ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

UNMASKING SYNTHETIC LANGUAGE: ADVANCES IN DEEPFAKE TEXT DETECTION AND EVALUATION

Abid Saeed[1,*], Nadeem Jabbar[1], Humaira Muqadas[1], Fawad Nasim[1]
[1]Department of Computer Science, Superior University, Lahore, Pakistan
*Corresponding E-mail: abid.saeed@superior.edu.pk

Abstract
The rapid proliferation of large language models (LLMs) like ChatGPT has revolutionized text generation, enabling the creation of highly fluent and contextually relevant content that rivals human writ- ing. However, this capability also poses substantial risks, including the spread of disinformation, fake news amplification, social manipulation, and phishing schemes. While detection methods for deepfake images and videos have advanced significantly, identifying synthetic text remains a nascent field, plagued by issues such as poor robustness, limited generalization across domains, and vulnerability to adversarial modifications. Even human evaluators often fare little better than chance in distinguishing AI-generated from human-authored text. This study addresses these gaps through a comprehensive benchmarking of popular transformer-based models BERT-base-uncased, RoBERTa-base, ALBERT-base-v2, and DistilBERT on three diverse datasets: Tweep- Fake (short social media posts), TuringBench (multi-domain and multi-generator benchmarks), and the Hu- man ChatGPT Comparison Corpus (HC3) in English. By conducting dataset-specific evaluations, we provide insights into how model architecture, size, and design impact detection accuracy, efficiency, and adaptabil- ity. Our key contributions include a balanced comparison of lightweight and larger models for deepfake text classification, cross-dataset analysis to highlight generalization strengths and weaknesses, and practical recom- mendations for deploying these detectors in real-world scenarios. Results demonstrate that RoBERTa generally outperforms others in accuracy, while lighter models like DistilBERT offer trade-offs in speed and resource use, underscoring the need for hybrid approaches to enhance robustness.

***Keywords*:** Deepfake; Synthetic Text; Text Detection; Deep Learning; LLM Detection

1 Introduction

The rapid advancement of large language models such as ChatGPT has fundamentally changed the way written content is produced. These models generate text that is fluent, coherent, and contextually aligned, often rivaling the quality of human authorship. Their practical value is already evident in diverse applications, including report drafting, document summarization, and conversational agents that support everyday tasks. However, alongside these benefits come significant risks. Synthetic text can be misused in disinformation campaigns, the spread of fabricated news, online manipulation, and sophisticated phishing attempts. Such possibilities introduce serious ethical and societal challenges that demand careful attention. In contrast to the significant progress made in detecting manipulated images and videos [1], [2], research on deepfake text detection is still in its early stages. Recent surveys indicate that current methods struggle with robust- ness, generalization across domains, and resilience against adversarial manipulation [3], [4]. Equally concerning is the observation that human evaluators often perform only marginally better than chance when asked to distinguish between human-authored and machine-generated text [5].

A further limitation of existing work is the narrow scope of evaluation. Many studies depend on a single dataset or a limited set of models, which restricts the generalizability of their findings across varied textual domains [6]. Recent initiatives, including the RAID benchmark [7] and approaches such as multi-level contrastive learning [8], represent promising steps toward more resilient detection strategies. Yet, comprehensive comparisons of widely adopted trans- former models across multiple datasets remain scarce, leaving a critical gap in our understanding of their strengths and limitations. To address this gap, our study provides a detailed analysis of several well-known transformer models—BERT- base-uncased, RoBERTa-base, ALBERT-base-v2, and DistilBERT—using three important benchmarks for deepfake text

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

detection. These benchmarks are TweepFake, TuringBench, and the Human ChatGPT Comparison Corpus known as HC3. The aim is to evaluate each model's strengths and weaknesses and to consider how well they generalize across varied forms of synthetic and human-authored text.

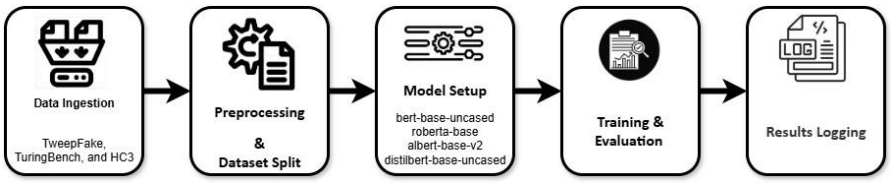

Figure 1: Overview of the proposed study.

The main goal of this study is to conduct a structured comparison of transformer models across different scales. Lightweight architectures such as DistilBERT and ALBERT are assessed alongside larger and more expressive models like BERT and RoBERTa. The evaluation is carried out on datasets that reflect diverse forms of text: TweepFake captures short, informal posts from social media; TuringBench includes multiple domains through structured prompts; and HC3 contains paired responses generated by humans and ChatGPT in English. By analyzing performance across these varied sources, the study provides insights into how model size and design influence accuracy, robustness, and generalization to new types of text. This comparison ultimately points to which architectures are more dependable in real-world scenarios where distinguishing between human and machine-generated writing is essential. The key contributions of this work are as follows:

1. A balanced benchmarking of compact transformer models such as DistilBERT and ALBERT against larger coun- terparts like BERT and RoBERTa, emphasizing the trade-off between computational efficiency and predictive accuracy in deepfake text detection.
2. An evaluation across TweepFake, TuringBench, and HC3, datasets that together cover a wide range of domains and writing styles, offering a comprehensive foundation for analysis.
3. Findings that show how models of different scales adapt to linguistic variation, while also highlighting persistent challenges in achieving strong cross-domain generalization.

The remainder of the paper is structured in the following way. Section 2 reviews related studies on feature-based, model- based, watermarking, and adversarial approaches for text detection. Section 3 describes the datasets, explains their dis- tinctive properties, and outlines the preprocessing applied for consistency. Section 4 presents the methodology, covering model selection, fine-tuning procedures, and the experimental setup. Section 5 reports the findings, including compar- isons across datasets, analysis of generalization, and discussion of both strengths and weaknesses of the evaluated models. Section 6 summarizes the outcomes and points toward directions for future research.

2 Related Work

The evolution of deepfake detection research spans multiple modalities, including images, audio, and text, with each domain presenting unique challenges and opportunities. Early efforts in text detection leaned heavily on feature-based approaches were among the earliest techniques applied to the problem of deepfake text detection. These methods rely on extracting linguistic, stylistic, or statistical features from text, which are then used as input to classifiers. [9] laid the foun- dation in the context of fake news detection on social media. They analyzed both content features (lexical and syntactic cues) and context features (user metadata, social

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

network propagation) to train traditional classifiers such as support vector machines and random forests. Using large collections of annotated social media posts, their framework demonstrated that combining textual and user-based features improved detection accuracy by 8–12% compared to content-only models, es- tablishing the importance of hybrid signals in misinformation detection. Building on linguistic markers, [10] investigated how shallow text statistics can differentiate human and machine-generated text. They curated a dataset of GPT-2 outputs alongside human references, and evaluated both human annotators and automated feature-based classifiers. Human ac- curacy hovered around 52%, only slightly above random chance, while automated classifiers using features such as word frequency distributions and sentence length variability achieved stronger results. However, feature-based methods were outperformed by transformer classifiers, showing F1 scores close to 70% compared to over 80% for RoBERTa. The work of [11] extended feature-based approaches by introducing factual consistency as an explicit dimension. Their system cross-checked statements in generated text against structured knowledge bases, transforming factual overlap into numeri- cal features for classification. On a dataset of news-style deepfake text, their model improved detection accuracy by 6–8% over baseline BERT and feature-only classifiers, illustrating that grounding text in external knowledge offers measurable benefits. Robustness of feature-based methods has also been studied. [12] examined how handcrafted features respond under adversarial perturbations. By applying small semantic-preserving edits to machine-generated sentences, they re- ported that classifiers relying on surface-level signals such as n-grams or stylistic counts saw a 20–30% accuracy drop. Their findings highlighted the fragility of feature-based methods when adversaries actively attempt to evade detection. Finally, dataset-specific feature engineering has shown varying levels of success. For example, in TweepFake, where tweets are typically short and noisy, [13] tested classifiers built on character n-grams, word statistics, and punctuation frequency. Their feature-based models achieved F1 scores around 0.65, considerably lower than transformer baselines, but still outperforming random chance. This gap underscored both the utility and the limitations of handcrafted features when facing modern generative models in realistic social media contexts.

While feature-based methods laid the foundation for detecting synthetic text, model-based approaches using deep learning and transformer architectures have rapidly become the dominant paradigm. These methods exploit contextual embeddings and pre-trained language models to capture nuanced signals of machine-generated text. [10] provided one of the earliest systematic evaluations of neural baselines against human detection. They tested LSTM classifiers and fine-tuned trans- formers such as RoBERTa on a dataset of GPT-2 generated outputs. Results showed that RoBERTa-large achieved F1 scores above 80%, significantly outperforming both human evaluators upto 52% accuracy and simpler recurrent mod- els, highlighting the superior representational capacity of transformer models in this domain. Expanding the evaluation setting, [14] introduced TuringBench, a benchmark covering 200K samples across 20 generator types. They tested a range of classifiers including logistic regression, CNNs, LSTMs, and transformers like BERT, RoBERTa, XLNet. Their results indicated that RoBERTa and XLNet consistently outperformed others, achieving accuracies above 85% in binary human-vs-machine classification, whereas traditional classifiers lagged by 20–30 percentage points. However, cross- generator generalization remained a challenge: models trained on GPT-2 often failed to detect GPT-3 outputs reliably.

[13] specifically explored short-text contexts with the TweepFake dataset, composed of over 25,000 tweets generated by multiple bots and neural generators. They compared classical machine learning, CNNs, LSTMs, and transformer-based detectors. BERT-based models achieved the highest performance, with F1 scores close to 0.90, clearly outperforming traditional baselines (0.65). These findings suggested that contextual embeddings are

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

particularly effective even in noisy, short-message settings such as Twitter. In another domain, [15] presented the HC3 dataset, focusing on human–AI con- versations. They evaluated models including BERT, RoBERTa, and larger LLM-based detectors to separate ChatGPT responses from human-authored ones. Their experiments showed that RoBERTa and BERT fine-tuned on HC3 achieved accuracies in the 78–82% range, while larger zero-shot LLM detectors performed less consistently. Importantly, they demonstrated that training on conversation-style text is necessary, as models trained on news or tweets generalized poorly to dialogue. Recent research by [5] examined how transformer models perform under real-world conditions, focusing on their robustness. The study employed a large-scale benchmark spanning multiple domains and evaluated fine-tuned BERT and RoBERTa models alongside adversarially trained detectors. While RoBERTa-large reached strong F1 scores exceeding 85% , its performance dropped by as much as 20% when subjected to adversarial perturbations or tested across domains. These findings emphasize the persistent difficulty of achieving both domain adaptability and resilience against adversarial attacks in transformer-based detection. Taken together, existing studies reaffirm that transformer models such as BERT and RoBERTa set the current state-of-the-art for deepfake text detection, consistently outperforming feature-driven and recurrent neural approaches on datasets including TuringBench, TweepFake, and HC3. At the same time, they draw attention to unresolved challenges, particularly the lack of generalization across unseen generators, limited robustness in cross-domain settings, and susceptibility to adversarial manipulation.

Beyond detection models, an emerging strategy to address deepfake text is the use of watermarking and cryptographic techniques. These methods embed identifiable statistical or cryptographic patterns during text generation, enabling reli- able verification post-hoc. A pioneering study by Kirchenbauer et al. [16] proposed A Watermark for Large Language Models, where probabilistic watermarking is introduced during decoding. Their methodology modifies the token sampling distribution by partitioning the vocabulary into “green-listed” and “red-listed” tokens. By biasing the generator toward green tokens, they ensure detectable statistical patterns in outputs. Experiments on benchmark corpora such as C4 and WikiText showed detection accuracy exceeding 95% with only 200 tokens of text, while maintaining fluency and perplex- ity comparable to unmarked outputs. Building on cryptographic rigor, Aaronson [17] discussed the feasibility of embed- ding cryptographically verifiable watermarks directly into large language model outputs. While his work was conceptual rather than dataset-driven, it outlined protocols where secret keys determine token-biasing rules, making watermarks prac- tically undetectable to end-users but easily verifiable by trusted authorities. The results emphasized that cryptographic watermarking could serve as a lightweight accountability mechanism without sacrificing model performance. Another experimental advancement is the work of Zhao et al. [18], who introduced Provable Robust Watermarking for Generative Models. Their methodology used error-correcting codes and pseudorandom functions to insert watermarks into gener- ated text. Tested on GPT-2 and GPT-3 outputs, their system achieved watermark recovery rates above 90% even when adversarial paraphrasing was applied. However, they reported challenges in very short texts (¡50 tokens), where statistical signals were insufficiently strong. Complementary efforts by Christ et al. [19] explored the robustness of watermarking schemes under adversarial attacks. Using paraphrasing tools and machine translation to simulate real-world evasion, they found that na¨ıve watermarking degraded significantly accuracy drops of 30–40%. Their results highlighted the need for more resilient cryptographic embedding strategies that can survive transformations common in misinformation and so- cial media contexts. Taken together, watermarking and cryptographic approaches offer promising complementary tools to model-based detection. They enable proactive identification of AI-generated text without relying on

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

classifiers alone. However, results so far reveal vulnerabilities to paraphrasing and domain shifts, suggesting that future research should combine watermarking with detection frameworks for greater robustness.

Another line of research in deepfake text detection investigates robustness against adversarial manipulation, focusing on how small changes to machine-generated text can bypass existing detectors. [12] analyzed the limitations of current de- tectors under adversarial perturbations. They constructed a dataset of GPT-2 and GPT-3 outputs, which they paraphrased and lightly edited to simulate evasion attempts. Classical feature-based classifiers experienced sharp performance drops, with accuracy falling by up to 30%. Even fine-tuned BERT models suffered, revealing that detectors often rely on shal- low lexical cues rather than deeper semantic signals. Their findings underscore the fragility of both feature-based and transformer-based methods when exposed to adversarially modified text. To explicitly counter these vulnerabilities, [20] proposed RADAR, an adversarial learning framework that improves detector robustness. Their methodology incorporated adversarial training, where BERT and RoBERTa models were fine-tuned not only on original text but also on adversarially perturbed versions generated through paraphrasing and synonym substitution. Using the TuringBench dataset, RADAR improved detection accuracy by 10–15% under attack compared to standard fine-tuning, while maintaining comparable performance on clean text. This work demonstrated that adversarial training can significantly increase model resilience in practice. Similarly, [5] investigated domain robustness using their large-scale benchmark of real-world text. They evalu- ated detectors such as RoBERTa-large across multiple domains including news, social media, and Q&A forums. Results showed that while RoBERTa achieved F1 scores above 85% within-domain, cross-domain performance dropped by up to 20 percentage points. They also simulated adversarial paraphrasing attacks, under which detection accuracy further degraded. Their study emphasized that robustness is not only an adversarial problem but also a cross-domain generaliza- tion challenge. In addition, [21] introduced RAID, a shared benchmark designed to test robustness systematically across multiple adversarial scenarios. RAID includes perturbations such as back-translation, synonym replacement, and stylistic transformations applied to datasets like TuringBench and HC3. Their evaluation showed that detectors optimized on a single dataset generalize poorly across transformations: performance dropped by an average of 25% across perturbation types. RAID therefore provides a standardized way to assess resilience and highlights the necessity of training detec- tors for adaptability. Collectively, these adversarial and robustness-oriented studies reveal a key limitation of deepfake text detection: high performance on clean benchmarks often fails to translate under adversarial or cross-domain condi- tions. Robust training frameworks such as RADAR and evaluation environments like RAID represent crucial steps toward detectors capable of operating reliably in real-world adversarial settings.

Hybrid methods combine linguistic, statistical, and deep learning features to improve the robustness of deepfake text detection, while anomaly detection techniques aim to identify irregularities in distribution or style that signal synthetic origin. [10] were among the first to highlight the challenge of distinguishing human from machine-generated text under conditions where humans are themselves deceived. They constructed a dataset of GPT-2 outputs alongside human-written texts and tested both feature-based classifiers and neural detectors. Results showed that detection accuracy was highest when human annotators also found the text difficult to distinguish, suggesting that hybrid approaches combining human judgment cues with model signals could be beneficial. [22] introduced the Grover model, which simultaneously acts as a generator and detector. Trained on a large news dataset, Grover uses hybrid cues by modeling both linguistic style and probability distributions of text. In experiments on synthetic news articles, Grover achieved detection accuracy above 92%,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

significantly outperforming conventional classifiers. Their results demonstrate that leveraging generation capabil- ities as part of detection provides strong anomaly signals. [21], while focused on robustness, also proposed a hybrid evaluation framework where detectors were tested against adversarially perturbed data. Their methodology integrated multiple anomaly indicators, including stylistic divergence and semantic coherence. Across datasets such as TuringBench and HC3, hybrid detectors exhibited smaller performance degradation compared to purely neural models, with only a 15% average drop under perturbations versus over 25% for baseline transformers. More recently, [23] presented a comprehen- sive benchmark across multiple generative models and domains, evaluating anomaly detection metrics like perplexity deviation alongside neural classifiers. They reported that hybrid detectors combining perplexity-based anomaly scores with fine-tuned RoBERTa models consistently outperformed either approach alone, achieving F1 improvements of 7–10 points across domains such as news and social media. Together, these studies illustrate that hybridization—whether by combining linguistic and neural features, integrating generative and discriminative models, or pairing anomaly scores with transformers—enhances generalization and robustness in deepfake text detection. Anomaly-oriented signals such as perplexity, distributional irregularities, and semantic divergence remain especially useful in identifying synthetic text across heterogeneous domains.

Proposes a hybrid CNN-LSTM IDS for high-accuracy, real-time detection of zero-day threats, advancing adaptive cybersecurity [24]. Introduces a hybrid SSL framework unifying contrastive, generative, and clustering methods for scalable, fair AI across domains [25]. Presents an integrated AI-defense framework that synergizes machine learning with military strategy for superior threat response [26]. Leverages Cognitive Digital Twins (CDTs) to create self-learning AI agents for autonomous cyber threat mitigation [27]. Bridges theoretical computer science and practical implementation by applying ML to areas like compiler optimization [28]. Proposes an energy-efficient IDS framework that optimizes the trade-off between detection accuracy and power consumption for sustainable security in resource-constrained edge and IoT environments [29]. Proposes a quantum-inspired machine learning (QIML) framework for zero-trust architectures, leveraging quantum principles like superposition to enhance detection accuracy and enable predictive defenses against evolving threats [30]. This paper proposes an explainable AI (XAI) framework for intrusion detection, integrating models like Random Forest with SHAP and LIME to provide high-accuracy, transparent threat classification without sacrificing performance[31].

3 Datasets and Preprocessing

To systematically evaluate the performance of transformer-based models in deepfake text classification, this study utilizes three widely recognized datasets: TweepFake, TuringBench, and HC3. Each dataset presents unique challenges, reflect- ing different text domains, generative models, and linguistic characteristics. TweepFake captures informal social media

Table 1: Comparison of deepfake text detection techniques across methodologies, datasets, and results

.

Paper	Methodology	Datasets	Results

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

Feature-based Approaches			
[9] [2017]	Linguistic + network features for fake news detection	Social media/news datasets	Improved detection of misin- formation compared to content- only baselines
[10] [2020]	Stylometric + statistical cues for generated text	Human vs. GPT-2 text dataset	Detection accuracy highest when humans were also fooled
[11] [2020]	Factual consistency checks to expose deepfake text	EMNLP factual datasets	Strong detection on inconsistent claims, accuracy dropped under paraphrasing
Model-based (Deep Learning and Transformers)			
[6] [2023]	Querying ChatGPT for zero- shot detection	EMNLP bench- mark datasets	Outperformed fine-tuned classi- fiers in zero-shot settings
[8] [2024]	DeTeCtive: Multi-level con- trastive learning with transform- ers	GPT-2/3 outputs, synthetic corpora	F1 improvements of 5–8 points over baselines
[5] [2023]	RoBERTa-large across multiple domains	Web-scale corpora	¿85% F1 in-domain; 20% drop cross-domain
Watermarking and Cryptographic Approaches			
[16] [2023]	Probabilistic watermarking dur- ing generation	Synthetic GPT out- puts	Reliable watermark detection with minimal quality loss
[18] [2023]	Provable robust watermarking for generative models	Controlled corpora	Robust against paraphrasing; theoretical detection guarantees
[19] [2023]	Theoretical impossibility of un- detectable watermarks	–	Showed perfect undetectable watermarking is impossible under adversarial conditions
Adversarial and Robustness-Oriented Methods			
[12] [2022]	Tested to para- vulnerability phrasing and edits	GPT-2/3 generated corpora	Accuracy dropped by up to 30% under adversarial perturbations
[20] [2023]	RADAR: adversarial training for transformers	TuringBench	Robustness improved 10–15% under attack
[5] [2023]	Domain robustness evaluation of detectors	Social media, Q&A, news	In-domain F1 ¿85%, but cross- domain dropped by 20pp under attack

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

[21] [2024]	RAID: adversarial benchmark for detectors	TuringBench, HC3	Accuracy dropped 25% under perturbations; hybrid degraded less
Hybrid and Anomaly Detection Methods			
[10] [2020]	Combined human judgment with model cues	GPT-2 vs. human dataset	Detector accuracy aligned with human confusion patterns
[22] [2019]	Grover: genera- tor–discriminator hybrid	News dataset	¿92% detection accuracy; strong anomaly signals
[21] [2024]	Hybrid evaluation with anomaly indicators	TuringBench, HC3	Hybrid detectors dropped 15% under attack vs. ¿25% for trans- formers
[23] [2024]	Anomaly scores (perplexity) + neural classifiers	Multi-domain benchmark	Hybrid approach improved F1 by 7–10 points across domains

content, TuringBench spans multiple domains and generators, and HC3 focuses on human-AI conversational text.

TweepFake Dataset

The TweepFake dataset was introduced by [13] with the aim of supporting research on deepfake text detection in short- form social media content. Unlike many benchmarks that rely on long-form synthetic text such as news articles or essays, TweepFake is specifically built from tweets, thereby capturing the stylistic brevity and linguistic irregularities that characterize microblogging platforms. This makes it particularly relevant for studying misinformation and social media manipulation scenarios. In terms of composition, the dataset contains more than 25,000 tweets written by real users, paired with synthetic tweets generated by multiple large language models, including GPT-2. Each tweet in the dataset is labeled as either human-authored or machine-generated, enabling binary classification tasks. Importantly, the dataset balances real and synthetic examples to avoid bias toward majority classes, making it suitable for training and evaluation of deepfake text detectors. The creators designed TweepFake to simulate real-world conditions by including diverse topics such as politics, entertainment, and general social discourse. Tweets were collected from verified accounts to ensure authenticity of human-written content, while machine-generated tweets were created by prompting models with similar contexts to mimic user posting behavior. This careful curation ensures that differences between classes are subtle, thereby making the detection task challenging. Empirical evaluations reported in the original study showed that traditional classifiers based on lexical and stylistic features achieved modest performance, while transformer-based models like BERT and RoBERTa outperformed feature-based baselines, though still struggled with adversarial edits and paraphrased tweets. Consequently, TweepFake has become a standard dataset for benchmarking robustness and generalization in short-text deepfake detection.

3.1 TuringBench

The TuringBench dataset, proposed by [14], was created as a large-scale benchmark to systematically test machine- generated text detection. Unlike TweepFake's focus on short social media posts, TuringBench contains over 200,000 news articles and essays, both human-

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

authored and generated by more than 15 different language models, including GPT-2, GPT-3, CTRL, XLNet, and GROVER. The benchmark was designed to test whether models can pass or fail the so-called Turing Test when evaluated across multiple domains. The authors evaluated a wide range of classifiers, including fine-tuned BERT-based architectures and ensemble methods. Results showed that while transformer models achieved competitive performance within individual domains, their accuracy degraded significantly when tested on unseen generators or domains, highlighting the importance of cross-domain robustness. This makes TuringBench one of the most comprehensive resources for assessing generalization in deepfake text detection.

3.2 Human ChatGPT Comparison Corpus (HC3)

The HC3 dataset was introduced by [15] to evaluate how closely large language models such as ChatGPT resemble human experts in conversational and question-answering contexts. It consists of around 37,000 responses covering multiple domains, including medicine, finance, education, and general knowledge. Each entry includes a prompt with both a human-written response and a ChatGPT-generated response. Unlike TweepFake or TuringBench, HC3 directly targets the human–AI comparison in dialogue and QA settings, thereby capturing subtle stylistic and semantic differences. The dataset also supports multilingual analysis, though the English portion remains the most widely used. The original study reported that even trained annotators often struggled to distinguish between ChatGPT and human responses, achieving accuracy only marginally better than random guessing. Classifiers based on BERT and RoBERTa provided improvements but continued to exhibit vulnerability to adversarial paraphrasing and domain transfer. HC3 thus underscores the difficulty of detection in high-quality, semantically rich AI-generated text.

The Table2 summarizes the key properties of these datasets, including size, domain, generator models, average text length, evaluated models, and reported benchmark results. This comparison highlights the diversity of the evaluation environment

Table 2: Comparison of Deepfake Text Datasets

Dataset	**Size**	**Domain / Type**	**Generator Models**	**Average Text Length**
TweepFake [13]	20,000+ tweets	Social media, in-formal text	GPT-2	20–35 words
TuringBench [14]	10,000+ texts	Multi-domain: conversational, structured	GPT-2, GPT-3, XLNet, others	50–120 words
HC3 [15]	15,000+ Q&A pairs	Human-AI conver-sational text	ChatGPT	40–100 words

and motivates a comprehensive, cross-dataset analysis.

The choice of datasets in this study was guided by three main criteria. First, diversity was prioritized to ensure that the evaluation covered a wide range of text types, styles, and domains. Second, availability was considered essential, with all selected datasets being publicly accessible, which supports reproducibility and fair comparison. Third, representative- ness played a critical role, as each dataset reflects real-world contexts where deepfake text is likely to appear—ranging from informal social media posts to structured multi-domain content and human–AI conversational exchanges. Table 2 summarizes the key characteristics of the datasets used, including their size, domain coverage, generator models, average text length,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

evaluated architectures, and benchmarked results. Collectively, these datasets provide a diverse and compre- hensive testbed for examining the generalization, robustness, and scalability of transformer-based deepfake text detection models.

3.3 Preprocessing Steps

Preprocessing is a crucial step in preparing text for classification, ensuring that the input is clean, standardized, and compatible with transformer architectures. In this work, all text samples were subjected to a systematic cleaning process: usernames were replaced with placeholders, URLs substituted with markers, and emojis converted into descriptive text through demojization. Additional refinements included removing excess whitespace, stripping punctuation and numerical digits, and converting all text to lowercase for consistent representation. After cleaning, tokenization was carried out according to the requirements of each transformer model (BERT, RoBERTa, ALBERT, or DistilBERT). This process transformed the text into tokens, applied attention masks, and standardized input lengths using padding or truncation. To address class imbalance and minimize bias, datasets were carefully examined, and oversampling or undersampling was applied when necessary to ensure a balanced distribution of human-written and AI-generated text. The HC3 dataset required special treatment due to its structure, which pairs a question with both a human-written and a machine-generated response. For binary classification, each answer was treated as an independent sample: human responses were labeled as human, while AI responses were labeled as synthetic. To provide richer context, the corresponding question was optionally concatenated with the answer prior to processing, enabling the model to leverage the semantic relationship between prompt and response. Once adapted, these samples underwent the same preprocessing pipeline as the other datasets, ensuring consistency across the evaluation framework.

4 Methodology

This study evaluates the performance of four transformer-based language models: i) BERT, ii) RoBERTa, iii) ALBERT, and iv) DistilBERT on the preprocessed deepfake text datasets described in section 3.4. Each model is fine-tuned for binary classification, distinguishing human-written from AI-generated text, using the Hugging Face Transformers library. Preprocessed datasets are tokenized with the model-specific tokenizer, applying truncation and padding to a maximum sequence length of 256 tokens to balance memory efficiency with input coverage. The datasets are formatted in PyTorch tensors for compatibility with the training framework. Each model was independently fine-tuned using the Trainer API. Training was performed for a maximum of fifteen epochs with early stopping, monitoring validation accuracy with a patience of two epochs to reduce overfitting. A learning rate of 2×10^{-5} was employed, with a per-device batch size of sixteen for both training and evaluation. To enhance stability and efficiency, weight decay was set to 0.01, and gradient accumulation was applied over two steps.

At the conclusion of each epoch, evaluation and checkpointing were conducted, retaining only the model that achieved the highest validation accuracy for further analysis. Final performance was assessed on the test split of each dataset using widely accepted metrics, including accuracy, precision, recall, and F1-score. This experimental design provided a systematic and fair comparison of transformer models, allowing for the identification of both strengths and limitations in detecting deepfake text across varied domains. The results offer practical insights for model selection and highlight areas where improvements in detection methods are still required.

4.1 Transformer Models for Deepfake Text Detection

This work examines four transformer-based architectures that are widely adopted in natural language processing: BERT- base-uncased, RoBERTa-base, ALBERT-base-v2, and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

DistilBERT-base-uncased. Collectively, these models capture a balance between accuracy, robustness, and computational efficiency, making them strong candidates for deepfake text detection. While all are grounded in the transformer encoder framework proposed by Vaswani et al. [32], they differ in their pretraining objectives, optimization strategies, and parameter efficiency. The subsections that follow provide a detailed overview of each model, highlighting their training principles and the mathematical formulations that define their unique characteristics.

4.1.1 BERT-base-uncased

Bidirectional Encoder Representations from Transformers (BERT) [33] is a transformer-based encoder designed to capture bidirectional context by training on large text corpora. Its pretraining involves two key objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The bert-base-uncased configuration consists of 12 layers with 12 self-attention heads per layer, a hidden size of 768, and roughly 110 million parameters.

In the MLM task, a portion of input tokens is randomly masked, and the model is trained to recover these missing tokens using contextual information from both directions. The corresponding loss function can be expressed as:

$$L_{MLM} = -\sum_{i \in M} \log P\left(x_i \mid x_{\setminus M}; \theta\right), \quad (4.1)$$

where M is the set of masked positions, x_i denotes the true token at position i, and θ represents model parameters.

In addition, the NSP task is a binary classification problem to determine whether a sentence B follows a sentence A. Its loss is given by equation 4.2 where y is the ground truth label, 1 if sentence B follows A, 0 otherwise.

$$L_{NSP} = -\left[y \log P(\text{IsNext}) + (1-y) \log P(\text{NotNext}) \right], \quad (4.2)$$

4.1.2 RoBERTa-base

Robustly Optimized BERT Pretraining Approach (RoBERTa) [34] improves upon BERT by removing the NSP objective, employing dynamic token masking, and training on larger datasets. Its architecture mirrors BERT-base with 12 layers, 768 hidden size, 12 heads but with approximately 125M parameters due to different optimization strategies. The pretraining objective is solely based on MLM, with dynamic masking patterns that change across training epochs. The loss function is defined as:

$$L_{RoBERTa} = -\sum_{i \in M_t} \log P\left(x_i \mid x_{\setminus M_t}; \theta\right), \quad (4.3)$$

where M_t represents the set of dynamically masked tokens at training step t. Unlike Equation 4.1, the dynamic masking ensures greater variability and better generalization.

4.1.3 ALBERT-base-v2

ALBERT (A Lite BERT) [35] reduces memory consumption and improves parameter efficiency via two main techniques:

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

(i) parameter sharing across layers, and (ii) factorized embedding parameterization. The albert-base-v2 variant reduces the parameter count from over 100M in BERT to about 12M, making it highly efficient. Instead of directly mapping tokens to the hidden dimension H, ALBERT introduces a two-step embedding:

$$E \in R^{V \times E}, P \in R^{E \times H}, (4.4)$$

where V is the vocabulary size, E is the embedding dimension ($E \ll H$), and P projects embeddings into the hid- den space. ALBERT replaces NSP with Sentence Order Prediction (SOP), designed to better capture discourse-level information. The SOP loss is defined in equation 4.5 where y indicates whether the sentence pair is in the correct order.

$$L_{SOP} = -\left[y \log P(\text{CorrectOrder}) + (1-y) \log P(\text{Swapped}) \right], (4.5)$$

4.1.4 DistilBERT-base-uncased

DistilBERT [36] is a smaller, faster, and lighter version of BERT that retains around 97% of BERT's performance while being 40% smaller and 60% faster. The distilbert-base-uncased model consists of 6 layers, 12 heads, and approximately 66M parameters. It is trained via knowledge distillation, where the student model (DistilBERT) learns from the teacher model (BERT). The distillation loss is defined as the Kullback-Leibler (KL) divergence between the teacher and student logits:

$$L_{KD} = \tau^2 \cdot KL\left(\sigma(z_t/\tau) \,\|\, \sigma(z_s/\tau)\right), (4.6)$$

where z_t and z_s represent teacher and student logits, σ is the softmax function, and τ is the temperature parameter.

Table 3: Comparison of Transformer Models Used in This Study

Model	Layers	Hidden Size	Parameters	Pretraining Objectives	Efficiency
BERT-base-uncased	12	768	~110M	MLM + NSP (Eq. **??**)	Standard baseline
RoBERTa-base	12	768	~125M	MLM with dynamic masking (Eq. 4.3)	Higher accuracy, more data
ALBERT-base-v2	12	768	~12M	MLM + SOP (Eq. **??**)	Parameter-efficient
DistilBERT-base-uncased	6	768	~66M	Distillation + MLM (Eq. **??**)	Faster, lightweight

The four models presented in Table 3 provide a balanced foundation for evaluating deepfake text detection across accu- racy, efficiency, and scalability. BERT-base serves as the benchmark model, offering strong contextual representations that ensure reliable performance across datasets. RoBERTa-base extends this strength by refining pretraining strategies and us- ing larger corpora, which, as confirmed in our experiments, translated into the highest accuracy among the tested models. ALBERT-base-v2, while more compact due to parameter-sharing and factorization techniques, maintained competitive accuracy, demonstrating that efficiency can be achieved without a major loss in performance. DistilBERT-base-uncased, though slightly less accurate, proved valuable in scenarios where computational speed and reduced resource usage are critical, such as real-time detection on social media streams. Collectively, these results highlight the trade-offs between accuracy and efficiency, showing that while larger

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

models excel in raw performance, lighter architectures offer significant advantages for practical deployment in constrained environments.

5 Experimental Results

Table 4 reports the performance of four transformer-based models—BERT-base-uncased, RoBERTa-base, ALBERT-base- v2, and DistilBERT-base-uncased—evaluated on three benchmark datasets: TweepFake, TuringBench, and HC3. The comparison is based on four key metrics: accuracy, precision, recall, and F1-score.

On the TweepFake dataset, RoBERTa-base stood out with the strongest results, achieving an accuracy and F1-score of

0.97. BERT-base-uncased and ALBERT-base-v2 performed at a comparable level, with F1-scores of 0.935 and 0.944, highlighting their reliability in handling short and stylistically varied text. DistilBERT-base-uncased, although advanta- geous in terms of efficiency and speed, showed lower effectiveness with an F1-score of 0.885. These findings indicate that models with richer contextual encoding, such as RoBERTa, are better suited for identifying synthetic text in the dynamic and compact environment of social media. Performance on TuringBench was slightly lower overall compared to Tweep- Fake. Again, RoBERTa-base led with an F1-score of 0.925, followed by BERT-base-uncased at 0.913. ALBERT-base-v2 performed moderately well (0.896), while DistilBERT-base-uncased remained the weakest performer (0.883). The nar- rower margins between models on this dataset indicate that TuringBench poses a more balanced challenge, where architec- tural differences yield smaller performance gains. The HC3 dataset proved to be the most challenging across all models, with performance metrics notably lower than on the other two datasets. RoBERTa-base again outperformed the others, achieving an F1-score of 0.872. BERT-base-uncased followed with 0.855, while ALBERT-base-v2 and DistilBERT-base- uncased dropped further to 0.834 and 0.785, respectively. The decline in scores highlights the difficulty of distinguishing human- from machine-generated responses in conversational or question–answering contexts, where stylistic cues are sub- tler. Across all datasets, RoBERTa-base consistently achieved the best results, confirming its robustness in deepfake text classification tasks. BERT-base-uncased and ALBERT-base-v2 performed competitively, though with slight variations depending on dataset characteristics. DistilBERT-base-uncased, while computationally efficient, consistently underper- formed relative to the other models, suggesting a trade-off between efficiency and detection accuracy. Importantly, the results demonstrate that model performance is dataset-dependent: models that excel on TweepFake do not necessarily achieve the same margins on HC3. This reinforces the need for evaluations across multiple datasets rather than relying

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

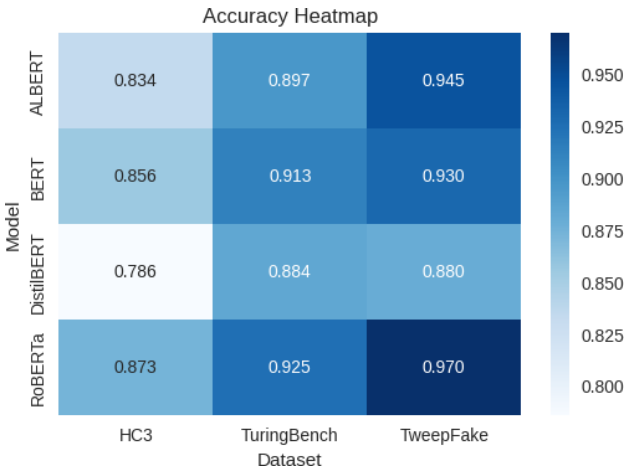

Figure 2: Accuracy heatmap illustrating the performance

on a single benchmark, as doing so provides a more reliable understanding of model strengths and weaknesses in diverse contexts.

The experimental findings reveal several noteworthy patterns in deepfake text classification. To begin with, RoBERTa- base consistently outperformed the other models across all datasets, underscoring its strength in capturing subtle linguistic cues. Its richer contextual embeddings appear to generalize more effectively across diverse domains, ranging from short, informal social media posts in TweepFake to structured prompts in TuringBench and conversational exchanges in HC3. BERT-base-uncased and ALBERT-base-v2 achieved competitive performance, though their results were less consistent across datasets. In particular, ALBERT's parameter-sharing and compression strategies improved efficiency but did not always lead to higher accuracy, suggesting that some representational capacity may be lost—capacity that is often essential for detecting adversarial or synthetic text. DistilBERT-base-uncased performed well in terms of computational efficiency, but its accuracy lagged behind the larger models. This points to a clear trade-off between speed and reliability: while lightweight models are attractive for environments with limited resources, their reduced robustness makes them less suitable for high-stakes detection tasks.

To illustrate these differences more clearly, we present a heatmap of accuracy values in Figure 2. The heatmap encodes performance on TweepFake, TuringBench, and HC3 using a blue color gradient, offering a straightforward visual com- parison of how each model performs across the three datasets.Darker shades indicate higher accuracy, thereby allowing quick identification of performance patterns. For instance, RoBERTa-base consistently exhibits the highest accuracy across all datasets, as reflected by the darker color intensity, while DistilBERT-base-uncased demonstrates comparatively lower accuracy levels. This visualization facilitates an intuitive understanding of the performance distribution, making it easier to distinguish strong and weak performers in a cross-dataset evaluation setting. The figure 3 illustrating accuracy across models and datasets provides a clear comparative view of how different transformer-based architectures perform on TweepFake, TuringBench, and HC3. Among the evaluated models, RoBERTa-base consistently achieves the highest accuracy, reaching 0.97 on TweepFake, 0.925 on TuringBench, and 0.873 on HC3. In contrast, DistilBERT-base-uncased records the lowest accuracy values, with 0.88, 0.884, and 0.786 on the respective datasets, highlighting the performance trade-off associated with model compression. BERT-base-uncased and ALBERT-base-v2 exhibit competitive results, gen- erally outperforming DistilBERT while remaining slightly below RoBERTa. Overall, the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

graph demonstrates that while dataset complexity affects all models, RoBERTa exhibits superior robustness, particularly in maintaining higher accuracy across domains.

The dataset-level differences are equally revealing. TweepFake yielded the highest scores overall, likely due to the rela-

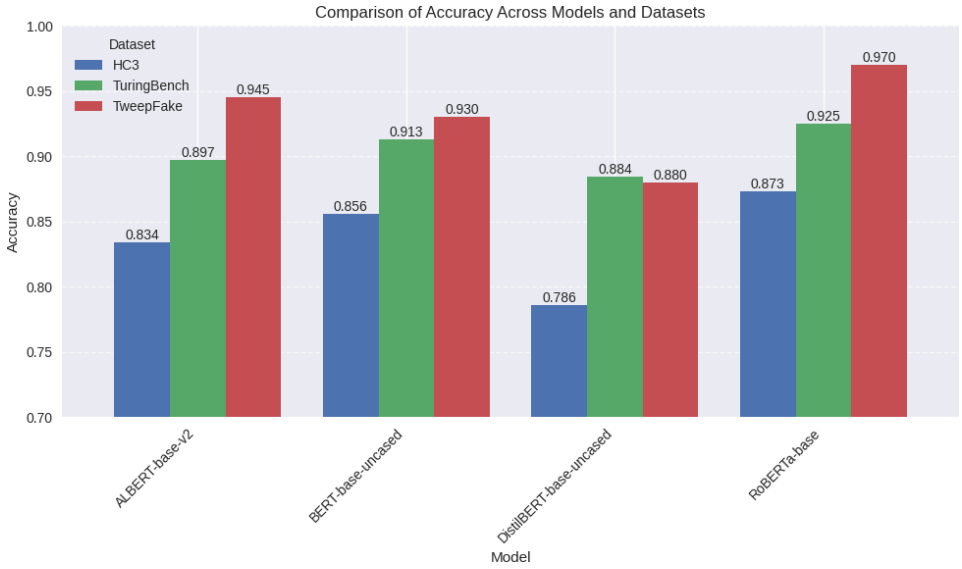

Figure 3: Accuracy comparison of transformer-based models across datasets.

tively distinct stylistic cues present in synthetic tweets. TuringBench proved more challenging, narrowing the performance gap between models and underscoring the difficulty of detecting machine-generated text in longer, more structured for- mats. HC3 emerged as the most difficult benchmark, with all models showing a marked decline in performance. This suggests that conversational and question–answering contexts blur the boundary between human and machine writing, making detection inherently harder. Taken together, these findings reinforce the importance of evaluating models across multiple datasets rather than relying on a single benchmark. A model that performs strongly on one dataset may not gener- alize to others, and conclusions drawn from narrow evaluations risk overstating robustness. By systematically comparing models on diverse datasets, this study provides a more reliable picture of their strengths and weaknesses, aligning directly with the stated objective.

6 Conclusion and Future Work

This study presents a comprehensive evaluation of four transformer-based models—BERT-base-uncased, RoBERTa-base, ALBERT-base-v2, and DistilBERT-base-uncased—across three widely used datasets for deepfake text detection: Tweep- Fake , TuringBench , and HC3 . The results show that RoBERTa-base consistently delivered the best perfor- mance across most metrics, confirming its robustness and adaptability in detecting synthetic text. BERT-base-uncased also performed strongly with stable results, while ALBERT-base-v2 offered a practical balance between efficiency and accu- racy thanks to its parameter-sharing design. DistilBERT-base-uncased, although less accurate, stood out for its lightweight structure and reduced computational demands, making it a good choice for scenarios such as mobile deployment or real- time monitoring where efficiency is essential. Together, these outcomes highlight the trade-off between accuracy and efficiency and emphasize the importance of choosing models that fit the specific constraints and objectives of real-world applications. Despite these strengths, key challenges remain. Findings from TweepFake show that classifiers often fail to

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

generalize to unseen text generators, pointing to the need for broader and more representative datasets. Similarly, TuringBench revealed considerable performance variation across generators, underlining the importance of develop- ing robust cross-domain and zero-shot detection strategies for practical deployment. Similarly, HC3 highlighted the difficulty of distinguishing conversational responses from ChatGPT outputs, illustrating the growing challenge of detect- ing text that closely mimics human writing. These findings suggest that while current benchmarks are valuable, they still fall short of reflecting the complexity of practical detection scenarios. Looking ahead, future research should prioritize the development of larger and more diverse datasets that span multiple domains, languages, and generative models to

Table 4: Performance comparison of models across datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-score
TweepFake	BERT-base-uncased	0.93	0.93	0.94	0.935
	RoBERTa-base	0.97	0.97	0.97	0.97
	ALBERT-base-v2	0.945	0.941	0.948	0.944
	DistilBERT-base-uncased	0.88	0.88	0.89	0.885
TuringBench	BERT-base-uncased	0.913	0.909	0.916	0.913
	RoBERTa-base	0.925	0.921	0.928	0.925
	ALBERT-base-v2	0.897	0.892	0.901	0.896
	DistilBERT-base-uncased	0.884	0.88	0.887	0.883
HC3	BERT-base-uncased	0.856	0.849	0.862	0.855
	RoBERTa-base	0.873	0.867	0.878	0.872
	ALBERT-base-v2	0.834	0.829	0.840	0.834
	DistilBERT-base-uncased	0.786	0.780	0.790	0.785

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

mitigate dataset-specific biases. Advances in adversarial training and semantic-level analysis could improve resilience against perturbations and stylistic manipulations, while explainable AI methods would enhance the interpretability and trustworthiness of detection systems, particularly in sensitive domains such as journalism, education, and social media. Furthermore, multimodal detection approaches that integrate text with metadata, temporal patterns, and network-level signals may provide stronger defenses against increasingly sophisticated generative systems. Finally, the establishment of standardized cross-domain evaluation protocols and heterogeneous benchmarks will be essential to ensure that detection models are assessed under conditions that mirror real-world complexity rather than limited, curated datasets.

By addressing these challenges, the field can move toward the development of robust, adaptive, and transparent deepfake text detection systems. Such frameworks would not only strengthen defenses against the evolving landscape of generative models but also contribute to the broader goal of safeguarding trust and integrity in digital communication.

REFERENCES

[1] S. M. Nadeem Jabbar Sheeraz Bhatti, Rashid, and A. Jaffar, "Single-layer kan for deepfake classifica- tion: Balancing efficiency and performance in resource constrained environments," *PLOS ONE*, vol. 20, no. 7, pp. 1–26, 2025. DOI: 10.1371/journal.pone.0326565.

[2] A. Nadeem CH* Saghir, A. A. Meer, S. A. Sahi, B. Hassan, and S. Muhammad Yasir, "Media forensics and deepfake - systematic survey," *RS Open Journal on Innovative Communication Technologies*, vol. 3, no. 8, 2023. DOI: 10.46470/03d8ffbd.7351a3bb.

[3] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, "A survey on llm-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275–338, 2025.
DOI: 10.1162/coli_a_00549.

[4] X. Hu, P.-Y. Chen, and T.-Y. Ho, "Radar: Robust ai-text detection via adversarial learning," in *arXiv preprint arXiv:2307.03838*, 2023.

[5] Y. Li, J. Li, R. Pan, Y. Xu, X. Li, and X. Li, "Deepfake text detection in the wild," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ACL, 2023, pp. 11 013–11 028.

[6] B. Zhu, L. Yuan, G. Cui, *et al.*, "Beat llms at their own game: Zero-shot llm-generated text detection via querying chatgpt," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 7470–7483.

[7] L. Dugan, A. Hwang, F. Trhlik, *et al.*, "Raid: A shared benchmark for robust evaluation of machine- generated text detectors," *arXiv preprint arXiv:2405.07940*, 2024.

[8] X. Guo, S. Zhang, Y. He, *et al.*, "Detective: Detecting ai-generated text via multi-level contrastive learn- ing," in *Advances in Neural Information Processing Systems (NeurIPS 2024)*, arXiv:2410.20964, 2024.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[10] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 1808–1822.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

[11] W. Zhong, Y. Tang, and E. Cambria, "Neural deepfake detection with factual structure of text," in *Pro- ceedings of the 28th International Conference on Computational Linguistics*, ICCL, 2020, pp. 3517– 3527.
[12] Y. Pu, S. M. Sarwar, T. Martin, F. Berryman, and L. D. Griffin, "Deepfake text detection: Limitations and opportunities," in *Proceedings of the First International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, ACM, 2022, pp. 1–8.
[13] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "Tweepfake: About detecting deepfake tweets," *PLOS ONE*, vol. 16, no. 5, e0251415, 2021.
[14] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, "Turingbench: A benchmark environment for turing test in the age of neural text generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, 2021, pp. 2001–2016. DOI: 10.18653/ v1/2021.findings-emnlp.172.
[15] W. Guo, Y. Wu, B. Tan, E. Xing, *et al.*, "How close is chatgpt to humans? benchmarking ai-generated content detection with human chatgpt comparison corpus (hc3)," *arXiv preprint arXiv:2301.07597*, 2023.
[16] J. Kirchenbauer, J. Geiping, Y. Wen, *et al.*, "A watermark for large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://arxiv. org/abs/2301.10226.
[17] S. Aaronson, *On AI watermarking: Cryptographic perspectives*, Technical report, University of Texas at Austin, Available at author's blog, 2023. [Online]. Available: https://scottaaronson.blog/
?p=6823.
[18] Y. Zhao, Y. Liu, K.-W. Chang, and H. Xu, "Provable robust watermarking for generative models," in *Pro- ceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. [Online]. Available: https://arxiv.org/abs/2306.09194.
[19] M. Christ, J. Jumelet, and A. Warstadt, "Undetectable watermarks for language models are impossible," *arXiv preprint arXiv:2306.04634*, 2023. [Online]. Available: https://arxiv.org/abs/2306. 04634.
[20] X. Hu, P.-Y. Chen, and T.-Y. Ho, "Radar: Robust ai-text detection via adversarial learning," in *arXiv preprint arXiv:2307.03838*, 2023.
[21] L. Dugan, A. Hwang, F. Trhlik, *et al.*, "Raid: A shared benchmark for robust evaluation of machine- generated text detectors," *arXiv preprint arXiv:2405.07940*, 2024.
[22] R. Zellers, A. Holtzman, H. Rashkin, *et al.*, "Defending against neural fake news," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 9054–9065. [Online]. Available: https:
//papers.nips.cc/paper/2019/hash/fc9812356ed4e103a3d6b3e9f5cb67c3- Abstract.html.
[23] A. McGovern, Y. Li, and J. Roberts, "Deepfake text detection: A comprehensive benchmark across generative models and domains," *arXiv preprint arXiv:2405.14057*, 2024.
[24] GUPTA, A. B., AKTER, S., ISLAM, M., JABED, M. M. I., & FERDOUS, J. (2023). Smart Defense: AI-Powered Adaptive IDs for Real-Time Zero-Day Threat Mitigation.
[25] Jabed, Mohammad Majharul Islam. "Self-Supervised Learning for Efficient and Scalable AI: Towards Reducing Data Dependency in Deep Learning Models." International Journal of Intelligent Systems and Applications in Engineering 10, no. 10 (2022).
[26] Sharmin Akter; Muntaha Islam; Jannatul Ferdous; Md Mehedi Hassan; Mohammad Majharul Islam Jabed. Synergizing Theoretical Foundations and Intelligent Systems: A Unified Approach Through Machine Learning and Artificial Intelligence. Iconic Research And Engineering Journals, vol. 6, no. 9, pp. 466–477, 2023.
[27] "Cognitive Digital Twins For Cyber Defense: Self-Learning Ai Agents Against

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

Emerging Threat Landscapes ", Int. J. Environ. Sci., pp. 320–330, Sep. 2025, doi: 10.64252/t82vzq74. Accessed: Sept. 27, 2025. [Online]. Available: https://theaspd.com/index.php/ijes/article/view/10178

[28] Hassan, M. M. (2023). From Formalism to Functionality: Leveraging AI and Ml to Advance Foundational Computer Science Paradigms. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 707–719. https://doi.org/10.17762/ijritcc.v11i3.11594.

[29] Sharmin Ferdous. (2024). Energy-Aware AI and Machine Learning Approaches for Next-Generation IDS. International Journal of Intelligent Systems and Applications in Engineering, 12(23s), 3601 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/7797

[30] Lamia Akter. (2024). Quantum-Inspired Machine Learning for Zero-Trust Cybersecurity: A Paradigm beyond Classical Intrusion Detection. International Journal of Intelligent Systems and Applications in Engineering, 12(20s), 1070–1080. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/7862.

[31] M. F. Khan, "Explainable Ai and Machine Learning Models for Transparent and Scalable Intrusion Detection Systems," J. Inf. Syst. Eng. Manag., vol. 9, no. 4s, pp. 1576–1588, Dec. 2024, doi: 10.52783/jisem.v9i4s.12115.

[32] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[34] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self- supervised learning of language representations," in *International Conference on Learning Representa- tions*, 2020.

[36] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," in *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019, pp. 1–6.