DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND RESEARCH

SETIF 1 UNIVERSITY-FERHAT ABBAS

FACULTY OF SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

DOCTORAL THESIS OF THE THIRD CYCLE

Option: Intelligent Systems and Machine Learning

Fighting Deepfake Text:

Towards Building Robust defences against Arabic AI-generated Text

---

By **Amal Boutadjine**

Supervisor:
Prof. **Fouzi Harrag**

Co-supervisor:
Prof. **Khaled Shaalan**

Committee in charge:

Houssam Mansouri, Prof, Univ Ferhat Abbas Setif 1,	President
Fouzi Harrag, Prof, Univ Ferhat Abbas Setif 1,	Supervisor
Khaled Shaalan, Prof, British University in Dubai, UAE,	Co-Supervisor
Mouhoub Belazoug, MCA, Univ Md Bachir Ibrahimi, BBA,	Examiner
Djamila Mohdeb, MCA, Univ Md Seddik Benyahia, Jijel,	Examiner
Samir Fenanir, MCA, Univ Ferhat Abbas Setif 1,	Examiner

Algeria, Setif, 2025

Abstract

The rapid advancement of Natural Language Processing (NLP), particularly through transformer-based architectures, has led to powerful large language models (LLMs) capable of generating human-like text for a variety of tasks, including question answering, content creation, and document completion. While these innovations bring transformative benefits, they also introduce ethical risks—most notably the potential for generating deceptive content at scale. Deepfake text poses a growing threat in digital ecosystems, enabling disinformation, academic fraud, and online manipulation.

In response, many efforts have been made and emerged to address the challenge of detecting AI-generated text. However, existing approaches overwhelmingly focus on English, overlooking the linguistic complexity and vulnerability of Arabic. This thesis addresses that gap by proposing a novel detection framework tailored to Arabic deepfake text. Leveraging state-of-the-art transformer models and curated Arabic corpora, we develop and evaluate scalable detection techniques that account for Arabic's richness. Our empirical results demonstrate high precision in distinguishing human-authored from machine-generated content, thereby contributing to the ethical deployment of generative AI in multilingual contexts and strengthening defences against AI-driven misinformation.

Keywords: Machine Learning, Transfer Learning, Generative AI, Large Language Models, Deepfake Text, Human-authored text, AI-Generated Text Detection.

ملخص:

أدى التقدم السريع في مجال معالجة اللغة الطبيعية (NLP) ، لاسيما من خلال البنى القائمة على المحولات (Transformers)، إلى ظهور نماذج لغوية ضخمة (LLMs) قادرة على توليد نصوص تُشبه اللغة البشرية في مجموعة واسعة من المهام، مثل الإجابة عن الأسئلة، وإنشاء المحتوى، واستكمال الوثائق. ورغم ما تحمله هذه الابتكارات من فوائد جمة، فإنها تنطوي أيضًا على مخاطر أخلاقية—أبرزها القدرة على توليد محتوى مُضلّل على نطاق واسع. إذ يُعدّ النص المُزيف (Deepfake text) تهديدًا متناميًا في البيئات الرقمية، حيث يُستخدم في نشر المعلومات المضللة، والاحتيال الأكاديمي، والتلاعب عبر الإنترنت.

وفي مواجهة ذلك، بُذلت جهود عديدة لمواجهة تحدي كشف النصوص المُولدة آليًا. غير أن معظم المقاربات الحالية تركز بشكل شبه حصري على اللغة الإنجليزية، متجاهلة التعقيد اللغوي والخصوصيات البنيوية للغة العربية، إلى جانب قابليتها للاستغلال. تسعى هذه الرسالة إلى سدّ هذه الفجوة من خلال اقتراح إطار جديد لرصد النصوص العربية المُزيفة المُولدة بالذكاء الاصطناعي. وبالاستفادة من أحدث نماذج المحولات (transformers) ومن مجموعات بيانات عربية مُولّدة ومُختارة بعناية، نطوّر تقنيات كشف قابلة للتوسع تأخذ في الحسبان ثراء اللغة العربية. وتُظهر نتائجنا التجريبية دقةً عالية في التمييز بين النصوص المكتوبة بواسطة البشر وتلك المُنتَجة آليًا، مما يسهم في الاستخدام الأخلاقي للذكاء الاصطناعي التوليدي في السياقات متعددة اللغات، ويُعزز سُبل التصدي للمعلومات المضللة الناتجة عن الذكاء الاصطناعي.

الكلمات المفتاحية: التعلم الآلي، التعلم الانتقالي، الذكاء الاصطناعي التوليدي، نماذج اللغة الكبيرة، النص المزيف، النص المحرر بواسطة إنسان، اكتشاف النص المحرر بواسطة الذكاء الاصطناعي.

Dedication

To my mother and my loving father.

Thanks for always being there for me.

Acknowledgments

I extend my deepest gratitude to my supervisor, Prof. Fouzi Harrag, whose unwavering guidance, intellectual rigor, and steadfast encouragement were instrumental in shaping this thesis. Your incisive feedback helped a lot in this academic contribution.

I am equally indebted to my co-supervisor, Prof. Khaled Shaalan, for their invaluable insights and unwavering support throughout this journey. Your specialized knowledge enriched the methodological framework of this work, and helped refine my approach to complex problems. Your accessibility, even amidst demanding schedules, underscored your commitment to my success and deepened my appreciation for collaborative scholarship.

Together, your complementary perspectives created an environment where curiosity thrived and innovation was nurtured. This thesis stands as a testament to your collective dedication to advancing knowledge and empowering the next generation of scholars.

Amal Boutadjine

boutadjine.amal@univ-setif.dz

Contents

List of Tables

List of Figures

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
GPT	Generative Pre-trained Transformer
LLMs	Large Language Models
ASV	Automatic Speaker Verification
GAN	Generative Adversarial Networks
NLG	Natural Language Generation
NLU	Natural Language Understanding
CFGs	context-free grammars
RNN	Recurrent Neural Network
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
RL	Reinforcement Learning
VAE	Variational AutoEncoder
BOW	Bag-Of-Words
PDF	Probability Density Function

GLTR	Giant Language model Test Room
DL	Deep Learning
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
GRU	Gated Recurrent Unit
GCN	Graph Convolutional Network
TCN	Temporal Convolution Network
DFTD	DeepFake Text Detector
RAG	Retrieval-Augmented Generation
GloVe	Global vectors for word representation
BERT	Bidirectional Encoder Representations from Transformers
AraBERT	Arabic Bidirectional Encoder Representations from Transformers
CAMeLBERT	Computational Approaches to Modeling Language (CAMeL) BERT
RBC	Retrieval-Based Classification
ChatGPT	Chat Generative Pre-trained Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
AR-ASAG	ARabic dataset for Automatic Short Answer Grading
MSA	Modern Standard Arabic
FEDM	Fake Essays Detection Model

seq2seq	sequence-to-sequence
FAISS	Facebook AI Similarity Search
k-NN	k-Nearest Neighbours
Trans-Detect	Transition Detector

CHAPTER I

Introduction

Misinformation, often spread rapidly and widely through social media, has emerged as a critical issue with far-reaching consequences [1, 148]. Instances of misinformation have been linked to public health crises [2, 3], political instability [4, 5], and social unrest [7]. For example, during the COVID-19 pandemic, the spread of false information about the virus and its treatment led to confusion, mistrust in health authorities, and, in some cases, harmful health practices [6]. Similarly, misinformation has been shown to influence election outcomes, as seen in various political campaigns where false narratives were disseminated to sway public opinion [8, 149].

The advent of advanced artificial intelligence (AI) and machine learning (ML) technologies has further complicated the issue of misinformation [9]. Generative models, such as those based on deep learning architectures, have achieved remarkable success in producing synthetic content that is increasingly difficult to distinguish from authentic material [10, 11, 12]. These models can generate realistic images, videos, audio recordings, and text, collectively referred to as "deepfakes." Among these, AI-generated text poses a unique and significant challenge [13], particularly in languages with complex structures and diverse morphology, such as Arabic.

The ability of AI models to generate coherent and contextually relevant text has raised concerns about the potential for misuse [159, 160]. In the Arab world, the implications are particularly profound. The Arabic language, with its complex script and rich morphology, presents unique challenges for natural language processing (NLP) tasks [14, 150]. The script's cursive nature, the presence of multiple forms for each letter depending on their position in a word, and the language's use of diacritics all contribute to the complexity of processing Arabic text.

Undetected AI-generated Arabic text has serious repercussions. In the realm of news media, synthetic text can be used to spread false information, manipulate public opinion, and undermine the credibility of genuine news sources. In education, the availability of AI-generated essays and academic content poses a threat to the integrity of educational institutions, as it can facilitate plagiarism and academic dishonesty [151]. On social media, AI-generated text can be employed to spread misinformation, fuel polarization, and incite social unrest.

Addressing these challenges requires the development of robust detection mechanisms capable of identifying AI-generated text with high accuracy. However, the task is not

straightforward. AI models, particularly those based on large language models like GPT (Generative Pre-trained Transformer), are designed to mimic human writing styles, making it increasingly difficult to distinguish between human-generated and machine-generated text [152]. The lack of discernible patterns or "fingerprints" in AI-generated text adds to the complexity of detection efforts.

This thesis aims to contribute to the ongoing efforts to combat AI-generated, or "deepfake text", focusing on the Arabic language. The research develops and evaluates models tailored to the unique characteristics of Arabic text. The work builds upon existing research in NLP and machine learning, exploring both detection and attribution tasks to effectively identify and differentiate AI-generated content from human-written text.

The contributions of this thesis are expected to have broader implications beyond the Arabic language. The methodologies and models developed can potentially be adapted for use in other languages, particularly those with similar linguistic complexities. Moreover, the research underscores the importance of maintaining the integrity of information in a digital society, where the proliferation of AI-generated content poses a threat to trust and reliability.

In summary, this thesis addresses a critical and timely issue in the field of NLP, with the potential to enhance the detection of AI-generated text in Arabic and beyond. The research not only advances the technical capabilities of NLP but also contributes to the broader goal of ensuring the authenticity and trustworthiness of digital content in an era where misinformation can have profound societal impacts.

1.1. Problem Statement

The problem addressed in this thesis is the detection of AI-generated Arabic text, which poses significant challenges due to the language's characteristics and the sophistication of AI models. Current detection techniques often fall short in accurately identifying synthetic text, especially in languages like Arabic, where cultural and linguistic nuances can complicate analysis. In order to address these limitations, the present research proposes novel approaches tailored to the specific case of the Arabic language.

Deepfake text detection is a challenging problem from NLP and AI perspectives. Similar to the concept of detecting deepfake videos in the computer vision domain, this problem revolves around the task of discerning between authentic textual content and artificially generated text created by advanced large language models. Given the increased skill of language models at producing grammatically correct and contextually coherent passages, the main difficulty lies in creating effective approaches to discriminate between authentic and synthesized textual content.

The problem can be framed as a classification task (with two distinguished classes: real text and deepfake text). Each input text sample x_i is assigned a label $y_i \in \{0,1\}$, with 0 denoting authentic text and 1 representing deepfake text. The goal is to train the detector

to learn a function: $f: x_i \mapsto y_i$ that generalises well to previously unseen text instances, while being resilient to the evolving sophistication of language models utilised in generating synthetic content. In this thesis, the text authored by a human is referred to as "real text".

1.2. Objectives and Research Questions

Aim & Objectives:

The ultimate goal is to develop more sophisticated methods for distinguishing AI-text from human produced text.
The objectives of this research are to:

1. Develop and evaluate detection models specifically for AI-generated Arabic text in the media and the educational environments.
2. Compare the performance of these models with human detection capabilities.
3. Explore methods for accurately attributing text segments to their true authors, whether human or AI.

Research Questions :

In particular, this dissertation will examine three main research questions:

RQ1: What are effective model-based approaches for detecting AI-generated Arabic text in media and academic domains?

RQ2: To what extent can human evaluators reliably detect deepfake text compared to automated detection systems?

RQ3: What techniques can be employed to attribute authorship of text segments accurately, distinguishing between human and AI origins in mixed text?

1.3. Main Contributions

This thesis contributes to the field of natural language processing (NLP) by:

- **Development of Arabic-specific Deepfake Text Detection Models:** This thesis introduces novel transformer-based models specifically optimized for detecting AI-generated Arabic text. These models address the morphological richness and syntactic specifics of the Arabic language—challenges that are often overlooked in existing detection systems predominantly designed for English.
- **Empirical Evaluation of Human Versus Machine Detection Capabilities:** A systematic comparative study is conducted to assess the effectiveness and

reliability of human evaluators in detecting deepfake text, benchmarked against state-of-the-art machine learning models. This evaluation provides critical insights into the limitations of human judgment in the context of AI-authored content.

- **Segment-Level Authorship Attribution Techniques:** The thesis proposes and validates new methodologies for attributing authorship at the segment level, distinguishing between human- and AI-generated portions within a single document. This granular approach enhances forensic linguistic capabilities and supports the development of explainable detection systems.
- **Curated Arabic Corpora for AI-Generated Text Research:** Purpose-built datasets comprising both human-authored and machine-generated Arabic texts are compiled, annotated, and made available for research use. These corpora fill a significant resource gap in the field and provide a foundation for future investigations into Arabic-language generative AI detection.

1.4. Thesis Overview

This thesis is structured to systematically address the challenge of detecting AI-generated ("deepfake") text in the Arabic language, progressing from foundational concepts to advanced methodologies and broader implications. The manuscript is organized into seven chapters, each building on the preceding ones to provide a cohesive exploration of the research problem. Below is an overview of the chapters and their contributions:

This current chapter established the context, motivation, and significance of the research. It outlined the general risks posed by AI-generated text in Arabic, particularly in domains such as news media, education, and social platforms. The chapter articulated the problem statement, research objectives, and key contributions, concluding with a roadmap of the thesis structure.

Chapter 2 provides the theoretical and contextual foundation for understanding deepfake text generation and detection. It introduces key concepts, including types of deepfakes, generative AI technologies for producing deepfake text, ethical implications, and a preliminary overview of deepfake text detection. The chapter synthesizes interdisciplinary perspectives to contextualize the technical and societal challenges of deepfake text.

The third chapter critically evaluates existing research on AI-generated text detection. It identifies gaps in the literature, particularly the lack of Arabic-specific detection frameworks. By situating the thesis within broader academic discourse, this chapter justifies the novelty and necessity of the proposed approaches.

Focused on the news media domain, chapter 4 presents a set of fine-tuned models for detecting AI-generated Arabic news content. It compares the model’s performance with human evaluators, highlighting the limitations of unaided human judgment and the importance of automated systems tailored to Arabic’s linguistic nuances.

Transitioning to the educational sector, the fifth chapter evaluates two methodologies for detecting AI-generated student essays: fine-tuned LLMs and retrieval-based classifier. It explores the trade-offs between accuracy and interpretability, proposing scalable solutions for maintaining academic integrity in Arabic-speaking institutions.

Addressing hybrid texts where authorship shifts between human and AI, chapter 6 introduces a segment-level detection framework. Using stylometric features, machine learning, deep learning, and transformer-based embeddings, the methodology identifies intra-textual inconsistencies, advancing the field beyond binary classification.

The final chapter synthesizes the thesis's contributions emphasizing its technical advancements, societal implications, and methodological innovations. It reflects on the broader impact of the research and explores emerging challenges and opportunities in deepfake text detection, and outlines avenues for future work in Arabic NLP and AI ethics.

Each chapter of the thesis logically builds upon the previous, offering a detailed examination of the research problem.

1.5. Publications

- A comprehensive study on multimedia DeepFakes [10]
- Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content [15]
- Detecting Human-to-AI Author Change in Arabic Text [52]

CHAPTER II

Background

Building on the foundational problem statement outlined in Chapter 1, which highlighted the urgent need to address deepfake text in Arabic, this chapter establishes the theoretical and contextual groundwork for the thesis. The previous chapter underscored the risks posed by deepfake text but did not delve into the technical and conceptual frameworks necessary to understand its generation, detection, and broader implications. To address this gap, this chapter synthesizes interdisciplinary perspectives to contextualize the problem within the broader landscape of generative AI and ethics.

The aim of this chapter is to provide a comprehensive overview of deepfake technologies, with a focus on text generation. By examining the evolution of generative models and their societal impacts, the reader gains critical insights into the field of deepfake text. This chapter advances the thesis by establishing a conceptual framework that will underpin the technical methodologies and analyses in subsequent chapters.

To achieve this aim, the current chapter begins by introducing deepfake technologies, distinguishing between text, image, audio, and video synthesis, and presenting the overall threats posed by these technologies. It then narrows its focus to text generation, exploring the role of large language models (LLMs) and their ethical implications.

2.1. Introduction to Deepfakes

Based on graphical methods or visual effects using computers, fake media were traditionally created only by experts. However, the recent advancements of deep learning models, namely autoencoders and generative adversarial networks, and their availability has facilitated the wide spread of deepfakes and their circulation especially on the social network.

The term deepfakes (stemmed from the two words « Deep Learning » and « fake ») refers to the stunning phenomenon that appeared in 2017 led by the blooming of AI-powered models and applications, which allowed the synthesis of high realistic digital media (images, audios, videos) and text, that can trick human to believe it is real. The name of this impressive phenomenon, “deepfake”, was first used by a Reddit user, who shared the first deepfakes by replacing celebrities’ faces into adult content clips [20].

While deepfakes started with the field of computer vision, it quickly emerged to include voice clips and even natural language text.

Deepfake Images

As a result of recent advances of deep learning and high-realistic image generation capabilities, permitted by means of variants of GANs (for instance: StarGAN [21], STGAN [22], StyleGAN [23] and MFF-GAN [24]), high-fidelity images are being created, with either realistic looking faces that does not exist in the real world; or identity falsification through face swap of the (source) face of someone in a photo with the (target) face of someone else.

Deepfakes technologies like SimSwap [25] and FaceSwap [26, 27] have demonstrated visualization algorithms' and computer graphics' great ability to manipulate pictures of people by replacing their face with the face of a different person.

According to [28] "There is likely no more controversial application of generative modelling than creating fake faces or applying the technology to swap faces" although [22] manifest that FaceSwap has ethical uses.

Deepfake Audio

While originally demonstrated for image synthesis, GANs have since shown super capabilities for other data types such as audio [29] [30], allowing to create highly realistic fake voice records.

Audio deepfakes, also known as voice spoofing, are fake voice clips that were automatically created by computer, either using machine learning approach (e.g. the Deep Voice method) or signal processing techniques (e.g. the Imitation method). These audio clips are human-sounding that they became a main threat to the automatic speaker verification (ASV) systems [31].

In mid-2019, Audio deepfakes gained some attention when cybersecurity firm Symantec reported some incidents where audio deepfakes were used to trick company finance officials to transfer large sums of money into the crooks' accounts [32]. A serious threat of audio deepfake is the simplification and ease of the communication-based crimes that have been difficult to commit over time. While technology is still improving, these types of crimes will continue.

Deepfake Videos

Over the last few years, the new technique of generating manipulated videos using deep learning models, and known as "deepfake", has taken hold on the social nets. This technology gives users the hand to alter videos content by replacing the face of a person in a clip with the face of a second person given a large amount of images.

Deepfake videos are manipulated videos that simulate the likeness of an individual through algorithmic-synthetic video of real people (mainly public figures). Where the individual appears in the video doing or saying things that they never did or say in reality [16]

In fact, deepfake videos gained large notoriety in the media very quickly as result of their applications, targeting famous actresses and politicians who were 'deepfaked' into adult content videos and shared on the internet.

The main points of distinction between deepfakes and other video manipulation techniques are: the possibility for obtaining realistic and very convincing results. Secondly, the availability of the technique to users with limited knowledge of programming and machine learning to create deepfakes [33].

Several techniques exist for creating deepfakes, the most used method is based on deep neural networks involving autoencoders using a face-swapping algorithm. Where a target video and a large collection of videos containing the person you want to insert in the target video; are presented as input to the neural net. Another widely used type of neural nets is Generative Adversarial Networks (GANs), which learn in an adversary way from large amounts of data to detect and improve any flaws in the generated deepfake, outputting highly accurate results and making the discrimination harder for deepfake detectors.

Deepfake Text

Deepfake text is text generated by AI (deep learning models) trained on large corpora of human language from multiple textual resources, and outputs highly realistic text indistinguishable from human-written language. Deepfake text can mimic specific individual style of writing, produce well written articles, documents and even poetry [34]. Public interest and discussion has been attracted by powerful and freely available natural language generation models due to their huge ability to produce high quality multi domains text. Though, no experimental evidence is performed to examine the people's capability to differentiate artificial text from human written one in the Arabic language at the time of conducting this work.

2.2. The Threat Posed by Deepfakes

General Threats

Studies such as [44] has shown that people are not able of reliably detect deepfakes, and despite raising awareness and introducing financial incentives, the results has not improve their detection accuracy.

A major risk of deepfakes is privacy threat. Besides, the darkest application on celebrity's face placed without their consent in extremely unethical situations [28]. Whereas, there are potential ethical flaws, even when deepfakes are used with active consent and these traps have been subject to ongoing philosophical evaluation [16].

As deepfake technologies are becoming more proficient in producing artificial manipulated visual content on one hand and social media continue to occupy more space in our lives in the other hand, the task of distinguishing authentic from fabricated content grows progressively more challenging [18]. This convergence has led to a phenomenon wherein AI-generated synthetic content increasingly blurs the boundaries between virtual and physical realities [16]. The contemporary digital landscape, characterized by its visual saturation, frequently serves as a conduit for hostile ideologies and extremist movements [18] that exploit deepfake technology to advance their objectives and disseminate propaganda. Deepfakes also pose significant threats to national interests and security through the weaponization of synthetic media [36].

Specific Threats of Deepfake Text

Fake text generated by language models also can have unfortunate and malicious uses, such as fake product reviews generation [54], fake news generation [43, 50, 53], spamming/phishing and generating misinformation, even if utilised by regular unskilled

adversaries. Besides, artefacts and synthetic data significantly decrease the trustworthiness social networks content. Although, users still appear vulnerable when encountering deepfakes [65].

Impact on News Media

The advent of deepfake text poses a significant threat to the integrity of news media. AI-generated text can be used to create convincing fake news articles that are difficult to distinguish from genuine content, thereby undermining public trust in journalism. According to the study [68], the proliferation of AI-generated misinformation can lead to information disorder, where the public is exposed to a mix of accurate and fabricated news, making it challenging to discern truth from fiction. This has the potential to erode the credibility of news outlets and exacerbate societal polarization. For instance, a case study by Singer and Cole (2019) [69] highlighted how AI-generated text was used to spread false information during political campaigns, demonstrating the real-world impact of this technology on democratic processes.

Impact on Education

In the educational sector, deepfake text presents a formidable challenge to academic integrity. Students may exploit AI-generated text to produce fraudulent academic papers or essays, leading to issues with plagiarism and the devaluation of genuine academic work. A recent study by Holmes and Bialik [67] explored the potential misuse of AI tools like ChatGPT in academic settings, revealing that existing plagiarism detection methods often fail to identify AI-generated texts effectively. This raises concerns about the authenticity of student submissions and the need for updated detection mechanisms to maintain educational standards.

Impact on Social Media

The impact of deepfake text on social media is profound, as it can be disseminated rapidly, influencing public opinion and potentially causing social unrest. Bots equipped with AI-generated text can spread false information, manipulate public sentiment, and affect everything from election outcomes to public health perceptions. A study by Benkler et al. [66] on computational propaganda highlighted the role of bots in spreading misinformation, emphasizing the need for robust strategies to counteract their influence. For example, during the COVID-19 pandemic, AI-generated text was used to spread vaccine misinformation, highlighting the potential health implications of unchecked deepfake content on social media platforms.

2.3. Deepfake Text Generation Technologies

Natural Language Generation

In general terms, Natural Language Generation (NLG) and Natural Language Understanding (NLU) are sub-categories of the more general domain Natural
Language Processing (NLP) that encompasses all systems interpreting or producing spoken or written natural human language.

• **NLU** uses human language as input and turns the unstructured data into a structured data in a representative form that can be understandable to computers. Some applications of NLU are: named entity recognition, sentiment analysis, relation extraction and semantic parsing.
• **NLG**, which refers to the process of producing natural language from computer-internal semantic representations by automatically transforming structured data into human-readable text. Several systems of language technology uses NLG as a key component, namely: question answering systems (IBM Watson, BERT), auto-reply for emails (SmartReply from Google), dialog & assistant systems (Alexa from Amazon), text summarization, story generation and others.
Text generation technology, like most NLP tasks, has taken a huge leap forward over the past few decades. Particularly, the overall quality of generated text is further improved by shifting from manual feature-extraction, rule-based and statistical methods to neural-network-based models.
The techniques and models for generating deepfake text have evolved rapidly over the last few years, with the advancement of transformers and NLP. Some of the main techniques and models are:

Rule-Based Approaches: These methods are based on predefined grammatical rules to generate synthetic text. Rule-based techniques can be represented by formal grammars, like regular expressions or context-free grammars ($CFGs$).

Statistical Methods: These techniques analyse existing human-written texts to learn statistical patterns. This category includes markov models, N-grams, and hidden Markov models. N-grams are often used to model the probabilistic of word sequences.

Recurrent Neural Networks (RNNs): A neural network that uses a hidden state to store the previous information in order to process sequential data, including text. RNNs can generate text samples by predicting the next word or character based on the previous words or characters. RNNs can generate coherent and fluent text samples, but they may suffer from problems such as vanishing or exploding gradients, repetition, or inconsistency.
An RNN computes hidden states recursively using the following equations:

$$h_t = f\,(W_{hh}h_{t-1} + W_{xh}x_t + b) \quad (1)$$

$$y_t = g(W_{hy}h_t + c) \quad (2)$$

Where h_t is the hidden state at time t, x_t is the input, b and c represents the bias vectors, y_t is the output, and g is an activation function that transforms the output.

Long Short-Term Memory (LSTM) Networks: An RNN type developed to address the vanishing gradient issue. These networks excel at capturing long-range dependencies thanks to their architecture. Unlike simple RNNs, which have a single hidden state, LSTMs maintain two states:

1. *Hidden state h:* Represents the memory or information stored in the LSTM cell. It captures context from previous time steps. It can be computed using the following equation:

$$h_t = o_t \odot tanh(c_t) \quad (3)$$

Where h_t is the hidden state at time t, o_t is the output gate, c_t is the Cell state, $\odot$ is the element-wise multiplication, and $tanh$ is the hyperbolic tangent activation function.

2. Cell state c: Represents the long-term memory which is updated using gates (input, forget, and output gates). The cell state, which is a combination of new information and the previous cell state, is modified by the gates.

3. Gates:

Input Gate i: Controls how much new information is added to the cell state.

Forget Gate f: Determines what information should be discarded from the cell state.

Output Gate o: Regulates how much of the cell state contributes to the hidden state.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

Where σ is the sigmoid activation function, $W's$ and b's are learnable weights and biases, x_t is the input vector at time t, h_{t-1} is the previous hidden state, and $U's$ are the weights for the previous hidden state h_{t-1}.

3. *Cell Update:* Combines the input gate, forget gate, and new candidate cell state $\tilde{c}_t$:

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (8)$$

5. Output: The hidden state modified by the output gate:

$$h_t = o_t \odot tanh(c_t) \quad (9)$$

Generative Adversarial Networks (GANs): A neural network based on two components: a generator and a discriminator. The generator tries to produce fake text samples that are indistinguishable from real text samples, while the discriminator tries to distinguish between real and fake text samples. The generator and the discriminator are trained in an adversarial manner, where the generator tries to fool the discriminator, and the discriminator tries to catch the generator. GANs can generate realistic and diverse text samples, but they also face some challenges, such as mode collapse, instability, and evaluation.

GANs have been implemented in various NLP applications including natural language generation [59, 60, 61, 62, 63], among others. The developed GAN-based models can be categorized according to the method they adopt in handling the discrete nature of text: Latent space based techniques, Reinforcement learning (RL) based methods, and continuous approximation of discrete sampling based approaches.

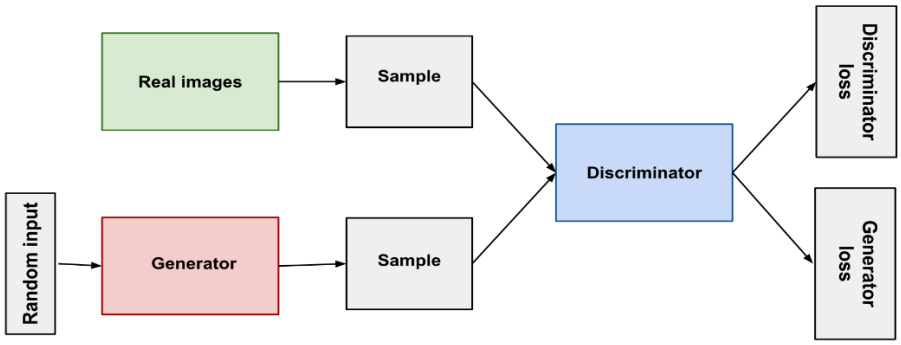


Figure 1. GANs architecture design[1]

Variational AutoEncoder (VAE)

Inspired by a computer vision work [35]; researchers in [161] proposed an innovative approach called Variational AutoEncoder (VAE) to takle natural text modelling. Like generative models, VAE estimates the Probability Density Function (PDF) of the training dataset samples. Variational autoencoders undertake autoencoders by their ability to avoid overfitting and to ensure good properties of the latent space that enable generative process. The encoder learns a conditional mean and standard deviation, which are the responsible for constructing the latent variables distribution. Using these conditional means and standard deviations as frame of reference, the decoder's job is to reconstruct the original input accurately from the developed latent variable distribution.

VAE is a probabilistic based approach for describing data samples x in the latent representation space R. Instead of learning a simple mapping of data distribution onto a fixed vector, the vital goal of variational auto-encoder (VAE) is to develop the explicit probabilistic distribution that can explain most of the input dataset samples. Each latent variable in the latent code R is described by a probability distribution. The architecture design of the model is shown in Figure 2 below.

[1] https://developers.google.com/machine-learning/gan/gan_structure

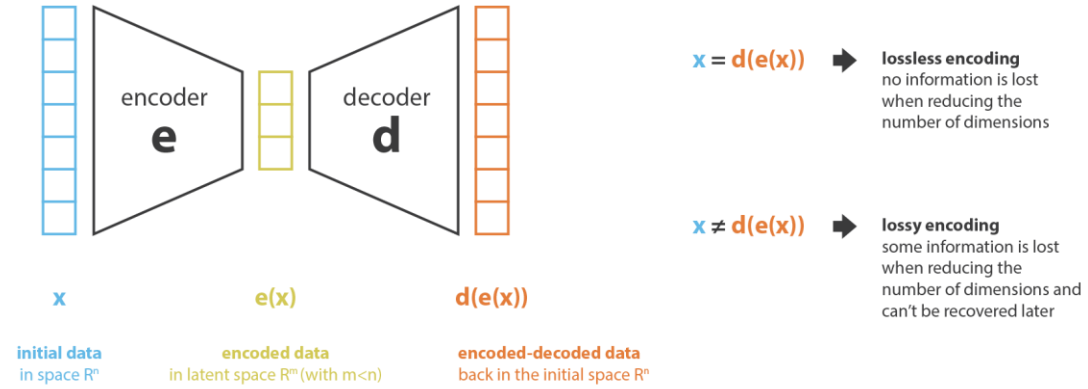


Figure 2. Variational auto-encoder architecture design[2]

Transformers

The Transformer architecture proposed by Vaswani et al. [57] relies on self-attention mechanisms instead of recurrent and convolutional components designed for processing sequential data. The attention mechanism emphasise the importance of all relevant tokens from the input text, in the encoding of a given input token. For instance, in a machine translation task, the attention mechanism allows the transformer to translate words like 'the' into a word of the correct gender in Spanish or French by attending to all the other words in the original sentence. Typically, transformers was a pivotal leap facilitating efficient parallel training, capturing long-range sequence features and showing better computations in both performance and training time. Therefore, this architecture has been widely implemented in several variants, and also in building big pre-trained models using one or both parts that a transformer is made up of (encoders and decoders). Instances of transformer models include: BERT [146], GROVER [43], GPT-2 [114], GPT-3 [50], XlNet, T5 and others. These language models, which has been trained on large amounts of textual data, can be tuned with texts for specific downstream tasks, including text classification. The architecture is used as the main building block to develop increasingly complex extension.

[2] https://www.cnblogs.com/wangxiaocvpr/p/11605989.html

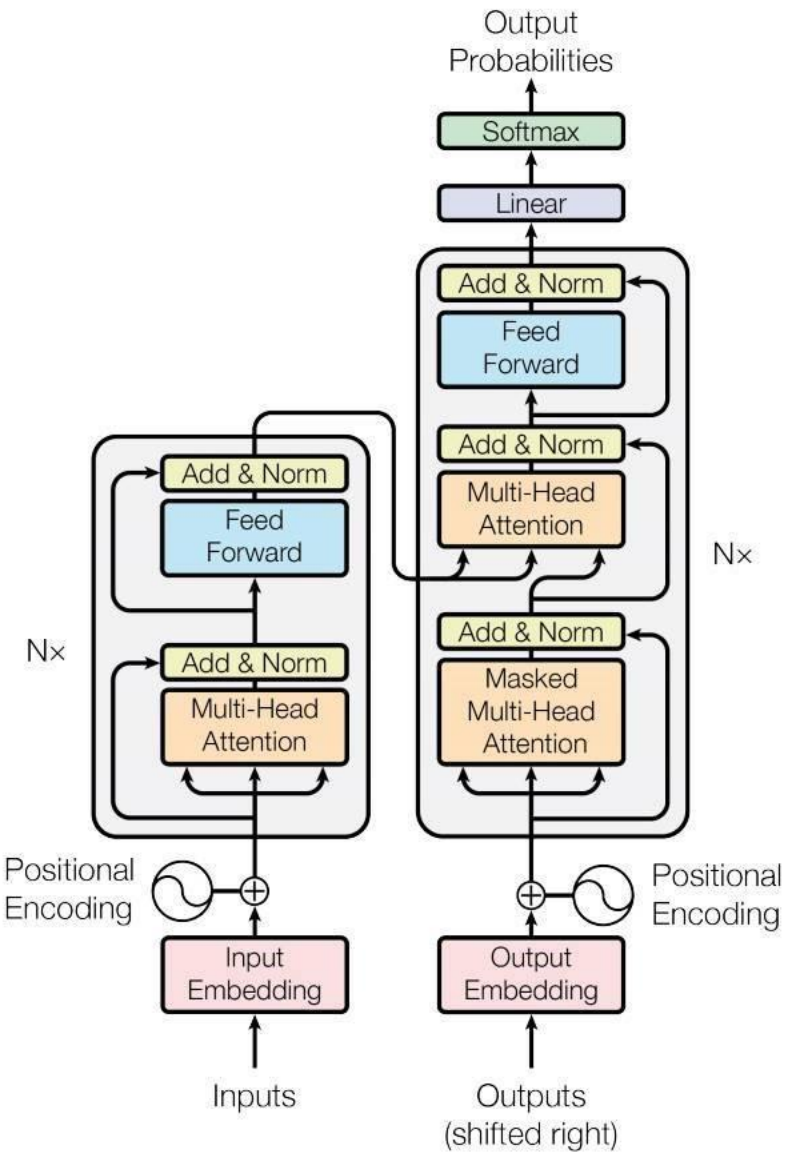


Figure 3. Transformer architecture design [57]

As the quality of machine-generated texts has improved significantly, these advances bring with them a wide range of practical problems and challenges that need to be addressed. It is possible today to use Transformers for abusive purposes. The ubiquitous use of these powerful big language models has provided an incentives to develop stronger machine-generated text detection systems.

LLMs

LLMs are sophisticated AI models trained on vast amounts of text data, enabling them to generate human-like text across various domains. Notable examples include OpenAI's GPT-3 and GPT-4, which have demonstrated remarkable prowess in language tasks [50, 70]. These models leverage deep learning techniques, specifically transformer architectures, to understand and generate text based on input prompts.

LLMs play a crucial role in the generation of deepfake text, where they can produce convincing narratives, articles, or even entire documents that mimic human writing styles. This capability has raised concerns about the potential misuse of these technologies in spreading misinformation [71].

The training process of LLMs involves exposing the model to extensive corpora of text data, allowing it to learn patterns and structures in language. During text generation, the model predicts the next word in a sequence based on the context provided, a process that can be fine-tuned for specific tasks [57]. This mechanism contributes to the effectiveness of LLMs in creating deepfake text.

While LLMs offer beneficial applications in content creation, customer service, and education, their potential for misuse is a growing concern [46]. For instance, they can be employed to generate fake news, impersonate individuals, or spread disinformation, posing significant risks to societal trust and security [71].

The ethical implications of LLMs in deepfake text generation are profound. Scholars have emphasized the need for regulations and ethical guidelines to prevent the misuse of these technologies [71]. Potential solutions include transparency requirements, user education, and the development of detection technologies to ensure accountability.

3.1. Ethical and Social Implications

In both good and harmful ways, deepfake text effects individuals and society in general.

Powerful developed language models are applied to generate text that approximately matches the style of human language, which can be used in several beneficial applications including code auto-completion [49], story generation [55, 64] and conversational response generation [56].

In the other hand, generative language models can capture biases present in their training data, amplifying extreme, racist, or overly generalized beliefs about a particular group of people, gender or religion that can be present in the training data and fed to the model. As demonstrated by Solaiman et al. [49] and Brown et al. [50], language generative models can generate text that favours males over females, white over black people, and Christians over Muslims reflecting social amplified biases in gender, race, and religion. These biases caused by language models pose a serious concern and threat to several groups of people [51] in varying degrees.

Since biases can be typically analysed just in the context of a particular use cases it is very difficult to mitigate it especially in large language models [155]. Some researchers argue that these detrimental biases should be encountered through a variety of steps (mitigation of a priori biases by purification of training data, reinforcement learning, training of a second model as a filter for the generated content of the first model of language, use trusted human partner’s feedback) [156], but not only big companies such as google and OpenAI

alone have the appropriate status and should not decide on behalf of society since the impact of the models goes beyond their boundaries to include society as a whole [157].

3.2. Deepfake Text Detection

Automated Detection of Deepfake Text

Deceptive automated text generated to impersonate humans have been successfully exploited for various kinds of abuse. Researchers have responded by developing AI based detectors to face this threat and fight it.

Recently, both natural language processing (NLP) and machine-learning (ML) communities have done significant work to build accurate detectors especially for English. In that, [39] proposed to automate the process of scoring and labelling the text samples rather than relying on human participants, by defining a differentiating procedure based on large pre-trained language models with their probability distributions.

Many of recently proposed text detection systems are based on pre-trained language models, namely: Grover of Zellers et al [43], GLTR (Giant Language model Test Room) of Gehrmann et al. [42] and others. More details about auto-detection of deepfake text will be discussed in Chapter II devoted to related work.

Human Detection of Deepfake

Human ability to measure how close auto-generated text is to human natural language is a critical issue. Authors in [38] assessed through their study the ability of non-experts to differentiate between human and machine-authored text (generated using GPT-2 and GPT-3) across three domains, where they found that evaluators' distinction between GPT-3 and human text was near random guess. To better spot GPT3-authored text, authors trained evaluators for the task although that did not lead to significant improvements in the three domains. This study [45] also showed that human annotators are easily tricked by generative language models.

Collaborative Human-machine Detection

Common sense knowledge and Human interpretation skills can be useful for building an automated classification system. Studies such as [42] and [48] takes advantage of both human and machine detection of auto-generated text. GLTR [42] is a tool that may help humans in the classification task providing a comprehensive visualization of text properties. This tool can facilitate untrained humans to locate synthetic text with an acceptable accuracy. Although, GLTR system performs well in identifying machine generated text, it is low confident when identifying human text or determining that the text is not machine generated. Therefore, a human-machine collaboration can improve the detection task [49]. Another tool, the RoFT, website designed by (Dugan et al., 2020)[48] invites human users to participate in defining a phrase boundary at which the text transitions from human written text to machine generated text.

3.3. Evaluation Methods

Different machine learning algorithms can be judged by their performance; therefore, evaluating classifiers is an essential part. The most common measures of a binary classifier's performance are accuracy, precision, recall, and F1-score. Although the model may perform satisfactorily when evaluated with a metric, and yet it performs poorly when evaluated against another metric.
For the calculation of the mentioned metrics, a confusion matrix is calculated, it contains the predicted and actual distribution of labels as shown below:

Table 1. Confusion matrix

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

True Positives (TP): text that is actually positive and estimated as positive.
True Negatives (TN): text that is actually negative and estimated as negative.
False Positives (FP): text that is actually negative and estimated as positive.
False Negatives (FN): text that is actually positive and estimated as negative.

For text generative models, an evaluation approach that is gaining rising popularity recently compares method outputs against human-authored references of a standard corpus using automatic metrics. This popular metrics used for text generative models evaluation include BLEU [40], ROUGE [41] and others.

Accuracy
The simplest metric used to evaluate a classifier's performance by measuring how often the algorithm classifies a data point correctly. It is the percentage of correctly identified data points out of the total data points (all observations).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

Precision

Precision is the attempt to measure how much we predict correctly. In other words, it is the proportion of positive identifications that was actually correct.

$$precision = \frac{TP}{TP + FP} \tag{11}$$

Recall

Recall or sensitivity evaluate a classifier by measuring the proportion of actual positives that was identified correctly.

$$recall = \frac{TP}{TP + FN} \tag{12}$$

F1-Score

F1-score combines precision and recall, it can be defined as the harmonic mean between recall and precision, or the weighted average of precision and recall in other words. This measurement takes into account both false positives and false negatives. The formula of calculation is the following:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{13}$$

3.4. Chapter Summary

This chapter has established the foundational knowledge necessary to contextualize the challenges of detecting deepfake Arabic text. By delineating the technical mechanisms of deepfake text generation—particularly through large language models like ChatGPT and Gemini—it has clarified the sophistication of modern generative technologies and their potential for misuse.

Furthermore, the chapter has synthesized interdisciplinary insights, demonstrating how the convergence of AI advancements exacerbates the risks of misinformation. The chapter also bridges theoretical concepts with practical social and ethical implications.

The contribution of this chapter is providing a rigorous conceptual framework for understanding deepfake text generation and detection. These insights set the stage for the subsequent chapters, which will build on this foundation to propose novel detection methodologies tailored to Arabic text. Having established the "why" and "what" of the problem, the thesis now transitions to the "how," beginning with a review of existing literature in Chapter 3.

CHAPTER III

Related Work

Following the contextual foundation laid in Chapter 2, which outlined the technological and ethical dimensions of deepfake text, this chapter synthesizes and analyses existing scholarly work on AI-generated text detection. While the earlier chapters established the urgency of addressing Arabic deepfake text, they did not systematically evaluate how prior research has approached similar challenges in other languages or domains. This chapter fills that gap by analysing advancements, limitations, and unresolved questions in the field, thereby positioning the thesis within the broader academic discourse and clarifying its novel contributions.

The aim of this chapter is: (1) to review the related work in Arabic NLP and the overall existing directions of AI-generated text detection, (2) to review state-of-the-art methods for detecting AI-generated text across languages, and (3) to identify gaps in the existing approaches. Readers will gain a nuanced understanding of how detection techniques—from statistical stylometry to transformer-based models—have evolved. By critically evaluating the literature, this chapter underscores the need for specific solutions for Arabic deepfake text, which the thesis addresses in its empirical studies.

To achieve this aim, the chapter is organized thematically. First, it examines the major approaches used for AI-generated text detection in high-resource languages. Next, it explores the datasets used and their characteristics. A dedicated section then highlights the strengths and critiques the limitations of existing studies.

Deepfake text detection is a challenging and important problem [72] for several reasons. Initially, deepfake text can be used for malicious purposes, such as disseminating false information, propaganda, or enabling cyberattacks. Second, because of the high linguistic quality and diversity of the generated text, deepfake text can be difficult to distinguish from human-written text, even for human specialists. Third, deepfake text can be adaptive and evasive, as the generative models can be fine-tuned or perturbed to bypass the detection methods [153, 154]. Therefore, developing effective and robust methods for deepfake text detection is essential for preserving the trustworthiness and integrity of online information [158]. Figure 4 summarizes the various types of auto-generated text that were detected in the included research works.

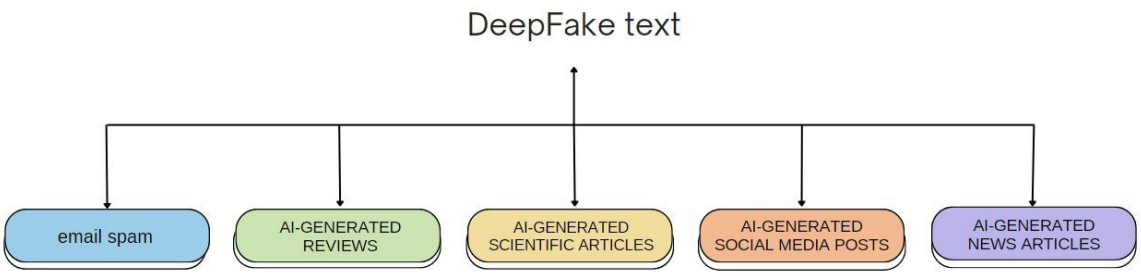


Figure 4. Forms of deepfake text types in the included studies

Deepfake text and fake news differ in scope, but they share characteristics: both can be misleading and manipulative, rely on technology for production and distribution, and can disseminate misinformation. False or misleading material that is presented as news in articles, headlines, or social media posts that is shared with the intention of misleading or deceiving readers through a variety of channels is referred to as fake news. Deepfake text, on the other hand, focuses on text that has been altered or manufactured through the use of artificial intelligence methods like deep learning and natural language processing. These methods create fictitious statements, quotes, and full articles by imitating the writing style of a real person. Figure 5 gives the relationship among the concepts of deepfake text, fake news, disinformation and misinformation.

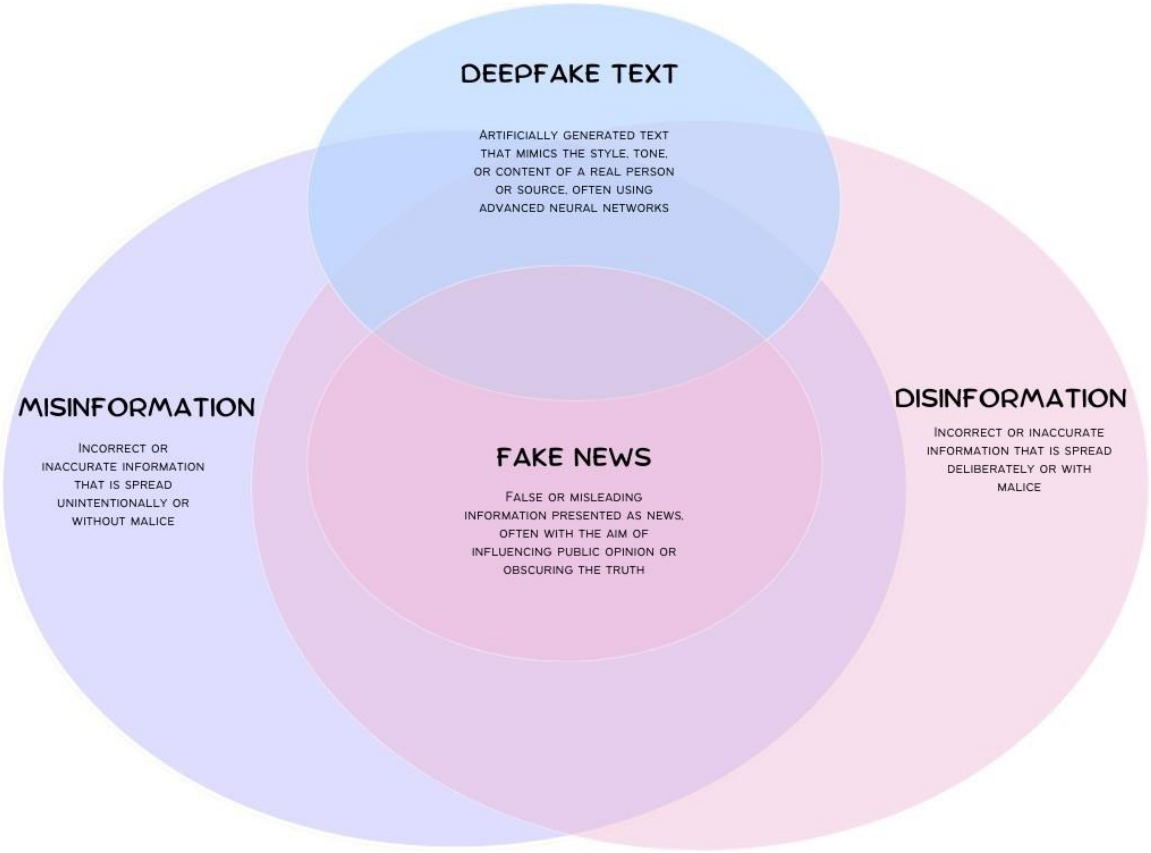


Figure 5. Relationship between deepfake text, fake news, disinformation and misinformation

3.1. Arabic Natural Language Processing

The Arabic language introduces a unique set of difficulties for researchers in Natural Language Processing (NLP). Arabic language suffers from common computational challenges such as lexical ambiguity, polysemy, and the ongoing need for large, high-quality annotated corpora to support effective model training. Contextual interpretation also plays a vital role, as meaning can shift considerably depending on surrounding text. Moreover, Arabic exhibits additional complexities rooted in its linguistic structure and historical development.

A major source of difficulty is Arabic's rich and highly productive morphology. Words are formed through root-and-pattern systems, with extensive inflection and derivational variety, unlike the comparatively simpler morphological structure of English [182, 183]. The language also presents a well-known phenomenon of diglossia: Modern Standard Arabic (MSA) functions as the formal written standard, while a wide range of dialects are used in everyday communication across different regions. These dialects differ in vocabulary, phonology, and even grammar, which complicates efforts to build universal NLP systems [184].

Another distinguishing challenge is the optional use of diacritical marks, which encode short vowels and can drastically alter meaning. When diacritics are absent—as is common

in most modern writing—words with identical spelling must be disambiguated solely from context [185]. Additionally, flexible word order, diverse syntactic constructions, and orthographic variations further increase the complexity of automated text processing [182]. Beyond linguistic characteristics, the Arabic NLP field continues to face a scarcity of labelled data and specialized resources. This lack of well-curated datasets limits the development and evaluation of robust models capable of capturing the full linguistic diversity of the Arabic-speaking world [186].

3.2. Framing Our Deepfake Text Detection Tasks

Binary classification

The investigation into identifying text generated by large language models has gained significant traction within the academic and media landscape, reflecting the pressing need to discern between human-authored content and machine-generated text in order to avoid misleading audience and various assessors.

Chaka [173] evaluated five AI-tools for detecting AI-generated content, including Copyleaks AI Content Detector and GPT-2 output detector, with Copyleaks AI Content Detector demonstrating superior accuracy in recognizing AI-generated responses from ChatGPT, YouChat, and Chatsonic. According to Chaka [173], all five AI content detectors seem to have the same drawback of being unable to accurately and convincingly identify AI-generated writings in various settings. Alamleh et al. [174] conducted a study that gathered responses from computer science students regarding essay and programming assignments to evaluate the efficacy of various machine learning (ML) algorithms, such as Support Vector Machines (SVM), Logistic Regression (LR), and Decision Trees (DT), in distinguishing between human-written and AI-generated text. Their work explored applications in content moderation and plagiarism detection, achieving high accuracy in distinguishing the two text sources. The study's stated results indicated that traditional machine learning methods (Random Forest and Extremely Randomized Trees) outperformed the neural classifier (LSTM).

Wu et al. [175] presented LLMDet, a tool that employs perplexity scores and self-watermarking to ascertain whether text originates from a large language model or is authored by a human. Given that the perplexity calculation necessitates transparent access to token-level log probabilities, which is unfeasible in practical applications, the authors suggested calculating a proxy perplexity for each target LLM utilizing standard n-gram probabilities. These probabilities serve as the LLM's writing signature to identify the closest source to the input text’s proxy perplexity.

Sadasivan et al. demonstrated in their study [176] the vulnerabilities of AI text detectors to paraphrase attacks, highlighting the necessity for resilient detection methods capable of withstanding such evasion strategies. Antoun et al. [177] developed ChatGPT detectors specifically for French text by translating existing English datasets and training classifiers. The authors showed how difficult it is for state-of-the-art classifiers trained on a combination of text produced by LLMs and human content to recognize hostile literature written in an academic, pedagogic, or encyclopedic style. Fagni et al. [13] introduced the TweepFake dataset, emphasizing the challenges in identifying DeepFake tweets.

Recent studies have also focused on ChatGPT-generated text detection in different domains. Katib et al. [178] presented the TSA-LSTMRNN model for distinguishing ChatGPT-generated text from human writing. Elkhatat et al. [179] investigated AI content detectors' efficacy in identifying ChatGPT-generated text on engineering topics. Zhenyu et al. [180] developed a technique for identifying ChatGPT-generated code using targeted masking perturbation, addressing the need for reliable detection methods in the context of programming. Liu et al. [181] treated AI text detection as authorship attribution problem, utilizing a stride sliding window approach based on GPT-2 for extracting perplexity features to distinguish text types. The magnitudes of perplexity features were compared to determine AI and human text.

Despite notable progress in AI-generated text detection, the field still faces significant challenges. Most research has focused on English or other well-resourced languages, leaving a gap in Arabic text detection. Additionally, the rapid evolution of AI models poses ongoing challenges to detection methods.

AI-generated student essays detection

In the educational sector, several studies have employed traditional machine learning algorithms combined with carefully engineered linguistic features to distinguish between human-written and AI-generated text. Researchers in [162] conducted a comprehensive assessment of existing plagiarism and AI-detection tools within the context of higher education, specifically examining English as a Second Language essays. Their investigation evaluated four commercial AI-detection tools alongside qualitative assessments provided by six ESL instructors, ultimately yielding recommendations for adaptive changes to educational practices and institutional policies in response to the challenges posed by generative AI technologies.

Tis study [163] developed a language model discrepancy classifier utilizing Support Vector Machines with n-gram bag-of-words feature representations to reliably differentiate be-tween essays produced by ChatGPT and those authored by humans. Their approach placed particular emphasis on minimizing false negative rates, thereby ensuring that authentic human-written work would not be erroneously flagged as AI-generated, a consideration of paramount importance for maintaining fairness in academic assessment con-texts.

This work [164] investigated the classification of student assignments through a comparative analysis of classical machine learning algorithms, including Support Vector Ma-chines, Random Forest, and K-Nearest Neighbours, alongside contextual models such as Long Short-Term Memory networks. Employing Term Frequency-Inverse Document Frequency feature extraction, their work underscored the critical importance of algorithm se-lection in upholding academic integrity standards and demonstrated varying performance characteristics across different algorithmic approaches.

In the study [165], detection systems for ChatGPT-produced essays based on text perplexity metrics and linguistic features derived from the Educational Testing Service e-rater engine were constructed. Leveraging large-scale data from Graduate Record Examinations writing assessments, their research explicitly investigated potential bias in detection accuracy between native and non-native English speakers, finding no

systematic disadvantage for non-native writers and thereby supporting the equitable application of automated detection systems across diverse student populations.

Moreover, this study [166] examined the adversarial attacks on LLMs-essays detection, where the suggestions for evasion strategies included word and phrase replacement. The AIG-ASAP dataset was introduced for detector assessment. AI-authored essays were readily identifiable when no perturbation was applied, but the application of perturbation methods substantially reduced detection accuracy. According to the study word substitution, among these techniques, proved particularly effective and potentially lowering detection accuracy to approximately 50%, and fine-tuning detection models can enhance their performance, but the effectiveness varies depending on the specific perturbation method employed. Interestingly, human evaluation reveals discrepancies in the perception of essay quality, suggesting that the impact of perturbation on essay coherence and readability may not be uniformly assessed by human evaluators.

The emergence of transformer-based architectures and deep neural networks has catalysed a significant shift toward more sophisticated detection methodologies. Recent surveys and empirical investigations have documented the increasing adoption of neural and transformer-based models for AI-generated text detection. A comprehensive survey of detection techniques synthesized state-of-the-art approaches across multiple paradigms, including watermarking, statistical and stylistic analysis, and machine learning classification methodologies. Authors in [167] introduced AI-Catcher, a deep learning architecture that combines a multilayer perceptron operating on linguistic and statistical features with a convolutional neural network analysing sequential text patterns. Evaluated on their newly developed AIGTxt dataset comprising scientific texts, this hybrid architecture demonstrated substantial performance improvements over baseline methods in distinguishing human-written content from ChatGPT-generated text.

Additional research has explored fine-tuning transformer models, including combinations of convolutional neural networks with bidirectional Long Short-Term Memory networks and RoBERTa architectures, on both linguistic and statistical feature representations. These investigations have consistently found that transformer-based models, particularly RoBERTa, achieve superior classification accuracy compared to earlier neural architectures [168]. However, empirical studies examining the practical deployment of detection tools have identified a critical limitation: while discrimination between human and AI-generated text is achievable under controlled conditions, detection reliability deteriorates markedly when text undergoes obfuscation or paraphrasing operations, raising concerns about real-world robustness [169].

All the above studies concluded that their detection approaches have a promising and efficient role in the essay's detection context. However, they have limitations that reduce their effectiveness in practice. Namely, they Lack diversity in datasets, potentially affecting generalizability, thus, findings may not apply across different contexts and may not detect a AI-essays generated by other models, lack the capability of understanding why certain features are predictive through addressing the models explainability. In addition, continuous evolution of AI-generated text necessitates frequent algorithm updates as the randomness in generative Language Models can produce outputs that are challenging to detect. Finally, existing research exhibits substantial linguistic concentration, with the vast majority of studies focusing exclusively on English-language contexts, and fewer still demonstrate robust performance across multilingual or non-

English educational environments. This limitation is particularly salient given the global proliferation of AI writing tools and the diverse linguistic contexts in which academic integrity concerns arise.

Segment level authorship attribution

Recent advancements in deep learning have fundamentally improved text origin identification. To improve the identification of synthetic text, Joy and Aishi [170] combined two transformer models, DeBERTaV3 and XLM-RoBERTa, using a feature-level ensemble learning technique. Other investigations have explored linguistic fingerprinting as a more nuanced detection strategy. Kumarage et al. [171] introduced StyloCPA algorithm to identify author changes in Twitter timelines and AI-generated tweets. In order to discover notable changes in writing style, the study used the Pruned Exact Linear Time (PELT) algorithm, and quantified stylistic changes using stylometric attributes.

While existing approaches predominantly focus on whole-text classification, our research builds upon emerging studies that explore more granular detection techniques. Recent advances in transformer-based models have opened new possibilities for detecting subtle text generation transitions Zellers et al. [43]. GPT-3 and subsequent large language models have demonstrated remarkable capabilities in generating contextually coherent text, simultaneously complicating detection efforts [50]. Forensic linguistic research has long employed stylometric techniques to identify textual authorship. Weber-Wulff et al. [169] developed test cases for a variety of document types, such as writings created by humans, AI- produced texts, and texts that were translated and manually edited. The study evaluated some detection tools based on accuracy and error type analysis and examined also how content obfuscation methods affected detection performance.

The ongoing combat between text generation and detection technologies necessitates continuous methodological innovation. Fu et al. [172] highlighted the critical challenges in developing robust watermarking mechanisms that can withstand rapidly evolving AI text generation capabilities.

The efforts in detecting AI-generated text in non-English languages are substantially lacking especially in Arabic [15, 52], and the extant literature primarily frames the issue as a binary classification task at the document level.

3.3. Major Existing Approaches for Combating Deepfake Text

To facilitate understanding, we group detectors according to their underlying approaches. Table 2 presents these approaches. The main detection approaches used in most of the analysed studies are: 1) Standard Machine-Learning and deep-learning Classifiers, 2) Feature-based Statistical classifiers, 3) transfer learning which has been considered a fast merging direction in auto-text detection during the last few years, and 4) graph-based detection.

Table 2. Sorting included studies by approach

Approach		Reference
Standard Machine-Learning and deep learning based detection		[13] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85] [86] [87]
Feature-based / hand-crafted statistical detection		[88] [89] [90] [91] [92] [93] [94] [95] [96] [81] [82] [97] [98] [42] [53] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111] [112] [85] [113]
Transfer learning	Zero-Shot detection	[43] [115]
	Fine-tuning detection	[116] [13] [74] [117] [118] [92] [119] [120] [53] [43] [121] [122]
Graph based detection		[92] [93] [78] [123] [124] [113]

Standard machine-learning and deep learning approach

Classical Machine learning (CML) and deep learning (DL) methods have been widely used in various NLP tasks including deepfake text detection. This approach relies on detecting deeply generated text based on classical machine learning algorithms and deep neural networks often built and trained from scratch. Analysing the corpus, deep neural networks are widely used to support detection because of their ability to detect automated text patterns. We analysed the studies to identify the neural network models and summarised them in Figure 10.

Machine learning algorithms are implemented with Bag-Of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) encoding [13]. The study [13] applied 13 detection methods across three phases. Two of the phases (nine methods in total) use machine-learning approaches; besides BOW and TF-IDF, the authors also used BERT for encoding in the second phase. Logistic Regression [91, 97, 98, 103, 108, 111], Random Forest [85, 86, 103, 107], and Support Vector Machine (SVM) [85, 97, 101, 112] are

widely applied as detectors of auto-generated text or baselines for the comparative purpose [53, 74, 110, 121].

Several works were devoted to content analysis and textual information. One of them is an article by Kudugunta and Ferrara [79]. They presented a deep learning network based on a contextual long short-term memory (LSTM) architecture that exploits content and metadata to detect auto-generated tweets using only one tweet and six account features. Furthermore, in this study, the authors showed that for models that leverage a small set of features in features –based-detection, using a minimal training set and then applying oversampling techniques to enhance the dataset, is a practical solution for dataset problems. Two blending ensemble techniques were performed in the paper [75], one based on LSTM and the other one on CNN, where LSTM outperformed CNN with a slight difference. Other articles used both CNN and LSTM together [76, 77]. In [76], an attention-aware deep neural network model based on CNN and BiLSTM is presented. The work [77] proposes a neural network ensemble of Text CNN and LSTM model with BERT embeddings to identify tweets as human or auto-generated. The combination of the two networks showed promising results. GloVe embeddings are used in [96] with an LSTM and dense layers, and BERT embeddings were used with Dense layers.

[83] puts focus on the identification of human and spambot content on Twitter, based on recurrent neural networks, specifically bidirectional Long Short-term Memory (BiLSTM), to efficiently extract tweets features.

[125] used BERT pre-trained model for producing vector representations of input tokens, followed by BiLSTM networks layer, and then employed a NeXtVLAD parametric pooling layer (which was proposed by Lin et al. [127] for video classification in computer vision) and assessed its usability in classification tasks. To classify a tweet they applied at last a fully connected layer. Luo et al. [84] proposed DeepBot detector that implements a Bi-directional LSTM-based model with the embedding of tweets as textual input into vectors using GloVe representation.

Active learning was employed to efficiently expand the labelled data in [80], which addressed Sina Weibo, a Chinese social network to recognize auto-generated text.

In the context of ML techniques in this literature review, the emphasis is placed on traditional machine learning methods (such as decision trees, rule-based systems, or linear models) based on feature engineering and labelled data.

Feature-based statistical approach

This approach is based on manually extracted features and applying statistical measures or classical machine learning algorithms. Focusing on different types of features and metadata, the goal of this investigation is to identify the text that meets the greatest number of automated text criteria. A combination of these feature criteria would provide a good model for deepfake text detection.

Using statistical analysis [42, 111, 82, 97], they tried to identify auto-generated text from real one. In [82] a detection algorithm is presented based on statistics and heuristics to optimize the error. Giant Language model Test Room (GLTR) [42] a tool by Gehrmann et al. offers a set of basic statistical methods that can reveal the differences between distributions of GPT-2 generated text and human-made text through visualizing the model's probability, the rank of each token in the distribution of the next predicted word, and the entropy of the distribution of the predicted words. In light of these visualizations, GLTR demonstrates that text generative models over-generate from a limited subset of the basic distribution of natural language and shows that GPT-2 generated texts have significantly fewer rare words than the human-authored text. The tool expects attackers to use sampling methods that favour high-likelihood tokens, which make machine-generated text detectable once histograms over per-token log-likelihoods are computed. The GLTR tool can help humans, including non-experts, to study texts. The main advantage of this tool is that it can aid untrained humans and non-experts to study texts, by visualizing the textual properties like unexpected and out-of-context words, in order to accurately detect synthetic text. However, GLTR can identify computer-generated text effectively but on the other hand, can not be confident that the text is not machine-generated. In the future, this tool might be less effective as language models become more and more capable of producing pieces of text that lack statistical anomalies. Shao, Uchendu, and Lee [112] proposed a reverse Turing test for detecting machine-made text and investigated the classification task of distinguishing between human-written and machine-generated texts. The dataset includes financial earnings reports, research articles, and chatbot dialogues. A supervised model using new stylometric features — e.g., text readability and "Digital DNA" — was developed by Pasricha and Hayes [111]. The authors assigned codes to each tweet, then converted them to ASCII characters, and a DNA sequence per user was created. They employed Statistical Measures for Text Richness and Diversity to extract the features from the Digital DNA and used publicly available datasets, by Cresci et al. [128] and Varol et al. [103] to train several machine learning classifiers: Gaussian Naive Bayes, Support Vector Machines (SVM), Logistic Regression, k-Nearest Neighbors (k-NN), Random Forest, and Gradient Boosting. Best performance scores were achieved when using unigram, bigram, and trigram features as input to the Random Forest Classifier.

Another study [104] proposed a machine learning-based classifier to detect spam auto-generated text from Arabic content on Twitter. The classifier adopts a semi-supervised learning method to label Twitter accounts according to their behaviour and profile information into spam or genuine account.

Figure 6. Features used for detecting bots and auto-generated text in the literature

Transfer learning based detection approach

Zero-Shot detection: An approach used to construct detection models for unseen target classes that have not been labelled for training. It uses class attributes as side information, transferring knowledge from labelled source classes to recognize unseen target classes. Researchers in [115] proposed GROVER language model, using semi-supervised and adaptive learning to detect deepfake articles generated using GROVER itself. The study [43] also performed zero-shot learning using GROVER model to detect fake articles generated by GROVER. A recent study by Adelani et al. [72] has demonstrated the possibility of using transfer learning with GPT-2 (124M) model for review generation purposes, producing thousands of fake Amazon and Yelp reviews. Automatically generated reviews were hardly flagged as fake since the produced reviews are considered as slick as human-written ones and can easily mislead online shoppers. For detection, authors set up an environment for identifying these reviews, in which the human annotator would be presented with one human-written review and three deepfake- generated reviews, and asked to pick out the human-written one. Some investigations have surprisingly found that smaller language models are capable of recognizing text created by LLM [91, 43]. Since LLM architectures are computationally demanding, smaller architectures may be effective for prediction instead of massive multi-billion parameter models.

Fine-tuning detection: An approach based on Fine-tuning a model with a small-to-medium-sized labelled dataset using a pre-trained model on a large, unlabelled dataset. Combining a pre-trained Gated Recurrent Unit (GRU) model with a CNN (with feature extraction), authors of [118] build a fine-tuned model for detecting machine- generated spam text. BERT transformer model was exploited and fine-tuned for deepfake text detection in numerous recent works [13, 121, 116, 43], as well as its multiple variants: RoBERTa [92], DistilBERT [13, 74], AraBERT [116], while [119] fine-tuned GPT-2 language model for identifying fake text on social media and [43] used a set of fine-tuned models: GPT-2, BERT, and FastText. [13] applied fine tuning on 4 models: BERT-FT, DistilBERT-FT, XLNET-FT, and RoBERTa-FT which achieved the highest accuracy. Using fine-tuned BERT and DistilBERT transformers, automatically generated headlines by language models are detected in [74] where BERT outperformed DistilBERT. A comparison between fine-tuning GPT-2 and BERT pre-trained language models [121] showed that the two models defer in the tweet representation level especially for hashtag representation, while in the embedding level, BERT generates embeddings that help to learn better classifiers than GPT-2. Although the embeddings alone are not considered a reliable indicator of the models' performance according to the authors. In [117] the researchers presented an end-to-end neural architecture (SC-Net) that learns the semantic coherence of text sequences using English and Chinese datasets. The architecture of the model contains an embedding layer, GPT-2 pre-trained language model, a convolutional neural net, and a fully connected layer for output prediction. By exploiting the bidirectional Transformer, the best performance was reached. The task of identification automation on Twitter was tackled in [116, 120] by adopting the fine-tuning approach and achieving high accuracies. The study [120] proposed Bot-DenseNet, which takes advantage of Transfer learning techniques through powerful state-of-the-art Transformers to extract compact multilingual representations of the text-based features associated with user accounts.

Graph-based detection approach

Graph-based modeling of social media accounts has emerged as a prominent approach for detecting bot activity on social networks and mitigating Sybil attacks [123, 93, 78, 92]. Models based on GCN (graph convolutional network) architecture are proposed in [123, 78, 92] to leverage user features and the Twittersphere structure. In the paper [78], the problem of detecting automated tweets based on user follow relationship is addressed, where the authors proposed BotRGCN. A classifier with Relational Graph Convolutional Networks, which constructs a heterogeneous graph from following relationships and applies relational graph convolutional networks to the Twittersphere. The study [93] proposed a method generating aggregate neighbourhood features in an unsupervised manner from unlabelled data of nodes adjacent to a user in the network's social graph. The study leverages the social graph's topology and differences in egographs of legitimate and fake user accounts to improve the identification of the latter. In the study [93], a number of

classical supervised and unsupervised machine learning algorithms were used with graph-based extracted features.

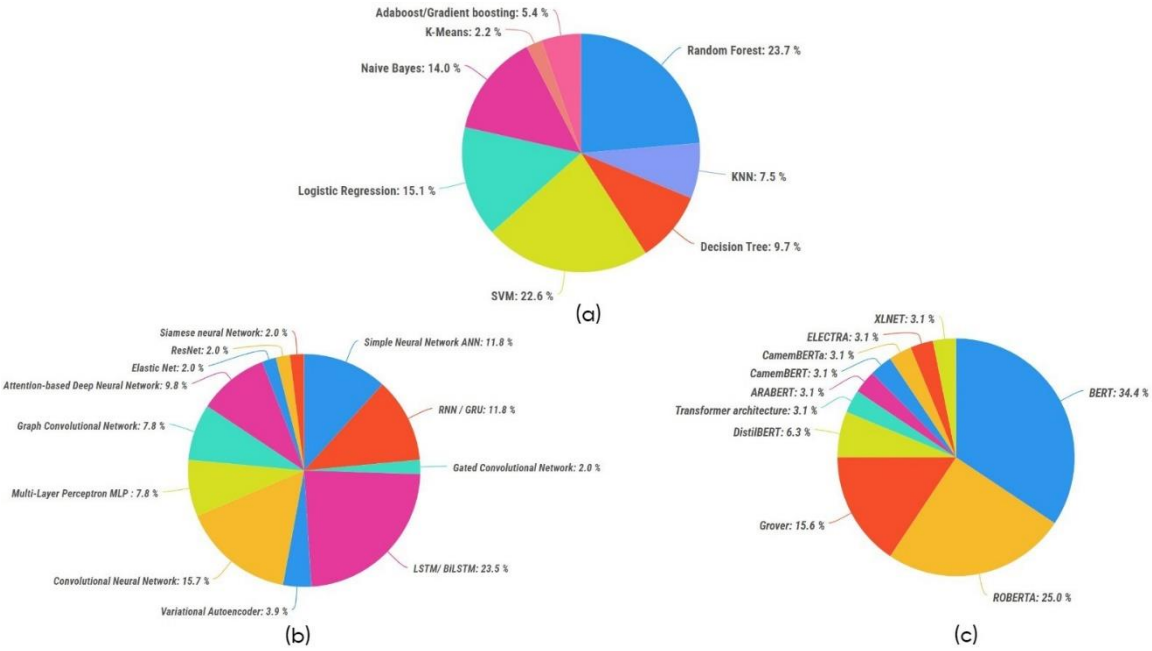


Figure 7. Different architectures used in the included studies: (a) machine learning algorithms, (b) deep learning networks, (c) Fine-tuned transformers

Detection techniques used in these approaches

With regard to the detection approaches used, we classified reviewed studies according to the techniques and algorithms that were applied for detection. Figure 8 summarises the existing state-of-the-art deepfake text detection approaches and their main methods. Figure 9 names different deep architectures applied for multiple tasks in the included studies.

Multilayer perceptron (MLP), the most traditional type of neural network architectures, despite these networks not being the most commonly used models in deep learning, they offer a soft introduction to the way neural networks function and are a robust option for incomplete tasks. Convolutional Neural Networks perform mathematical transformations called convolutions to the input text and apply a fully connected layer to predict the output. In Recurrent neural networks (RNNs), the downstream output of a layer is fed back upstream as inputs into the next layer. RNNs are powerful choices and are considered appropriate for sequential data such as text because of their potential to capture sequential correlations in data. Long-Short-Term memory (LSTM), is a special kind of recurrent

neural nets, where cells are a prominent example of a gated neural unit that is widely used for text and sequence classification tasks.

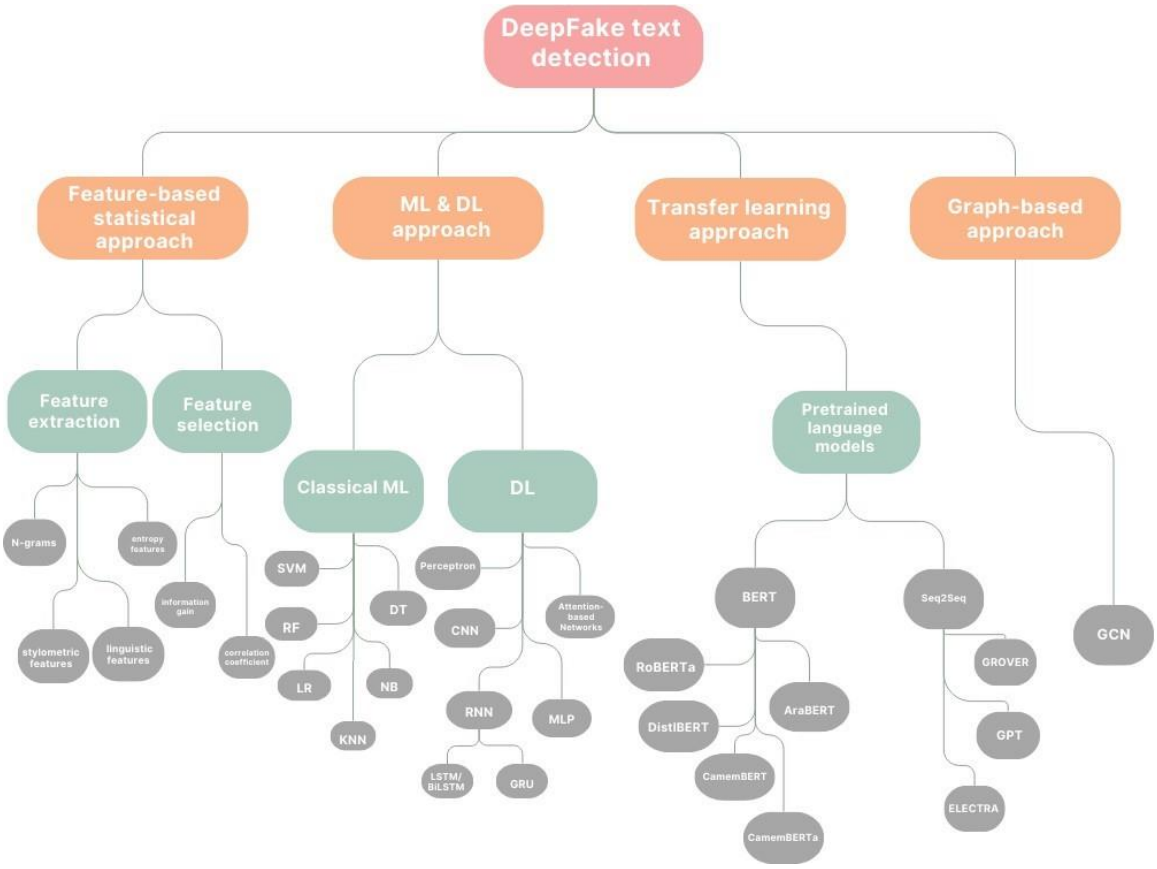


Figure 8. Summary of deepfake text detection approaches and their most used models

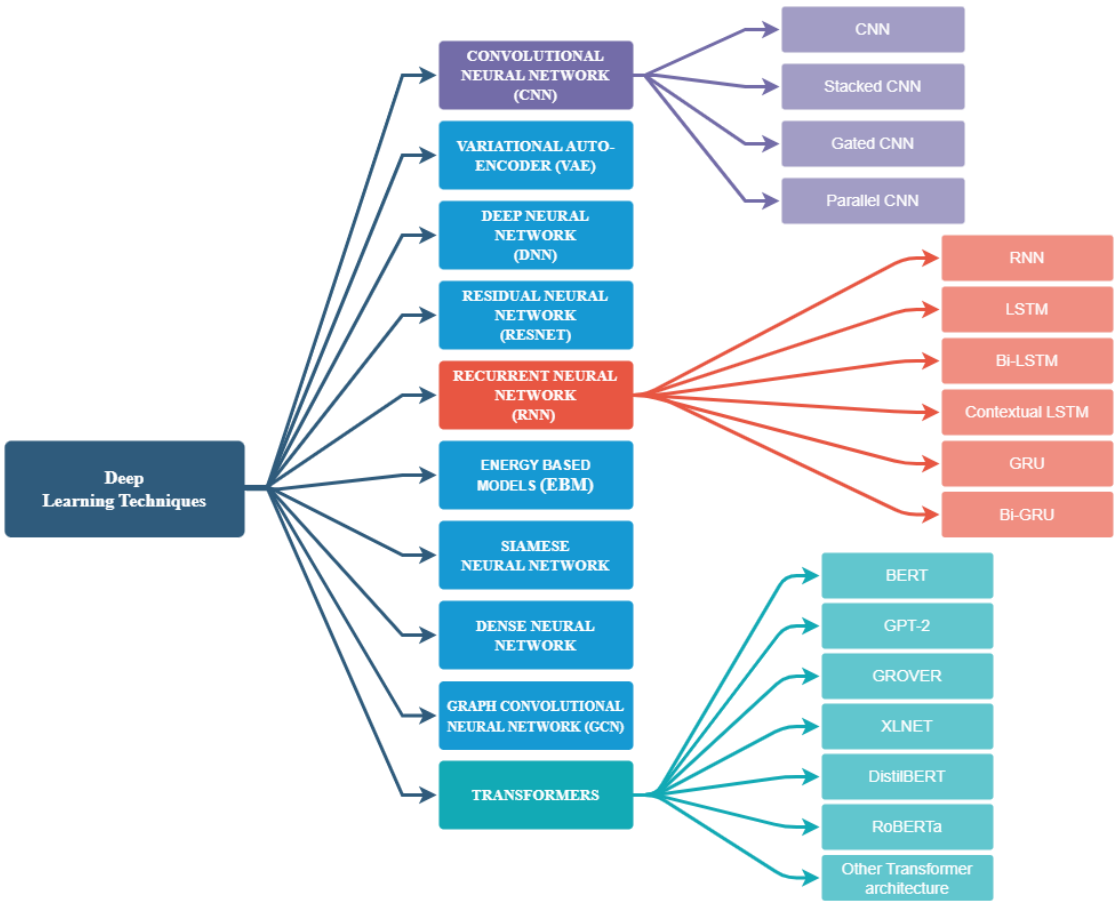


Figure 9. Types of Deep Learning architectures used in the included studies

3.4. Datasets used for Deepfake Text Detection

Cleaning and filtering data are among the most important steps in any social-media analysis workflow. In this section, the datasets utilised in existing studies for evaluating the performance of their model are listed in Table 3. They have utilised benchmark datasets for both training and testing. A major challenge is the lack of large, labelled benchmark datasets with reliable ground truth. For example, some of the datasets are constructed only with political statements like PolitiFact, LIAR, Weibo, etc. The Twitter dataset consists of social media posts, whereas the FNC-1 dataset is built based on news articles. Moreover, datasets can be varied through size, labels, and modalities. Similarly, most of the studies use self-collected data from either news articles or any social media platforms. The details and characteristics of the datasets utilised are shown in Table 3.

The data imbalance problem is tackled in a number of included works. Although these techniques assure an equal amount of examples from each class of text in order to reduce bias, studies that handle data imbalance issues in the dataset like [77, 80], generally

generate repetitive sentences and uninformative examples that are not new, but just slight modifications of the sentences already present in the dataset. Besides, handling the data imbalance leads to early overfit during the training process. Figure 10 describes the distribution of Embeddings and Data Encoding techniques utilised for fake text detection.

Table 3. The description of publically available datasets used in the studies included

Dataset	Quality	Used in
Almerekhi H, Elsayed T (2015)	3,503 tweets: manually labelled as "human content" (1,559 tweets) or "bot content" (1,944 tweets)	[116] (used only human content which was expanded to 4,196 tweets + 3,512 GPT-2- Small-Arabic auto-generated sentences). [101]
TweepFake	23 bots and 17 human accounts with 25,836 tweets; half human and half bots	[13] [119] [125]
Cresci-2017	Total: 8386 user accounts, 11.834.866 tweets Social Spambots #2: 3457 social spambots (428,542 total tweets). Genuine accounts: 3474, 8.377.522 tweet	[89] [75] [120] [76] [77] [79] [94] [83] [111]
QiHoo Generated News (QHGN)	Positive:420.632, Negative: 599.531, Avg token:605, Max token:2.618	[118]
Modified New York Times (MNYT)	Positive: 599.994, Negative: 417.521, Avg token: 1.306, Max token: 69.517	[118]
RealNews [43]	News-style GROVER-generated dataset 10000 real news articles	[92] [119] [43]
Caver lee	15,483 bots	[89] [94]
varol-icwsm [103]	15k manually verified social Bots 2.6 million tweets by bots	[120] [94] [103] [109] [111]
Pronbots Yang-2019	17,882 bots	[120] [94]
Yang-2013	1000 bots with 220.90 tweet	[123]
vendor-purchased Yang-2019	1,087 bots	[120] [94]
The botometer-feedback Yang-2019	139 bots	[120] [94]
political-bots yang-2019	62 bots	[120] [94]
gilani-17	1,090 bots	[120] [94] [121]
cresci-rtbust [144]	353 bots	[120] [94]
botwiki-2019 [94]	698 bots	[120] [94]
midterm-2018 [94]	42,446 bots	[120] [94]
Cresci-2018	25,987 user accounts : 18.508 bots	[93] [94] [120]
TwiBot-20 [124]	229,573 Twitter users. 8,723,736 user property items.	[78] [124]

Swld-300k	150,000 normal users 10,779,129 posts	[80]
CLEF2019 [143]	Maximum length of tweet: 933 4120 accounts	[95] [96]

Another dimension considered during data extraction is the source of the datasets. We identified the sources and tools used to collect datasets, such as platform APIs and web crawlers. Recent studies rely on Twitter, although other social networks have been utilised such as Sina Weibo and Facebook.

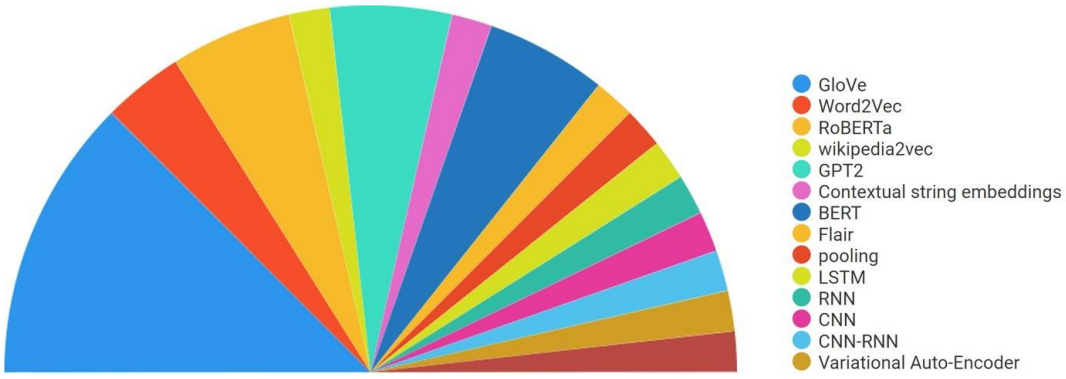


Figure 10. Embeddings and data encoding techniques utilised for fake text detection

Based on the synthesis of the findings, we can conclude that the existing datasets for deepfake text detection have different characteristics and challenges, as well as different advantages and disadvantages, for deepfake text detection. Some of the main factors that affect the quality and suitability of the datasets are:

Language: We noticed a severe dearth of language diversity, where studies that counter-attack deepfake text using a language other than English in the processed data are nearly rare. Despite the fact that some works targeted other languages such as Spanish [84, 95], Chinese [80], Arabic [104, 116] and Italian [94], as shown previously in Figure 5, there is still a considerable gap in research combating deepfake text in non-English languages. While in actuality, information operations, which include machine-generated text, are likely to involve a wide range of languages tailored to the intended target audiences, not just English. Moreover, the impact of various linguistic or cultural context on the identification of machine-generated text is not taken into account in most of the literature.

The suggested methods' efficacy can differ depending on the language and cultural contexts, so more study is required to confirm the suitability of detection methods in diverse settings.

Data availability and quality: The availability and quality of the data used for training and testing the detection methods are crucial for the accuracy and generalization of the detection methods. However, the data for deepfake text detection is often scarce, noisy, imbalanced, or domain-specific, which can pose challenges for the data collection, annotation, processing, and augmentation methods. Moreover, the data for deepfake text detection may vary in terms of format, style, topic, language, etc., which can require different features, models, and metrics for the detection methods. The remarkably restricted data availability and the lack of publicly available well-organized deepfake text datasets, and missing or imprecise ground truths, especially datasets with text generated using more recent large language models.

Data diversity and representativeness: The diversity and representativeness of the data used for training and testing the detection methods are important for the performance and robustness of the detection methods. However, the data for deepfake text detection may not capture the full spectrum and complexity of the real and synthetic text content, especially when the data is collected from specific sources, domains, or topics, or when the data is generated by specific models, techniques, or parameters. For instance, each work in the literature focuses on a specific social media platform for detection and may not generalize well to detection in other platforms, limiting its applicability in a broader context. Therefore, developing diverse and representative datasets is a significant challenge for researchers and essential to improve the validity and reliability of detection methods.

Data collection and annotation: Analysing patterns and labelling large amounts of data, on platforms such as social networks, can be a time and effort-consuming task. This is where semi-supervised learning approach can be largely useful for labelling the datasets [72]. Obtaining large amounts of labelled data for AI-generated text detection can be costly, time-consuming, or difficult, as it requires human experts to manually annotate the texts. Due to its ability to enhance the performance and generalization of detection models by utilizing the abundant unlabelled data that are accessible on platforms like social networks, semi-supervised learning (SSL) presents itself as a promising human-interference-free solution for AI-generated text recognition. SSL aims to learn useful representations from unlabelled data and augment labelled datasets with confident model predictions to balance classes. In semi-supervised learning, both labelled and unlabelled data are used to train the detection models. Several studies have explored SSL for constructing datasets for detection, such as [43, 80, 104, 115]. However, there is still room for developing more efficient and effective data collection and annotation methods, such as crowdsourcing, active learning, or self-supervision.

Data processing and augmentation: Data processing and augmentation are methods that aim to improve the data availability and quality for deepfake text detection, by cleaning, preprocessing, balancing, sampling, or augmenting the data. However, data processing and augmentation are also challenging for deepfake text detection, as they may introduce noise, bias, or inconsistency in the data, or they may not capture the full spectrum and complexity of the real and synthetic text content. Therefore, developing more robust and adaptable data processing and augmentation methods, such as noise reduction, feature extraction, data synthesis, or data transformation, is a future direction and challenge for deepfake text detection.

3.5. Strengths and Limitations of Existing Techniques

Figure 11 depicts the taxonomy of all the mentioned detection approaches based on their positives and weaknesses. The "+" denotes advantages and the "–" denotes disadvantages for each category studied.

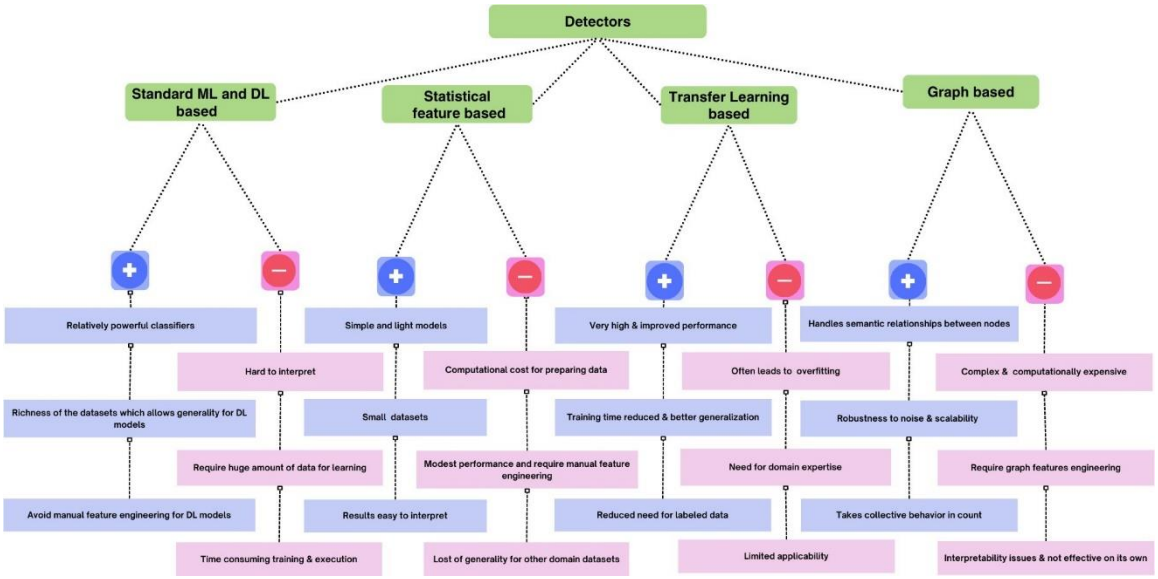


Figure 11. General view of the pros and cons of the approaches used

Strengths of the existing detectors

Statistical feature based approach: This approach extracts various statistical features from texts, such as n-gram frequencies, perplexity, entropy, etc., and uses them as inputs for detection models. The strength of this approach is that it can capture some intrinsic characteristics of texts that are hard to mimic by language models, such as word usage, style, and diversity. Moreover, these features can provide additional adversarial robustness, as they are less sensitive to slight modifications of the texts that aim to evade detection [135]. Statistical and hand-crafted feature-based methods are considered to be robust to noisy data and can handle missing or erroneous words. Furthermore, these methods are computationally efficient and very fast at training and inference, as they require less processing time and power compared to deep learning methods [136]. Statistical methods also have low data requirements and yet can achieve acceptable results [42, 111, 105]. These methods are also suitable for detecting machine-generated text in different domains and scenarios, as they have low data requirements and yet can achieve acceptable results [42, 111, 133, 105]. This approach can also be more robust to adversarial attacks that modify the texts slightly to evade detection.

Standard ML and DL based approach: This approach can achieve high accuracy on specific domains or language models, as long as the training and testing data are from the same distribution [131, 43]. However, this approach may not generalize well to new data that differ from the training data in terms of patterns, quality, quantity, complexity, or diversity [117, 49]. CML and DL methods have the potential to generalize well to new data because they learn from patterns in the data, but this also depends on the quality and quantity of the training data, and the complexity and diversity of the test data [96, 76, 75, 77], as this approach may require careful tuning of the models and their parameters to achieve optimal performance [132]. CML methods provide straightforward feature-based classifiers that can recognize machine-generated text and sometimes outperform more expensive techniques [91], providing a readily available "first line of defence" against the misuse of LLMs. ML methods can also provide interpretable results that can help understand why the model made a particular prediction by showing the features, rules, or coefficients that contribute to the prediction [42], although the interpretability of these methods may vary depending on the complexity and size of the model, the number and type of features, and the specific domain of application. DL methods are flexible and are able to handle a variety of input types, such as raw text, word embeddings, or other types of feature representations. DL methods can also capture complex and non-linear relationships between the input and the output, which may be useful for distinguishing subtle differences between human and machine-generated text.

Transfer learning based approach: The strength of this approach is that it can leverage the large-scale and diverse data that the language models are trained on, and thus generalize

better to unseen domains or language models [138]. This approach can also benefit from the continuous improvement of language models, as new models can be easily integrated into the detection pipeline [50]. Transfer learning allows faster training of models because the pre-trained model has already learned the underlying features and patterns in the data. The new model only needs to learn the task-specific features. This can also significantly improve the performance of the new model since it is based on a pre-trained model that has already learned to recognize the relevant features and patterns. Other strengths of TL lie in the fact that it can help the model generalize better and reduce the need for labelled data. Fine-tuned models don't need a large dataset for training [72, 101, 13, 119, 121, 116, 110, 74, 53], and yet can achieve high state-of-the-art accuracies.

Graph based approach: This approach represents texts as graphs, where nodes are words or sentences, and edges are syntactic or semantic relations. The approach then uses graph neural networks [123, 92] or other graph algorithms [123, 78] to perform detection. The strength of this approach is that it can exploit the structural and relational information in texts, which are often overlooked by other approaches. This approach can also handle long and complex texts better, as graphs can capture the global and local coherence of texts. Graph- based approaches can capture semantic links between words such as synonymy, antonymy, and hypernymy. By modeling these relationships as edges in the graph, the method can better capture the meaning of the text. This approach also allows robustness to noisy data because it uses a graph representation that can handle missing or erroneous words. The graph structure can help in dealing with ambiguities and uncertainties in the data. Graph-based methods allow scalability to large datasets and high-dimensional feature spaces because they only require pairwise similarity computations between the nodes in the graph, rather than considering all possible combinations of words. In graph-based studies, like [78], the multi-modal user semantic and property information are used to avoid feature engineering and enhance its ability to capture bots with diversified disguises.

Weakness of the existing detectors

Statistical feature based approach: This approach has several limitations that may affect its performance and applicability as it may not capture the semantic and contextual information in texts, which are important for human perception and understanding [133]. This approach also has difficulty in selecting and combining the optimal features for detection, as different features may have different discriminative power and correlation [139]. Furthermore, this approach may be affected by the noise or variation in the texts, such as spelling errors, slang, or dialects [135]. Moreover, this approach may fail to catch all the variations and nuances in text written by machines, which could result in false positives or false negatives during the identification process [91].
Methods based on statistical and hand-crafted features often face an issue of having a limited feature set of the data, and lack generalization to new data because they rely on predefined rules and features. These methods require domain expertise to design and

engineer features, which can be time-consuming and expensive [72]. Additional drawbacks include the fact that they may not perform as well on complex tasks and can be limited in the types of input they can handle, such as raw text or word embeddings, and may be vulnerable to adaptive attacks that modify the texts slightly to evade detection [131].

Standard ML and DL based approach: The weakness of this approach is that it is highly dependent on the quality and quantity of the training data, and it may suffer from overfitting or underfitting problems [132]. This approach may also fail to generalize to new domains or language models, as the distribution of the texts may change significantly [49, 131]. Moreover, this approach may be vulnerable to adversarial attacks that exploit the gradient information or the model architecture to generate undetectable texts [117]. Because DL approaches rely on big, complex models with many layers of nonlinear transformations, they can be less interpretable than ML methods and the features learned by these models can be difficult to interpret [42]. To train, deep learning methods can be computationally expensive, requiring enormous quantities of memory and computing power, which can make training these models on low-resource devices problematic [140].

Transfer learning based approach: The weakness of this approach is that it may inherit the biases or limitations of the pre-trained language models, such as factual errors, ethical issues, or domain specificity [49, 71]. This approach may also face the challenge of fine-tuning the language models for detection, as the objective and data of detection may differ from those of pre-training [42, 138]. Additionally, this approach may incur high computational and memory costs, as the language models are often large and complex [140]. This approach also can make the model less interpretable since the pre-trained model's features may be difficult to interpret, and the fine-tuned model's features may rely on both the pre-trained and task-specific characteristics [72]. Transfer learning requires domain expertise to choose an appropriate pre-trained model and fine-tuning strategy, and to interpret the results of the fine-tuned model [142]. Detectors of this approach may be prone to false positives, as some detectors utilizing information from pre-trained language models such as GLTR and Longformer have shown a propensity to mistakenly classify human-authored texts as machine-generated, particularly in scenarios outside of their trained data distribution [42, 138].

Graph based approach: The weakness of this approach is that it may require sophisticated and reliable methods to construct and analyse the graphs, which are not always available or easy to implement [133]. This approach may also encounter the problem of graph sparsity or inconsistency, as the texts may not have enough or consistent nodes and edges to form meaningful graphs [139]. Moreover, this approach may have scalability issues, as the graphs may grow exponentially with the length and complexity of the texts [140]. Graph-based approaches can be computationally expensive because they need the construction of the graph as well as the computation of pairwise similarities between nodes [72], and also less interpretable than traditional feature-based methods, as it relies on the graph structure to capture the relationships between words, which may be difficult to explain [42]. Moreover, this approach is also highly dependent on a number of hyperparameters, including the similarity measure and the graph creation process. Hence the performance of these methods can be sensitive to the choice of these hyperparameters.

This approach may not be sufficient to discriminate real text from deepfake text on its own, and it needs to be combined with other approaches like ML or DL to increase the overall performance [93]. In addition, it may lack a detailed exposition of the computational complexity of the method used, which may hamper the understanding of its scalability and efficiency [78].

3.6. Chapter Summary

This chapter has systematically analysed the strengths and limitations of existing research on AI-generated text detection, revealing critical gaps that this thesis addresses. The comprehensive examination of current literature demonstrates that while significant progress has been made in detecting synthetic content in high-resource languages like English—primarily through transformer-based models and stylometric analysis—Arabic remains severely underexplored, representing a substantial opportunity for scholarly contribution.

The literature review process has illuminated several fundamental deficiencies in the current research landscape. Most notably, the literature reveals a pronounced lack of Arabic-specific detection frameworks, despite Arabic being one of the world's most widely spoken languages with over 400 million native speakers. This linguistic bias in existing research creates significant vulnerabilities in detecting AI-generated content across diverse digital ecosystems where Arabic content proliferates.

Furthermore, the review identifies a concerning limitation in the scope of current detection methodologies, which predominantly focus on binary classification approaches. This narrow focus fails to address the growing complexity of hybrid texts—content that combines both human-authored and AI-generated segments—which represents an increasingly common phenomenon in real-world scenarios. The absence of segment-level analysis capabilities in existing detection systems creates blind spots that malicious actors can exploit to circumvent detection mechanisms.

The findings of this review align closely with broader observations in the field of deepfake text detection research. As established in recent review studies, social media platforms have emerged as fertile breeding grounds for online misinformation, with machine-generated text receiving unprecedented attention from both academic and industrial communities. This heightened interest stems from the remarkable efficiency of contemporary language generative models in emulating human-written text, creating an urgent need for detection mechanisms that can maintain pace with rapidly evolving natural language generation technologies.

The development of accurate detection systems has become a crucial necessity as the sophistication of generative models continues to advance. Current research indicates that while humans struggle to differentiate between authentic and generated text, automatic detection systems demonstrate superior performance in this critical task. However, the

review reveals that existing detection approaches—encompassing classical machine learning algorithms, deep learning frameworks, manual feature extraction methods, statistical approaches, and graph-based methodologies—exhibit significant gaps and limitations that must be addressed through targeted research efforts.

The literature analysis reveals that current detection methodologies primarily rely on transformer-based architectures and stylometric features, which have proven effective for English-language content but lack validation in Arabic linguistic contexts. The unique morphological, syntactic, and semantic characteristics of Arabic present distinct challenges that existing detection frameworks have not adequately addressed. This gap necessitates the development of language-specific approaches that account for Arabic's rich morphological system and complex orthographic conventions.

Additionally, the predominant focus on binary classification in existing literature overlooks the nuanced reality of contemporary content creation, where human authors increasingly collaborate with AI systems to produce hybrid texts. The inability of current detection systems to perform segment-level analysis represents a critical limitation that undermines their practical utility in real-world applications where partial AI assistance is becoming commonplace.

These findings comprehensively validate the thesis's strategic focus on Arabic-specific detection frameworks and segment-level analysis methodologies. The identified gaps directly support the necessity for developing novel detection approaches that can address the unique challenges presented by Arabic-language content while simultaneously providing the granular analysis capabilities required for hybrid text detection.

The review also underscores the importance of incorporating interpretability mechanisms and ethical accountability frameworks into detection systems. As these technologies become increasingly integrated into content moderation and academic integrity enforcement, the ability to explain detection decisions and ensure fair, unbiased operation becomes paramount. The thesis's emphasis on these aspects addresses critical concerns raised in the broader literature regarding the responsible deployment of automated detection systems.

By positioning the thesis within these identified research gaps, this chapter establishes a robust foundation for the empirical contributions presented in subsequent chapters. The proposed novel methods for detecting AI-generated news articles, student essays, and hybrid texts in Arabic directly respond to the deficiencies revealed through systematic literature analysis. The technical design of these methods draws extensively from the reviewed literature while addressing its limitations through innovative approaches tailored to Arabic linguistic characteristics and hybrid text detection requirements.

The societal relevance of these contributions extends beyond technical advancement to encompass critical applications in combating misinformation and preserving academic integrity. As the review demonstrates, the current inability to effectively detect AI-generated Arabic content creates vulnerabilities that undermine information

trustworthiness and educational standards. The thesis's focus on these applications reflects the urgent need for detection capabilities that can protect the integrity of digital information ecosystems in Arabic-speaking communities.

The comprehensive literature analysis reveals that deepfake text detection represents a challenging and multifaceted problem with far-reaching implications across diverse domains and application scenarios. While existing methodologies have demonstrated promising results in controlled environments, their limitations become apparent when applied to real-world scenarios involving low-resource languages and complex content structures.

This review process has illuminated the need for continued research that addresses current gaps while anticipating future challenges posed by rapidly advancing generative technologies. The thesis contributes to this ongoing effort by providing targeted solutions for Arabic content detection and hybrid text analysis, thereby advancing the broader goal of maintaining information trustworthiness and integrity in increasingly complex digital environments.

This literature review not only informs the technical design decisions underlying the thesis's proposed methodologies but also reinforces their critical importance in addressing pressing societal challenges. The convergence of identified research gaps with the thesis's specific focus areas validates the scholarly contribution while establishing clear pathways for future research development in this rapidly evolving field.

CHAPTER IV

Detection of AI-Generated News Articles

Following the foundational exploration of deepfake technologies and their implications for Arabic text in Chapter 2, this chapter transitions from conceptual frameworks to practical solutions by addressing the detection of AI-generated news articles in Arabic. While the previous chapter established the linguistic and technical challenges, it did not evaluate specific methodologies for identifying synthetic content in real-world applications. This chapter bridges that gap by presenting a focused investigation into the efficacy of automated detection systems and their comparison with human evaluators, thereby advancing the thesis's goal of combating Arabic deepfake text.

The aim of this chapter is to rigorously evaluate the performance of fine-tuned detection models in distinguishing AI-generated Arabic news articles from human-authored ones, while also benchmarking its accuracy against human evaluators. Readers will learn how state-of-the-art natural language processing (NLP) techniques can be adapted to Arabic, a language often underrepresented in AI research, and gain insights into the limitations of human judgment in detecting synthetic text. By quantifying the gap between machine and human detection capabilities, this chapter contributes novel empirical evidence to the field, reinforcing the urgency of developing language-specific detection tools.

To achieve this aim, the chapter is structured as follows: First, the methodology for dataset construction is detailed, including the collection of human-authored Arabic news articles and their AI-generated counterparts using models like ChatGPT. The design and fine-tuning process of the detection models are explained. The experimental setup is then outlined, covering evaluation metrics such as precision, recall, and F1-score. Next, a comparative analysis of the model's performance against human evaluators is presented.

4.1. Methodology

In this chapter, we employ transfer learning on extensive pre-trained models to establish a method for detecting Arabic deepfake text. Our methodology consists of three primary phases: the initial phase involves dataset generation, the second phase entails deepfake text detection utilizing the proposed classifiers, and the third phase, detailed in Section 4.6, explores a comprehensive comparison between automated detection and human discernment of deepfake text. We developed a novel dataset, Ara-Deep, utilizing the ChatGPT model (version 3.5) for the training data. Our dataset was utilised to train our four models, DFTD1, DFTD2, DFTD3, and DFTD4, which were subsequently assessed on alternative datasets, establishing our study as one of the initial pioneering efforts in this domain to tackle the identification of Arabic deepfake text generated using ChatGPT. Sections 3.2 and 3.3 illustrate the dataset generation methodology and the comprehensive architecture of the proposed models.

Data Collection and Preprocessing

Our custom dataset, so-called Ara-Deep, represents a fusion of artificially generated and authentic text data. The process of generating artificial text involves a meticulous prompt engineering approach, wherein real examples from the SANAD dataset were employed [145]. These real examples served as input to ChatGPT (version 3.5), guiding the model in rephrasing and completing text to produce coherent, natural, and human-like deepfake text.

As outlined in Table 4, we provide insights into our generated dataset's composition in terms of sentence count, word count, and unique words count.

Figure 12 provides a visual representation of the word frequency distribution within our dataset. For a more lucid global visualization, the dataset is split into ten distinct and equal segments in these renderings.

Table 4. Characteristics and statistical information of the produced dataset

Property	Value
#sentences	19,807
#words	2,065,670
#Unique_words	55,153
Average_words_per_sentence	≈ 104

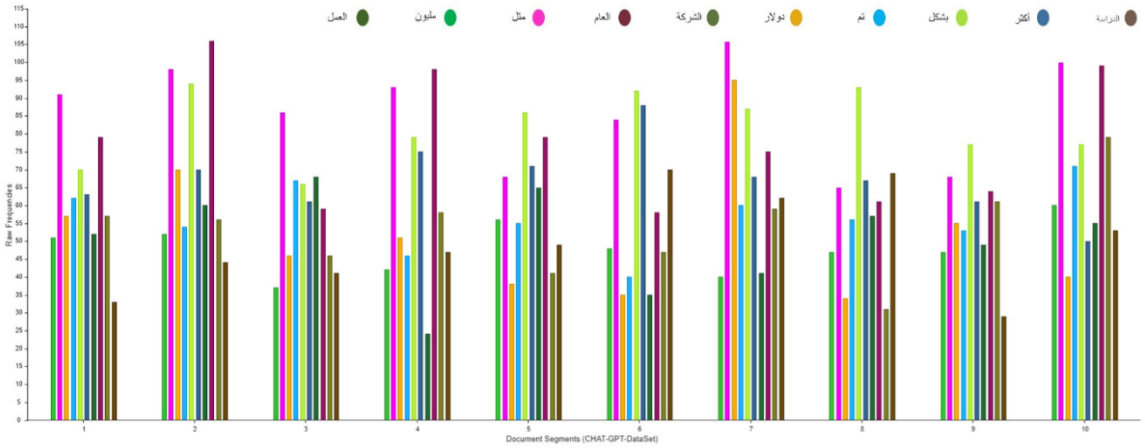


Figure 12. Frequencies distribution of the top 10 words in our dataset. The whole dataset is segmented into10 segments

As shown in Figure 14, during the prompt engineering phase, text reformulation and text completion were employed, with the text being entered either at the beginning or the end of the prompt. The dataset generation pipeline are visually expounded upon in Figure 13. Table 5 offers an overview of the real text that was given as input and the corresponding deepfake generated text samples.

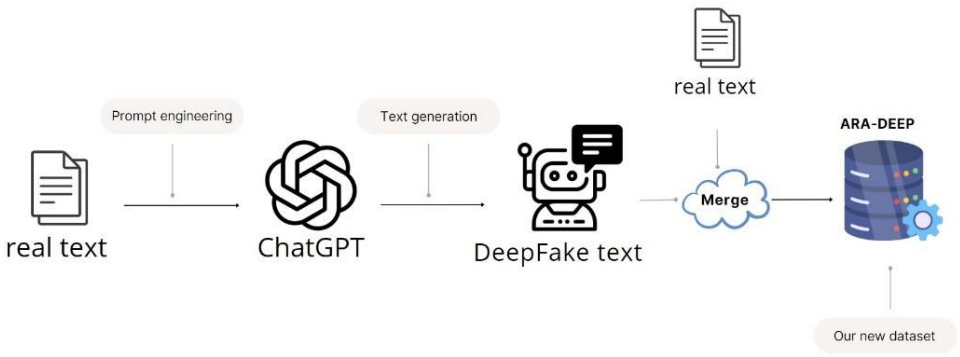


Figure 13. The dataset construction pipeline

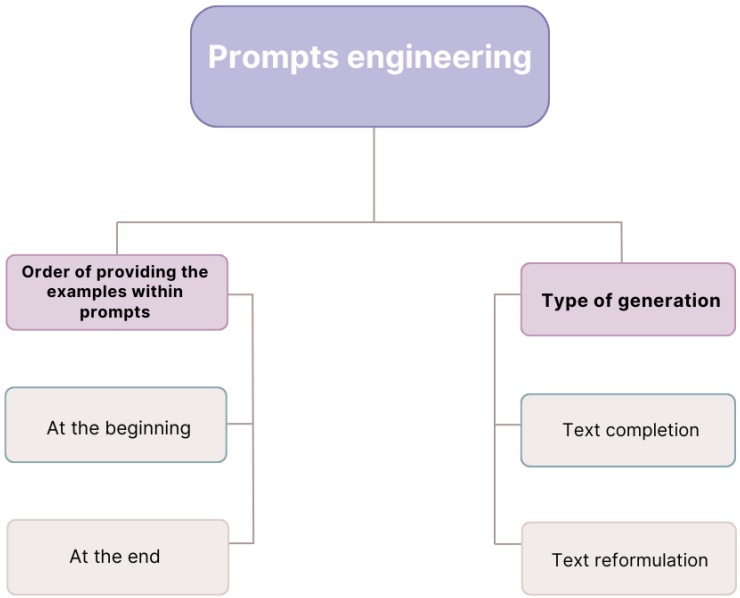


Figure 14. Prompts engineering techniques used with ChatGPT for generating AI-produced text

To improve data quality before using the dataset for the training, we clean the data by removing URLs, hashtags, and symbols. The process was performed using Python with Pandas libraries, yielding a dataset containing 2,065,670 terms and 19,807 phrases, with a total of 55,153 unique words among these terms.
A word cloud representation of the most meaningful and visible words included in the Ara-Deep dataset can be seen in Figure 15.

Table 5. Two examples of the instances included in our dataset generated by ChatGPT. The first example is generated using text completion, while the second is produced based on a text paraphrasing prompt.

	Input text		Output text (deepfake text)
	Prompt	real text	
text completion	أكمل النص الآتي:	شهدت أرباح شركة إعمار المدينة الاقتصادية	شهدت أرباح شركة إعمار المدينة الاقتصادية نموًا قويًا خلال الفترة الماضية. وفقًا للتقارير المالية الأخيرة، سجلت الشركة أداءًا ممتازًا، حيث ارتفعت أرباحها بشكل ملحوظ. تعزى هذه الزيادة في الأرباح إلى نجاح استراتيجية الشركة في تنمية وإدارة مشاريعها العقارية المتنوعة. حيث قامت إعمار المدينة الاقتصادية بتطوير مجموعة من المشاريع الكبيرة والمبتكرة في قطاعات مختلفة مثل العقارات السكنية والتجارية والضيافة والترفيه. تعكس نتائج الأرباح الإيجابية لشركة إعمار المدينة الاقتصادية قوة الاقتصاد المحلي والاستقرار العام للسوق العقاري. كما تعزز هذه النتائج ثقة المستثمرين وتجذب المزيد من الاستثمارات إلى المشاريع العقارية في المدينة. توجهت إعمار المدينة الاقتصادية نحو تنويع مصادر الدخل وتعزيز الابتكار والتكنولوجيا في إطار رؤية استراتيجية طويلة الأجل. كما تسعى الشركة لتعزيز التعاون الاقتصادي والتجاري مع شركات ومؤسسات دولية، بهدف تحقيق النمو والاستدامة.
	Complete the following text:	Emaar The Economic City’s profits have witnessed	“Emaar The Economic City”’s profits have witnessed strong growth over the past period. According to recent financial reports, the company has recorded excellent performance, as its profits have increased significantly. This increase in profits is due to the success of the company's strategy in developing and managing its various real estate projects. Emaar The Economic City has developed a group of large and innovative projects in various sectors such as residential and commercial real estate, hospitality and entertainment. The positive earnings results of Emaar The Economic City reflect the strength of the local economy and the general stability of the real estate market. These results also enhance investor confidence and attract more investments to real estate projects in the city. Emaar The Economic City has moved towards diversifying sources of income and enhancing innovation and technology within the framework of a long-term strategic vision. The company also seeks to enhance economic and commercial cooperation with international companies and institutions, with the aim of achieving growth and sustainability. …
text reformulation	أعد صياغة النص الآتي:	حصلت شركة سامسونغ على حق براءة اختراع جديدة لآلية ميكانيكية تسهل من عملية طي الشاشات للهواتف الذكية، من المتوقع رؤية أول نموذج لها بالأسواق خلال عام 2017، كانت قد تقدمت بها في يونيو (حزيران) 2015. وتعمل هذه الآلية على طي شاشة الهاتف عبر أجزاء ميكانيكية مثبتة خلف الشاشة، والتي من شأنها أن تدعم عملية فتح وغلق نصفي الشاشة على بعضهما البعض، كما تعمل على تثبيت الشاشة في الوضع المفتوح عند ولا تعد هذه براءة الاختراع الأولى من الاستخدام. نوعها التي تسجلها سامسونغ، حيث سبق وأن سجلت	تمكّنت شركة سامسونغ من الحصول على براءة اختراع جديدة لآلية ميكانيكية تسهل عملية طي شاشات الهواتف الذكية. من المتوقع أن نشهد أول نموذج يستخدم هذه التقنية في الأسواق خلال عام 2017. تعتمد هذه الآلية على أجزاء ميكانيكية مثبتة خلف الشاشة لتمكين فتح وإغلاق الشاشة بشكل مطوّى وتأمين استقرارها عند الاستخدام. وليست هذه هي أول براءة اختراع تُسجّلها سامسونغ في هذا النوع، فقد قدّمت الشركة الكورية عدة براءات اختراع لتقنيات وآليات تسهّل صنع الهواتف النقالة ذات الشاشات القابلة للطي. وما زالت سامسونغ تدرس جدوى هذه الفكرة من الناحية الاقتصادية، خاصة مع وجود بعض المستهلكين الذين لا يفضلون هذه الفكرة.

	Input text		Output text (deepfake text)
	Prompt	real text	
		الشركة الكورية عدة براءات اختراع لتقنيات وآليات تسهل من صنع الهواتف النقالة المزودة بشاشات قابلة ولا تزال تبحث سامسونغ مدى الجدوى للطي. الاقتصادية من الفكرة، لاسيما وأنها وجدت بعض المستهلكين لا يحبذون الفكرة.	
	Reformulate the following text:	Samsung has obtained a new patent for a mechanism that facilitates the process of folding screens for smartphones. It is expected to see its first model on the market during 2017, which it filed in June 2015. This mechanism works to fold the phone screen through mechanical parts installed behind the screen, which will support the process of opening and closing the two halves of the screen to each other, and also works to hold the screen in the open position when in use. This is not the first patent of its kind registered by Samsung, as the Korean company has previously registered several patents for technologies and mechanisms that facilitate the manufacture of mobile phones equipped with foldable screens. Samsung is still examining the economic feasibility of the idea, especially since it found that some consumers do not like the idea.	Samsung was able to obtain a new patent for a mechanism that facilitates the process of folding smartphone screens. It is expected that we will see the first model using this technology on the market during 2017. This mechanism relies on mechanical parts installed behind the screen to enable the screen to be opened and closed in a folded manner and to ensure its stability when in use. This is not the first patent registered by Samsung in this type, as the Korean company has filed several patents for technologies and mechanisms that facilitate the manufacture of mobile phones with foldable screens. Samsung is still studying the economic feasibility of this idea, especially with some consumers who do not prefer this idea.

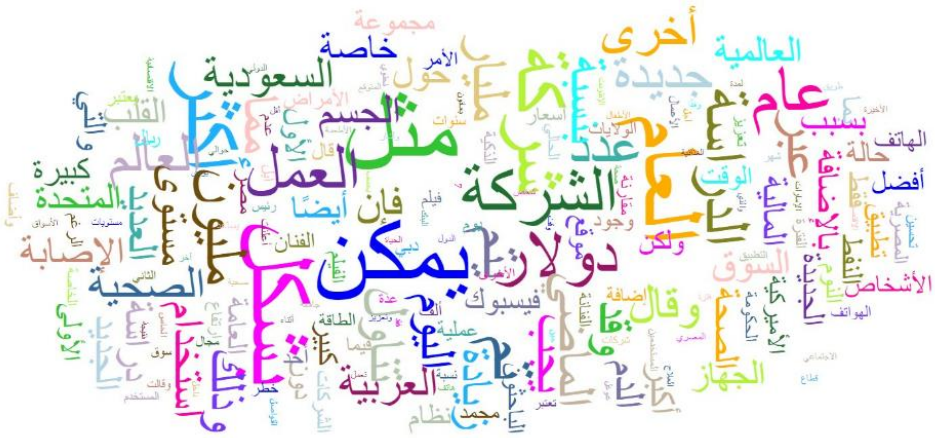


Figure 15. Word cloud of the top words included in the dataset

Model Architecture

As previously introduced, the current task of deepfake text detection is approached as a binary classification problem. In our pursuit of delving into the ability to distinguish text generated by large language models, our study conducted a meticulous investigation involving a cohort of four distinct pre-trained models, trained on multi-languages including Arabic and that performed well on NLU tasks. Notably, the models encompassed Multilingual-BERT (mBERT) [146], XLM-RoBERTa-large, XLM-RoBERTa-base, and XLM-RoBERTa-large-XNLI [147]. We judiciously selected these models for their established prominence and demonstrated capabilities in NLU and linguistic feature extraction [73, 126, 129, 134].

Our transfer learning approach leveraged these pre-trained models' rich linguistic knowledge acquired from their extensive pre-training on multilingual corpora. Specifically:

DFTD1 transferred knowledge from fine-tuning Multilingual-BERT, an embodiment of the BERT architecture, which has pretrained deep bidirectional representations from unlabelled text by conditioning on both left and right context in all layers and serves as a universal transformer model that exhibits proficiency in various languages.

DFTD2, DFTD3, and DFTD4 transferred knowledge from fine-tuning XLM-RoBERTa-base, XLM-RoBERTa-large, and XLM-RoBERTa-XNLI, respectively. These models leveraged the knowledge learned by XLM-RoBERTa, a multilingual variant of Facebook's RoBERTa model released in 2019. It is a large multi-lingual language model, trained on 2.5TB of filtered CommonCrawl data on one hundred languages including Arabic.

The transfer learning process involves a meticulous fine-tuning procedure that was diligently undertaken. The pre-trained models were further trained on Arabic corpus using our Ara-deep dataset for our specific downstream task of deepfake text detection. In the fine-tuning process, we maintained the architecture and weights of the pre-trained model, incorporated a task-specific classification head for binary classification, and modified the model's parameters utilizing our Ara-deep dataset to tailor the pre-trained knowledge to our specific task (refer to Section 4.3.2 for the detailed fine-tuning configuration and implementation setup).

This methodological orchestration allowed for a thorough and insightful evaluation of the models' detection capabilities. In Section 4, we provide a detailed presentation of the experiments designed to assess the defensive performance of these detectors.

As shown in Figure 16, the process can be divided into three main phases. The first phase is dedicated to the task of dataset building, the second phase is for fine-tuning pretrained multilingual models, and the third phase is to study the human capacity for distinguishing between deepfake and genuine text. In general, the second phase is the same for the four employed models. After completing the three phases, results of both human and automatic detection are analytically compared and evaluated.

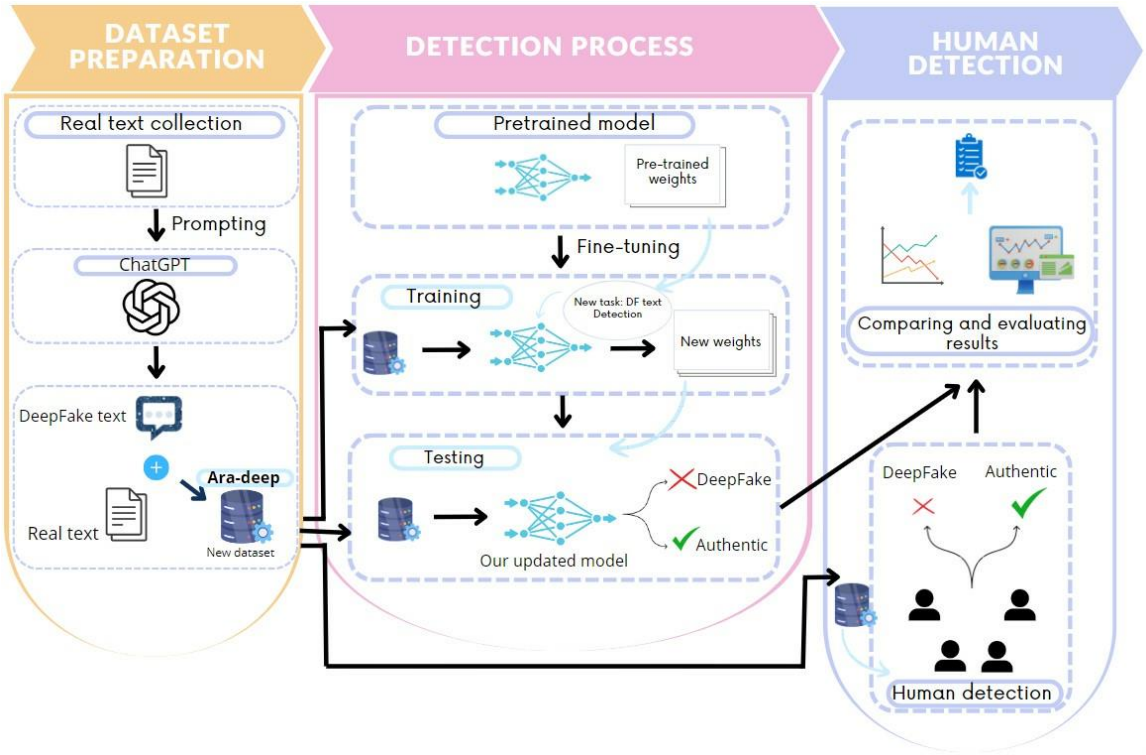


Figure 16. The overall architecture of the proposed model

The implementation details of the proposed detectors, including the specific fine-tuning hyperparameters (learning rate, number of epochs, batch size) and experimental results, are presented in the next section.

Experimental Settings and Results

In this section, we executed a number of experiments in order to evaluate the efficacy of our proposed models. To analyse the performance, an in-depth examination of the results yielded by each individual experiment is conducted. We carried out comparative experiments to evaluate the performance of the proposed models.

Baseline

Due to the absence of published baseline models for Arabic deepfake text recognition, this study constructed a temporal convolution network (TCN) architecture to serve as a baseline for comparison with the proposed models.

Experiment I

Objective of the Experiment

This experiment seeks to evaluate the efficacy of the constructed models in detecting Arabic deepfake text utilizing the developed dataset, specifically examining how effective are the designed models in identifying Arabic deepfake text using the developed gold standard dataset.

Setting the Experiment

As outlined earlier in Section 3.3, we set up our framework with four multilingual pre-trained models: (1) DFTD1 (Multilingual-BERT), (2) DFTD2 (XLM-ROBERTA-base), (3) DFTD3 (XLM-ROBERTA-large), and (4) DFTD4 (XLM-ROBERTA-large-XNLI). For all experiments, these models underwent optimization with AdamW (Adam with decoupled weight decay) [141], employing a learning rate of 2e-5. We conducted trials over 50 epochs since extending the training beyond this point did not significantly impact the training error. To expedite convergence and reduce memory consumption, a batch size of 10 was adopted during the training process.
Table 6 describes the combinations of hyperparameters used to train the four architectures. The proposed models and their variants are trained and tested utilizing the Ara-Deep Arabic dataset.

To avoid overfitting, we implemented early stopping based on validation performance. To demonstrate the model's capacity for generalization, we report findings on a hold-out validation set. To ensure reproducibility, we randomly divided the dataset into training and testing sets, allocating 80% for training and reserving the remaining 20% for testing. This partitioning remained consistent across all experiments.

During fine-tuning, we meticulously selected optimal hyperparameters through an extensive search encompassing combination of learning rates, batch sizes, and epoch numbers from the following sets {2e-5, 9e-5, 1e-4, 1e-3}, {10, 16, 32}, and {30, 50, 60}, respectively. All of the experiments are carried out using Python 3.10.12, transformers 4.33.1, with Cuda 11.8, and trained on a single NVIDIA Tesla T4 GPU.

Table 6. Hyperparameters used for training our DFTD1, DFTD2, DFTD3, & DFTD4 models

Hyperparameter	Value
optimizer	AdamW
learning rate	{2e-5, 9e-5, 1e-4, 1e-3}
epochs	{30, 50, 60}
batch size	{10, 16, 32}

Results

In this subsection, the results from the experimental investigation I are presented. In addition to the accuracy, the following metrics were also evaluated: Precision, Recall, and F1 score. The precision is the fraction of predicted human-written texts that are actually human-written. The recall is the fraction of actual human-written texts that are predicted to be human-written. The F1 score is the harmonic mean of precision and recall. Table 7 reports the main results of our approach and the comparison with the baseline TCN network’s results on the detection task. The accuracy, precision, recall, and F1 scores for each model are shown in Figure 17.

Table 7. Results and performance of the fine-tuned models Vs baseline

Model	Accuracy	F1-score	Precision	Recall
DFTD_1	0.9970	0.9970	0.9950	0.9990
DFTD_2	0.9980	0.9980	0.9980	0.9980
DFTD_3	0.9880	0.9881	0.9775	0.9990
DFTD_4	0.9975	0.9975	0.9970	0.9980
TCN	0.8938	0.9100	0.9000	0.9000

Figure 17 shows the main experimental results of our proposed models, and Table 8 describes in detail the results of the TCN network.

Table 8. Performance reports of baseline TCN on each class

Label	Precision	Recall	F1-Score	Accuracy
0: Real text	0.9900	0.7200	0.8400	0.8900
1: Deepfake text	0.8000	1.0000	0.8900	

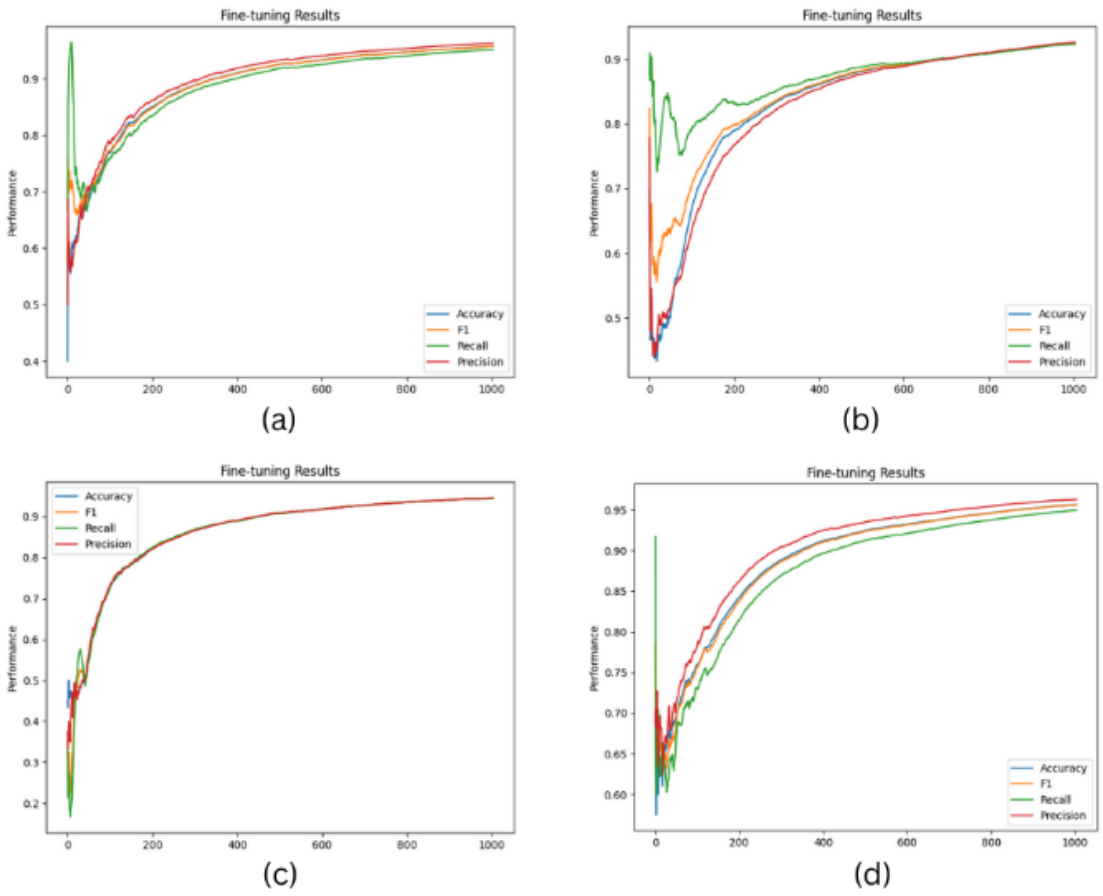


Figure 17. Performance of the fine-tuned models. Figures (a), (b), (c), and (d) represent the results of evaluating DFTD1, DFTD2, DFTD3, and DFTD4 respectively

Discussion

This subsection discusses the outcomes of the initial experimental inquiry. The comprehensive experimental results presented in Table 7 indicate that fine-tuning

strategies routinely outperform the model trained from scratch. This is attributable to the significantly higher size and increased number of parameters in large language models.

The differences in the performances of the LLMs and the baseline may stem from the varying capacities of each model to extract structural information from the text:

- DFTD1: Fine-tuned Multilingual-BERT (mBERT) (179M params): This model displayed outstanding accuracy of 99.70% on the test set. This is a further improvement over the DFTD3 model and suggests that the multilingual-BERT base model is able to learn to identify the principal linguistic features that are characteristic of deepfake text.
- DFTD2: Fine-tuned XLM-RoBERTa-base (279M params): This model achieved an accuracy of 99.80% on the test set. This is the highest accuracy achieved by any of the models. As can be seen from table 7, DFTD2 model achieved also the highest precision, and F1 scores. The DFTD2 model is able to learn to identify the most subtle linguistic features of human-written and deepfake text, and is therefore able to make the most accurate predictions.
- DFTD3: Fine-tuned XLM-RoBERTa-large (561M params): This model achieved an accuracy of 98.80% on the test set. This is a significant improvement over the baseline accuracy of 91.0%, illustrating the models adeptness in the detection task. The model has demonstrated its aptitude for navigating the challenges presented by managing complex linguistics of AI-generated text.
- DFTD4: Fine-tuned XLM-RoBERTa-large-XNLI (561M params): The performance of DFTD4, which boasted a 99.75% accuracy rate on the test set, was consistent with its enhanced architecture. This model showcased heightened performance, effectively capturing the patterns of ChatGPT-generated text.

The performance analysis of multilingual language models revealed an impressive sensitivity in identifying text produced by ChatGPT. With accuracies rates over 0.98, their cross-lingual prowess underscored the potential as robust detectors in the Arabic language. The culmination of these findings delineates the efficacy of the four models in detecting text originating from ChatGPT. The precision of their classifications highlights their potential for contextually nuanced detection tasks. Analysing the performance of each model, the findings show the nuanced interplay between model architecture, depth, and its performance.

The enlarged architecture of DFTD3 is reflected in its higher sensitivity, supporting the claim that increased model complexity is associated with a stronger ability to grasp contextual complexities. However, DFTD2 outperformed DFTD3 although its architecture is smaller. DFTD2 revealed its high aptitude for accurate text detection by being the best performing model with an accuracy of 99.80%.

DFTD1’s remarkable accuracy highlights its innate ability to recognize a wide range of linguistic variations. The fact that DFTD1 (based on the 179M params Multilingual-

BERT) performed better than the much larger DFTD3 was an intriguing outcome. DFTD1 was outperformed by DFTD4 in terms of overall performance.

Experimental findings demonstrate that all our developed models outperform the baseline method.

Experiment II

Objective of the Experiment

The aim of this assessment phase is twofold: first, to assess the models' performance, and second, to examine the influence of differing input parameters on decision-making, performance, and the capacity to generalize effectively to unseen examples from diverse domains, sources, and LLMs.

Setting the Experiment

As Figure 18 illustrates, this experiment is designed to examine whether the modification of the input X, the internal architecture of Unit A (the auto-text generation unit), or the output Y generated by Unit A (which serves as the input for Unit B, the detection unit) have a significant influence on the decision processes and overall performance metrics of Unit B (our proposed models). The input X represents the initial data (prompt + text) fed into the LLM, while Unit A is responsible for generating deepfake text as its output (Y) based on its internal parameters. Unit B, conversely, which encompasses our four proposed detectors: DFTD1, DFTD2, DFTD3, and DFTD4, is responsible for analyzing the outputs from Unit A to determine if they are human-produced or AI-generated.

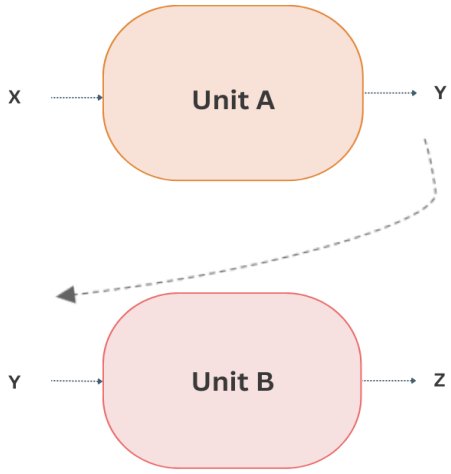


Figure 18. Rational underpinning experiment II

To validate the performance of our proposed models comprehensively, further evaluation on three different datasets were conducted. These datasets permitted a systematic manipulation of the key variables (X, Unit A, and Y), enabling a comprehensive analysis of their individual and combined effects on the models' behaviour. Specifically, we used:

- M4 dataset: introduced by Wang et al. [129], which serves as a substantial benchmark for machine-generated text detection. Notably, the dataset incorporates a diverse array of examples, ensuring a robust evaluation. In our analysis, we utilised Arabic text generated by ChatGPT. The dataset used for evaluation contains both human text and AI-generated text.
- LLM Question-Answer Dataset : includes AI-generated texts and prompts produced in 32 different languages by Large Language Models (LLMs). The model is given prompts to generate text. The length and complexity of the writings that the LLM produced in response to these prompts are varied. The dataset used for evaluation consists of Arabic deepfake text generated using the models GPT-3.5, and GPT-4.
- BLOOM dataset: We created another dataset for evaluating our models using the BLOOM large language model. The dataset contains both human text and deepfake text.

Results

The results of our assessment, displayed in Tables 8, 9, and 10, validate the encouraging findings of our proposed models. In the M4 and LLM Question-Answer Datasets, DFTD1 surpassed other models in accuracy, precision, and F1-score, attaining scores of 99.65%, 99.60%, 99.65%, and 99.80%, 100%, and 99.79% respectively. DFTD3 surpasses the other models in both datasets for F1-score, achieving 99.90%.

Table 9. Evaluation results on the M4 dataset

	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9965	0.9960	0.9980	0.9965
DFTD_2	0.9940	0.9679	0.9970	0.9940
DFTD_3	0.9810	0.9819	0.9800	0.9810
DFTD_4	0.9995	1.0000	0.9990	0.9995

Table 10. Evaluation results on LLM Question-Answer dataset

	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9980	1.0000	0.9960	0.9979
DFTD_2	0.9935	0.9900	0.9970	0.9935
DFTD_3	0.9410	0.8952	0.9990	0.9442
DFTD_4	0.9900	0.9832	0.9970	0.9900

Table 11. Evaluation results on BLOOM dataset

Model	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9935	0.9900	0.9970	0.9935
DFTD_2	0.9940	0.9900	0.9980	0.9940
DFTD_3	0.9950	0.9910	0.9990	0.9950
DFTD_4	0.9555	0.9198	0.9980	0.9573

Discussion

The results underscore their efficacy in not only analysing machine-generated text from diverse sources but also in demonstrating a capacity to generalize proficiently across different domains and LLMs.

According to the logic illustrated in Fig. 8, the assessment of M4 facilitated the validation of the model's performance when Unit A remains constant (Chat-GPT), while X and Y are altered in terms of source and domain. The assessment of the LLM Question-Answer and BLOOM Datasets facilitated the validation of the models' performance when all three critical factors: Unit A (GPT-3.5, GPT-4, and BLOOM), X, and Y deviate from those encountered during training. This experiment demonstrates that the suggested models can effectively generalize to fresh data from diverse areas and sources supplied by various LLMs. This performance across the three evaluation datasets indicates the models' adaptability and highlights their utility.

Experiment III

Objective of the Experiment

The aim of this experiment is to investigate the performance of our proposed model in comparison to other baselines and state-of-the-art models in terms of key metrics such as precision, recall, F1-score, and accuracy.

Setting the Experiment

Our best performing model in term of accuracy, precision, and F1-score, DFTD4 model, was used in this experiment.

To the best of our knowledge, there are no detectors specifically trained for Arabic deepfake detection. However, a recent study by Wang et al. [129] focused on detecting

ChatGPT-generated text using a diverse dataset that includes multiple languages, including Arabic. Although the study did not explicitly specify detection performance for each language, it provides valuable insights. In our experiment, we compare our top-performing model with the detectors used in the previously mentioned study. This comparison is based on the same dataset M4, which contains Wikipedia of ChatGPT vs Human data, and closely aligns with our custom dataset used for training and testing our models.

Results

In this subsection, the results from the experimental investigation III are presented. The comparison results between our top-performing model (DFTD4) with the detectors from [129] in term of accuracy, precision, recall, and F1-score are shown in the following table:

Table 12. Comparison of our best performing model on our test data with other state-of-the-art detectors of ChatGPT generations

Model	Accuracy	Precision	Recall	F1_Score
RoBERTa	0.9970	0.9940	1.0000	0.9970
LR-GLTR	0.9740	0.9760	0.9720	0.9740
Stylistic	0.9740	0.9760	0.9720	0.9740
NELA	0.9560	0.9670	0.9430	0.9550
DFTD_4	**0.9995**	**1.0000**	**0.9995**	**0.9995**

Discussion

Overall, our XLM-RoBERTa-large-XNLI based model, DFTD4, outperforms the four ChatGPT detectors, the evaluation quantifying the performance of our model relative to state-of-the-art models is encapsulated in Table 12.

Among the existing state-of-the-art models, the highest performance was achieved by the model based on the RoBERTa architecture, yielding an accuracy of 99.70%, as reported in [129].

A scrutiny of the results presented in Table 12 reveals that our proposed methodology exhibits superior performance and efficacy, as evinced by the higher accuracy of 99.95%, coupled with an impressive precision and F1-score of 100% and 99.95% respectively, when juxtaposed with the existing state-of-the-art approaches.

4.2. Human Baseline Study

As online content undergoes a profound transformation driven by the widespread proliferation and the unprecedented accessibility of deepfake text, internet and social media users are increasingly exposed to synthetically generated textual information. In light of this, it becomes imperative to undertake a critical assessment of human capabilities in discerning Arabic deepfake text. This experiment was designed to tackle this problem by comparing human abilities for detection against auto-detection with the primary objective of this evaluative examination being to assess the judgment abilities of typical internet users. The research question addressed here is RQ4: How accurately can internet users distinguish between authentic and deepfake textual content, and how reliable are their judgments?

Study Design

To conduct this examination in this experiment, a comprehensive annotation procedure was meticulously crafted, involving the participation of a cohort comprising eight individuals whose native language is Arabic. These human annotators were deliberately devoid of any specialized training in the domain of deepfake text detection. The annotators were assigned distinct subsets, each comprising 20 data instances drawn from the Ara-Deep dataset, their task encompassed the identification of both deepfake and authentic text examples within their allocated samples.

4.3. Results and Analysis

Upon collating the feedback provided by all these human annotators, the computed detection accuracy exhibited a mere 51.00%, which was closely resembling random classification with little discriminatory capacity evident in their assessment of text veracity. Figure 19 shows a visualisation of human performance in the detection.

ROC Curve: As the classification decision threshold changes, the ROC curve shows the trade-off between the true positive rate (sensitivity) and the false positive rate. It aids in the visualization of human annotators' performance at various operating points.

AUC: The AUC measures how well human annotators perform overall. It stands for the ROC curve's area under the curve. Better discriminative power is indicated by a larger AUC, with a perfect classifier having an AUC of 1.

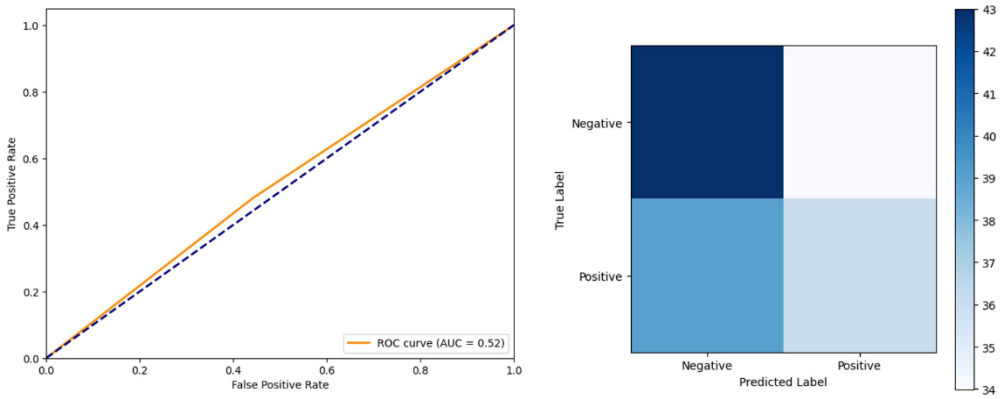


Figure 19. Visualisation of human performance in the annotation process

The results of human annotation on the Ara-Deep dataset were compared to the gold standard dataset that consists of the true-labelled instances, reflecting the ground truth classifications. Results are illustrated on Table 13 and Figure 20. Table 13 Compares the results of different models used with human performance in terms of the average accuracy, precision, recall, and F1-score values. The best results are highlighted in bold.

Table 13. Evaluation results of several deepfake text detectors in terms of accuracy, precision, recall, and F1-score

Detector	Accuracy	precision	Recall	F1-score
DFTD_1	0.9970	0.9970	0.9950	0.9990
DFTD_2	0.9980	0.9980	0.9980	0.9980
DFTD_3	0.9880	0.9881	0.9775	0.9990
DFTD_4	0.9975	0.9975	0.9970	0.9980
Human annotators	0.5200	0.5100	0.4800	0.5000

4.4. Discussion

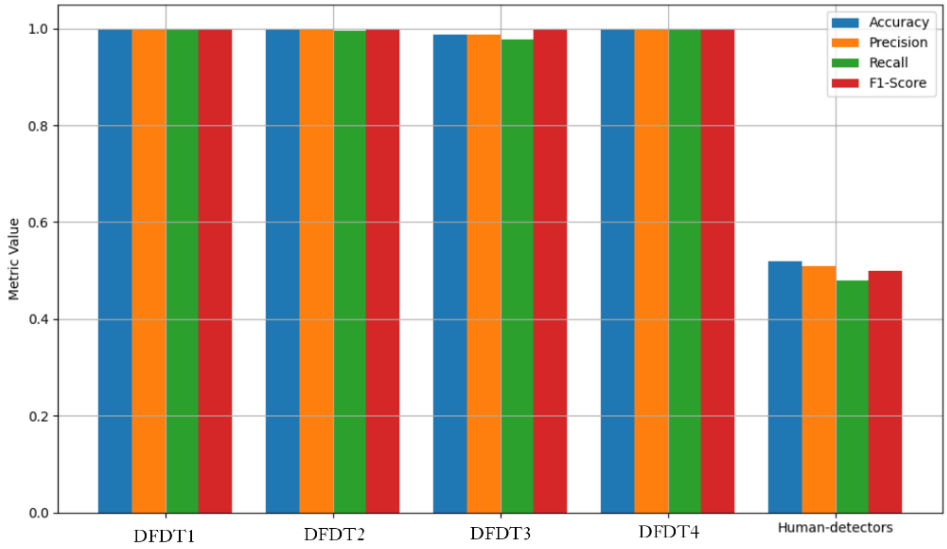


Figure 20. Performance comparisons of LLM-based deepfake text detection models with human detection results

The huge difference between the efficacies of the aforementioned models with human detection abilities unveils critical concerns within the realm of large language models' text detection. The comprehensive analysis of the four models —Multilingual-BERT, XLM-RoBERTa-base, XLM-RoBERTa-large, and XLM-RoBERTa-large-XNLI— evinces their remarkable prowess in identifying text generated by such models. The models' capability to harness intricate contextual relationships and cross-lingual comprehension stands out as a testament to the advancements in DL technology. This suggests that the models are able to learn how to identify the linguistic features of human-written text more effectively than humans.

Nevertheless, the comparison with human detection elicits multifaceted considerations. It is important to recognize that LLM text easily tricks the human cognitive capacity for contextual understanding and nuanced interpretation, even with the human's capability to make subtle inferences beyond the immediate linguistic constructs. However, the models' performance, as evidenced by their high accuracy rates, mirrors their superiority and proves the importance of automating the deepfake text detection tasks.

Furthermore, while models excel in processing voluminous data and executing repetitive tasks with speed and precision, human intuition and domain knowledge fall short in deciphering complex linguistic nuances and ambiguities between real and coherent LLM

generated text when it comes to large amounts of data, especially in lengthy text samples with no syntax or grammar errors.

The LLMs' superior performance over humans may have been influenced by a variety of reasons. The models can process text far more quickly than people can. They can therefore quickly analyse big amounts of data. The second advantage is that the models can also gain knowledge from a large body of data. This indicates that they are able to spot patterns that people would miss. Finally, the models can be fine-tuned on certain datasets, which enables them to be improved on text generated by a particular language model or text relevant to a given area.

The comparison's findings imply that the proposed models are a promising technique for identifying text produced by LLMs. It is crucial to note that the models are not flawless. Nonetheless, they are susceptible to errors, especially when the text is well-written or has a lot of noise. Besides, the quality of the models depends on the data which they were trained on. The models cannot output reliable results if the data is not indicative of reality. Overall, this comparison's findings are encouraging. They assert that the models can be used to accurately identify text produced by ChatGPT. However, to increase the models' precision and provide more reliable detection systems, additional study is necessary. Table 14 shows the detection result of each model with the average of human detection on two distinct samples from the two classes (Real and Deepfake text). These instances were extracted from the test set.

Table 14. Analysis of typical sample cases and their classification results

	Text	Method	Prediction	Label
Example 01	يسعى بنك فالكون الذي تملكه أبوظبي ويتخذ من سويسرا مقراً له إلى صفقات استحواذ، حيث يتوقع عملية اندماجات في الصناعة المصرفية، وفق ما قاله أريك فيستر، رئيس الصيرفة الخاصة في البنك. وأضاف فيستر الذي تولى منصبه في يناير الماضي، أن البنك سوف يكون مستعداً لعقد صفقات استحواذ تغطي أسواق نشاطه الأساسية المستهدفة في الخليج وأوروبا الشرقية وإفريقيا وجنوب شرق آسيا، بحسب صحيفة "البيان" الإماراتية. وأوضح فيستر أن البنك يهتم بالصفقات التي تشمل إدارة البنوك التي تتراوح بين 200 مليون فرانك وملياري فرانك سويسري من حيث قيمة الأصول. وقال فيستر في مقابلة في دبي: "لا بد أن تخدم الصفقات أهدافنا وما نركز عليه، ما يعني أننا لن نتساهل في استراتيجيتنا، ولا بد أن تكون أي صفقة مستهدفة متماشية مع أسواقنا الرئيسة".	**DFTD_1**	Real-text	Real-text
		DFTD_2	Real-text	
		DFTD_3	Real-text	
		DFTD_4	Real-text	
		Human-annotators	Deepfake-text	
Example 02	في الوقت الذي تكثر فيه بعض المعتقدات الشعبية "الغريبة" عن العطس، يجب أن نفهم الحقائق العلمية حول هذه الظاهرة الطبيعية. العطس هو استجابة غير إرادية لجسمنا للتخلص من المهيجات التي تدخل أجسامنا، مثل الغبار أو الحبوب اللقاحية.على الرغم من أن العطس يعتبر طبيعيًا وشائعًا، إلا أن بعض المعتقدات الشعبية تنتشر حوله. ومعظم هذه المعتقدات ليس لها أساس علمي وتعتبر مجرد خرافات. على سبيل المثال، هناك من يعتقد أن العطس يمكن أن يؤدي إلى تشوه الوجه أو أنه يعني وجود أحد الأرواح الشريرة في الجسم. هذه الافتراضات لا تمت للواقع بصلة ولا تستند إلى أي دليل علمي.الحقيقة العلمية هي أن العطس ليس خطيرًا ولا يتسبب عادة في مشاكل صحية جوهرية. ومع ذلك، يجب أن نتخذ بعض الاحتياطات الأساسية أثناء العطس لمنع انتشار الجراثيم والحفاظ على	**DFTD_1**	Deepfake-text	Deepfake-text
		DFTD_2	Real-text	
		DFTD_3	Deepfake-text	

النظافة الشخصية. يُنصح بتغطية الفم والأنف بمنديل ورقي أو مرفق الكوع عند العطس، وذلك لمنع انتشار الجسيمات المتطايرة وتقليل انتقال الأمراض المعدية.إذا كانت لديك مخاوف مرتبطة بالعطس أو لاحظت تغيرات غير طبيعية في عادات العطس الخاصة بك، فمن الأفضل التشاور مع مقدم الرعاية الصحية. يمكنه تقييم حالتك وتوفير المشورة المناسبة بناءً على أعراضك وتاريخك الصحي.	**DFTD_4**	Deepfake-text
	Human-annotators	Real-text

4.5. Chapter Summary

This chapter has provided compelling empirical evidence demonstrating that automated systems can substantially outperform human evaluators in detecting AI-generated news articles, establishing a critical foundation for understanding the current state of synthetic text detection capabilities. The comprehensive evaluation revealed that fine-tuned language models achieved significantly higher performance metrics compared to the accuracy levels observed in human participants, thereby highlighting fundamental limitations of unaided human judgment when confronting sophisticated synthetic text generation systems.

The chapter's findings are grounded in rigorous experimental methodology employing adaptive fine-tuning strategies specifically developed for Arabic text detection. The research implemented a comprehensive detection framework utilizing four distinct transformer-based models, including mBERT, XLM-RoBERTa-base, XLM-RoBERTa-large, and XLM-RoBERTa-large-XNLI, each subjected to extensive hyperparameter optimization and performance evaluation protocols.

The development of the Ara-Deep dataset represents a significant methodological contribution, providing a robust foundation for training and evaluation that addresses the previously identified gap in Arabic-language detection resources. This dataset creation effort ensures that the experimental findings are based on linguistically representative content that captures the complexity and nuance of Arabic text generation scenarios.

The chapter presents particularly noteworthy findings regarding the substantial performance gap between human evaluators and automated detection systems. The empirical evidence reveals that regular internet users without specialized training frequently struggle to identify ChatGPT-generated text, often failing to recognize sophisticated synthetic content that exhibits high stylistic coherence and linguistic fluency. This human performance limitation can be attributed to the remarkable advancement in modern generative models, which have been optimized to produce text that closely mimics human writing patterns and stylistic conventions. The susceptibility of human evaluators to this stylistic coherence represents a critical vulnerability in manual content verification processes, particularly in contexts where rapid identification of synthetic content is essential for maintaining information integrity.

The comparative analysis between human and automated detection capabilities underscores the fundamental importance of developing robust language model-based detection systems. While human evaluators demonstrate consistent difficulty in

distinguishing authentic from AI-generated content, the implemented transformer-based models prove highly effective in this classification task, achieving detection accuracies that far exceed human performance levels.

The research findings carry profound implications for addressing the proliferation of misinformation and disinformation, particularly in multilingual digital environments where synthetic content can be deployed to manipulate public discourse. The demonstrated effectiveness of automated detection systems provides a viable technological solution for identifying potentially harmful AI-generated content before it achieves widespread distribution across social media platforms and news aggregation services.

The linguistic specificity of the developed detection framework addresses a critical gap in current misinformation mitigation strategies, which have historically focused on English-language content while neglecting other major world languages. The successful implementation of Arabic-specific detection capabilities establishes a methodological precedent for developing similar systems across other under-resourced languages, thereby expanding the global reach of automated content verification systems.

The chapter's emphasis on language-specific detection tools reflects the recognition that effective misinformation mitigation requires nuanced understanding of linguistic and cultural contexts. Generic detection approaches often fail to capture the subtle linguistic markers that characterize synthetic text in morphologically complex languages, necessitating the development of targeted solutions such as those presented in this research.

The empirical evidence presented establishes automated AI-driven detection systems as not only feasible but demonstrably superior to human evaluation methods in controlled experimental conditions. This validation provides essential groundwork for extending these detection capabilities beyond news media into other critical domains where synthetic content poses significant risks to institutional integrity and public trust.

The methodological framework developed for news article detection serves as a robust foundation for subsequent investigations into academic content verification, where similar principles can be applied to address emerging challenges related to AI-generated student submissions and academic misconduct. The transferability of the core detection approach across different content types demonstrates the versatility and scalability of transformer-based detection systems.

The comprehensive evaluation methodology employed in this chapter establishes clear protocols for assessing detection system performance across diverse linguistic and contextual scenarios. These evaluation frameworks will prove essential for validating the effectiveness of detection systems as they are adapted for application in educational environments and other specialized domains.

The chapter's contributions extend beyond immediate practical applications to encompass broader academic discourse on large language model performance, adaptability, and societal impact. The research advances theoretical understanding of synthetic text detection

while simultaneously providing concrete solutions for addressing real-world challenges posed by increasingly sophisticated content generation systems.

The comprehensive model evaluation serves as a methodological stepping stone toward developing more robust and nuanced language understanding applications. The insights generated through this research contribute valuable knowledge to the academic community while establishing practical frameworks for implementing detection systems in operational environments.

The research facilitates comprehensive and discerning evaluation of detection capabilities, contributing meaningfully to ongoing academic discourse on language-model-generated content identification and verification. This contribution proves particularly valuable as the field continues to evolve in response to rapid advancements in generative AI technologies.

Having conclusively established the efficacy of automated detection systems in news media contexts, the thesis now transitions to examining similar challenges within educational environments. The demonstrated superiority of machine-based detection over human evaluation provides strong justification for implementing automated systems in academic integrity enforcement, where the stakes of undetected synthetic content are equally significant.

The methodological precedent established through news article detection research provides a solid foundation for addressing academic integrity challenges posed by AI-generated student essays and assignments. The principles validated in this chapter will be systematically extended and adapted to accommodate the unique characteristics and requirements of educational content verification, ensuring comprehensive coverage of synthetic text detection across multiple critical application domains.

CHAPTER V

Dual-Method Approach for Student Essay Authentication

Building on the methodologies and findings presented in the preceding chapter, which focused on detecting AI-generated news articles in Arabic, this chapter shifts the analytical lens to the educational domain. While news media and education both face threats from synthetic text, the latter introduces unique challenges, such as the need to preserve academic integrity and evaluate student originality. The previous chapter established the efficacy of automated systems in detecting synthetic content in journalistic contexts; however, educational settings demand tailored approaches due to the stylistic diversity of student writing, the ethical imperative to safeguard learning outcomes, and the need for interpretability in academic settings. This chapter addresses these nuances by proposing and evaluating two distinct methods for distinguishing AI-generated student essays from authentic ones.

The aim of this chapter is twofold: first, to evaluate the performance of fine-tuned pretrained language models (LLMs) in detecting AI-generated essays, and second, to explore the viability of a retrieval-based approach combined with machine learning classifier for the same task. By comparing the two approaches, this chapter advances the thesis by identifying scalable, domain-specific solutions for maintaining academic integrity in an era of ubiquitous generative AI.

5.1. Methodology

This section outlines the specifics of our proposed methods illustrated in Figure 21, where we utilise a binary classification framework to differentiate between answer passages produced by actual students ("0") and those created by AI/machine ("1").

The first detection approach involves fine-tuning pretrained LLMs. Considering the proven capabilities of BERT as a leading model for tackling diverse NLP tasks with remarkable efficiency, we implemented fine-tuning techniques on CAMeLBERT models for Modern Standard Arabic (MSA) [137], which is a set of BERT-based models that had been previously trained on Arabic data.

Our second detection method utilises an approach inspired by RAG by implementing RAG principles and utilizing them in more straightforward models. The method deviates from the conventional RAG methodology by incorporating the retrieval component of RAG while streamlining the "augmentation" and "generation" phases.

Sections 5.1 and 5.2 provide a comprehensive overview of the two methods. The performance of our techniques for recognizing essays generated by the two leading generation models, ChatGPT and Gemini, is evaluated through a series of experiments.

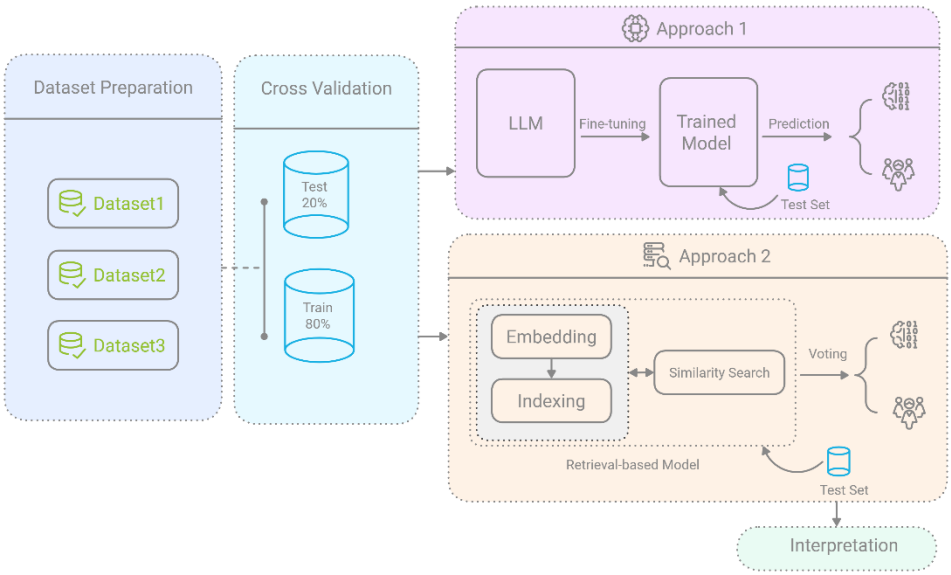


Figure 21. Overall workflow of the approaches used for AI-generated essays detection

Datasets

AI-generated Data:

To provide a wider representation of AI writing styles, in this study, we created two different datasets specifically curated to benchmark the detection of AI-generated text in the context of student essays, containing auto-generated essays produced by the two state-of-the-art LLMs, i.e. ChatGPT and Gemini. The LLMs-generated answers were intended to emulate the intricacy and style of student writing.

For each dataset, we included AI-generated answers and student essays to 48 questions in computer science, ensuring an equal number of examples per class in each dataset. The used datasets were split into a training set (80%) and a test set (20%) to assess the model performance, and we used cross-validation with folds number = 5.
A summary of the datasets used in our study is given in Table 16, while Table 17 shows examples extracted from our datasets representing: real student essay, ChatGPT generated text, and Gemini generated text with the prompt question used to generate them.

Human Data:

For real student essays, we collect samples from AR-ASAG (ARabic dataset for Automatic Short Answer Grading Evaluation V1. ISLRN 529-005-230-448-6) dataset, c.f. [100], these were included in both datasets. In the AR-ASAG, responses to three separate exams given to three different student classes are included. Exams were administered in an evaluation-friendly environment. There are sixteen short answer questions on each test (a total of 48 questions). Students submitted their responses to these questions; a sample answer is suggested for each question. The total number of responses varies depending on the question.

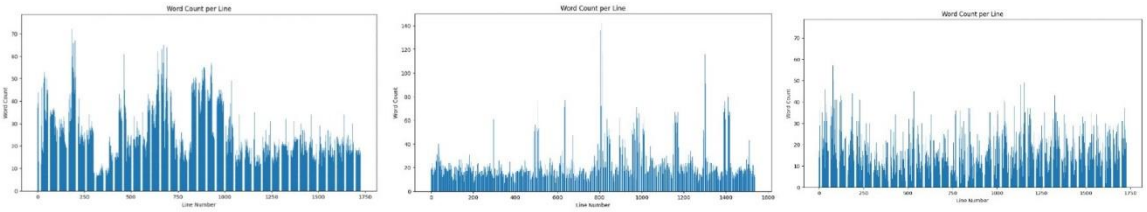
There are five different categories of questions in the dataset as outlined in Table 15. Table 16 shows the text distribution within each of the three datasets: dataset 1 (human written + ChatGPT-generated essays), dataset 2 (human written + Gemini-generated essays), and dataset 3 (containing human written + ChatGPT-generated + Gemini-generated essays). Table 17 compares between three examples from the three different sources, generated in response to the given prompt.

Table 15. Different question types included in the dataset

	Question type
1	• "عرف ": Define?
2	• "إشرح": Explain?
3	• "ما النتائج المترتبة على": What consequences?
4	• "علل": Justify?
5	• "ما الفرق": What is the difference?

Figures 22 represents visualizations on word count per line. Figure (22.a) illustrates that ChatGPT's generated essays exhibit more regular oscillation with higher average word counts, primarily ranging between 20-40 words, and appear to maintain the most consistent upper bound. Figure (22.b) displays that Gemini's generated essays show more sporadic spikes with generally lower baseline counts and shows obvious outliers of word

counts in each line of the generated-essays, while Figure (22.c) exhibits variation that is more natural as typical human writing patterns. This figure suggests that both models exhibit distinct "signatures" that differentiate them from human-authored essays.



(a) (b) (c)

Figure 22. Word count per line in each category of essays: (a) ChatGPT-generated text,

(b) Gemini-generated text, and (c) real student-authored essays

Table 16. Data distribution

Datsets	**Number of instances**		
	ChatGPT essays	real student essays	Gemini
Dataset1 (ChatGPT-essays-dataset)	1727	1727	\
Dataset2 (Gemini-essays-dataset)	\	1539	1539
Dataset3 (ChatGPT+Gemini-essays-dataset)	1539	3078	1539

Table 17. An example comparison between real student essays, ChatGPT generated text, and Gemini generated text. We used the general form of prompt as illustrated bellow with different questions

Prompt: Write a short essay in the style of a high school student in response to the following question [question= Define the term 'Information Security'].		
Human Real Essays	**ChatGPT-written Text**	**Gemini-written Text**
هو حماية وتامين المعلومات من التخريب (سرقة، تغيير)، نظام المعلومات حيث كل منشأة تقوم بحماية نفسها وحماية الموظفين العاملين عليها وحماية الأجهزة والحاسبات الآلية وحماية البيانات المخزنة فيها في كل الأوساط من المخترقين وأي وسيلة تخريب في مختلف مراحل المعلومة (تخزين نقل معالجة). It is protecting and securing information from sabotage (theft, alteration), the information system where every facility protects itself, protects the employees working on it, protects the devices and computers, and protects the data stored in them in all circles from hackers and any means of sabotage at various stages of the information (storage, transmission, processing).	أمان المعلومات يشير إلى مجموعة من التدابير والسياسات والتقنيات التي تهدف إلى حماية سرية وسلامة وتوفر الوصول للمعلومات. يتضمن ذلك تأمين البيانات من الاختراق، والتحقق من الهوية، وتشفير البيانات، وتطبيق سياسات الوصول الصارمة لمنع الوصول غير المصرح به. Information security refers to a set of measures, policies, and technologies aimed at protecting the confidentiality, integrity, and availability of information. This includes securing data from hacking, verifying identity, encrypting data, and enforcing strict access policies to prevent unauthorized access.	يُعرَّف أمن المعلومات ببساطة على أنه مجموعة من الممارسات والتقنيات التي تهدف إلى حماية المعلومات من الوصول غير المصرح به أو الاستخدام أو الكشف أو التغيير أو التدمير. ويشمل ذلك حماية البيانات والبرامج والأجهزة والشبكات من الهجمات الإلكترونية والتهديدات الأخرى. Information security is simply defined as a set of practices and technologies aimed at protecting information from unauthorized access, use, disclosure, alteration or destruction. This includes protecting data, software, hardware and networks from cyberattacks and other threats.

Figure 23 depicts the process of how the three datasets were created and Figure 24 highlights their readability of two generated-essay examples.

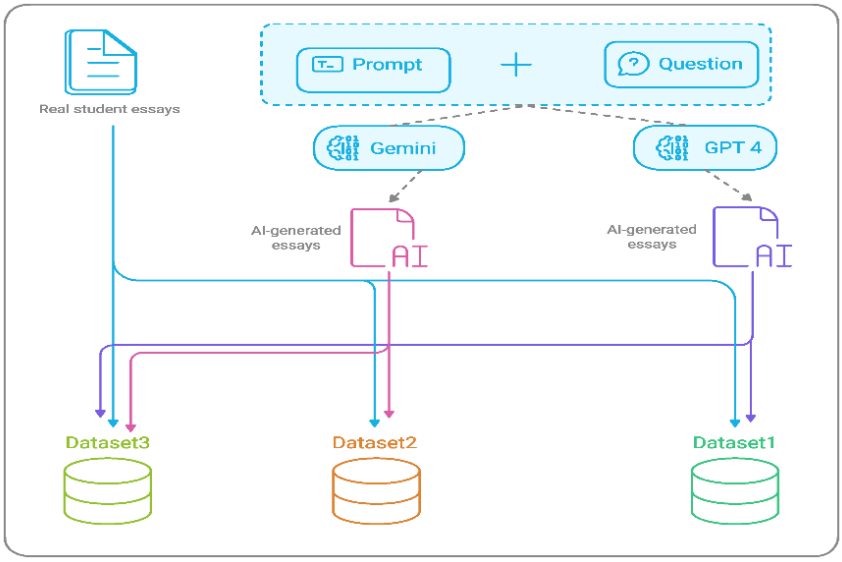


Figure 23. General schematic diagram of the three datasets creation

Figure 24. Two samples of the readability of the generated essays

Fine-tuning Approach

We use four state-of-the-art large language models, each pretrained on Arabic corpora. The fine-tuning process involves adjusting the models' parameters to optimize their detection capabilities for the nuances of student essays. This process is incremental, with multiple rounds of training and validation to ensure the models' robustness and reliability in identifying AI-generated text. Table 18 summarizes the configurations of the original models used.

Namely, we introduce FEDM1 (Fake Essays Detection Model 1), FEDM2, FEDM3, and FEDM4, based on fine-tuning CAMeLBERT models, which were pre-trained on Arabic texts with different sizes and variants. Our models are the following:

FEDM1: Based on fine-tuning CAMeLBERT-MSA-half (bert-base-arabic-camelbert-msa-half), a model that has undergone pre-training using half of the complete MSA dataset.

FEDM2: Based on fine-tuning CAMeLBERT-MSA-quarter (bert-base-arabic-camelbert-msa-quarter), a model that has undergone pre-training using a quarter of the complete MSA dataseta model pre-trained on a quarter of the full MSA dataset.

FEDM3: Based on fine-tuning CAMeLBERT-MSA-eighth (bert-base-arabic-camelbert-msa-eighth), a model that has undergone pre-training using one-eighth of the complete MSA dataset

FEDM4: Based on fine-tuning CAMeLBERT-MSA-sixteenth (bert-base-arabic-camelbert-msa-sixteenth), a model that has undergone pre-training on one-sixteenth of the complete MSA dataset.

Table 18. Configurations of CAMeLBERT models used

Model	**Variant**	**Size**	**#Word**
bert-base-arabic-camelbert-msa-half	MSA	53GB	6.3B
bert-base-arabic-camelbert-msa-quarter	MSA	27GB	3.1B
bert-base-arabic-camelbert-msa-eighth	MSA	14GB	1.6B
bert-base-arabic-camelbert-msa-sixteen	MSA	6GB	746M

Retrieval-based Approach

Our second method focuses on identifying AI-generated essays in an accurate, scalable and explainable approach. For that, we propose a retrieval-based detection model inspired by some RAG (which stands for Retrieval Augmented Generation) principles. RAG was first presented in this work [25], which aimed to handle both hallucinations and out-of-date knowledge. In [25], the authors suggested pairing a pre-trained seq2seq generative model with a pre-trained retriever, and shown how the generative model performs better when both are combined. Although RAG (Retrieval-Augmented Generation) was initially developed to improve the performance of LLMs through the integration of pertinent external information, we utilise RAG indexing and retrieval principles for classification purposes. In its purest form, RAG involves using an LLM for the final task, while we use a simpler ML Algorithm: K-Nearest-Neighbour for the final classification task. Our proposed method, RBC (Retrieval-Based Classification), includes the following steps:

1. Data Collection, Cleaning, and Preprocessing:

In this phase, we use our same datasets described in section 4 where we compiled our pertinent essays, both student-authored and AI-generated, eliminated any irrelevant or redundant information, standardized the text format by removing special characters and diacritics, and addressed any missing data or inconsistencies.

2. Chunking:

In this step, we analyse the gathered data to facilitate storage in chunks instead of complete documents. Ensuring that every segment provides sufficient context to stand alone effectively.

3. Embedding Generation:

Let $\phi : e \rightarrow \mathbb{R}^d$be an embedding function that maps Arabic essay to a d-dimensional dense vector space. Here we use AraBERT, a pre-trained language model on Arabic, to define ϕ, for generating vector representations of each chunk. For each essay e_i, we compute its embedding as: $x_i = \phi(e_i)$
The knowledge base is structured as a vector database. Each entry in the database consists of a text snippet and its corresponding vector embedding.

4. Index Construction:

This step facilitates the development of an effective index for similarity search, enabling rapid access to pertinent information throughout the classification process.
We construct an index $\mathcal{I}$ of all essay embeddings in the training set: $\mathcal{I} = (x_i, y_i)|(e_i, y_i) \in \mathcal{D}$. We used FAISS (Facebook AI Similarity Search) [99] to create our index.

5. Retrieval:

In this step, we retrieve the relevant information given a query essay q, we define a similarity function as follows:

$$sim : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}.$$
$$\mathcal{I}: Nk(q) = argmaxksim(\phi(q), x_i)|(x_i, y_i) \in \mathcal{I} \quad (14)$$

6. Classification:

For the classification layer, we employ a k-Nearest Neighbours (k-NN) classifier. The predicted class $\hat{y}$ for query q is determined by majority voting among the retrieved k neighbours $Nk(q)$:

$$\hat{y} = mode(y_i)|(e_i, y_i) \in Nk(q) \quad (15)$$

7. Performance Evaluation:

We assess the model's performance using cross-validation with 5 folds. Let $\mathcal{D}$ be partitioned into 5 sub-sets: $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$,and $\mathcal{D}_5$. For each fold i, we define:

$$f_i = train(\mathcal{D} \setminus \mathcal{D}_i) \quad (16)$$

$$score_i = evaluate(f_i, \mathcal{D}_i) \quad (17)$$

The overall performance is then estimated as:

$$performance = \frac{1}{5}\sum_{1}^{5} score_i \quad (18)$$

This proposed approach, as can be seen in Figure 25, blends the simplicity and interpretability of k-NN algorithm with the advantages of dense retrieval, exploiting AraBERT's semantic understanding capabilities. The retrieval step permits effective scalability to huge datasets, while the classification stage facilitates simple decision-making based on small neighbourhood patterns in the embedding space.

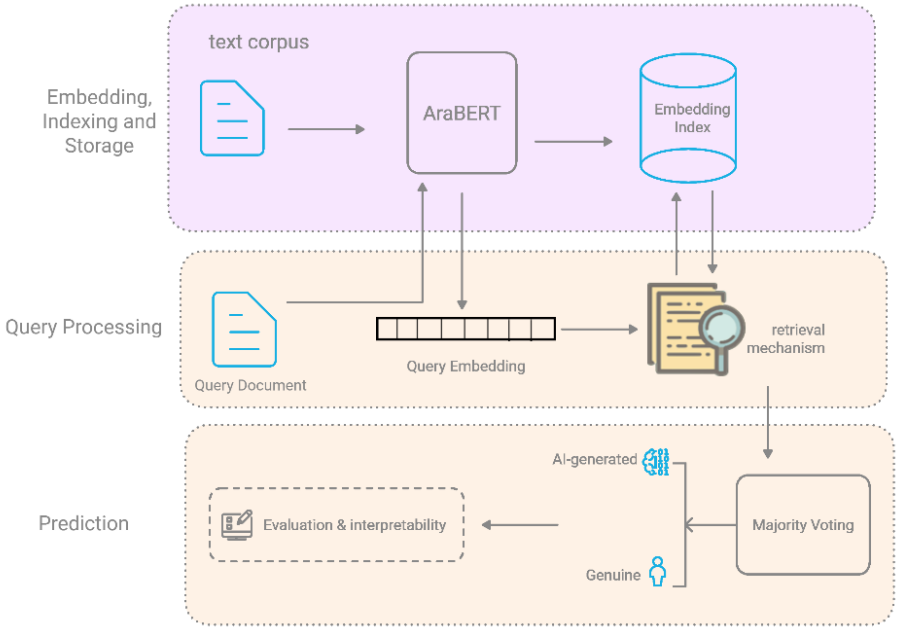


Figure 25. RBC model workflow for Arabic fake essays detection

The corresponding algorithm that describes our second proposed method is as follow:

Algorithm: RAG-Inspired Classification

Input: Training data D = {(x_1, y_1), ..., (x_n, y_n)}, x_{test}, k, embedding model E, vector database V
Output: Predicted class label y_{pred} for x_{test}
For each x_i in D **do**: // Preprocessing and Embedding
1: Preprocess x_i (e.g., tokenization, lowercasing)
2: $e_i \leftarrow E(x_i)$ // Generate embedding using model E
3: Add all e_i to vector database V
4: **end for**

// Build Index:
Create an efficient index I on V for similarity search
// Training:
Store all (e_i, y_i) pairs for retrieval
For test instance x_{test} **do**: // Classification:
    Preprocess x_{test}
    $e_{test} \leftarrow E(x_{test})$ // Generate embedding for test instance
    $N \leftarrow I.search(e_{test}, k)$ // Retrieve k nearest neighbors
    $Y \leftarrow \{y_i \text{ for } i \text{ in } N\}$ // Get labels of nearest neighbors
    $y_{pred} \leftarrow mode(Y)$ // Predict label by majority voting
end for
Return y_{pred}

Training data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ where x_i is the text and y_i is the label
test instance x_{test}
number of neighbors k
embedding model E
vector database V

The strength points of this algorithm lie in: 1) Efficiency: The FAISS index allows for fast similarity search, crucial when dealing with large amounts of data. 2) Scalability: This approach can handle large numbers of examples, limited mainly by memory and storage capacity. 3) Interpretability: training examples that influenced a particular decision are relatively easy to inspect. 4) Flexibility: The algorithm can be modified very easily to use different embedding models, distance metrics, or classification rules.

5.2. Experimental Setup

We designed our experiments to evaluate and validate the feasibility of the suggested approaches as follows:

1) Experiment 1: Fine-tune the pretrained models for each dataset centrally and evaluate its performance in terms of accuracy using cross-validation over all the datasets.

2) Experiment 2: implement the proposed retrieval-based classifier and evaluate the built model using a cross-validation fashion over all the datasets.

Experiment 1

In this experiment, the four previously presented models (FEDM1, FEDM2, FEDM3, and FEDM4) were trained over each mentioned dataset to evaluate the performance of the first AI-generated essays detection approach in order to answer our first research question (RQ1). The models are implemented using the public transformers framework from Hugging Face. For all the models, a custom dataset class is implemented to handle text tokenization and encoding. Hyperparameters were fixed across fine-tuning runs: Adam optimizer with initial learning rate 2e-5 using a 5-fold cross-validation strategy to assess

model performance across different data subsets, with each fold undergoing a training process of 10 epochs and a fixed batch size of 10.

Experiment 2

In the second experiment, we aim to address the second research question (RQ2). The core of our proposed model uses the AraBERT (bert-base-arabertv02) architecture for text embedding. A custom embedding function is implemented to convert text samples into dense vector representations, which are then indexed using FAISS (Facebook AI Similarity Search) for efficient nearest neighbour retrieval. In the experiment, we employ a 5-fold cross-validation strategy to ensure robust evaluation, and for each fold the FAISS index is created from the training data. The classification process of he samples from the validation set involves retrieving the k-nearest neighbours (with k=5) from the training set and then employing a majority-voting-based mechanism to determine the label.

5.3. Results

Experiment 1

Tables 19, 20, 21, and 22 report the models performance metrics per dataset. We noticed that the models have demonstrated exceptionally high performance across all evaluation metrics over all datasets. For instance, FEDM1 exhibits remarkable consistency across all five folds. In four out of five folds (2-5), the model achieves perfect performance with accuracy, F1 score, recall, and precision all at 1.0. The slight deviation in Fold 1 is noteworthy. While still achieving an impressive accuracy of 0.9980, it is the only fold where we observe a misclassification. This resulted in a marginally lower precision (0.9960) compared to the perfect recall (1.0). This high recall score indicates that the model successfully identified all positive cases, while the slightly lower precision suggests a minimal tendency towards false positives in this particular fold.

FEDM2 and FEDM4 demonstrate near-perfect performance across all datasets and folds. FEDM3 Exhibits the most variability, especially on the Gemini dataset.

For the datasets, on the Gemini dataset FEDM1 and FEDM3 show minor fluctuations across folds while FEDM2 and FEDM4 maintain perfect scores throughout. Meanwhile, all models achieve perfect scores (1.0) across all folds and metrics on the ChatGPT dataset. However, dataset 3 (which contains a mixture of AI-generated essays: ChatGPT-generated and Gemini-generated essays) shows the most variability among the datasets. FEDM1 and FEDM3 have lower scores in some folds, particularly in precision and F1-score, while FEDM2 and FEDM4 maintain near-perfect performance. The results show that ChatGPT seems to be the easiest dataset, with all models achieving perfect scores. Dataset 3 appears to be the most challenging, revealing differences in the models'

performance. The Gemini dataset (Dataset 2) falls between these two in terms of detection difficulty.

Precision scores show the most variation, especially for FEDM and FEDM3 on dataset 3. Recall scores show consistently high across all models and datasets. F1-scores reflect the balance between precision and recall, showing minor fluctuations for FEDM1 and FEDM3. All models have high accuracy scores, which indicates excellent overall classification performance.

FEDM2 and FEDM4 show remarkable stability and appear to be the most robust models, maintaining perfect or near-perfect scores across all scenarios. FEDM1 and FEDM3 exhibit some variability and show slight vulnerabilities, particularly on dataset 3.

Table 19. The performance achieved by the fine-tuned FEDM1 per dataset

Metric		**precision**			**recall**			**f1-score**			**accuracy**		
Datasets		**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**
FEDM1	**Fold 1**	1.000	0.996	0.994	1.000	1.000	0.994	1.000	0.998	0.994	1.000	0.998	0.994
	Fold 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 4	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.997
	Fold 5	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.997

Table 20. The performance achieved by the fine-tuned FEDM2 per dataset

Metric		**precision**			**recall**			**f1-score**			**accuracy**		
Datasets		**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**
FEDM2	**Fold 1**	1.000	0.980	0.985	1.000	1.000	0.991	1.000	0.990	0.988	1.000	0.989	0.988
	Fold 2	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.998
	Fold 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 21. The performance achieved by the fine-tuned FEDM3 per dataset

Metric		**precision**			**recall**			**f1-score**			**accuracy**		
Datasets		**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**
FEDM3	**Fold 1**	0.952	0.995	1.000	0.992	0.991	0.955	0.972	0.993	0.977	0.972	0.993	0.978
	Fold 2	1.000	1.000	0.970	1.000	1.000	1.000	1.000	1.000	0.984	1.000	1.000	0.985

	Fold 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 22. The performance achieved by the fine-tuned FEDM4 per dataset

Metric		**precision**			**recall**			**f1-score**			**accuracy**		
Datasets		Dataset1	Dataset2	Dataset3	Dataset1	Dataset2	Dataset3	Dataset1	Dataset2	Dataset3	Dataset1	Dataset2	Dataset3
FEDM4	**Fold 1**	0.996	1.000	0.994	1.000	1.000	0.997	0.998	1.000	0.995	0.998	1.000	0.995
	Fold 2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Fold 5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

In summary, the experiment's findings reveal that the fine-tuned models, especially FEDM2 and FEDM4, consistently achieve near-perfect scores across all the evaluation datasets, exhibiting an exceptional ability to recognize AI-generated essays accurately. These models' strong balance between precision and recall, together with their high and consistent performance, indicate that the fine-tuning approach is a significantly effective and robust solution for detecting AI-authored student essays.

Experiment 2

Figure 26 illustrates the learning process evolution of the RBC model on the Gemini dataset and the changes with varying amounts of training data.

The results presented in Table 23 show the performance of the RBC model across the three datasets (Gemini, ChatGPT, and Mix), demonstrating strong performance across all datasets and metrics, with scores consistently above 0.95.

Over the ChatGPT dataset, the model consistently shows the highest overall performance (0.98 for precision, recall, F1-score, and accuracy). Gemini dataset performs slightly lower than ChatGPT but still maintains high scores (0.97 across all metrics). Dataset 3 shows comparable performance to Gemini (0.97 across all metrics), which suggest that the model handles the combined dataset well.

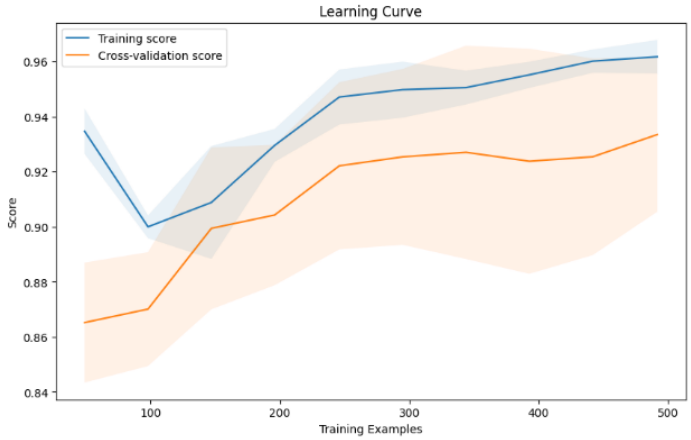


Figure 26. Learning Curve of the RBC model performance on the Gemini dataset

Precision scores range from 0.96 to 0.98 across datasets and folds. Recall has similar range to precision, 0.96 to 0.98. F1-scores consistently matches or closely follows precision and recall, indicating a good balance between the two. The accuracy aligns closely with the other metrics, which indicates an overall consistent performance of the classification.

The model shows good stability across folds, with only minor fluctuations. Specifically, fold 3 shows slightly lower performance for Gemini and ChatGPT but higher for Mix, while Fold 1 has the lowest scores for dataset 3.

Gemini dataset showed the most consistent performance across folds with a slight dip in Fold 3 (0.96 for precision and recall). For the ChatGPT dataset, the highest overall performance among the other datasets is reported with minor fluctuations in precision and recall across folds (0.97 to 0.98). Dataset 3 shows more variability across folds compared to individual datasets (lowest performance in Fold 1: 0.955 for precision and recall, while highest performance in Folds 3 and 5 for precision: 0.98).

The RBC setup demonstrates a thoughtful balance between leveraging pre-trained language understanding and adapting to the specific characteristics of the dataset, making it particularly suitable for tasks where contextual similarity plays a crucial role in classification. The model's ability to maintain performance on the third dataset (dataset 3) indicates relatively good generalization capabilities.

Table 23. The performance achieved by the retrieval-based model per dataset

	Precision			**recall**			**f1-score**			**accuracy**		
	Dataset1	**Datase2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset1**	**Dataset2**	**Dataset3**

Fold 1	**0.975**	**0.970**	**0.955**	**0.975**	**0.970**	**0.955**	**0.975**	**0.960**	**0.960**	**0.970**	**0.960**	**0.960**
Fold 2	**0.970**	**0.970**	**0.960**	**0.970**	**0.960**	**0.960**	**0.970**	**0.960**	**0.960**	**0.970**	**0.960**	**0.960**
Fold 3	**0.960**	**0.960**	**0.980**	**0.960**	**0.960**	**0.980**	**0.960**	**0.950**	**0.980**	**0.960**	**0.950**	**0.980**
Fold 4	**0.975**	**0.970**	**0.970**	**0.970**	**0.970**	**0.970**	**0.970**	**0.970**	**0.960**	**0.970**	**0.970**	**0.960**
Fold 5	**0.975**	**0.970**	**0.980**	**0.975**	**0.960**	**0.970**	**0.975**	**0.960**	**0.970**	**0.980**	**0.960**	**0.970**
Overall performance	**0.980**	**0.970**	**0.970**	**0.980**	**0.970**	**0.970**	**0.980**	**0.970**	**0.970**	**0.980**	**0.970**	**0.970**

Figure 27 depicts the model's confidence in its predictions across different samples. This visualization shows three distinct visible confidence levels: near 1.0, around 0.8, and around 0.6, where the majority of predictions have very high confidence (near 1.0), a significant number of predictions fall in the 0.8 range and a smaller set of predictions have confidence around 0.6. The model demonstrates high confidence in most of its predictions, the lack of very low-confidence predictions suggests the model rarely encounters samples for which it is completely uncertain. The combination of well-separated classes in the embedding space with the high confidence predictions supports the strong performance metrics, which aligns with the high performance metrics we presented earlier.

Moreover, Figure 28 shows the silhouette plot which helps visualize the quality of clustering in the embedding space, particularly for understanding how well-separated the classes are in the AraBERT embedding space. The majority of samples in both classes have positive silhouette coefficient values, class "1" (fake essays) appears to be more cohesive and well-defined in the embedding space and shows generally higher silhouette scores, indicating better separation and cohesion within this class. Class "0" (real student essays) has a wider range of silhouette scores showing more variability, including some negative values reflecting samples potentially closer to the fake essays class. The overall positive silhouette scores suggest that the RBC approach using AraBERT embeddings is effectively separating the fake and real essays.

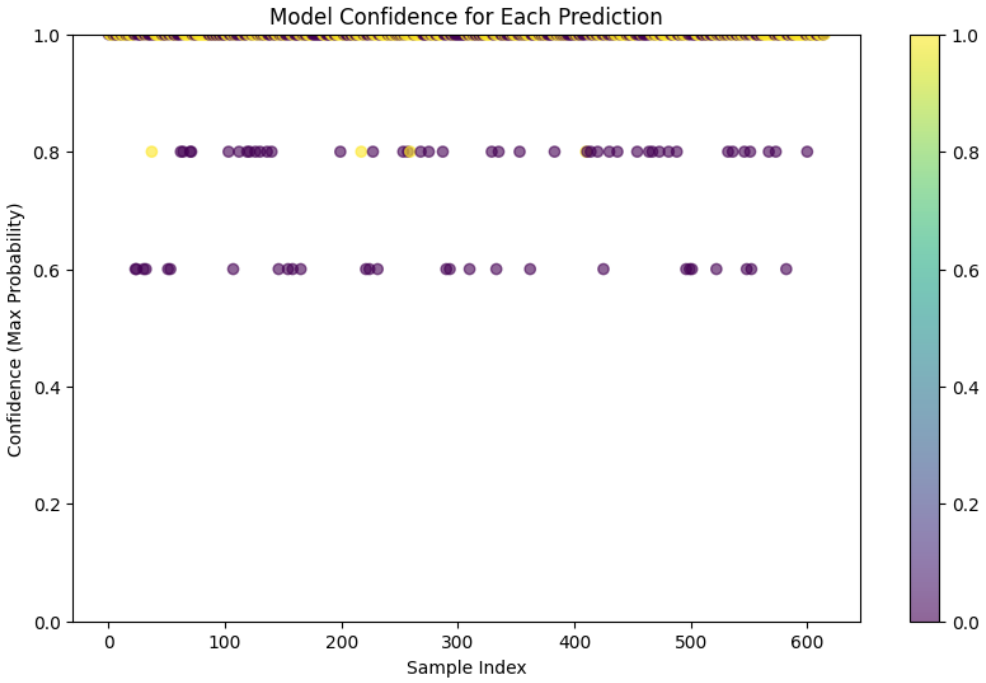


Figure 27. Model Confidence Visualization

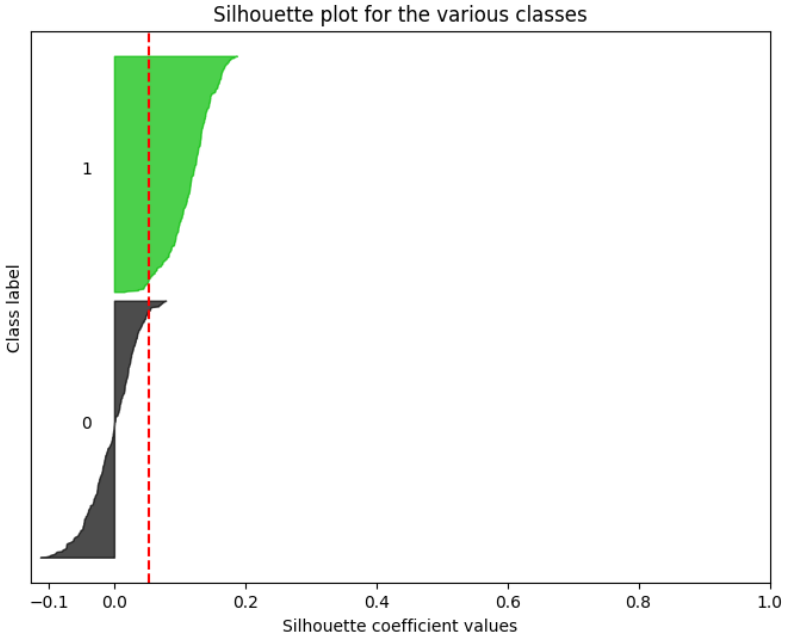


Figure 28. Silhouette Plot for real/fake essays classes separation

In conclusion, the previous results demonstrate strong and robust capabilities performed by the retrieval-based classification (RBC) model in detecting AI-generated student

essays, making it a compelling approach for real-world deployment in scenarios where the flexibility, the interpretability, and the generalization are critical factors.

5.4. Comparative Analysis

This subsection tries to answer the third research question (RQ3). The fine-tuned models, particularly FEDM2 and FEDM4, outperform the RAG-based model in terms of raw scores. However, the RBC model's performance is still excellent and more consistent across datasets, especially considering its performance on dataset 3. This suggests that it might be more robust to varied or unseen data.

Both the RBC model and fine-tuned models maintain a good balance between precision and recall, as evidenced by F1-scores closely matching precision and recall values.

While the performance of the RBC model is significantly strong, it is slightly lower than the excellent scores reported for the fine-tuned models from the first detection approach, which is expected, as retrieval-based models often trade some performance for potential advantages in: 1)Flexibility as it can adapt to new information without retraining, 2) Interpretability as it allows examination of the retrieved passages for decision-making insights, and 3) Generalization and ability to handle evolving datasets.

The consistent performance of the RBC across metrics indicates a well-balanced model suitable for deployment in real-world classification tasks.

For an efficient model choice, fine-tuned models (especially FEDM2 and FEDM4) would be preferable when dealing with data that is very similar to the training set, absolute maximum performance is required and computational resources at inference time are limited. While the RBC model would be advantageous for cases where interpretability and transparency of decisions are important, flexibility to incorporate new information without retraining is required, and when dealing with evolving or diverse datasets.

Based on the above, in general, we find that the fine-tuned models from the first detection approach offer superior performance in terms of raw metrics, making them the preferred choice when maximum accuracy is the primary concern and the data closely matches the training distribution. However, the RBC model's consistent performance, flexibility, and interpretability make it a compelling alternative, especially in scenarios where adaptability and decision transparency are critical factors. Ultimately, the choice between the two approaches should be guided by the specific requirements and constraints of the detection task.

5.5. Chapter Summary

This chapter has established compelling evidence that both fine-tuned large language models and retrieval-based machine learning classifiers demonstrate substantial

effectiveness in distinguishing AI-generated student essays from authentic academic writing, though their respective performance characteristics reveal important contextual dependencies that inform practical deployment decisions. The comprehensive evaluation framework reveals that fine-tuned models consistently outperformed retrieval-based approaches in terms of overall accuracy metrics, yet the latter methodology provides significantly enhanced interpretability capabilities that enable educators to identify specific stylistic patterns and linguistic markers indicative of synthetic content generation.

The research addresses a critical gap in natural language processing applications within educational contexts, particularly concerning the integrity of student essay assessment in Arabic-language academic environments. The investigation implemented dual methodological approaches grounded in transfer learning and retrieval-based classification paradigms, each designed to address distinct operational requirements within educational institutions.

The transfer learning approach leveraged advanced transformer architectures to develop detection capabilities through domain-specific fine-tuning processes. This methodology demonstrated superior classification performance across three distinct datasets, achieving accuracy levels ranging from 95% to 98% depending on dataset characteristics and content complexity. The high-performance outcomes validate the efficacy of transfer learning strategies when applied to Arabic essay detection tasks, establishing robust baseline capabilities for automated academic integrity enforcement.

The retrieval-based classification methodology, while achieving somewhat lower raw accuracy scores, provided substantial advantages in terms of system transparency and educator utility. This approach enables instructional staff to examine specific textual features and stylistic elements that contribute to classification decisions, thereby supporting informed pedagogical responses to suspected academic misconduct. The interpretability advantages prove particularly valuable in educational settings where understanding the rationale behind detection decisions directly impacts student evaluation and disciplinary procedures.

The creation of a comprehensive Arabic essay dataset represents a foundational contribution to natural language processing research in educational applications. This dataset addresses the pronounced scarcity of Arabic-language resources for academic integrity research while providing standardized evaluation frameworks for future investigations in this domain. The dataset encompasses authentic student writing samples alongside AI-generated content produced by state-of-the-art language models, including ChatGPT and Gemini, thereby capturing realistic scenarios that educators encounter in contemporary academic environments.

The dataset development process incorporated rigorous quality assurance protocols to ensure linguistic authenticity and representative coverage of typical student writing characteristics. This methodological approach establishes reliable benchmarks for

evaluating detection system performance while supporting reproducible research outcomes across different institutional contexts and academic disciplines.

The availability of this Arabic essay dataset fills a substantial research gap that has previously limited investigation into AI detection capabilities for non-English educational content. This contribution enables expanded research into multilingual academic integrity solutions while supporting the development of culturally and linguistically appropriate detection systems for Arabic-speaking educational communities.

The comparative evaluation reveals fundamental trade-offs between detection accuracy and system interpretability that have significant implications for practical deployment in educational settings. Fine-tuned language models demonstrated superior classification performance through their ability to capture subtle linguistic patterns and contextual relationships that characterize authentic versus synthetic academic writing. However, these models operate as complex black-box systems that provide limited insight into the specific features driving classification decisions.

Conversely, retrieval-based classification systems offer substantial transparency advantages that prove essential for educational applications requiring explainable automated decision-making. Educators can examine specific textual elements, stylistic patterns, and linguistic features that contribute to detection outcomes, thereby supporting informed responses to suspected academic misconduct cases. This interpretability proves particularly valuable when communicating detection rationale to students, academic administrators, and institutional review processes.

The performance differential between approaches highlights the necessity for domain-specific optimization strategies that balance accuracy requirements against interpretability needs. Educational institutions must carefully consider operational priorities when selecting detection methodologies, weighing the benefits of maximum classification accuracy against the practical advantages of transparent, explainable detection systems.

The research findings provide actionable strategies for educational institutions seeking to address academic dishonesty challenges while maintaining pedagogical trust and supporting legitimate student learning processes. The demonstrated effectiveness of both detection approaches offers institutional flexibility in selecting methodologies that align with specific educational contexts, student populations, and administrative requirements.

The availability of interpretable detection capabilities through retrieval-based methods addresses critical concerns regarding automated academic assessment systems. Educators can examine detection rationale and engage in informed discussions with students regarding writing authenticity, thereby maintaining the educational value inherent in academic integrity enforcement processes. This transparency supports constructive pedagogical interventions rather than purely punitive responses to suspected misconduct.

The research establishes robust frameworks for implementing automated detection systems while preserving essential human oversight and educational judgment. The combination of high-accuracy automated screening with interpretable result explanation enables efficient

processing of large student submission volumes while maintaining appropriate educational standards and supporting student development objectives.

The methodological frameworks and empirical findings presented establish essential groundwork for addressing more sophisticated detection challenges that emerge as AI-generated content becomes increasingly prevalent in educational settings. The binary classification capabilities validated through this research provide foundational tools that support expansion into more granular authorship analysis tasks, including the detection of hybrid human-AI collaborative writing and multi-author attribution scenarios.

The technical methodologies developed for distinguishing authentic student essays from AI-generated content translate directly to more complex detection challenges involving partial AI assistance and collaborative writing scenarios. The retrieval-based classification framework proves particularly valuable for identifying specific text segments that exhibit characteristics indicative of AI generation, thereby supporting segment-level analysis capabilities essential for hybrid text detection.

The research establishes a comprehensive foundation for continued investigation into ethical and innovative detection systems capable of real-world deployment in educational environments. Future research directions include expanded evaluation of additional state-of-the-art language models as sources of AI-generated content, comparative analysis of alternative classification algorithms for retrieval-based detection, and integration of human expert judgment into automated detection processes.

The methodological precedents established through this research support the development of practical detection systems that can effectively combat academic dishonesty while preserving pedagogical relationships and supporting legitimate educational objectives. The balance achieved between detection accuracy and system interpretability provides a template for developing ethical AI detection tools that serve educational communities while maintaining transparency and accountability in automated decision-making processes.

The transition to addressing authorship shifts within individual documents represents a natural progression from the binary classification frameworks established in this chapter. The robust methodological foundations and empirical validation achieved through essay-level detection provide essential building blocks for tackling the more granular challenge of identifying transitions between human and AI-authored content within single documents, thereby extending detection capabilities to address increasingly sophisticated academic misconduct scenarios.

CHAPTER VI

Segment-Level Detection of Mixed-Authorship

Building on the preceding chapter, which focused on distinguishing fully AI-generated essays from human-authored ones, this chapter addresses a more granular and underexplored challenge: identifying hybrid texts where authorship shifts between human and AI within the same document. While binary classification methods, as discussed earlier, are effective for detecting entirely synthetic content, they fall short in real-world scenarios where human writers may intersperse AI-generated segments—either intentionally or inadvertently—into their work. This chapter responds to the growing prevalence of collaborative human-AI writing practices where partial reliance on generative tools blurs traditional boundaries of authorship.

The aim of this chapter is to propose and evaluate a methodology for detecting intra-textual authorship shifts, thereby advancing the thesis's goal of ensuring content authenticity in hybrid human-AI environments. By moving beyond binary classification, this chapter contributes a novel framework for verifying text integrity in contexts where human and machine contributions coexist, such as collaborative writing platforms or edited AI drafts.

4.2. Methodology

The proposed methodology consists of the following key steps to detect writing style changes in Arabic text generated by humans and AI generative models:

Hybrid Dataset Construction

We set out to create a hybrid text dataset in Arabic since, as far as we know, there exist no datasets that contain hybrid Arabic texts that are appropriate for examining our research question (stated in Section 1). Human-written texts were collected from the online news site Al-Arabiya; AI text was generated using the Gemini model. Then to

generate hybrid texts, we combined segments from human-written and AI-generated sources. Our text-processing pipeline consisted of two main components: text segmentation and hybrid text generation.

Text Segmentation: We developed two distinct text segmentation methods to process the input corpora. The first method ("split into windows") used a fixed-size windowing approach, segmenting texts into consecutive chunks of a fixed number of words. The second method (split by punctuation) employed a linguistically-motivated segmentation strategy, splitting texts at natural boundaries marked by punctuation marks (periods, commas, semicolons, colons, question marks, and exclamation points). This approach preserved the semantic coherence of text segments by respecting syntactic boundaries and therefore we used this punctuation-based segmentation.

Hybrid Text Generation: Using these segmented texts, we generated hybrid samples through a controlled randomization process. For each hybrid text, we:

1) Selected random windows from the human-authored corpus and from the AI-generated corpus.
2) Assigned binary labels to each window ('0' = human-authored, '1' = AI-generated).
3) Implemented an alternating selection algorithm that ensured consecutive windows would not originate from the same source, thereby avoiding extended sequences of either human or AI-generated text.
4) Merged the selected windows while maintaining word-level source tracking through paired identifiers.

The resulting hybrid texts comprised the concatenated text and a corresponding sequence of binary identifiers indicating the source (human or AI) of each word, as illustrated in Figure 29. This methodology enabled the creation of natural-seeming hybrid texts while maintaining precise tracking of the source of each constituent word, facilitating subsequent analysis of linguistic patterns and detection tasks.

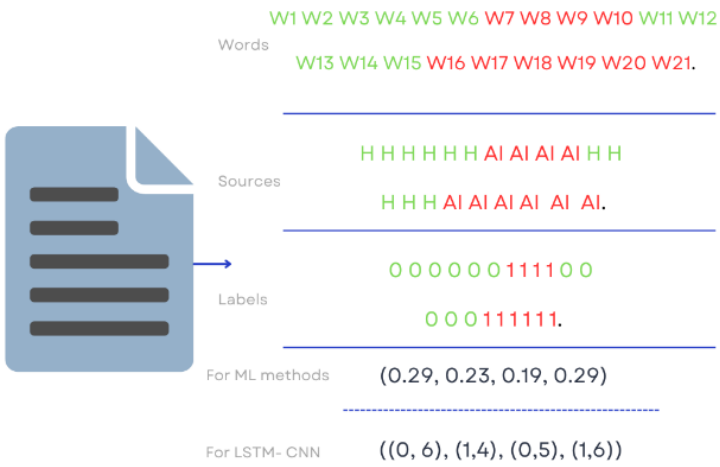


Figure 29. Example of segment-based representations used

Data Preprocessing and Feature Extraction

The preprocessing step is crucial to prepare the data for the subsequent analysis in order to identify the transition points between human-authored and AI-generated text segments. The hybrid texts was cleaned, then split into training (80%) and testing (20%) sets using stratified sampling (random_state=42).

For feature extraction, we calculated several stylometric features from the cleaned text, including linguistic, statistical, and semantic features. These features capture essential characteristics of the writing style that can be used to differentiate between human and AI-generated text. We processed texts in batches with a window size of 128 tokens. The extracted features were then standardized to ensure uniform scale across all feature dimensions.

Detection of Human-AI Transition Points

ML-based methods: For each text, we developed a novel label encoding scheme that converted the binary source labels (human/AI) into continuous percentage distributions based on consecutive sequences of the same class. This approach enabled us to capture the proportional length of continuous segments from each source within the text. The percentage-based labels were padded to a uniform length (determined by the maximum number of transitions in any text) to ensure consistent dimensionality across all samples. This transformation converted the discrete classification problem into a regression task

aimed at predicting the relative positions of transitions between human and AI-generated content.

For the transition points' detection between authentic segments and AI-generated segments, we leveraged two ML-based algorithms to predict the writing style changes: XGBoost regressor and Random Forest (RF). The models were trained to predict the percentage distribution of human and AI-generated segments within each text, effectively learning to identify transition points between different sources. The models were initialized with specific hyperparameters, then fitted to the training data, and predictions were made on the test set. The classifiers were trained to predict binary labels (human/AI) for each text segment. The predictions were then post-processed to identify transition points in the text.

We evaluated models' performance with macro F1 (F1-macro) score to account for both precision and recall across classes, and with Mean Squared Error (MSE) to measure deviation between predicted and ground-truth transition percentages. MSE is important to assess the accuracy of transition-position estimates.

DL-based Methods

LSTM-CNN: We developed another segment-based representation approach that improves upon the word-level binary classification as shown in Figure 29. Rather than maintaining individual binary indicators for each word, we aggregated consecutive words with the same classification (human-authored or AI-generated) into coherent segments. Each segment is represented as a tuple (l, n), where $l \in \{0, 1\}$ denotes the segment label (0 = human-authored, 1 = AI-generated) and n is the segment length in words. This representation offers several advantages. First, it significantly reduces the dimensionality of the classification problem while preserving the sequential nature of the text. Second, it better captured the natural boundaries between human-authored and AI-generated content, as these typically occur in meaningful segments rather than at arbitrary word boundaries. Third, it facilitates more efficient model training by focusing on segment-level rather than word-level transitions.

In this approach, we combined Long Short-Term Memory networks (LSTM) and Convolutional Neural Networks (CNN) to leverage both sequential dependencies and local feature patterns in the text. The architecture consists of two parallel branches that merge before final classification: 1) The CNN branch employs three parallel convolutional layers with different kernel sizes to capture various n-gram features. Each convolutional layer is followed by a max-pooling layer for feature selection. This multi-scale approach enables the model to identify distinctive patterns at different granularities within the text. 2) The LSTM branch comprises two stacked LSTM layers to capture long-range dependencies and contextual information. The first LSTM layer returns sequences, while the second produces a fixed-size representation of the entire input sequence.

The outputs from both branches are concatenated and passed through two dense layers with ReLU activation and dropout for regularization. The final layer employs sigmoid activation for binary classification of segments. The model were trained using binary cross-entropy loss and the Adam optimizer.
For the evaluation, we used precision, recall, and F1-score per class, and stratified k-fold cross-validation to obtain robust estimates across data splits. Additionally, we analyse confusion matrices to understand the model's error patterns and potential biases in classification.

Trans-Detect: We proposed another model architecture, Trans-Detect, comprising a multi-layered neural network system. At its core, the model utilises the AraBERT [130] as the primary encoder, augmented with additional components optimized for the specific challenges of hybrid text detection. The training data consists of hybrid Arabic texts, where each segment is labelled with binary indicators (0 for human-written, 1 for AI-generated). The preprocessing pipeline includes: 1) Segmentation of hybrid texts based on authorship boundaries. 2) Tokenization using AraBERT's specialized tokenizer. 3) Dynamic padding to accommodate varying text lengths. 4) Attention mask generation for efficient processing.
The transformer-based encoder consists of 12 attention layers with 768-dimensional hidden states and 12 attention heads. The model has been pre-trained on a substantial corpus of Arabic text, enabling it to capture intricate linguistic patterns and contextual relationships specific to the Arabic language. Following the base encoder, we implement a bidirectional Long Short-Term Memory (BiLSTM) network to capture long-range dependencies and sequential patterns within the text. The BiLSTM layer is configured with: input dimension: 768 (matching AraBERT's hidden state size), hidden dimension: 256, and number of layers: 2, with Bidirectional processing: enabled. This configuration allows the model to process contextual information in both forward and backward directions, essential for identifying subtle variations in writing style and structure. Then we incorporate a novel attention mechanism designed to focus on stylometric features that distinguish between human and AI-generated text. The attention layer consists of: a dimension reduction layer, a non-linear transformation using hyperbolic tangent activation, and a final projection layer. This attention mechanism enables the model to assign varying importance to different parts of the text, particularly focusing on linguistic patterns and structural elements that are characteristic of AI-generated content. The overall architecture of this model is summarized in Figure 30.

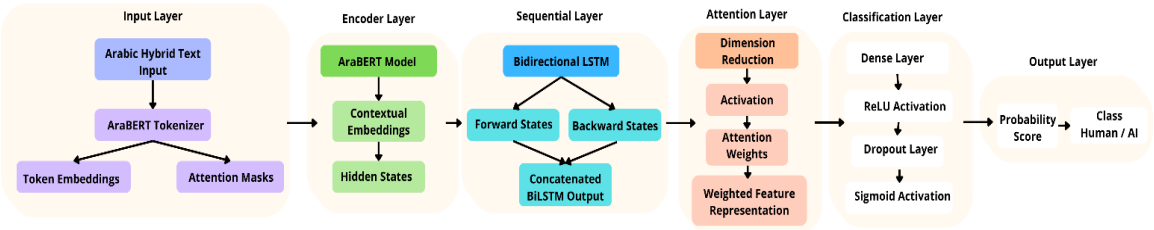


Figure 30. Overall architecture of Trans-Detect model

We implemented early stopping to prevent overfitting. For evaluation, we employed a comprehensive set of metrics including precision, recall, and F1-score for both classes. We utilised stratified k-fold cross-validation to ensure robust performance estimation across different data splits. Additionally, we analyse confusion matrices to understand the model's error patterns and potential biases in classification.

4.3. Results

Our experiments from the ML-based approaches (Random Forest and XGBoost) for detecting transitions between human and AI-generated text segments revealed distinct performance characteristics for each method.

The Random Forest detector demonstrated strong predictive accuracy with a Mean Squared Error (MSE) of 0.0052 on the test set. This exceptionally low MSE highlights the model's ability in predicting the proportional distribution of human and AI-generated content within the texts. The model particularly excelled at identifying the granular transitions between different sources while maintaining consistent performance across various text lengths.

The XGBoost model showed comparable but slightly different performance metrics: Mean Squared Error of 0.0105, indicating strong predictive accuracy though marginally higher than the Random Forest approach, and F1-macro Score of 0.4869, reflecting moderate success in discriminating between human and AI-generated segments. Both models demonstrated consistent behaviour in handling the sequential nature of the text. However, the XGBoost model's lower F1-macro score (0.4869) demonstrates that it was less effective at maintaining balanced performance across different classes.

Table 24 highlight the results reported from the CNN-LSTM and Trans-Detect detectors. The LSTM-CNN model exhibits commendable proficiency in distinguishing between human-authored and AI-generated text segments within hybrid texts. The precision and recall metrics indicate a nuanced balance, where the model's ability to correctly identify AI-generated segments is robust (precision of 0.85), while maintaining a substantial recall for human-authored segments (0.87). The relatively low number of false positives (85) and false negatives (48) underscores the model's reliability in practical scenarios.

Table 24. Performance results of LSTM-CNN and Trans-Detect

		Precision	Recall	F1-score	Accuracy
LSTM-CNN	0	0.7900	0.8700	0.8200	0.8200
	1	0.8500	0.7700	0.8100	
Trans-Detect	0	0.9800	0.9700	0.9800	0.9800
	1	0.9700	0.9800	0.9800	

The classification report summarized in Table 24, demonstrates that our model achieves a high level of accuracy in this task. Specifically, the model exhibits an overall accuracy of 98%, with precision, recall, and F1-score values consistently above 97% for both classes (AI-generated and human-written). These results indicate that the model is highly effective in correctly identifying the origin of text segments within hybrid documents.

Table 25. Comparison between the four detectors

	XGBoost	RF	CNN-LSTM	*Trans-Detect*
F1-macro score	0.4869	0.010	**0.8200**	**0.98**
MSE	0.0105	**0.0052**	0.1724	0.0218

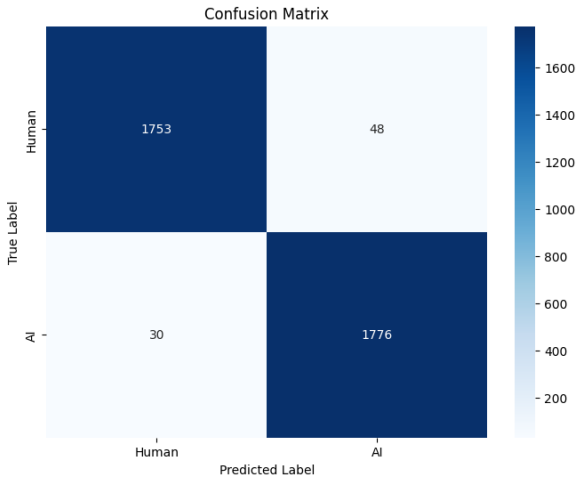


Figure 31. Confusion matrix of Trans-Detect model performance

4.4. Analysis and Discussion

These results suggest that while both models perform well for transition detection, the RF model offers marginally better performance for this specific task. The lower MSE and higher F1-macro score indicate that it may be more suitable for applications requiring precise identification of transition points between human and AI-generated content. The XGBoost-based approach provided complementary insights to the RF regression method. The combination of both approaches may enable a more robust analysis of hybrid text composition.

The slightly higher precision for AI segments suggests that the model is adept at recognizing the more deterministic patterns typically associated with AI-generated texts, while the higher recall for human segments indicates an effective capture of the more complex and varied nature of human-authored content. This performance balance is crucial for applications where both false positives and false negatives carry significant implications. Furthermore, the high overall accuracy reinforces the model's effectiveness. The Mean Squared Error value complements these findings, confirming the model's precision in predictions. The F1-macro Score of 0.82 aligns with the overall harmony in the model's classification capability.

Our proposed transformer-based model (Trans-Detect) demonstrated superior performance in distinguishing between human-authored and AI-generated text segments compared to traditional machine learning and deep learning approaches. Table 25 compares between all the four detectors in terms of F1-macro score and MSE. Trans-Detect achieved an impressive macro F1-score of 0.98, substantially outperforming other approaches including XGBoost (0.4869), Random Forest (0.010), and CNN-LSTM (0.8200). This performance differential highlights the effectiveness of our architecture in capturing the characteristics of hybrid text indicating robust generalization capabilities. The confusion matrix presented in Figure 31 highlights the classification performance of Trans-Detect, delineating the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix analysis reveals exceptional classification accuracy, with 1,753 true negatives (correctly identified human-authored segments) and 1,776 true positives (correctly identified AI-generated segments), resulting in a total of 3,529 correct classifications out of 3,607 samples where only 48 human-authored segments were incorrectly classified as AI-generated and 30 AI-generated segments were misidentified as human-authored. This relatively balanced error distribution suggests that the model does not exhibit significant bias toward either class, though it shows a slight tendency toward AI-generated classification in ambiguous cases.

The lower MSE achieved by RF compared to XGBoost, suggests superior accuracy in predicting transition points. This difference, while small in absolute terms, represents approximately a 50% improvement in prediction accuracy. Both models showed stable performance across the test set, maintaining consistent prediction lengths and demonstrating reliable scaling of features. The Random Forest approach appeared to provide more balanced predictions across different text segments, as indicated by its

superior F1-macro score. These results suggest that while both models perform well for transition detection, the RF model offers marginally better performance for this specific task. The lower MSE and higher F1-macro score indicate that it may be more suitable for applications requiring precise identification of transition points between human and AI-generated content. The XGBoost-based approach provided complementary insights to the RF regression method. The combination of both approaches may enable a more robust analysis of hybrid text composition.

When we examine MSE across architectures, deep learning models show higher MSE values (Trans-Detect: 0.0218; CNN-LSTM: 0.1724). This suggests a trade-off: while classical models (e.g., Random Forest with lower MSE) may achieve lower point-wise errors, deep models can provide superior classification (F1) and better generalization. While traditional machine learning approaches like Random Forest achieved lower MSE (0.0052), they significantly underperformed in terms of classification accuracy and F1-scores. This apparent paradox can be attributed to several factors such as prediction confidence where Deep learning models tend to produce more nuanced probability distributions, reflecting their uncertainty in borderline cases. While this may result in higher MSE, it often leads to better generalization and real-world performance. Another factor may be the feature complexity where the higher MSE in deep learning models suggests they are capturing more complex, non-linear relationships in the text. This complexity, while increasing point-wise error, enables better overall classification performance. Furthermore, traditional models may achieve lower MSE by making "safer" predictions closer to the decision boundary, but this conservative approach compromises their ability to distinguish between subtle variations in text characteristics.

4.5. Chapter Summary

This chapter has established definitive evidence that segment-level authorship attribution methodologies can effectively detect human-to-AI transitions within individual documents, achieving superior F1-scores on hybrid Arabic texts while demonstrating substantial improvements over traditional binary classification approaches. The comprehensive evaluation demonstrates that granular analysis of authorship patterns provides essential capabilities for addressing the evolving landscape of collaborative human-AI content creation, where conventional detection systems fail to identify partial synthetic content integration.

The research presents a comprehensive methodological framework that combines machine learning and deep learning techniques to address the intricate dynamics of text generation within Arabic language contexts. The investigation implemented four distinct detection architectures designed to identify writing style transitions between human authors and AI

systems through strategic extraction of stylometric features and advanced neural network architectures.

The core technical contribution focuses on Trans-Detect, a transformer-based architecture that demonstrates exceptional performance in detecting transitions between human-authored and AI-generated segments within single documents. This deep learning approach achieved balanced performance across precision, recall, and F1-score metrics while maintaining high overall accuracy levels that position the methodology as a robust solution for practical natural language processing applications.

The strategic integration of stylometric feature extraction with machine learning models provides complementary detection capabilities that enhance overall system reliability. The dual approach leverages both traditional linguistic analysis techniques and modern neural network architectures to capture subtle transitions in writing patterns that characterize shifts between human and artificial authorship within continuous text.

The research directly confronts a fundamental limitation in existing deepfake text detection systems, which predominantly focus on binary classification of entire documents while struggling to identify partially synthetic content. Current detection tools exhibit significant blind spots when confronting hybrid texts that combine authentic human writing with AI-generated segments, creating vulnerabilities that sophisticated content manipulation can exploit.

The segment-level analysis framework developed through this research provides essential capabilities for detecting these subtle authorship transitions that occur within single documents. This granular approach enables identification of specific text regions where stylistic patterns shift from human to AI characteristics, thereby addressing scenarios where traditional binary classifiers prove inadequate.

The practical implications of these detection capabilities extend across numerous real-world applications where hybrid content creation is increasingly prevalent. The methodology provides essential tools for validating edited AI drafts, auditing collaborative writing platforms, and ensuring content authenticity in environments where human authors integrate AI assistance into their writing processes.

The creation of a comprehensive dataset containing hybrid Arabic texts represents a substantial contribution to natural language processing resources for low-resource languages. This dataset addresses a notable void in available linguistic resources while providing standardized evaluation frameworks for future research in Arabic text analysis and authorship attribution.

The dataset development process incorporated rigorous quality control measures to ensure representative coverage of authentic human-AI writing transitions while maintaining linguistic authenticity across diverse content domains. The resulting resource supports reproducible research outcomes and provides benchmarking capabilities for evaluating alternative detection methodologies in Arabic language contexts.

The availability of this specialized dataset enables expanded investigation into multilingual authorship detection while supporting the development of culturally and linguistically appropriate analysis tools for Arabic-speaking communities. This contribution establishes essential infrastructure for continued research advancement in hybrid text detection across underrepresented linguistic contexts.

The comprehensive evaluation demonstrates the effectiveness of the proposed detection methodologies through rigorous testing across diverse hybrid text scenarios. The LSTM-CNN architecture emerged as the most reliable solution for identifying human-authored and AI-generated segments, exhibiting consistent performance characteristics that support deployment in operational environments.

The balanced performance metrics achieved across precision, recall, and F1-score indicators demonstrate the robustness of the detection framework under varying content conditions and authorship transition patterns. This consistent performance profile indicates that the methodology can reliably identify stylistic shifts without exhibiting bias toward either human or AI content classification.

The high overall accuracy levels attained through the detection system validate the technical approach while demonstrating practical utility for real-world applications. The performance characteristics position the methodology as a viable solution for content moderation, plagiarism detection, and forensic analysis scenarios where accurate identification of authorship transitions proves essential.

The research provides essential tools for maintaining textual communication integrity and trustworthiness in environments where AI assistance is increasingly integrated into human writing processes. The granular analysis capabilities enable nuanced evaluation of content authenticity while supporting informed decision-making regarding text provenance and authorship verification.

The interpretable framework developed through this research addresses critical transparency requirements for automated content analysis systems deployed in high-stakes environments. Stakeholders can examine specific textual segments and stylistic patterns that contribute to authorship attribution decisions, thereby supporting accountability and trust in automated analysis outcomes.

The practical applications span diverse domains including academic integrity enforcement, journalistic fact-checking, legal document analysis, and social media content moderation. The methodology provides essential capabilities for identifying sophisticated content manipulation attempts that traditional binary classification systems cannot detect.

The methodological innovations established through this research provide essential groundwork for extending similar analytical approaches to other languages and writing systems. The underlying detection principles demonstrate potential for generalization beyond Arabic contexts while maintaining effectiveness across diverse linguistic and cultural environments.

The research establishes critical precedents for investigating cultural and linguistic factors that influence style change detection accuracy and reliability. Future research directions include comprehensive evaluation of the methodology across multiple language families while examining how cultural writing conventions impact authorship transition identification.

The integration potential with advanced transformer-based models presents opportunities for further enhancing detection accuracy and system robustness. The foundational approach developed through this research provides essential building blocks for incorporating state-of-the-art natural language processing techniques into hybrid text analysis frameworks.

This work represents a significant advancement in hybrid text authorship detection, particularly within Arabic language contexts where limited research resources have previously constrained investigation. The proposed methodologies demonstrate clear implications for content moderation, plagiarism detection, and forensic analysis applications while establishing robust foundations for continued research development.

The comprehensive approach addresses the evolving field of text generation technologies while providing valuable insights into the intricate dynamics that characterize human-AI collaborative writing. As text generation capabilities continue advancing, this research establishes essential frameworks for ensuring content integrity and trustworthiness across diverse communication contexts.

The successful demonstration of segment-level analysis capabilities in Arabic texts establishes critical precedents for global applications where hybrid human-AI authorship is becoming increasingly common. The methodological foundations developed through this research provide essential tools for addressing contemporary challenges in content authenticity while supporting the development of ethical and transparent AI integration practices in human communication systems.

Chapter VII

Conclusion

The rapid advancement of artificial intelligence (AI) and natural language processing (NLP) technologies has brought about transformative changes in how we generate, consume, and interact with textual content. While these advancements have unlocked unprecedented opportunities for innovation, they have also introduced significant challenges, particularly in the form of AI-generated or "deepfake" text. This thesis has focused on addressing these challenges in the context of the Arabic language. Through a series of studies, this research has contributed to the development of robust methods for detecting and attributing of AI-generated Arabic text, offering both practical solutions and theoretical insights.

One of the key contributions of this thesis lies in the development of fine-tuned models for detecting AI-generated text in specific domains, such as news articles and student essays. The first study demonstrated that machine-based detection systems can outperform human evaluators in identifying AI-generated news articles, highlighting the potential of AI to combat its own misuse. This finding underscores the importance of leveraging advanced computational techniques to address the growing threat of misinformation. The second study expanded on this by proposing two distinct approaches for distinguishing fake student essays from authentic ones. By fine-tuning pre-trained language models and employing retrieval-based methods with machine learning classifier, this research provided effective tools for maintaining academic integrity in an era where AI-generated content is increasingly accessible.

A particularly innovative aspect of this thesis is its exploration of segment-level authorship attribution, as demonstrated in the third study. Unlike traditional binary classification approaches, this method focuses on identifying shifts in authorship within a single text, enabling the detection of human-to-AI transitions. This approach not only enhances the granularity of detection but also opens new avenues for verifying the authenticity of complex documents. By addressing the problem at a more nuanced level, this research has advanced the field of text verification, offering a framework that can be adapted to other languages and contexts.

The implications of this research extend beyond the technical realm, touching on broader societal and ethical considerations. The proliferation of AI-generated text poses a significant threat to the integrity of information, with potential consequences for public trust, democratic processes, and social cohesion. By developing tools to detect and mitigate

deepfake text, this thesis contributes to the broader effort of safeguarding the authenticity of digital content. Moreover, the findings underscore the need for interdisciplinary collaboration, combining insights from computer science, linguistics, and social sciences to address the multifaceted challenges posed by AI-generated content.

However, the work presented in this thesis is not without its limitations. While the proposed models have shown promising results, their performance is inherently tied to the quality and diversity of the training data. The Arabic language, with its richness and variations presents a particularly challenging landscape for data collection and model generalization. Future research should focus on expanding the datasets to include a wider range of dialects and text types, ensuring that the models remain robust across different contexts. Additionally, the rapid evolution of AI technologies necessitates continuous updates to detection methods, as newer generative models (e.g., DeepSeek[3]) may produce text that evades current detection mechanisms.

Another area for future exploration is the integration of multimodal approaches, combining text analysis with other forms of media, such as images and audio, to enhance detection accuracy. As deepfake technologies increasingly blur the boundaries between different media types, a holistic approach to detection will become essential. Furthermore, the ethical implications of AI-generated text warrant further investigation, particularly in terms of developing guidelines and policies to regulate its use. This includes exploring methods for watermarking AI-generated content to ensure transparency and accountability.

The findings of this thesis also highlight the importance of public awareness and education in combating the spread of misinformation. While technological solutions are critical, they must be complemented by efforts to equip individuals with the skills to critically evaluate digital content. Educational initiatives that promote media literacy and digital citizenship can play a vital role in reducing the impact of AI-generated text, empowering users to discern credible information from falsehoods.

In conclusion, this thesis has made significant strides in addressing the challenge of AI-generated Arabic text, offering innovative solutions that advance the field of NLP while addressing pressing societal concerns. By developing models for the detection in different levels and contexts, this research has provided a foundation for ensuring the authenticity and reliability of digital content in an era dominated by AI. The work underscores the importance of interdisciplinary collaboration, continuous innovation, and ethical considerations in navigating the complex landscape of AI-generated text. As the capabilities of generative models continue to evolve, so too must our approaches to detecting and mitigating their misuse. This thesis represents a step forward in that ongoing journey, contributing to a future where the benefits of AI can be harnessed without compromising the integrity of information.

Ultimately, the fight against deepfake text is not merely a technical challenge but a societal imperative. It requires the collective efforts of researchers, policymakers, educators, and

[3] https://chat.deepseek.com/

the public to create a digital ecosystem that prioritizes truth, transparency, and trust. This thesis serves as a call to action, urging stakeholders to recognize the urgency of this issue and to work together in developing solutions that safeguard the integrity of information in the digital age. The journey is far from over, but with continued innovation and collaboration, we can build a future where the authenticity of content is preserved, and the potential of AI is realized for the greater good.

References

[1] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, *13*(1), 30.

[2] Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, *23*(1), e17187.

[3] Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, *240*, 112552.

[4] Berinsky, A. J. (2012). Rumors, truths, and reality: A study of political misinformation. *Unpublished manuscript, Massachusetts Institute of Technology, Cambridge, MA*.

[5] Jerit, J., & Zhao, Y. (2020). Political misinformation. *Annual Review of Political Science*, *23*(1), 77-94.

[6] Barnard, M., Iyer, R., Del Valle, S. Y., & Daughton, A. R. (2021). Impact of COVID-19 policies and misinformation on social unrest. *arXiv preprint arXiv:2110.09234*.

[7] Rahman, T., & Jahan, I. (2020). The Role of Social Media Rumors in Social unrest of Bangladesh. *International Journal for Studies on Children, Women, Elderly and Disabled*, *11*.

[8] Kušen, E., & Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, *5*, 37-50.

[9] Benzie, A., & Montasari, R. (2022). Artificial intelligence and the spread of mis-and disinformation. In Artificial intelligence and national security (pp. 1-18). Cham: Springer International Publishing.

[10] Boutadjine, A., Harrag, F., Shaalan, K., & Karboua, S. (2023, March). A comprehensive study on multimedia DeepFakes. In *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)* (pp. 1-6). IEEE.

[11] Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119.

[12] Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, *14*(1), 26133.

[13] Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *Plos one*, *16*(5), e0251415.

[14] Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. (2019). Challenges in Arabic natural language processing. In *Computational linguistics, speech and image processing for Arabic language* (pp. 59-83).

[15] Boutadjine, A., Harrag, F., & Shaalan, K. (2024). Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
[16] Cotton, M. (2021). Virtual Reality, Empathy and Ethics.
[17] Doorn, M. Van, & Duivestein, S. (n.d.). The Synthetic Generation.
[18] Zuev, D., & Bratchford, G. (2020). Visual Sociology Practices and Politics in Contested Spaces.
[19] Li, M., Wang, X., Gao, K., & Zhang, S. (2017). A Survey on Information Diffusion in Online Social Networks : Models and Methods. https://doi.org/10.3390/info8040118
[20] Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006
[21] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-toimage translation. In CVPR, pages 8789–8797, 2018.
[22] Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: A unified selective transfer network for arbitrary image attribute editing. In CVPR, pages 3673–3682, 2019.
[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR, 2019.
[24] Zhang, H., Le, Z., Shao, Z., Xu, H., & Ma, J. (2021). MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Information Fusion, 66(June 2020), 40–53. https://doi.org/10.1016/j.inffus.2020.08.022
[25] SimSwap. GitHub - neuralchen/SimSwap: The official project of SimSwap (ACM MM 2020). Accessed: 2021-11-15.
[26] Deepfakes github. https://github.com/deepfakes/faceswap. Accessed: 2021-11-15.
[27] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2021-11-15.
[28] Lanham, M. (2021). Generating a New Reality. In Generating a New Reality. https://doi.org/10.1007/978-1-4842-7092-9
[29] Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., & Simonyan, K. (2019). High Fidelity Speech Synthesis with Adversarial Networks. 1–15. http://arxiv.org/abs/1909.11646
[30] Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 2020-December(NeurIPS).
[31] Alzantot, M., Wang, Z., & Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, (September 15–19, 2019), 1078–1082. https://doi.org/10.21437/Interspeech.2019-3174
[32] Jesse, D. "A Voice Deepfake Was Used To Scam A CEO Out Of $243,000," Forbes (Sept 3, 2019), https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of243000/#353235912241
[33] Rodriguez, A. M., Koopman, M., Macarulla Rodriguez, A., & Geradts, Z. (2018). Detection of Deepfake Video Manipulation. Imvip, December, 133–136. http://www.cost.eu/COST_Actions/ca/CA17124

[34] Zhao, J., & Lee, H. J. (2021). Classical Chinese Poetry Generation based on Transformer-XL. 57–61. https://doi.org/10.1109/icceai52939.2021.00011

[35] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[36] Bonfanti, M. E. (2020). The weaponisation of synthetic media : what threat does this pose to national security ? 19, 1–9.

[37] A. Calderwood, V. Qiu, K. I. Gero, and L. B. Chilton, "How novelists use generative language models: An exploratory user study," CEUR Workshop Proc., vol. 2848, 2020.

[38] Clark, E., August, T., Serrano, S., & Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. 2007.

[39] Çano, E., & Bojar, O. (2020). Automating Text Naturalness Evaluation of NLG Systems. 19. http://arxiv.org/abs/2006.13268

[40] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002), http://aclweb.org/anthology/P02-1040

[41] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proc. ACL workshop on Text Summarization Branches Out. p. 10 (2004), http://aclweb.org/anthology/W04-1013

[42] Gehrmann, S., Strobelt, H., Rush, A.: GLTR: Statistical detection and visualization of generated text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 111–116. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-3019, https://www.aclweb.org/anthology/P19-3019

[43] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 32, 1–21.

[44] Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. IScience, 24(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

[45] Kirubarajan, A. (n.d.). Learning to Trick Humans : Evaluation Criteria for Human-Written and Machine-Generated Text.

[46] Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. 1–8. http://arxiv.org/abs/2102.02503

[47] Saha, B. (2025). Generative AI for Text Generation: Advances and Applications in Natural Language Processing. Journal of Computer Allied Intelligence (JCAI, ISSN: 2584-2676), 3(1), 77-91.

[48] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text. CoRR, abs/2010.03070.

[49] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. CoRR, abs/1908.09203.

[50] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et

al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[51] Kate Crawford. 2017. The trouble with bias. NIPS 2017 Keynote.

[52] Boutadjine, A., Harrag, F., Bensouilah, M., Karboua, S., & Deriche, M. (2025, February). Detecting Human-to-AI Author Change in Arabic Text. In *2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 348-353). IEEE.

[53] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

[54] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human and Machine-Based Detection. In Proceedings of the 34th International Conference on Advanced Information Networking and Applications, AINA-2020, volume 1151, pages 1341–1354.

[55] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898.

[56] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 270–278.

[57] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[58] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[59] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)

[60] Rajeswar, S., Subramanian, S., Dutil, F., Pal, C., Courville, A.: Adversarial generation of natural language. arXiv preprint arXiv:1705.10929 (2017)

[61] Fedus, W., Goodfellow, I., & Dai, A. M. (2018). Maskgan: better text generation via filling in the_. arXiv preprint arXiv:1801.07736.

[62] Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., Carin, L.: Adversarial feature matching for text generation. arXiv preprint arXiv:1706.03850 (2017)

[63] Zhu, Y., Lu, S., Zheng, L., Jiaxian, G., Weinan, Z., Jun, W., Yong, Y.: Texygen: A benchmarking platform for text generation models. arXiv preprint arXiv:1802.01886. (2018)

[64] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In Proceedings of the 23rd Conference on Computational Natural Lang.

[65] Mink, J., Luo, L., Barbosa, N. M., Figueira, O., Wang, Y., & Wang, G. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks.

[66] Benkler, Y., Faris, R., & Roberts, H. (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press.

[67] Holmes, W., & Bialik, M. (2021). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.

[68] Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe.

[69] Singer, P. W., & Brooking, E. T. (2018). LikeWar: The weaponization of social media. Eamon Dolan Books.

[70] Radford, A. (2018). Improving language understanding by generative pre-training.

[71] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[72] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314* (2020).

[73] Srinivasan, A. et al. (2021) 'Predicting the Performance of Multilingual NLP Models', 1(1), pp. 1–17. Available at: http://arxiv.org/abs/2110.08875.

[74] Antonis Maronikolakis, Hinrich Schutze, and Mark Stevenson. 2020. Identifying Automatically Generated Headlines using Transformers. *arXiv preprint arXiv:2009.13375* (2020).

[75] Sanjay Kumar, Ryan Bansal, and Raghav Mehta. 2021. A Study of Blending Ensembles for Detecting Bots on Twitter. In *Innovative Data Communication Technologies and Application*. Springer, 29–40.

[76] Mohd Fazil, Amit Kumar Sah, and Muhammad Abulaish. 2021. DeepSBD: A Deep Neural Network Model With Attention Mechanism for SocialBot Detection. *IEEE Transactions on Information Forensics and Security* 16 (2021), 4211–4223.

[77] Shubham Kumar, Shivang Garg, Yatharth Vats, and Anil Singh Parihar. 2021. Content Based Bot Detection using Bot Language Model and BERT Embeddings. In *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, 285–289.

[78] Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2021. BotRGCN: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 236–239.

[79] Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Information Sciences* 467 (2018), 312–322.

[80] Yuhao Wu, Yuzhou Fang, Shuaikang Shang, Jing Jin, Lai Wei, and Haizhou Wang. 2021. A novel framework for detecting social bots with deep neural networks and active learning. *Knowledge-Based Systems* 211 (2021), 106525.

[81] Jonas Lundberg, Jonas Nordqvist, and Mikko Laitinen. 2019. Towards a language independent Twitter bot detector.. In *DHN*. 308–319.

[82] Fred Morstatter, Liang Wu, Tahora H Nazer, Kathleen M Carley, and Huan Liu. 2016. A new approach to bot detection: striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 533–540.

[83] Feng Wei and Uyen Trang Nguyen. 2019. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems*

and Applications (TPS-ISA). IEEE, 101–109.

[84] Linhao Luo, Xiaofeng Zhang, Xiaofei Yang, and Weihuang Yang. 2020. Deepbot: a deep neural network based approach for detecting Twitter bots. In *IOP Conference Series: Materials Science and Engineering*, Vol. 719. IOP Publishing, 012063.

[85] Suruchi Gera and Adwitiya Sinha. 2022. T-Bot: AI-based social media bot detection model for trend-centric twitter network. *Social Network Analysis and Mining* 12, 1 (2022), 1–19.

[86] Yeyang Chen, Mondher Bouazizi, and Tomoaki Ohtsuki. 2022. A Comprehensive System for Social Robot Detection Using RoBERTa Classifier and Random Forest Regressor with Similarity Analysis. *IEICE Technical Report; IEICE Tech. Rep.* 122, 278 (2022), 12–17.

[87] Brandon Wood and Khaled Slhoub. 2022. Detecting Amazon Bot Reviewers Using Unsupervised and Supervised Learning. In *2022 IEEE World AI IoT Congress (AIIoT)*. IEEE, 01–08.

[88] A Ramalingaiah, S Hussaini, and S Chaudhari. 2021. Twitter bot detection using supervised machine learning. In *Journal of Physics: Conference Series*, Vol. 1950. IOP Publishing, 012006.

[89] Salvatore Giorgi, Lyle Ungar, and H Andrew Schwartz. 2021. Characterizing Social Spambots by their Human Traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 5148–5158.

[90] Spencer Lee Kirn and Mark K Hinders. 2021. Bayesian identification of bots using temporal analysis of tweet storms. *Social Network Analysis and Mining* 11, 1 (2021), 1–17.

[91] Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science* 7 (2021), e443.

[92] Wanjun Zhong, Duyu Tang, Zenan Xu, Ruize Wang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Neural deepfake detection with factual structure of text. *arXiv preprint arXiv:2010.07475* (2020).

[93] Björn Bebensee, Nagmat Nazarov, and Byoung-Tak Zhang. 2021. Leveraging node neighborhoods and egograph topology for better bot detection in social graphs. *Social Network Analysis and Mining* 11, 1 (2021), 1–14.

[94] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 1096–1103.

[95] Xiujuan Wang, Qianqian Zheng, Kangfeng Zheng, Yi Sui, Siwei Cao, and Yutong Shi. 2021. Detecting social media bots with variational autoencoder and k-nearest neighbor. *Applied Sciences* 11, 12 (2021), 5482.

[96] Ivan Anic, Ivan Bilic, and Silvije Škudar. [n. d.]. Bot Detection From a Single Tweet. *Text Analysis and Retrieval 2020 Course Project Reports* ([n. d.]), 5.

[97] Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echi Zen. 2017. Identifying computer- generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1504–1511.

[98] Jürgen Knauth. 2019. Language-agnostic twitter-bot detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 550–558.

[99] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., ... & Jégou, H. (2024). The faiss library. arXiv preprint arXiv:2401.08281.

[100] Leila Ouahrani and Djamal Bennouar. 2020. AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 2634–2643, Marseille, France. European Language Resources Association.

[101] Hind Almerekhi and Tamer Elsayed. 2015. Detecting automatically-generated arabic tweets. In *AIRS*. Springer, 123–134.

[102] Giovanni C Santia, Munif Ishad Mujib, and Jake Ryland Williams. 2019. Detecting social bots on facebook in an information veracity context. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 463–472.

[103] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11.

[104] Reem Alharthi, Areej Alhothali, and Kawthar Moria. 2019. Detecting and characterizing arab spammers campaigns in Twitter. *Procedia Computer Science* 163 (2019), 248–256.

[105] Nguyen Minh Tien and Cyril Labbé. 2017. Curious cases of automatically generated text and detecting probabilistic context free grammar sentences with grammatical structure similarity. In *Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017)*.

[106] GIOVANNI LUCA FAVUZZI. 2020. Deep-fake review detection. (2020).

[107] Bayan Boreggah, Arwa Alrazooq, Muna Al-Razgan, and Hana AlShabib. 2018. Analysis of Arabic Bot Behaviors. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*. IEEE, 1–6.

[108] David M Beskow and Kathleen M Carley. 2019. Its all in a name: detecting and labeling bots by their name. *Computational and Mathematical Organization Theory* 25, 1 (2019), 24–35.

[109] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*. 273–274.

[110] Peter Kowalczyk, Marco Röder, Alexander Dürr, and Frédéric Thiesse. 2022. Detecting and understanding textual deepfakes in online reviews. (2022).

[111] Dijana Kosmajac and Vlado Keselj. 2019. Twitter bot detection using diversity measures. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*. 1–8.

[112] Jialin Shao, Adaku Uchendu, and Dongwon Lee. 2019. A reverse turing test for detecting machine-made texts. In *Proceedings of the 10th ACM Conference on Web Science*. 275–279.

[113] Alexander Shevtsov, Maria Oikonomidou, Despoina Antonakaki, Polyvios Pratikakis, Alexandros Kanterakis, Paraskevi Fragopoulou, and Sotiris Ioannidis. 2022. Discovery and classification of Twitter bots. *SN Computer Science* 3, 3 (2022), 1–29.

[114] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[115] Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational*

Linguistics 46, 2 (2020), 499–510.

[116] Fouzi Harrag, Maria Debbah, Kareem Darwish, and Ahmed Abdelali. 2021. Bert transformer model for detecting Arabic GPT2 auto-generated tweets. *arXiv preprint arXiv:2101.09345* (2021).

[117] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351* (2019).

[118] Mengjiao Bao, Jianxin Li, Jian Zhang, Hao Peng, and Xudong Liu. 2019. Learning semantic coherence for machine generated spam text detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[119] Margherita Gambini. 2020. Developing and Experimenting Approaches for DeepFake Text Detection on Social Media. (2020).

[120] David Martín-Gutiérrez, Gustavo Hernández-Peñaloza, Alberto Belmonte Hernández, Alicia Lozano-Diez, and Federico Álvarez. 2021. A deep learning approach for robust detection of bots in twitter using transformers. *IEEE Access* 9 (2021), 54591–54601.

[121] Andres Garcia-Silva, Cristian Berrio, and Jose Manuel Gomez-Perez. 2021. Understanding Transformers for Bot Detection in Twitter. *arXiv preprint arXiv:2104.06182* (2021).

[122] Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics* 13, 4 (2022), 363–383.

[123] Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. 2019. Detect me if you can: Spam bot detection using inductive representation learning. In *Companion Proceedings of The 2019 World Wide Web Conference*. 148–153.

[124] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4485–4494.

[125] Sina Mahdipour Saravani, Indrajit Ray, and Indrakshi Ray. 2021. Automated Identification of Social Media Bots Using Deepfake Text Detection. In *International Conference on Information Systems Security*. Springer, 111–123.

[126] Li, B., He, Y. and Xu, W. (2021) 'Cross-Lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-RoBERTa Alignment'. Available at: http://arxiv.org/abs/2101.11112.

[127] Rongcheng Lin, Jing Xiao, and Jianping Fan. 2018. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[128] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*. 963–972.

[129] Wang, Y. et al. (2023). "M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection". Available at: *http://arxiv.org/abs/2305.14902*.

[130] Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.

[131] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023.

Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1613–1630.

[132] Alejandro Morales-Hernández, Inneke Van Nieuwenhuyse, and Sebastian Rojas Gonzalez. 2023. A survey on multi-objective hyperpa- rameter optimization algorithms for machine learning. *Artificial Intelligence Review* 56, 8 (2023), 8043–8093.

[133] Hong Liang, Xiao Sun, Yunlei Sun, and Yuan Gao. 2017. Text feature extraction based on deep learning: a review. *EURASIP journal on wireless communications and networking* 2017 (2017), 1–12.

[134] Bharathi Mohan, G. et al. (2023). "Cross-lingual Machine Translation: An Analysis Model for Low Resource Languages". In International Conference on Emerging Trends and Technologies on Intelligent Systems. Singapore: Springer Nature Singapore. doi: 10.1007/978-981-99-3963-3_7.

[135] Evan Crothers, Nathalie Japkowicz, Herna Viktor, and Paula Branco. 2022. Adversarial robustness of neural-statistical features in detection of generative transformers. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[136] Daria Beresneva. 2016. 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*. Springer, 421–426.

[137] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. arXiv preprint arXiv:2103.06678.

[138] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake Text Detection in the Wild. *arXiv preprint arXiv:2305.13242* (2023).

[139] Julliano Trindade Pintas, Leandro AF Fernandes, and Ana Cristina Bicharra Garcia. 2021. Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review* 54, 8 (2021), 6149–6200.

[140] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)* 54, 3 (2021), 1–40.

[141] Loshchilov, I. and Hutter, F. (2019) 'Decoupled weight decay regularization', 7th International Conference on Learning Representations, ICLR 2019.

[142] Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Learning to transfer prompts for text generation. *arXiv preprint arXiv:2205.01543* (2022).

[143] Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter. *In Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop.*

[144] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. 2019. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*. 183–192.

[145] Einea, O. et al. (2019) "SANAD: Single-label Arabic News Articles Dataset for automatic text categorization", Data in Brief, 25, p. 104076. doi: 10.1016/j.dib.2019.104076.

[146] Devlin, J. et al. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', NAACL HLT 2019 - 2019 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), pp. 4171–4186.

[147] Conneau, A. et al. (2020) 'Unsupervised cross-lingual representation learning at scale', Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[148] Wang, J., Zhai, Y., & Shahzad, F. (2025). Mapping the terrain of social media misinformation: A scientometric exploration of global research. *Acta Psychologica*, *252*, 104691.

[149] Khazhomia, S. (2025). Using Trolls and Bots in Social Media: Propagandistic Influence on Public Opinion: A Literature Review. *ESI Preprints (European Scientific Journal, ESJ)*, 21(39), 103-103.

[150] Hamzaoui, B., Bouchiha, D., & Bouziane, A. (2025). A comprehensive survey on arabic text classification: progress, challenges, and techniques. *Brazilian Journal of Technology*, *8*(1), e77611-e77611.

[151] Balalle, H., & Pannilage, S. (2025). Reassessing academic integrity in the age of AI: A systematic literature review on AI and academic integrity. *Social Sciences & Humanities Open*, *11*, 101299.

[152] Fariello, S., Fenza, G., Forte, F., Gallo, M., & Marotta, M. (2025). Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection. *International Journal of Interactive Multimedia & Artificial Intelligence*, *9*(3).

[153] Cheng, Y., Sadasivan, V. S., Saberi, M., Saha, S., & Feizi, S. (2025). Adversarial Paraphrasing: A Universal Attack for Humanizing AI-Generated Text. *arXiv preprint arXiv:2506.07001*.

[154] Kadhim, A. K., Jiao, L., Shafik, R., & Granmo, O. C. (2025). Adversarial attacks on AI-generated text detection models: A token probability-based approach using embeddings. *arXiv preprint arXiv:2501.18998*.

[155] Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, *122*(8), e2416228122.

[156] Gallegos, I. O., Rossi, R. A., Barrow, J., Lee, J., Zhou, J., & Tran, T. (2023). Bias and fairness in large language models: A survey. *arXiv. https://arxiv.org/abs/2309.00770*.

[157] Taeihagh, A. (2025). Governance of generative AI. *Policy and society*, *44*(1), 1-22.

[158] Rosca, C. M., Stancu, A., & Iovanovici, E. M. (2025). The New Paradigm of Deepfake Detection at the Text Level. *Applied Sciences*, *15*(5), 2560.

[159] Mohamed, Y. A., Mohamed, A. H., Kannan, A., Bashir, M., Adiel, M. A., & Elsadig, M. A. (2024). Navigating the ethical terrain of ai-generated text tools: a review. *IEEE Access*.

[160] Cao, L. (2025). A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content. *arXiv preprint arXiv:2504.02898*.

[161] Bowman, Samuel, et al. "Generating sentences from a continuous space." *Proceedings of the 20th SIGNLL conference on computational natural language learning*. 2016.

[162] Alexander, Katarzyna, Christine Savvidou, and Chris Alexander. "Who wrote this essay? Detecting AI-generated writing in second language education in higher education." *Teaching English with Technology* 23.2 (2023): 25-43.

[163] Cingillioglu, Ilker. "Detecting AI-generated essays: the ChatGPT challenge." *The*

International Journal of Information and Learning Technology 40.3 (2023): 259-268.

[164] Tariq, Rasikh, et al. "Detecting Generative Artificial Intelligence Essays using Large Language Models: Machine and Deep Learning Approaches." 2024 *International Conference on Engineering & Computing Technologies (ICECT)*. IEEE, 2024.

[165] Jiang, Yang, et al. "Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers?." *Computers & Education* 217 (2024): 105070.

[166] Peng, Xinlin, et al. "Hidding the ghostwriters: An adversarial evaluation of AI-generated student essay detection." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.

[167] Alhijawi, Bushra, et al. "Deep learning detection method for large language models-generated scientific content." *Neural Computing and Applications* 37.1 (2025): 91-104.

[168] Yadagiri, Annepaka, et al. "Detecting AI-generated text with pre-trained models using linguistic features." *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. 2024.

[169] Weber-Wulff, Debora, et al. "Testing of detection tools for AI-generated text." *International Journal for Educational Integrity* 19.1 (2023): 1-39.

[170] Joy, S. S., & Aishi, T. D. (2023, November). Feature-level ensemble learning for robust synthetic text detection with DeBERTaV3 and XLM-RoBERTa. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association* (pp. 169-172).

[171] Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). Stylometric detection of ai-generated text in twitter timelines. arXiv preprint *arXiv:2303.03697*.

[172] Fu, Y., Xiong, D., & Dong, Y. (2024, March). Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 18003-18011).

[173] Chaka, C. (2023). "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.

[174] Alamleh, H., Alqahtani, A. A. S. and Elsaid, A. (2023) 'Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning', *2023 Systems and Information Engineering Design Symposium*, SIEDS 2023, pp. 154–158. doi: 10.1109/SIEDS58326.2023.10137767.

[175] Wu, K. et al. (2023) 'LLMDet: A Large Language Models Detection Tool'. Available at: *http://arxiv.org/abs/2305.15004*.

[176] Sadasivan, V. S. et al. (2023) 'Can AI-Generated Text be Reliably Detected?'. Available at: *http://arxiv.org/abs/2303.11156*.

[177] Antoun, W. et al. (2023) 'Towards a Robust Detection of Language Model Generated Text: Is ChatGPT that Easy to Detect?', pp. 1–15. Available at: *http://arxiv.org/abs/2306.05871*.

[178] Katib, I., Assiri, F. Y., Abdushkour, H. A., Hamed, D., & Ragab, M. (2023). Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning. *Mathematics*, 11(15), 3400.

[179] Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17.

[180] Xu, Z., Xu, R., & Sheng, V. S. (2024, March). ChatGPT-Generated Code Assignment Detection Using Perplexity of Large Language Models (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 21, pp. 23688-23689).

[181] Liu, X., & Kong, L. (2024). AI text detection method based on perplexity features with strided sliding window. *Working notes of clef*.

[182] Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 1-22.

[183] Al-Sughaiyer, I. A., & Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American society for information science and technology*, *55*(3), 189-213.

[184] Habash, N., Bouamor, H., & Oflazer, K. (2014). A multidialectal parallel corpus of Arabic.

[185] Boulesnam, I., & Boucetti, R. (2025). Arabic Language Characteristics that Make its Automatic Processing Challenging. *International Arab Journal of Information Technology (IAJIT)*, *22*(4).

[186] Alayba, A. M. (2025). Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends. *Computers*, *14*(11), 497.