Selected Papers 15*, 2018: Springer, pp. 288-300.

[45] H.-Q. Nguyen-Son, N.-D. T. Tieu, H. H. Nguyen, J. Yamagishi, and I. E. Zen, "Identifying computer-generated text using statistical analysis," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017: IEEE, pp. 1504-1511, doi: https://doi.org/10.1109/APSIPA.2017.8282270.

[46] M. Dhaini, W. Poelman, and E. Erdogan, "Detecting chatgpt: A survey of the state of detecting chatgpt-generated text," *arXiv preprint arXiv:2309.07689*, 2023, doi: https://doi.org/10.48550/arXiv.2309.07689.

[47] R. Safi and A. J. Naini, "The Work of Students and ChatGPT Compared: Using Machine Learning to Detect and Characterize AI-Generated Text," in *Twenty-ninth Americas Conference on Information Systems*, Panama, 2023.

[48] C. Chaka, "Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review," *Journal of Applied Learning and Teaching*, vol. 7, no. 1, 2024, doi: https://doi.org/10.37074/jalt.2024.7.1.14.

[49] G. P. Georgiou, "Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool," 2024.

[50] C. Opara, "StyloAI: Distinguishing AI-generated content with stylometric analysis," in *International conference on artificial intelligence in education*, 2024: Springer, pp. 105-114, doi: https://doi.org/10.1007/978-3-031-64312-5_13.

[51] T. T. Nguyen, A. Hatua, and A. H. Sung, "How to Detect AI-Generated Texts?," in *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2023: IEEE, pp. 0464-0471.

[52] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[53] L. Yang, F. Jiang, and H. Li, "Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 2, 2023, doi: 10.1561/116.00000250.

[54] H. M. Fadhil, Z. O. Dawood, and A. Al Mhdawi, "Enhancing Intrusion Detection Systems Using Metaheuristic Algorithms," *Diyala Journal of Engineering Sciences*, pp. 15-31, 2024.

[55] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, "Mgtbench: Benchmarking machine-generated text detection," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 2251-2265, doi: https://doi.org/10.1145/3658644.3670344.

[56] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[57] R. J. Kolaib and J. Waleed, "Crime Activity Detection in Surveillance Videos Based on Developed Deep Learning Approach," *Diyala Journal of Engineering Sciences*, pp. 98-114, 2024.

[58] B. Mann *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, vol. 1, 2020.

[59] Y. Xie, A. Rawal, Y. Cen, D. Zhao, S. K. Narang, and S. Sushmita, "MUGC: Machine Generated versus User Generated Content Detection," *arXiv preprint arXiv:2403.19725*, 2024.

[60] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, and C. Callison-Burch, "Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 11, pp. 12763-12771.

[61] N. Islam, D. Sutradhar, H. Noor, J. Raya, M. Maisha, and D. Farid, "Distinguishing Human Generated Text from ChatGPT Generated Text Using Machine Learning. arXiv," *arXiv preprint arXiv:2306.01761*, 2023.

[62] Y. Wang *et al.*, "M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection," *arXiv preprint arXiv:2305.14902*, 2023.

[63] C. A. Gao *et al.*, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *NPJ Digital*

Medicine, vol. 6, no. 1, p. 75, 2023, doi: https://doi.org/10.1038/s41746-023-00819-6.

[64] Y. Zhang *et al.*, "Enhancing Text Authenticity: A Novel Hybrid Approach for AI-Generated Text Detection," *arXiv preprint arXiv:2406.06558*, 2024.

[65] N. Prova, "Detecting AI Generated Text Based on NLP and Machine Learning Approaches," *arXiv preprint arXiv:2404.10032*, 2024.

[66] B. Alhijawi, R. Jarrar, A. AbuAlRub, and A. Bader, "Deep Learning Detection Method for Large Language Models-Generated Scientific Content," *arXiv preprint arXiv:2403.00828*, 2024.

[67] Y. Hui, "Using generative adversarial network to improve the accuracy of detecting AI-generated tweets," *Scientific Reports*, vol. 14, no. 1, p. 29322, 2024.

[68] H. Wang, J. Li, and Z. Li, "AI-Generated Text Detection and Classification Based on BERT Deep Learning Algorithm," *arXiv preprint arXiv:2405.16422*, 2024.