M-DAIGT-ST 2025

Proceedings of the Multi-Domain Detection of AI-Generated Text Shared Task

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Salima Lamsiyah, Saad Ezzini, Abdelkader El Mahdaouy,
Hamza Alami, Abdessamad Benlahbib, Samir El Amrany,
Salmane Chafik and Hicham Hammouchi

11 September, 2025
Varna, Bulgaria

The Multi-Domain Detection of AI-Generated Text Shared Task
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
11 September 2025

Online ISBN 978-954-452-108-0

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

Welcome to the Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task, held on September 11, 2025, in Varna, Bulgaria, as part of the Recent Advances in Natural Language Processing (RANLP) conference.

The last few years have witnessed an extraordinary leap in the fluency and versatility of text produced by Large Language Models (LLMs). These developments have not only opened new opportunities in natural language generation but have also raised critical concerns regarding information integrity, authorship verification, and the reliability of academic research. Detecting AI-generated text, particularly in domains where accuracy and trust are paramount, has therefore become an urgent research priority.

The M-DAIGT shared task was created to address this challenge by focusing on the detection of AI-generated text across multiple domains, with an emphasis on two particularly sensitive genres: news articles and academic writing. The task comprised two binary classification subtasks: News Article Detection (NAD) (Subtask 1) and Academic Writing Detection (AWD) (Subtask 2). To support participants, we developed and released a large-scale benchmark dataset containing 30,000 samples, balanced between human-written and AI-generated texts. The AI-generated texts were produced using a variety of modern LLMs (e.g., GPT-4, Claude) and diverse prompting strategies to ensure data diversity and robustness.

A total of 40 unique teams registered for M-DAIGT, of which four submitted final results. All four teams took part in both subtasks, bringing forward a diverse range of methodologies, from transformer-based deep learning models to feature-engineered and hybrid approaches. The proceedings of this shared task present the datasets, evaluation methodology, system descriptions, and results, offering insights into the current state of AI-generated text detection.

We hope that M-DAIGT will serve as a valuable step toward more reliable and domain-adaptive detection methods, and that it will inspire further research addressing the rapidly evolving capabilities of generative AI.

We thank the participating teams for their innovative contributions, the reviewers for their constructive feedback, and the organizing committee for their dedication to making this shared task possible. We look forward to seeing how the outcomes of M-DAIGT will shape future work in this important and dynamic area.

Salima Lamsiyah, General Chair, on behalf of the M-DAIGT organizing committee.

Organizing Committee

Salima Lamsiyah, University of Luxembourg, Luxembourg
Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia
Abdelkader El Mahdaouy, Mohammed VI Polytechnic University, Morocco
Hamza Alami, Sidi Mohamed Ben Abdellah University, Morocco
Abdessamad Benlahbib, Sidi Mohamed Ben Abdellah University, Morocco
Samir El Amrani, University of Luxembourg, Luxembourg
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Hicham Hammouchi, University of Luxembourg, Luxembourg

Table of Contents

Shared Task Program

Thursday, September 11, 2025

09:00–09:10 ***Welcome and Opening Remarks***

09:10–09:25 *M-DAIGT: A Shared Task on Multi-Domain Detection of AI-Generated Text*
salima lamsiyah, Saad Ezzini, Abdelkader El Mahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El amrany, Salmane Chafik and Hicham Hammouchi

09:25–09:40 *AI-Generated Text Detection Using DeBERTa with Auxiliary Stylometric Features*
ANNEPAKA YADAGIRI, L. D. M. S Sai Teja, PARTHA PAKRAY and Chukhu Chunka

09:40–09:55 *Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT)*
Sareem Farooqui, Ali Zain and Dr Muhammad Rafi

09:55–10:10 *A Multimodal Transformer-based Approach for Cross-Domain Detection of Machine-Generated Text*
Mohammad AL-Smadi

10:10–10:25 *Inside the Box: A Streamlined Model for AI-Generated News Article Detection*
Nsrin Ashraf, Mariam Labib and Hamada Nayel

10:25–10:30 ***Closing Remarks, and Wrap-Up by Dr Salima Lamsiyah***

M-DAIGT: A Shared Task on Multi-Domain Detection of AI-Generated Text

Salima Lamsiyah[1], Saad Ezzini[2], Abdelkader El Mahdaouy[3], Hamza Alami[4], Abdessamad Benlahbib[4], Samir El Amrany[1], Salmane Chafik[3], Hicham Hammouchi[1]

[1]University of Luxembourg, Luxembourg
[2]King Fahd University of Petroleum and Minerals, Saudi Arabia
[3]Mohammed VI Polytechnic University, Morocco
[4]Sidi Mohamed Ben Abdellah University, Morocco

Abstract

The generation of highly fluent text by Large Language Models (LLMs) poses a significant challenge to information integrity and academic research. In this paper, we introduce the Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task, which focuses on detecting AI-generated text across multiple domains, particularly in news articles and academic writing. M-DAIGT comprises two binary classification subtasks: News Article Detection (NAD) (Subtask 1) and Academic Writing Detection (AWD) (Subtask 2). To support this task, we developed and released a new large-scale benchmark dataset of 30,000 samples, balanced between human-written and AI-generated texts. The AI-generated content was produced using a variety of modern LLMs (e.g., GPT-4, Claude) and diverse prompting strategies. A total of 46 unique teams registered for the shared task, of which four teams submitted final results. All four teams participated in both Subtask 1 and Subtask 2. We describe the methods employed by these participating teams and briefly discuss future directions for M-DAIGT.

1 Introduction

The recent advancements in large language models have created a paradigm shift in content generation (Naveed et al., 2023; Chang et al., 2024). These models offer numerous opportunities to improve a wide range of applications, including academic research and journalism (Chung et al., 2023). However, their powerful capabilities also raise critical concerns regarding the integrity of the information ecosystem (Wu et al., 2025). In journalism, the potential for large-scale automated generation of misinformation and fake news represents a serious societal threat, with AI-generated articles already appearing on both mainstream and disinformation websites (Wu et al., 2025; Ali et al., 2025). In academia, LLMs challenge the fundamental principles of academic honesty (Bittle and El-Gayar, 2025), and the accessibility of these tools has made it easier for students to generate ghostwritten assignments, contributing to a noticeable rise in academic misconduct (Bittle and El-Gayar, 2025; Go et al., 2025). Research indicates that a significant number of students acknowledge using such tools for their coursework, making it increasingly difficult to distinguish between appropriate academic support and plagiarism (Kovari, 2025).

Distinguishing AI-generated text from human writing is a non-trivial scientific challenge. Modern LLMs produce text that is grammatically correct, stylistically coherent, and often factually plausible, making it difficult to differentiate from human output (Brown et al., 2020; Urlana et al., 2024; Mitchell et al., 2023). Empirical studies have shown that humans, including experienced educators with high confidence in their judgment, perform only marginally better than random chance when attempting to distinguish AI-generated text from human-written content (Urlana et al., 2024). Moreover, recent detection approaches, such as entropy-based statistical methods (Shen et al., 2023), syntactic pattern analysis (Tassopoulou et al., 2021), and neural classifiers (Ippolito et al., 2020; Li et al., 2025), show promise yet remain vulnerable to paraphrasing and prompt variation (Rivera Soto et al., 2025; Kirchenbauer et al., 2023). The field is effectively locked in a technological "arms race": as detection tools improve, so do generative models and the methods used to evade them, including paraphrase attacks and text "humanizers" (Wu et al., 2025; Sadasivan et al., 2023).

Therefore, this rapidly evolving landscape underscores the need for ongoing research and rigorous evaluation methods for AI content detection. The motivation for advancing detection methodologies extends beyond a reactive approach aimed solely at identifying academic dishonesty. Rather, it serves as a proactive strategy to preserve the integrity of the digital information ecosystem. One key con-

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 1–9, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-001

cern is the phenomenon of recursive degradation, where future language models may be trained on vast amounts of unlabeled AI-generated text collected from the internet. This process risks diminishing the quality, originality, and diversity of training data, potentially leading to a degradation of model performance over time (Wang et al., 2024b). Given that news articles and academic publications constitute essential sources of high-quality training data, maintaining their authenticity is crucial for ensuring the long-term robustness, reliability, and generalization capabilities of future AI systems.

To address some of these challenges and to further encourage work on AI-generated text detection, we organized the **Multi-Domain Detection of AI-Generated Text (M-DAIGT)** shared task. M-DAIGT focuses on two domains where the authenticity of text is particularly vital: news articles and academic writing. Specifically, the task is structured into two binary classification subtasks:

- **Subtask 1. News Article Detection (NAD):** Classifying news content as human-written or AI-generated.
- **Subtask 2. Academic Writing Detection (AWD):** Classifying academic texts as human-written or AI-generated.

The key contributions of this work are as follows: (1) the creation and public release of a large and diverse dataset of 30,000 samples specifically designed for AI-generated text detection in the domains of news and academia, featuring outputs from models like GPT-4 and Claude using varied prompts (Wang et al., 2024b); and (2) a comprehensive analysis of participating systems that range from statistical methods to transformer-based detectors (Li et al., 2025; Kuznetsov et al., 2025), offering insights into the current state-of-the-art and highlighting key challenges for future research.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI-generated text detection. Section 3 presents the dataset creation process and evaluation metrics. Section 4 presents the baseline and participant models, along with the evaluation methodology and results. Finally, Sections 5 and 5 conclude the paper and discuss the limitations of the shared task.

2 Related Work

AI-Generated Text Detection Methods. The detection of AI-generated text is a rapidly evolving research domain, with increasing attention due to the widespread development of large language models (Wu et al., 2025). Several methods have been proposed for AI-generated text detection, which can be broadly classified into statistics-based methods, neural-based methods, watermarking, and the use of LLMs as detectors.

Statistics-based approaches aim to exploit intrinsic differences in linguistic features between human and machine-generated texts. Early efforts, such as those Shen et al. (2023) and Tassopoulou et al. (2021), leveraged entropy measures and n-gram frequency analysis to differentiate between text origins. Krishna et al. (2022) utilized sentence repetition patterns, noting that LLMs often assign high probability to repetitive content. DetectGPT (Mitchell et al., 2023) proposed a perturbation-based method to identify whether text lies in negatively curved regions of the log-likelihood landscape.

Neural-based methods dominate recent advances in AI-generated text detection due to their high accuracy and adaptability. Early methods adopt fine-tuned models like BERT (Devlin et al., 2019) and RoBERTa (Solaiman et al., 2019). Furthermore, Ippolito et al. (2020) demonstrated that training on outputs generated using diverse decoding strategies (e.g., top-k sampling, nucleus sampling, temperature control) significantly improves detection robustness. Recently, Li et al. (2025) proposed IRON, a robust adversarial training framework that improves resilience against attacks designed to evade detection systems. Jiao et al. (2025) introduced M-RangeDetector, which enhances model generalization via multi-range attention masks. Similarly, Kuznetsov et al. (2025) provided feature-level interpretability through sparse autoencoders, offering insights into which patterns distinguish AI and human text. Tong et al. (2025) combined reinforced sampling with LLM augmentation for improved fake news detection, while Ali et al. (2025) extended neural classifiers to low-resource languages, specifically addressing Urdu fake news detection. These efforts reflect a growing focus on robustness, explainability, and multilingual applicability in neural detection research.

Watermarking-based approaches offer proactive detection capabilities by embedding or identifying implicit signals in generated text. Early methods include synonym replacement, lexical substitution (Li et al., 2023; Sadasivan et al., 2023), and soft watermarking using curated token lists

(Kirchenbauer et al., 2023). Hidden-space watermarking approaches (Zhao et al., 2023) manipulate token-level probability vectors to introduce tamper-resistant signatures. Some methods, like Bhattacharjee and Liu (2024), aim to exploit surface-level word randomness as a trigger for detection. Recently, Rivera Soto et al. (2025) proposed Paraphrase Inversion, a novel technique to counter paraphrase attacks that aim to remove watermark signals by recovering semantic intent. This approach highlights the challenges posed by adversaries seeking to bypass detection through surface-level text alterations. While many watermarking techniques rely on controlled generation, this method contributes a defensive post-processing solution that does not depend on direct access to generation mechanisms.

Lastly, LLMs themselves are increasingly used as detectors. Tools such as GPTZero[1], ZeroGPT[2], and OpenAI's[3] AI text classifier exemplify this trend. Sadasivan et al. (2023) proposed a zero-shot framework using clustering to differentiate between watermarked and unwatermarked text. Wang et al. (2024b) proposed M4, a comprehensive black-box framework for machine-generated text detection that operates across multiple generators, domains, and languages. Their approach focuses on generalization under realistic, diverse conditions by evaluating detectors on unseen generators and multilingual datasets, setting a new benchmark for robust and scalable AI text detection. More recently, Su et al. (2025) introduced HACo-Det, which focuses on fine-grained detection of human-AI coauthored text, a challenging scenario due to subtle stylistic blending. Go et al. (2025) proposed XDAC, a detection and attribution framework using explainable AI for Korean-language content. Li and Wan (2025) examined how authorial intent and role influence AI-text detectability, emphasizing the social and cognitive dimensions of authorship. These works represent a trend toward leveraging LLMs as meta-models that interpret, explain, and critique textual content.

Benchmark Datasets and Shared Tasks. Standardized evaluation frameworks through both benchmark datasets and shared tasks play a crucial role in advancing AI-generated text detection by providing standardized, diverse, and challenging evaluation settings. A variety of benchmarks have been proposed to test generalization across languages, domains, modalities, and attack scenarios. The MultiSocial dataset (Macko et al., 2025) supports multilingual detection on social media content, while XDAC (Go et al., 2025) introduces explainable detection and attribution for LLM-generated news comments in Korean. Double Entendre (Frohmann et al., 2025) expands detection tasks beyond pure text through a multimodal benchmark focused on audio-based AI-generated lyrics. To assess robustness under adversarial conditions, IRON (Li et al., 2025) incorporates adversarially perturbed examples, and Stress-Testing (Pedrotti et al., 2025) manipulates LLM writing styles to mislead detectors. In parallel, feature-level datasets (Kuznetsov et al., 2025) offer interpretable benchmarks using sparse autoencoders, while M4GT-Bench (Wang et al., 2024b) evaluates black-box detectors across multiple generators, domains, and languages, which is critical for real-world deployment. Additional public resources, such as the *AI-and-Human-Generated-Text* dataset available on Hugging Face[4] and the *GPT-generated Text Detection: Benchmark Dataset and Tensor-based Detection Method* (Qazi et al., 2024), further enrich the landscape of available datasets.

Complementing benchmark datasets, shared tasks have emerged as key drivers of progress in AI-generated text detection by offering standardized, competitive, and collaborative evaluation platforms. The SemEval-2024 Task 8 (Wang et al., 2024a) was specifically designed to evaluate detection systems under multimodal, multidomain, and multilingual settings in a black-box scenario, challenging participants to detect text generated by unseen language models across diverse languages and content types. The task highlighted major real-world concerns such as domain shift, lack of training-time transparency, and linguistic variability. Among the participating systems, TrustAI (Urlana et al., 2024) provided a comprehensive analysis of multi-domain machine-generated text detection techniques, implementing various strategies across statistical, neural, and ensemble approaches. Their findings underscored the importance of domain-specific fine-tuning and robust feature extraction in black-box detection contexts. In parallel, the 1st Workshop on GenAI Content De-

[1] https://gptzero.me/
[2] https://www.zerogpt.com/
[3] https://platform.openai.com/ai-text-classifier
[4] https://huggingface.co/datasets/Ateeqq/AI-and-Human-Generated-Text

tection (GenAIDetect) (Alam et al., 2025), held at COLING 2025, provided a dedicated forum for advancing research on generative content detection. It addressed key challenges such as multilingual robustness, adversarial evasion, and watermarking techniques, fostering discussion around emerging benchmarks and methodological innovation. Building on prior efforts, M-DAIGT focused on AI-generated text detection in two critical domains: news journalism and academic writing. It features two binary classification subtasks, News Article Detection (NAD) and Academic Writing Detection (AWD), supported by a newly released dataset.

3 Datasets and Evaluation Metrics

3.1 Datasets Collection

To support the M-DAIGT shared task, we curated a dataset tailored to evaluate systems on detecting AI-generated news and academic texts.

3.1.1 News dataset:

We gathered 7,000 manually written news articles from the CNN Daily News website, covering more than 40 categories. To create the AI-generated counterparts, we used the titles extracted from these human-written articles as input prompts. Multiple language models were employed to generate news content, including LLaMA3.2-3B-Instruct, Qwen2.5-3B-Instruct, Mistral-7B-Instruct-v2.0, and various models from the GPT family (GPT-4o, GPT-3.5, GPT-4o-mini). Each model was prompted using the standardized prompt shown in Listing 1, with the role definition randomly selected at runtime to encourage stylistic diversity in the generated outputs.

3.1.2 Academic texts dataset:

We collected 7,000 abstracts from published papers on ArXiv, covering a range of categories. To minimize the likelihood of including AI-generated content, only papers published before 2019 were selected. For each human-written abstract, we extracted the corresponding paper title and used it as a prompt to generate an AI-written counterpart. The same models described earlier were employed for this task. Each model was prompted using one of the two prompts shown in Listing 2,

In conclusion, we compiled balanced datasets of manually written and AI-generated texts for both news articles and academic abstracts, totaling 14,000 examples per task. Each dataset was

Listing 1: Prompt's Key Components for Generating News Articles

```
-- Each time this prompt is used, a role is
    randomly selected to influence the
    assistant writing style.

-- Randomly select one of the following
    journalist roles:

Role Definition:
    - "You are an expert journalist."
    - "You are a professional news writer with
        a focus on clear, unbiased reporting."
    - "You are a friendly and engaging
        journalist, writing in an
        easy-to-understand style."
    - "You are an opinion writer, focusing on
        offering personal insights on current
        news."

-- Instructions:

Generate a news article of approximately
    '{article_length}'-words on the following
    topic: '{Title}'

Write only the article content. Do not
    include a title or any additional
    commentary.
```

Listing 2: Prompts for Generating Scientific Abstracts

```
-- Prompt 1:

You are a researcher working on a research
    paper. Your English proficiency level is
    '{english_proficiency}'.
Your task is to write a well-structured
    abstract of approximately 250 words for
    your research paper in response to the
    given topic: '{title}'.
Ensure your abstract is clear and concise,
    following the standard format:
    'background', 'objective', 'methodology',
    'key findings', and 'conclusion'.
The response should contain only the abstract
    text, without titles or introductory
    phrases.

-- Prompt 2:

Generate a 250-word abstract for work with
    the given topic: '{title}'.
Describe the 'results obtained', the
    'problem' the work attempts to solve, and
    the 'key ideas' and 'methodology' in a
    formal academic and scientific writing
    voice.
Use the first plural person form. Use active
    voice.
Please provide only the abstract text,
    excluding any titles or introductory
    phrases.
```

randomly divided into 10,000 samples for training, 2,000 for development, and 2,000 for testing, providing a robust foundation for evaluating models performance across different stages.

3.2 Evaluation Metrics

The performance of the participating systems in both the News Article Detection (NAD) and Academic Writing Detection (AWD) subtasks was evaluated based on standard classification metrics. The official ranking of the teams was determined by the F1-score. The primary metrics used for evaluation were:

- **Accuracy:** The proportion of correctly classified instances.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of a model's performance.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall:** The ratio of correctly predicted positive observations to all observations in the actual class.

In addition to these primary metrics, a secondary analysis was planned to assess model robustness across different text lengths, writing styles, topic domains, and the various generation models used to create the dataset.

4 Shared Task Teams & Results

In this section, we present the shared task baseline models, participating systems descriptions, and their obtained results.

4.1 Baselines

We evaluate three simple baselines on both subtasks:

- **ARBERTv2:** A transformer-based model pretrained on large-scale Arabic text (Abdul-Mageed et al., 2021), fine-tuned on each task (5 epochs, learning rate 2×10^{-5}).
- **LogReg (char 2–5):** Logistic Regression using character-level n-grams (2–5) with TF–IDF features, designed to capture fine-grained morphological patterns.

Model	P	R	F_1	Supp.
human				
ARBERTv2	0.9979	0.9410	0.9686	1000
LogReg (char 2–5)	0.9791	0.9820	0.9805	1000
LogReg (word 1–2)	0.9679	0.9940	0.9808	1000
machine				
ARBERTv2	0.9442	0.9980	0.9703	1000
LogReg (char 2–5)	0.9819	0.9790	0.9805	1000
LogReg (word 1–2)	0.9938	0.9670	0.9802	1000
accuracy			0.9695	2000
macro avg	0.9710	0.9695	0.9695	2000
weighted avg	0.9710	0.9695	0.9695	2000

Table 1: Task 1 (NAD) development set.

Model	P	R	F_1	Supp.
human				
ARBERTv2	0.9946	0.9240	0.9580	1000
LogReg (char 2–5)	0.9759	0.9700	0.9729	1000
LogReg (word 1–2)	0.9529	0.9920	0.9721	1000
machine				
ARBERTv2	0.9290	0.9950	0.9609	1000
LogReg (char 2–5)	0.9702	0.9760	0.9731	1000
LogReg (word 1–2)	0.9917	0.9510	0.9709	1000
accuracy			0.9595	2000
macro avg	0.9618	0.9595	0.9594	2000
weighted avg	0.9618	0.9595	0.9594	2000

Table 2: Task 1 (NAD) test set.

- **LogReg (word 1–2):** Logistic Regression using word-level n-grams (1–2) with TF–IDF features, providing simple but effective word co-occurrence representations.

Tables 1–4 report the full per-class metrics on the development and test splits.

On the news domain sub-task, both logistic-regression baselines outperform ARBERTv2, achieving 98.05 F_1 on dev (vs. 96.95) and 97.15 F_1 on test (vs. 95.95), indicating that simple n-gram features capture domain-specific style cues very effectively.

For the academic domain sub-task, ARBERTv2 reaches near-perfect performance (99.85 F_1 dev, 99.75 F_1 test), slightly outperforming the n-gram baselines. These results set a high bar for future participants: transformer fine-tuning excels on formal academic text, while lightweight n-gram classifiers remain surprisingly competitive, especially on news.

4.2 Participants Systems

Four teams submitted system description papers, and their approaches are summarized as follows.

Zain et al. team explored three different architectures: a fine-tuned RoBERTa-base model, a TF-IDF based system with a Linear SVM classifier, and an experimental system named Candace that

Model	P	R	F_1	Supp.
human				
ARBERTv2	0.9980	0.9990	0.9985	1000
LogReg (char 2–5)	0.9950	0.9990	0.9970	1000
LogReg (word 1–2)	0.9881	0.9970	0.9925	1000
machine				
ARBERTv2	0.9990	0.9980	0.9985	1000
LogReg (char 2–5)	0.9990	0.9950	0.9970	1000
LogReg (word 1–2)	0.9970	0.9880	0.9925	1000
accuracy			0.9985	2000
macro avg	0.9985	0.9985	0.9985	2000
weighted avg	0.9985	0.9985	0.9985	2000

Table 3: Task 2 (AWD) development set.

Model	P	R	F_1	Supp.
human				
ARBERTv2	1.0000	0.9950	0.9975	1000
LogReg (char 2–5)	0.9950	0.9980	0.9965	1000
LogReg (word 1–2)	0.9920	0.9960	0.9940	1000
machine				
ARBERTv2	0.9950	1.0000	0.9975	1000
LogReg (char 2–5)	0.9980	0.9950	0.9965	1000
LogReg (word 1–2)	0.9960	0.9920	0.9940	1000
accuracy			0.9975	2000
macro avg	0.9975	0.9975	0.9975	2000
weighted avg	0.9975	0.9975	0.9975	2000

Table 4: Task 2 (AWD) test set.

used probabilistic features from multiple Llama-3.2 models (Zain et al., 2025). Their final submission was based on the fine-tuned **RoBERTa-base** model, which yielded the highest performance on the development sets.

IntegrityAI team proposed a multimodal architecture that combines textual features from a pre-trained **ELECTRA** model with four handcrafted stylometric features: word count, average sentence length, vocabulary richness (TTR), and average word length (IntegrityAI, 2025). For the news subtask, they also employed a pseudo-labeling technique to augment their training data.

Hamada Nayel team focused on classical machine learning algorithms, submitting a system based on a **Linear Support Vector Machine (SVM)** classifier with **TF-IDF** features (Ashraf et al., 2025). Their approach emphasized efficiency and interpretability, demonstrating that traditional methods can achieve competitive performance without the need for resource-intensive deep learning models.

CNLP-NITS-PP team developed a hybrid model that fine-tuned a **DeBERTa-base** transformer and augmented it with nine auxiliary stylometric features, such as Unique Word Count, Stop Word Count, and Type-Token Ratio (Yadagiri et al., 2025). The contextual embedding from DeBERTa was concatenated with the feature vector before being passed to a final classification layer.

4.3 Results

The official results for both subtasks are presented in Tables 5 and 6. All participating teams achieved exceptionally high scores, indicating the high quality of the submitted systems.

Team	F1	Acc.	Prec.	Rec.
Zain et al.	1.000	1.000	1.000	1.000
IntegrityAI	0.996	0.996	0.996	0.996
Hamada Nayel	0.990	0.990	0.980	0.990
CNLP-NITS-PP	0.898	0.898	0.898	0.898

Table 5: Official results for Subtask 1 (NAD).

Team	F1	Acc.	Prec.	Rec.
Zain et al.	1.000	1.000	1.000	1.000
CNLP-NITS-PP	1.000	1.000	1.000	1.000
IntegrityAI	0.999	0.999	0.999	0.999

Table 6: Official results for Subtask 2 (AWD). The team Hamada Nayel focused their paper on Subtask 1.

In Subtask 1 (NAD), the top-performing systems were all based on transformer architectures. The winning system from **Zain et al.**, a fine-tuned RoBERTa model, achieved a perfect F1-score of 1.000. The **IntegrityAI** team, using ELECTRA with stylometric features, also achieved a near-perfect score of 0.996. Notably, the classical SVM-based system from **Hamada Nayel** secured the third rank with an F1-score of 0.990, outperforming one of the transformer-based systems and demonstrating the viability of simpler models.

In Subtask 2 (AWD), the performance was even higher across the board, with two teams, **Zain et al.** (RoBERTa) and **CNLP-NITS-PP** (DeBERTa + features), achieving perfect scores. The **IntegrityAI** system was just behind with an F1-score of 0.999. The near-perfect results from all teams suggest that detecting AI-generated text in the academic writing domain, at least with the data provided, was a less challenging task compared to the news domain. The structured and formal nature of academic abstracts may provide more distinct signals for classifiers to distinguish between human and machine-generated content. The general trend indicates that while fine-tuned transformers are dominant, augmenting them with stylometric features is a popular and effective strategy.

5 Conclusion

The M-DAIGT shared task aimed to advance the detection of AI-generated text in the critical domains of news and academic writing. The results demonstrate the remarkable effectiveness of current state-of-the-art models, with participating systems achieving near-perfect to perfect scores on both subtasks. The primary findings indicate that fine-tuned transformer models, such as RoBERTa, ELECTRA, and DeBERTa, are highly proficient at this task. Furthermore, the integration of stylometric features proved to be a valuable strategy for several teams, suggesting that a hybrid approach combining deep contextual understanding with traditional linguistic analysis is robust. The strong performance of a classical TF-IDF+SVM model in the news subtask also highlights that resource-efficient methods remain highly competitive. Overall, this shared task provides a valuable benchmark and dataset for the community, confirming the strength of existing methods while also pointing to the nuanced challenges posed by different domains.

Limitations

Despite the success of the shared task, several limitations should be acknowledged. First, the dataset, while diverse in its use of generator models and prompts, represents a static snapshot of LLM capabilities. The rapid evolution of generative models means that detectors trained on this data may not generalize well to text produced by future, more sophisticated LLMs. Second, the task was framed as a binary classification problem (human vs. AI), which does not capture the increasingly common scenario of human-AI collaborative writing, where text is partially generated and then edited by a human. Detecting such mixed-authorship content remains a significant open challenge. Third, the task did not explicitly evaluate the robustness of systems against adversarial attacks, such as paraphrasing or "humanization" techniques designed to evade detection. The exceptionally high scores, particularly in the academic subtask, might also indicate that the detection task within our dataset's parameters was not sufficiently challenging to fully differentiate the capabilities of the top systems. Finally, our study was confined to the English language, and the findings may not be directly applicable to other languages with different linguistic structures. Future iterations of this shared task could address these limitations by incorporating more recent LLMs, including co-authored text, introducing adversarial evaluation tracks, and expanding to multilingual contexts.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros, editors. 2025. *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*. International Conference on Computational Linguistics, Abu Dhabi, UAE.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in Urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428, Vienna, Austria. Association for Computational Linguistics.

Nsrin Ashraf, Mariam Labib, and Hamada Nayel. 2025. Inside the box: A streamlined model for AI-generated news article detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21.

Kyle Bittle and Omar El-Gayar. 2025. Generative ai and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4):296.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and

human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Frohmann, Gabriel Meseguer-Brocal, Markus Schedl, and Elena V. Epure. 2025. Double entendre: Robust audio-based AI-generated lyrics detection via multi-view fusion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1914–1926, Vienna, Austria. Association for Computational Linguistics.

Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. 2025. XDAC: XAI-driven detection and attribution of LLM-generated news comments in Korean. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22728–22750, Vienna, Austria. Association for Computational Linguistics.

IntegrityAI. 2025. A multimodal transformer-based approach for cross-domain detection of machine-generated text. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. 2025. M-RangeDetector: Enhancing generalization in machine-generated text detection through multi-range attention masks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8971–8983, Vienna, Austria. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Attila Kovari. 2025. Ethical use of chatgpt in education—best practices to combat ai-induced plagiarism. In *Frontiers in Education*, volume 9, page 1465703. Frontiers Media SA.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics.

Jiatao Li and Xiaojun Wan. 2025. Who writes what: Unveiling the impact of author roles on AI-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26620–26658, Vienna, Austria. Association for Computational Linguistics.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. 2025. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3091–3113, Vienna, Austria. Association for Computational Linguistics.

Zongjie Li, Chaozheng Wang, Shuai Wang, and Cuiyun Gao. 2023. Protecting intellectual property of large language model-based code generation apis via watermarks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2336–2350.

Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 727–752, Vienna, Austria. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Andrea Pedrotti, Michele Papucci, Cristiano Ciaccio, Alessio Miaschi, Giovanni Puccetti, Felice Dell'Orletta, and Andrea Esuli. 2025. Stress-testing

machine generated text detection: Shifting language models writing style to fool detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3010–3031, Vienna, Austria. Association for Computational Linguistics.

Zubair Qazi, William Shiao, and Evangelos E Papalexakis. 2024. Gpt-generated text detection: Benchmark dataset and tensor-based detection method. In *Companion Proceedings of the ACM Web Conference 2024*, pages 842–846.

Rafael Alberto Rivera Soto, Barry Y. Chen, and Nicholas Andrews. 2025. Mitigating paraphrase attacks on machine-text detection via paraphrase inversion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4421–4433, Vienna, Austria. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Lujia Shen, Xuhong Zhang, Shouling Ji, Yuwen Pu, Chunpeng Ge, Xing Yang, and Yanghe Feng. 2023. Textdefense: Adversarial text detection based on word importance entropy. *arXiv preprint arXiv:2302.05892*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22015–22036, Vienna, Austria. Association for Computational Linguistics.

Vasiliki Tassopoulou, George Retsinas, and Petros Maragos. 2021. Enhancing handwritten text recognition with n-gram sequence decomposition and multitask learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10555–10560. IEEE.

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290, Vienna, Austria. Association for Computational Linguistics.

Ashok Urlana, Aditya Saibewar, Bala Mallikarjunarao Garlapati, Charaka Vinayak Kumar, Ajeet Singh, and Srinivasa Rao Chalamala. 2024. TrustAI at SemEval-2024 task 8: A comprehensive analysis of multi-domain machine generated text detection techniques. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 927–934, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024a. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Annepaka Yadagiri, L. D. M. S. Sai Teja, Partha Pakray, and Chukhu Chunka. 2025. AI-generated text detection using DeBERTa with auxiliary stylometric features. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Ali Zain, Sareem Farooqui, and Muhammad Rafi. 2025. A multi-strategy approach for AI-generated text detection. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria. INCOMA Ltd.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR.

AI-Generated Text Detection Using DeBERTa with Auxiliary Stylometric Features

Annepaka Yadagiri, L. D. M. S. Sai Teja, Partha Pakray and Chukhu Chunka
Department of Computer Science & Engineering
National Institute of Technology Silchar, Assam, India - 788010
{annepaka22_rs, lekkalad_ug_22, partha, chukhu}@cse.nits.ac.in

Abstract

The global proliferation of Generative Artificial Intelligence (GenAI) has led to the increasing presence of AI-generated text across a wide spectrum of topics, ranging from everyday content to critical and specialized domains. Often, individuals are unaware that the text they interact with was produced by AI systems rather than human authors, leading to instances where AI-generated content is unintentionally combined with human-written material. In response to this growing concern, we propose a novel approach as part of the Multi-Domain AI-Generated Text Detection (M-DAIGT) shared task, which aims to accurately identify AI-generated content across multiple domains, particularly in news reporting and academic writing. Given the rapid evolution of large language models (LLMs), distinguishing between human-authored and AI-generated text has become increasingly challenging. To address this, our method employs fine-tuning strategies using transformer-based language models for binary text classification. We focus on two specific domains news and scholarly writing and demonstrate that our approach, based on the DeBERTa transformer model, achieves superior performance in identifying AI-generated text. Our team CNLP-NITS-PP achieved 5th position in Subtask 1 and 3rd position in Subtask 2.

1 Introduction

The rapid advancement and widespread adoption of Large Language Models (LLMs) have contributed to a significant increase in the generation of artificial content through Generative AI (GenAI). This technology is now integrated into various facets of everyday life. However, its pervasive use has raised important concerns, particularly regarding the authenticity of student work and the dissemination of misleading or fabricated information (Wang et al., 2023). As LLMs become more sophisticated, distinguishing between human-written and AI-generated text has become increasingly challenging for end users. In response to these issues, there is a pressing need for reliable detection methods. To address this, we introduce our work at the Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task, which aims to identify AI-generated content across multiple domains, with a focus on news articles and academic writing.

2 Related Work

This section discusses prior work about machine-generated text detection methods, datasets, and shared task.

2.1 Detection Methods

Approaches for detecting machine-generated texts (MGTs) can generally be categorized into two main types: training-free and training-based methods. Training-free techniques rely on the statistical properties of text to identify content produced by AI systems (Solaiman et al., 2019; Gehrmann et al., 2019). A range of features have been investigated in this context, including perplexity scores (Vasilatos et al., 2023), perplexity curvature (Mitchell et al., 2023), log-rank metrics (Su et al., 2023), intrinsic dimensionality (Tulchinskii et al., 2023), and N-gram frequency analysis (Yang et al., 2023). One such method, Revise-Detect, is based on the assumption that AI-generated text undergoes fewer edits when processed by LLMs compared to human-written text (Zhu et al., 2023). Another method, Binoculars, introduced by (Hans et al., 2024), utilizes two LLMs to compute the ratio of perplexity to cross-perplexity, effectively measuring how one model interprets the next-token predictions of another.

In contrast, training-based detection approaches typically involve fine-tuning pre-trained models to perform binary classification of text as either human- or machine-authored (Yu et al., 2023). These models may also employ advanced strate-

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 10–14, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-002

gies such as adversarial training (Hu et al., 2023) or abstention-based decision making (Tian et al., 2023). Additionally, (Verma et al., 2023) proposes fine-tuning a linear classifier atop the learned feature representations extracted from language models.

2.2 Task

The Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task (Lamsiyah et al., 2025) focuses on identifying AI-generated content across different domains, particularly news articles and academic writing. With the rapid advancement of LLMs, distinguishing human-written and AI-generated text has become a critical challenge. This task aims to contribute to research on information integrity and academic honesty.

Subtask 1 News Article Detection: Binary classification of news articles as human-written or AI-generated. Evaluation on both full articles and snippets. Covers various genres: politics, technology, sports, etc.

Subtask 2 Academic Writing Detection: Binary classification of academic texts as human-written or AI-generated. Evaluation of student coursework and research papers covers multiple academic disciplines and writing styles.

2.3 Dataset Statistics

This task provides two datasets presenting one for each subtask. **Human-written content:** Sourced from verified news websites and academic papers with proper permissions. **AI-generated content:** Created using multiple LLMs (GPT-3.5, GPT-4, Claude, etc.) with different prompting strategies and generation settings.

Split	Human	AI-generated
Train	5,000	5,000
Dev	1,000	1,000
Test	1500	1500

Table 1: Dataset split by Human and AI-generated labels for both the subtasks.

Both tasks, subtask 1 and subtask 2, there is a balanced distribution of human-written and AI-generated text.

3 Evaluation Metrics

In this study, we employed standard evaluation metrics to assess model performance, including Accuracy, Precision, Recall, F1-Score, and Matthews Correlation Coefficient (MCC). Additionally, we considered the fundamental classification components True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to provide a comprehensive analysis of the models predictive capabilities.

4 System Description

The system architecture for fine-tuning DeBERTa (Decoding-enhanced BERT with Disentangled Attention) with Linguistic Features for the Multi-Domain Detection of AI-Generated Text (M-DAIGT) task (Lamsiyah et al., 2025) consists of several key components that work together to process input text and classify it as human-written or AI-generated.

4.1 System Architecture

We propose a hybrid architecture that integrates both deep contextual language representations and handcrafted linguistic features to detect AI-generated text. The backbone of our model is the Decoding-enhanced BERT with Disentangled Attention DeBERTa-base transformer (He et al., 2020), which has shown strong performance on various natural language understanding tasks. To incorporate external linguistic cues, we extracted nine handcrafted features from the text, including metrics such as Unique Word Count, Stop Word Count, Type-Token Ratio, Hapax Legomenon Rate, and Burstiness.

The architecture consists of three main components:

- **Transformer Backbone:** We use the pre-trained `microsoft/deberta-base` model to encode the input text. Specifically, we extract the hidden state corresponding to the [CLS] token from the final layer to represent the sentence-level semantics.
- **Feature Encoder:** A linear layer is applied to the handcrafted features to project them into a 64-dimensional space, followed by a ReLU activation.
- **Fusion and Classification:** The contextual embedding corresponding to the [CLS] token from DeBERTa-base (a 768-dimensional vector) is concatenated with the handcrafted feature representation. The feature vector, originally 9-dimensional, is first passed through

Dataset	Model ↓ Metrics →	Accuracy	Precision	Recall	F1-Score	MCC
Subtask-1	*FastDetectGPT (Falcon)*	60.42	61.35	60.42	55.12	33.27
	FastDetectGPT (GPT-Neo)	58.10	59.00	58.10	52.85	31.75
	Binoculars	61.33	62.70	61.33	54.90	32.94
	DeBERTa	**89.75**	**89.78**	**89.75**	**89.75**	**79.53**
	ModernBERT	62.80	77.93	62.80	56.97	37.81
	RoBERTa	86.00	87.60	86.00	85.85	73.58
	DistilBERT	85.91	86.21	85.90	84.72	72.61
Subtask-2	*FastDetectGPT (Falcon)*	81.75	83.20	81.75	80.85	76.30
	FastDetectGPT (GPT-Neo)	75.90	77.60	75.90	74.30	70.85
	Binoculars	84.01	84.99	84.01	83.50	78.95
	DeBERTa	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
	ModernBERT	100.00	100.00	100.00	100.00	100.00
	RoBERTa	100.00	100.00	100.00	100.00	100.00
	DistilBERT	100.00	100.00	100.00	100.00	100.00

Table 2: Performance metrics of various models along with the zero-shot approaches on Subtask-1 and Subtask-2.

a fully connected layer that maps it to a 64-dimensional vector using a ReLU activation. This results in a combined vector of size 768+64=832. A dropout layer with a rate of 0.3 is applied to the concatenated vector to reduce overfitting. Finally, the output is fed into a fully connected classification layer that maps the 832-dimensional input to 2 output logits corresponding to the binary classification task (human-written vs. AI-generated).

This design enables the model to benefit from both the deep contextual understanding of language offered by transformers and the interpretable, statistically motivated handcrafted features.

4.2 Training Method

Models are trained on Amazon Web Services (*AWS*) Cloud server, Amazon Elastic Compute Cloud (*EC2*) instance. In the EC2 instance, we initiated an instance for Accelerated Computing. The specifications are **g6e.xlarge** instance, which provides **3rd generation AMD EPYC processors (*AMD EPYC 7R13*)**, with a **NVIDIA L40S Tensor Core GPU with 48 GB GPU memory**, and 4x vCPU with 32 GiB memory and a network bandwidth of 20GBps, and our OS type is **Ubuntu Server 24.04 LTS (*HVM*), EBS General Purpose (*SSD*) Volume Type.**

Models are trained on a CUDA-enabled GPU, and for all the models the hyperparameter settings are as follows: the batch-size is 32, the maximum sequence length is 512, AdamW optimizer with a learning rate of 1e-5 and weight decay of 0.01, Cross-entropy loss, `ReduceLROnPlateau` reduces the learning rate by a factor of 0.1 if validation loss plateaus for 1 epoch, up to 3 epochs with early stopping, with a loss as the main metric.

5 Results

For subtask 1 and Task-2, as shown in Table 2, the performance of various transformer-based models, evaluated using standard metrics: Accuracy, Precision, Recall, F1-Score, and Matthews Correlation Coefficient (MCC). For experimental purposes, we have used open-source zero-shot AI detectors like *FastDetectGPT* (Bao et al., 2023) and *Binoculras* (Hans et al., 2024) and four HuggingFace base models: DeBERTa (He et al., 2020), ModernBERT (Warner et al., 2024), RoBERTa (Liu, 2019), and DistilBERT (Sanh et al., 2019).

For Subtask-1, which involved distinguishing between AI-generated and human-written text, DeBERTa achieved the highest performance among all models, with a test accuracy of 89.75%, precision of 89.78%, recall of 89.75%, F1-score of 89.75%, and an MCC of 79.53% and the corresponding confusion matrix for the DeBERTa model can be seen in the Fig 1. This demonstrates DeBERTa's strong ability to generalize in binary classification tasks with nuanced language distinctions. RoBERTa and DistilBERT followed closely, achieving F1-scores of 85.85% and 84.72%, respectively, and MCC scores above 70%, indicating stable and reliable predictions.

In contrast, all models achieved perfect scores across all metrics in Subtask-2, indicating that this task was comparatively easier or more separable. The models reached 100% on accuracy, precision, recall, F1-score, and MCC. This suggests that the task structure, data distribution, or underlying linguistic features in Subtask-2 allowed the models to learn and generalize with very high confidence.

For both subtask datasets, after checking the classification with zero-shot methods, their performance is not above the mark, as we can see in Table 2. Here, the variations of FastDetectGPT are the scores models, and those scorer models are Falcon and GPT-Neo, and Binoculars is based on the perplexity values of the sentence.

These results highlight the robustness of DeBERTa in handling nuanced AI vs. human text classification and also underscore the importance of selecting appropriate architectures and feature representations based on task difficulty and data characteristics.

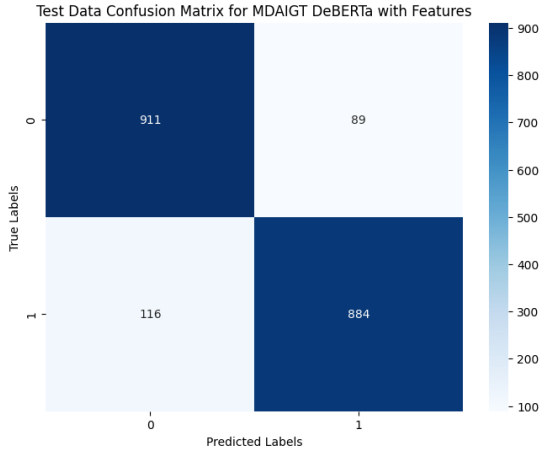

Figure 1: Confusion matrix of DeBERTa model with proposed approach on Subtask 1.

5.1 Error Analysis

The DeBERTa-base model demonstrated robust performance, achieving an accuracy of **89.75%**, precision of **89.78%**, recall of **89.75%**, F1-score of **89.75%**, and a MCC of **79.53** in the subtask-1. Despite these strong results, we identify the following key error patterns and limitations:

- **Contextual Ambiguities:**
  - Errors persist in cases involving **complex syntax** like nested negations, long-range dependencies or **figurative language** like sarcasm, where DeBERTa's disentangled attention may not fully resolve ambiguity.
- **Tokenization Challenges:**
  - Subword tokenization struggles with **rare terms** or **noisy inputs** (e.g., social media typos), leading to suboptimal representations for domain-specific jargon.
- **MCC Interpretation:**
  - The MCC score of **79.53** reflects strong classification, but its divergence from F1 suggests residual bias in edge cases, possibly due to class skew.

Mitigation Strategies: To address the limitations like the misclassification, ambiguity, etc, we recommend a few techniques that we expect to do as future work, that are: 1) Data Augmentation, 2) Fine-tuning on error cases to reduce systematic misclassifications.
This analysis highlights DeBERTa's strengths while pinpointing avenues for improvement, particularly in handling nuanced linguistic constructs.

6 Conclusion

In this paper, we presented our approach for the Multi-Domain Detection of AI-Generated Text (M-DAIGT) 2025 shared task, which focuses on identifying AI-generated content across diverse domains, including news articles and academic writing. We proposed a comparative evaluation of multiple transformer-based language models like DeBERTa, RoBERTa, DistilBERT, and ModernBERT on two subtasks aimed at detecting synthetic text. Our experiments demonstrated that DeBERTa and DistilBERT consistently achieved strong performance, with DeBERTa yielding the highest overall metrics by our team CNLP-NITS-PP with a value of 89.75% recall standing in the Top-5 among the participants on Subtask-1, and all models attaining perfect scores and standing on Top-3 on Subtask-2 and also outperforming all the zero-shot training free methods with a significant differences of evaluation metrics.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. *arXiv preprint arXiv:2305.18149*.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36:39257–39276.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Nenghai Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. *CoRR*.

Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. 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*, pages 7470–7483.

A Multi-Strategy Approach for AI-Generated Text Detection

Ali Zain
vin.alizain@gmail.com

Sareem Farooqui
sareemfarooqui10@gmail.com

Muhammad Rafi
muhammad.rafi@nu.edu.pk

National University of Computer and Emerging Sciences, FAST
Karachi, Pakistan

Abstract

This paper presents presents three distinct systems developed for the M-DAIGT shared task on detecting AI generated content in news articles and academic abstracts. The systems includes: (1) A fine-tuned RoBERTa-base classifier, (2) A classical TF-IDF + Support Vector Machine (SVM) classifier , and (3) An Innovative ensemble model named Candace, leveraging probabilistic features extracted from multiple Llama-3.2 models processed by a custom Transformer encoder.

The RoBERTa-based system emerged as the most performant, achieving near-perfect results on both development and test sets.

1 Introduction

The proliferation of sophisticated large language models (LLMs) has led to a surge in AI-generated text, making its detection a critical area of research (Jawahar et al., 2020). Identifying machine-generated content is crucial for maintaining information integrity, combating misinformation (Pan et al., 2023), and ensuring academic honesty. The M-DAIGT (Multi-domain DAIGT) shared task (Lamsiyah et al., 2025) aims to foster research in this domain by providing datasets for two distinct scenarios: news articles (Subtask 1) and academic abstracts (Subtask 2). Participants are tasked with building systems to classify given texts as either human-written or machine-generated.

In response to this challenge, our team developed and evaluated three different systems:

1. **RoBERTa-based Classifier:** A fine-tuned RoBERTa-base model, a widely successful approach for text classification tasks.

2. **TF-IDF + SVM Classifier:** A traditional machine learning pipeline combining Term Frequency-Inverse Document Frequency (TF-IDF) features with a Linear Support Vector Machine (SVM) (Joachims, 1998). This served as a strong baseline, particularly for Subtask 1.

3. **Llama-Feature Ensemble with Transformer Classifier (Candace):** An experimental system designed to capture nuanced signals from multiple LLMs. It extracts probabilistic features (Sarvazyan et al., 2024) (e.g., token log-probabilities, entropy) from a suite of Llama-3.2 models (Meta AI, 2024) and uses a custom Transformer Encoder-based model for final classification.

This paper details the architecture, data handling, implementation, and experimental results of these systems on the provided test datasets. Our RoBERTa-based approach yielded the most consistent and high-performing results on the development and test sets and was selected for our final submissions for both subtasks.

2 System Architectures

We developed three distinct systems, each employing a different strategy for AI-generated text detection.

2.1 System 1: RoBERTa-based Classifier

This system (Figure 1) fine-tunes a pre-trained RoBERTa-base model (Liu et al., 2019). The input text is tokenized, and the RoBERTa model processes these tokens. The final hidden state corresponding to the special ‘[CLS]‘ token is then passed through a linear classification layer to produce a binary prediction (human or machine).

2.2 System 2: TF-IDF + SVM Classifier

Our second system (Figure 2) follows a traditional machine learning pipeline. Textual input is first

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 15–19, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-003

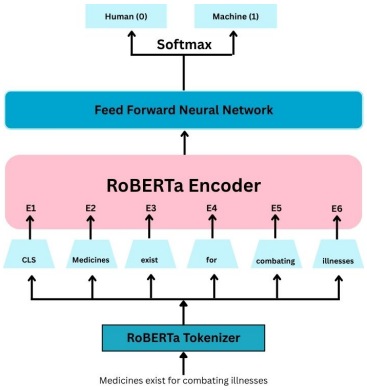

Figure 1: Architecture of System 1: RoBERTa-based Classifier.

converted into a numerical representation using TF-IDF vectorization, capturing n-grams. These TF-IDF features are then fed into a Linear Support Vector Machine (SVM) for classification.

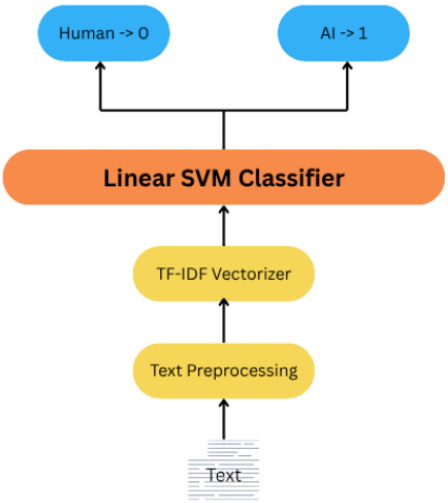

Figure 2: Architecture of System 2: TF-IDF + SVM Classifier.

2.3 System 3: Llama-Feature Ensemble with Transformer Classifier (Candace)

The third system (Figure 3), named Candace, is more experimental. It involves a two-stage process. First, probabilistic features (alpha, beta, gamma, as described in Section 4.3) are extracted from each token of the input text using multiple Llama-3.2 models. These feature vectors are concatenated. Second, this sequence of combined Llama-derived features is processed by a custom Transformer Encoder-based classification head, which then makes the final human/machine prediction.

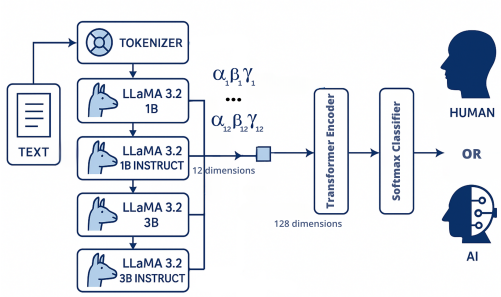

Figure 3: Architecture of System 3: Candace - Llama-Feature Ensemble with Transformer Classifier.

3 Data and Resources

The M-DAIGT shared task provided datasets for two subtasks:

- **Subtask 1 (News):** Comprised of ‘T1_train.csv‘ (10,000 samples), ‘T1_dev.csv‘ (2,000 samples), and ‘T1_test_unlabeled.csv‘ (2,000 samples).
- **Subtask 2 (Academic Abstracts):** Comprised of ‘T2_train.csv‘ (10,000 samples), ‘T2_dev.csv‘ (2,000 samples), and ‘T2_test_unlabeled.csv‘ (2,000 samples).

Each labeled dataset contained an ‘id‘, ‘text‘, and a ‘label‘ column, where labels were either ‘human‘ or ‘machine‘. For training, labels were mapped to 0 (human) and 1 (machine). Minimal preprocessing was applied for the RoBERTa and TF-IDF systems, primarily consisting of standard tokenization handled by the respective libraries. The Candace system’s feature extraction used raw text. External resources included:

- Pre-trained ‘roberta-base‘ model and tokenizer from Hugging Face Transformers (Wolf et al., 2020).
- Pre-trained Llama-3.2 models (Meta AI, 2024) (‘meta-llama/Llama-3.2-1B‘, ‘meta-llama/Llama-3.2-1B-Instruct‘, ‘meta-llama/Llama-3.2-3B‘, ‘meta-llama/Llama-3.2-3B-Instruct‘) and the ‘meta-llama/Llama-3.2-1B‘ tokenizer.

4 Methodology

4.1 System 1: RoBERTa-based Classifier

Model Architecture: We used the ‘Roberta-Model‘ from Hugging Face Transformers, pre-

trained on ‘roberta-base‘. A linear classification layer was added on top of the pooled output (representation of the ‘[CLS]‘ token) from the RoBERTa model. The output layer predicts a score for the two classes (human vs. machine).

Input Representation: Texts were tokenized using ‘RobertaTokenizerFast‘ with a maximum sequence length of 512 tokens. Padding was applied to shorter sequences, and longer sequences were truncated.

Training: The model was fine-tuned for 4 epochs using the Adam optimizer with a learning rate of 1×10^{-5}. We used ‘CrossEntropyLoss‘ as the loss function. The batch size was set to 16. This setup was applied independently for both Subtask 1 and Subtask 2, using their respective training and development datasets.

4.2 System 2: TF-IDF + SVM Classifier

This system was developed primarily as a baseline for Subtask 1 (News).

Feature Extraction: We used `TfidfVectorizer` from scikit-learn to convert text into numerical features. We configured it to use n-grams of range (2, 3) and limited the maximum number of features to 5,000.

Classifier: A Linear Support Vector Machine (LinearSVC) was employed for classification. The hyperparameters were set as follows: C (regularization parameter) = 0.5, ‘class_weight=’balanced’‘ to handle potential class imbalance, ‘dual=False‘ (as n_samples were greater than n_features), and ‘max_iter=5000‘ to ensure convergence.

4.3 System 3: Llama-Feature Ensemble with Transformer Classifier (Candace)

This experimental system explores the utility of probabilistic features derived from multiple instruction-tuned and base Llama-3.2 models.

Feature Extraction: For each input text and for each of the four Llama models (‘meta-llama/Llama-3.2-1B‘, ‘meta-llama/Llama-3.2-1B-Instruct‘, ‘meta-llama/Llama-3.2-3B‘, ‘meta-llama/Llama-3.2-3B-Instruct‘), we extracted three features per token up to a maximum sequence length of 256:

- **Alpha (α):** The maximum log-probability assigned by the Llama model to any token at that position, given the preceding tokens.
- **Beta (β):** The entropy of the Llama model's predicted probability distribution over the vocabulary at that position.
- **Gamma (γ):** The log-probability assigned by the Llama model to the actual observed token at that position.

The Llama models were loaded with 8-bit quantization to manage memory. Features from all four Llama models were concatenated token-wise, resulting in 4 models $\times$ 3 features from each model $=$ 12 features per token.

Classifier Architecture (CandaceClassifier): The sequence of aggregated indicators was then processed by a custom classification architecture. This architecture begins with a projection of the indicator sequence into a higher-dimensional space. This transformed sequence is then passed through a Transformer Encoder block, designed to capture contextual relationships between the token-level indicators. The output of the Transformer Encoder is subsequently pooled across the sequence dimension, and a final linear layer produces the binary classification.

Training: The CandaceClassifier was trained for 10 epochs using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-4} and ‘CrossEntropyLoss‘. The batch size was 8. This architecture was trained separately for Subtask 1 and Subtask 2.

5 Experiments and Results

All systems were trained and evaluated on the M-DAIGT development sets for their respective subtasks. The primary evaluation metrics were Accuracy and F1-score.

RoBERTa-based System (System 1): For Subtask 1 (News), our fine-tuned RoBERTa model achieved an accuracy of 99.95% and an F1-score of 99.95% on the development set (best at epoch 4). For Subtask 2 (Academic Abstracts), the RoBERTa model achieved 100.00% accuracy and 100.00% F1-score on the development set (stable from epoch 1 onwards). Given its strong and consistent performance, this system was chosen for our official submissions for both subtasks.

TF-IDF + SVM System (System 2): This system was evaluated on Subtask 1 and Subtask 2. On

System	Subtask 1 (News) - Test Set				Subtask 2 (Academic) - Test Set			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
RoBERTa-base (System 1)	**99.99%**	**99.99%**	**99.80%**	**100.0%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
TF-IDF + SVM (System 2)	97.90%	97.91%	97.52%	98.30%	99.85%	99.85%	100.0%	99.70%
Candace (System 3)	99.75%	99.75%	99.60%	99.90%	99.95%†	99.95%	100.00%	99.90%

Table 1: Test set performance, RoBERTa base model with Fast tokenizer outperforming all models

its internal training data (as dev metrics were not explicitly separated in its notebook), it achieved an accuracy of 99.81% and an F1-score of 0.9981. While competitive, it was slightly outperformed by the RoBERTa model on the development set.

Candace System (System 3): For Subtask 1 (News), the Candace system achieved a development accuracy of 99.80% (best at epoch 6). The same architectural design and training procedure were applied to Subtask 2, and similar development accuracy (99.80%) was observed during its separate training run. While promising, this system is more computationally intensive due to the multi-LLM feature extraction step. The RoBERTa system offered slightly better or comparable performance with significantly less overhead for these specific datasets.

6 Discussion

Our experiments highlight the continued effectiveness of fine-tuned transformer models like RoBERTa for text classification tasks, achieving near-perfect scores on the development sets for both news and academic abstract domains. The RoBERTa model's ability to capture subtle linguistic cues makes it highly suitable for distinguishing between human and AI-generated text.

The TF-IDF + SVM approach, while simpler, provided a very strong baseline for Subtask 1, underscoring the utility of traditional methods, especially when coupled with robust feature engineering like n-grams.

The Candace system, which extracts features from multiple Llama-3.2 models, also showed excellent performance. This approach is interesting as it attempts to distill knowledge from several powerful LLMs into a smaller, specialized classifier. However, the feature extraction process is computationally expensive. For the M-DAIGT datasets, the gains over a well-tuned RoBERTa model were not substantial enough to justify the additional complexity and computational cost as the primary submission.

Runtime for RoBERTa inference is efficient, while Candace inference is slower due to the initial pass through multiple Llama models.

7 Conclusion

We presented three distinct systems for detecting AI-generated text in news articles and academic abstracts. Our fine-tuned RoBERTa-base model demonstrated exceptional performance on the development and test sets for both subtasks, achieving near-perfect accuracy and F1-scores, and was selected as our primary submission. The TF-IDF+SVM system served as a strong baseline, and the experimental Candace system, leveraging features from multiple Llama models, also showed high efficacy. Future work could involve ensembling these diverse models, exploring more sophisticated feature fusion techniques for the Candace system, and investigating the robustness of these models against adversarial attacks or text generated by newer, more advanced language models.

Acknowledgments

We thank the organizers of the M-DAIGT shared task for providing the dataset and the evaluation platform. BEsides that, the research is also supported by the provisional award under the National Research Program for Universities (NRPU), Higher Education Commission (HEC) Pakistan, with the title "NRPU: Automatic Multi-Model Classification of Religious Hate Content from Social Media" (Reference Research Project No. 16153).

References

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant

features. In *European conference on machine learning*, pages 137–142. Springer.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. ArXiv:1711.05101.

Meta AI. 2024. Llama 3.2 Model Family. https://www.llama.com/models/llama-3/.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Arsen M. Sarvazyan, Jorge Á. González, and Marc Franco-Salvador. 2024. Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Multimodal Transformer-based Approach for Cross-Domain Detection of Machine-Generated Text

Mohammad AL-Smadi
Qatar University
Doha, Qatar
malsmadi@qu.edu.qa

Abstract

The rapid advancement of large language models (LLMs) has made it increasingly challenging to distinguish between human-written and machine-generated content. This paper presents IntegrityAI, a multimodal ELECTRA-based model for the detection of AI-generated text across multiple domains. Our approach combines textual features processed through a pre-trained ELECTRA model with handcrafted stylometric features to create a robust classifier. We evaluate our system on the Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task, which focuses on identifying AI-generated content in news articles and academic writing. IntegrityAI achieves exceptional performance and ranked 1st in both subtasks, with F1-scores of 99.6% and 99.9% on the news article detection and academic writing detection subtasks respectively. Our results demonstrate the effectiveness of combining transformer-based models with stylometric analysis for detecting AI-generated content across diverse domains and writing styles.

1 Introduction

The rapid increase in the development of large language models (LLMs) has revolutionized natural language processing and generation capabilities. Models such as GPT-4, Claude, and others can now produce text that is increasingly difficult to distinguish from human-written content (Fagni et al., 2021; Wee and Reimer, 2023; Liang et al., 2023). While these advancements offer numerous benefits across various domains, they also present significant challenges related to information integrity, academic honesty, and the potential for misuse in spreading misinformation (Al-Smadi, 2023).

The ability to reliably detect AI-generated content has become a critical research area with implications for journalism, academia, and online information ecosystems (Weber-Wulff et al., 2023). The "Multi-Domain Detection of AI-Generated Text (M-DAIGT)" shared task was established to support addressing this challenge by evaluating systems designed to identify machine-generated content across different domains, specifically news articles and academic writing.

In this paper, we evaluate our model IntegrityAI (ALSmadi, 2025) on the M-DAIGT shared task (Lamsiyah et al., 2025). Our approach leverages a domain-agnostic architecture that combines the contextual understanding capabilities of the ELECTRA transformer model (Clark et al., 2020) with handcrafted stylometric features that capture linguistic patterns often present in machine-generated text. This hybrid approach allows our system to identify subtle differences between human and AI-written content across diverse domains and writing styles.

The main contributions of this paper are:

- A multimodal architecture that effectively combines transformer-based text representations with stylometric features for AI-generated text detection
- Empirical evaluation demonstrating the effectiveness of our approach across multiple domains, including news articles and academic writing
- Analysis of the impact of pseudo-labeling techniques on detection performance

2 Related Work

The detection of AI-generated text has become an increasingly important research area as language models continue to advance in their generation capabilities. Several approaches have been proposed in recent literature.

Zellers et al. (2019) presented GROVER, a model that can both generate and detect neural fake

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 20–25, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-004

news. Their work demonstrated that models trained to generate text can also be effective at detecting text generated by similar architectures.

Uchendu et al. (2020) explored authorship attribution techniques for detecting machine-generated text. They found that stylometric features combined with deep learning approaches could effectively identify different "authors," including various language models.

Ippolito et al. (2020) investigated methods for automatic detection of machine-generated text. They found that hybrid approaches combining multiple detection signals outperformed single-method approaches.

Jawahar et al. (2020) presented one of the early approaches to detecting text generated by neural language models. Their work demonstrated that statistical features of text could be used to distinguish between human and machine-generated content, though with limitations as generation models improved.

Fagni et al. (2021) presented a benchmark dataset and detection methods for machine-generated tweets. Their work highlighted the challenges of detecting short-form AI-generated content on social media platforms.

More recently, Mitchell et al. (2023) introduced DetectGPT, a zero-shot approach that leverages the curvature of the model's log probability function to identify text generated by that same model. Their work showed promising results without requiring extensive training data specific to each generation model.

Guo et al. (2023) conducted a comprehensive analysis of detection methods for large language model. They found that while supervised methods can achieve high accuracy on in-domain data, their performance degrades significantly when tested on outputs from unseen models or domains, highlighting the challenge of generalization.

In the academic domain, Markov et al. (2023) proposed a holistic approach combining linguistic features with neural representations to detect AI-generated academic writing. Their work, published in the Journal of Artificial Intelligence Research, demonstrated the importance of domain-specific features for academic text.

Focusing on the capabilities of stylometric features in boosting models abilities in detecting machine-generated content, the work of (Kutbi et al., 2024; Opara, 2024; ALSmadi, 2025) developed machine learning models with stylometry for machine-genereated content detection.

Our work builds upon these foundations while addressing the specific challenges of cross-domain detection. Unlike many previous approaches that focus on a single domain or generation model, IntegrityAI is designed to detect AI-generated content across multiple domains and from various generation models, making it more applicable to real-world scenarios.

3 Research Methodology

3.1 Task Description

The Multi-Domain Detection of AI-Generated Text (M-DAIGT) shared task focuses on identifying AI-generated content across different domains (Lamsiyah et al., 2025). The task is divided into two subtasks:

1. **News Article Detection (NAD)**: Binary classification of news articles as human-written or AI-generated, with evaluation on both full articles and snippets covering various genres including politics, technology, and sports.

2. **Academic Writing Detection (AWD)**: Binary classification of academic texts as human-written or AI-generated, with evaluation on student coursework and research papers across multiple academic disciplines and writing styles.

The primary evaluation metric for both subtasks is the F1-score, which balances precision and recall.

3.2 Dataset

The M-DAIGT dataset consists of both human-written and AI-generated texts across the two domains. Human-written content was sourced from verified news websites and academic papers with proper permissions. AI-generated content was created using multiple LLMs, including GPT-3.5, GPT-4, and Claude, with different prompting strategies and generation settings to ensure diversity.

The dataset is divided into training (10,000 samples per subtask), development (2,000 samples per subtask), and test (3,000 samples per subtask) splits, with a balanced distribution of human-written and AI-generated text in each split.

Feature	Description
Word Count	Total count of alphabetic tokens in the text
Average Sentence Length	Mean number of words per sentence
Vocabulary Richness	Measured using Type-Token Ratio (TTR)
Average Word Length	Mean number of characters per word

Table 1: Stylometric features used in IntegrityAI

Subtask	Without Pseudo-labeling	With Pseudo-labeling
News Article Detection (NAD)	0.993	0.996
Academic Writing Detection (AWD)	0.999	-

Table 2: F1-scores achieved by IntegrityAI on the M-DAIGT test set

3.3 Model Architecture

IntegrityAI employs a multimodal architecture that combines textual features processed through a pre-trained transformer model with handcrafted stylometric features. The upcoming sections explain the model architecture in more detail.

3.3.1 Features

Our model utilizes two types of features:

Text Embeddings: Raw text sequences are tokenized using Google's ELECTRA tokenizer [1]. The tokenized inputs include input IDs and attention masks, which are then processed by the ELECTRA model (Clark et al., 2020).

Stylometric Features: We extract four numerical features from each text using NLP techniques[2] , as detailed in Table 1. We standardize these features to ensure comparability and faster convergence during training [3].

3.3.2 Model Components

IntegrityAI is a multimodal deep learning model with the following components:

Textual Encoding: We use the pre-trained ELECTRA model from HuggingFace Transformers. The output of ELECTRA's encoder (last_hidden_state[:, 0, :]) is processed through a dropout layer, followed by a linear layer that reduces the dimensionality from the ELECTRA hidden size to 192, and finally a Rectified Linear Unit (ReLU) activation Function (Glorot et al., 2011) to help the model better learn complex patters in the data.

Numerical Feature Processing: The four stylometric features are processed through a linear layer that expands their dimensionality from 4 to 64, followed by batch normalization (Ioffe and Szegedy, 2015), ReLU activation, and Dropout layer (Srivastava et al., 2014).

Fusion Layer: The outputs from the textual encoding and numerical feature processing components are concatenated to form a 256-dimensional feature vector. This combined representation is then passed through a fully connected layer that maps to the number of output classes (2 for binary classification).

Final Output: The classification logits are passed to CrossEntropyLoss for supervised classification during training.

3.4 Training Setup

Training Pipeline: We use CrossEntropyLoss as our loss function and AdamW as our optimizer with a learning rate of 2e-5 and weight decay of 0.01. The model is trained for up to 5 epochs with early stopping after 2 epochs of no validation improvement.

Model Checkpointing: The best model (with the lowest validation loss) is saved and restored at the end of training.

Evaluation Metrics: We evaluate our model using accuracy and weighted F1-score, with the latter being the primary metric for the shared task.

Hardware: Training was performed on GPU (CUDA) machine of NVIDIA A10 with 22G RAM.

[1] We used (google/electra-base-discriminator) from huggingface https://huggingface.co/google/electra-base-discriminator

[2] We used NLTK Library for this purpose https://www.nltk.org/

[3] Features are standardized using StandardScaler from SciKit Learn https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

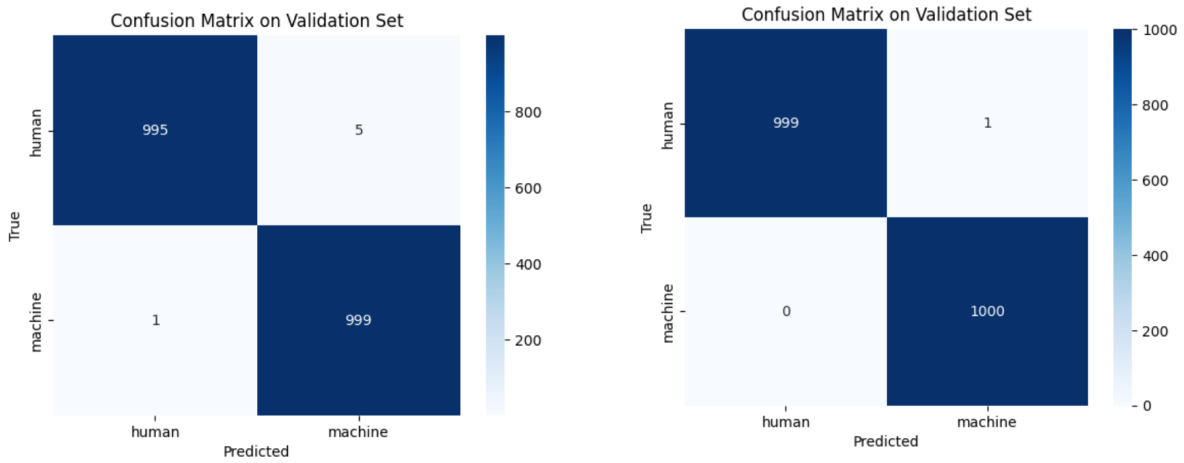

Figure 1: Confusion matrices on the validation sets (News subtask on the left).

We used seed initialization for reproducibility.

Pseudo-Labeling: For the subtask: News Article Detection (NAD) only, we employ a pseudo-labeling technique where we use our trained model to generate predictions on unlabeled data, and then incorporate high-confidence predictions into our training set for a second round of training.

4 Results

IntegrityAI achieved exceptional performance securing the 1st rank in both subtasks of the M-DAIGT shared task. Table 2 presents the F1-scores for our model on the test set, both with and without the pseudo-labeling technique.

Our model performed exceptionally well on both subtasks, with F1-scores of 0.993 and 0.999 for news article detection and academic writing detection, respectively, even without pseudo-labeling. The application of pseudo-labeling further improved our performance on the news article detection subtask, increasing the F1-score to 0.996.

The near-perfect performance on the academic writing detection subtask suggests that our model is particularly effective at identifying patterns that distinguish between human and AI-generated academic writing. This may be due to the more structured and formal nature of academic writing, which could make AI-generated content more distinguishable from human-written text in this domain.

The slightly lower (though still exceptional) performance on the news article detection subtask may reflect the greater diversity and variability in news writing styles, which could make the distinction between human and AI-generated content more challenging in this domain. Figure 1 depicts the confusion matrix for the two classification subtasks with relatively higher challenge in classifying human-written news articles.

The improvement in performance with pseudo-labeling on the news article detection subtask indicates that our model benefits from additional training data in this more diverse domain.

5 Discussion

The exceptional performance of IntegrityAI on both subtasks of the M-DAIGT shared task demonstrates the effectiveness of our multimodal approach to detecting AI-generated text across different domains. Several key factors contribute to this success:

Multimodal Architecture: The combination of transformer-based textual representations with handcrafted stylometric features allows our model to capture both contextual semantic information and statistical linguistic patterns. This multimodal approach provides a more comprehensive view of the text than either approach alone would offer.

ELECTRA's Discriminative Pre-training: Unlike many other transformer models that are pre-trained using generative objectives, ELECTRA is pre-trained using a discriminative approach, where it learns to distinguish between original and replaced tokens. This pre-training objective aligns well with the task of distinguishing between human and AI-generated text, potentially giving ELECTRA an advantage for this specific application.

Stylometric Features: The inclusion of stylometric features captures statistical patterns in text that may not be fully represented in the contextual embeddings. Features such as vocabulary richness

and sentence length distribution have long been used in authorship attribution and can help identify subtle differences between human and AI writing styles.

Pseudo-labeling: The improvement in performance with pseudo-labeling on the news article detection subtask highlights the value of semi-supervised learning approaches for this task. By leveraging unlabeled data, we can expand our training set and improve model robustness, particularly in more diverse and challenging domains.

Domain Differences: The near-perfect performance on academic writing detection compared to the slightly lower (though still exceptional) performance on news article detection suggests that there may be more distinctive patterns that separate human and AI-generated content in academic writing. This could be due to: the input text length, the more structured and formal nature of academic writing, or it could reflect differences in how the AI models were prompted when generating content for the dataset.

6 Conclusion and Future Work

In this paper, we presented IntegrityAI, a multimodal ELECTRA-based approach for detecting AI-generated text across multiple domains. Our system combines transformer-based textual representations with handcrafted stylometric features to create a robust classifier that achieves exceptional performance on the M-DAIGT shared task.

The results demonstrate the effectiveness of our approach, with F1-scores of 0.996 and 0.999 on the news article detection and academic writing detection subtasks, respectively. These results highlight the potential of multimodal approaches that leverage both deep learning and traditional stylometric analysis for detecting AI-generated content.

As language models continue to advance, the ability to reliably detect AI-generated content will become increasingly important for maintaining information integrity and academic honesty. Our work contributes to this goal by providing a robust and effective approach to cross-domain detection of AI-generated text.

Future work will focus on improving the generalization of our approach to new language models and domains, enhancing adversarial robustness, and addressing the ethical considerations associated with AI text detection technologies misapplication(Wee and Reimer, 2023; Liang et al., 2023; Weber-Wulff et al., 2023). We believe that continued research in this area is essential for ensuring that the benefits of advanced language models can be realized while mitigating potential risks and harms.

References

Mohammad Al-Smadi. 2023. Chatgpt and beyond: The generative ai revolution in education. *arXiv preprint arXiv:2311.15198*.

Mohammad ALSmadi. 2025. Integrityai at genai detection task 2: Detecting machine-generated academic essays in english and arabic using electra and stylometry. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 284–289.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Daphne Ippolito, Daniel Duckworth, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822. Association for Computational Linguistics.

Ganesh Jawahar, Muhammad Abdul-Mageed, and VS Laks Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309. Association for Computational Linguistics.

Mohammed Kutbi, Ali H. Al-Hoorie, and Abbas H. Al-Shammari. 2024. Detecting contract cheating through linguistic fingerprint. *Humanities and Social Sciences Communications*, 11:1–9.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7):100779.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pages 15009–15018.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 105–114, Cham. Springer Nature Switzerland.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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 (EMNLP)*, pages 8384–8395.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):26.

Hin Boo Wee and James D Reimer. 2023. Non-english academics face inequality via ai-generated essays and countermeasure tools. *BioScience*, 73(7):476–478.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Inside the Box: A Streamlined Model for AI-Generated News Article Detection

Nsrin Ashraf[1,2], Mariam Labib[2,3], Hamada Nayel[1,4]
[1]Department of Computer Science, Faculty of Artificial Intelligence, Benha University, Egypt
[2]Computer Engineering, Elsewedy University of Technology, Cairo, Egypt
[3]Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Egypt
[4]Department of Computer Engineering and Information, College of Engineering, Prince Sattam Bin Abdulaziz University, Wadi Addawasir, Saudi Arabia
Correspondence: hamada.ali@fci.bu.edu.eg

Abstract

The rapid proliferation of AI-generated text has raised concerns. With the increasing prevalence of AI-generated content, concerns have grown regarding authenticity, authorship, and the spread of misinformation. Detecting such content accurately and efficiently has become a pressing challenge. In this study, we propose a simple yet effective system for classifying AI-generated versus human-written text. Rather than relying on complex or resource-intensive deep learning architectures, our approach leverages classical machine learning algorithms combined with the TF-IDF text representation technique. Evaluated on the M-DAIGT shared task dataset, our Support Vector Machine (SVM) based system achieved strong results, ranking second on the official leaderboard and demonstrating competitive performance across all evaluation metrics. These findings highlight the potential of traditional lightweight models to address modern challenges in text authenticity detection, particularly in low-resource or real-time applications where interpretability and efficiency are essential.

1 Introduction

The emergence of advanced language models such as GPT, BERT, and other generative AI systems has revolutionized the way text is produced, enabling machines to generate coherent, context-aware, and human-like language (Cingillioglu, 2023). While these technologies offer immense benefits across industries from customer service automation to educational tools, they also pose significant challenges. One of the most pressing issues is the detection of AI-generated text, a task that has grown in importance due to its implications for academic integrity, information authenticity, cybersecurity, and digital content moderation.

The ability to distinguish between human-written and machine-generated content is essential in various contexts. For example, educational institutions need tools to verify the originality of student submissions. Social media platforms and news outlets must identify and limit the spread of synthetic misinformation. Similarly, cybersecurity frameworks may leverage such detection to prevent automated phishing or spam campaigns crafted by generative models.

Numerous methodologies have been explored for this task. Deep learning-based approaches, such as fine-tuning transformers or using binary classifiers trained on large datasets, have shown high accuracy. However, these methods are often resource-intensive, require large labeled datasets, and are not always interpretable. In contrast, traditional machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression (LR), k-Nearest Neighbors (KNN), and Stochastic Gradient Descent (SGD) provide a lightweight and interpretable alternative (Nayel and Amer, 2021). When paired with effective text representation techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) or n-gram analysis, these models can yield strong performance while remaining computationally efficient (Shehab et al., 2024; Fetouh and Nayel, 2023; Nayel, 2020).

This paper investigates the use of simple yet efficient machine learning models for the detection of AI-generated text. We evaluate their performance on benchmark datasets and analyze their potential for real-world deployment in low-resource environments. Our findings suggest that classical models, despite their simplicity, can offer competitive accuracy and practical advantages over more complex deep learning systems.

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 26–30, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-005

2 Related Work

The detection of AI-generated text has become a critical task in natural language processing (NLP), driven by the proliferation of large-scale generative language models such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and ChatGPT. These models are capable of producing text that is often indistinguishable from human writing, raising concerns about misinformation, plagiarism, and the integrity of online discourse.

Early approaches to detecting machine-generated text focused on statistical irregularities and text perplexity. Ippolito et al. (2020) evaluated the effectiveness of humans and models in distinguishing human-written from machine-generated outputs, showing that even humans struggle with high-quality generations. Similarly, Solaiman et al. (2019) used output probability distributions and likelihood scores to develop classifiers that identify synthetic text based on model uncertainty and overconfidence.

More recent work has turned to supervised machine learning, where classifiers are trained on labeled datasets of human vs. machine-generated text. Notably, Jawahar et al. (2019) explored a fine-tuned BERT-based classifier, showing strong performance on multiple generation sources.

To address issues of generalization and efficiency, other studies have investigated traditional machine learning algorithms. Zhang et al. (2011) applied logistic regression and SVMs on TF-IDF and n-gram features, demonstrating that simpler models can achieve competitive performance, particularly when interpretability and low latency are prioritized.

Despite these advances, there remains a trade-off between accuracy, interpretability, and computational cost. This work builds on the latter line of research by systematically comparing several classical models—SVM, Logistic Regression, KNN, and SGD—for the task of AI-generated text detection. In doing so, we aim to evaluate whether these models, when paired with strong feature engineering, can offer a practical and scalable solution for real-world applications.

3 Methodology

Our proposed architecture presents a machine learning pipeline designed to classify textual content as either AI-generated or human-written using the M-DAIGT dataset as shown in Figure 1. The dataset comprises labeled samples from verified human sources and outputs from various large language models (LLMs) prompted with diverse instructions. The data is partitioned into three subsets: training, development (dev), and testing.

Both the training and dev sets undergo a comprehensive feature engineering process to extract informative attributes that characterize the text. These features are subsequently utilized in model training via `GridSearchCV`, which facilitates exhaustive hyperparameter tuning and selection of the best-performing model configuration. The test set, kept unseen during training and tuning, is independently subjected to the same feature engineering steps and used to evaluate the final model's performance. This structured pipeline ensures the development of a robust and generalizable classifier through careful preparation, tuning, and evaluation.

As outlined earlier, this study performs a comparative analysis of three traditional machine learning algorithms for binary text classification: Logistic Regression (LR), SVM, and Decision Trees (DT). The methodology encompasses the full pipeline—data preprocessing, feature extraction, model training, hyperparameter optimization, and evaluation—with the objective of maximizing classification accuracy, particularly in terms of the f1-score.

`GridSearchCV` plays a critical role in this process by systematically exploring multiple hyperparameter combinations for each model. This enables:

- Automated and exhaustive search for the optimal hyperparameters
- Enhanced model generalization via cross-validation
- Fair and consistent comparison across different classification algorithms

Through this methodology, we aim to identify the most effective model and configuration to detect AI-generated content with high reliability and generalizability.

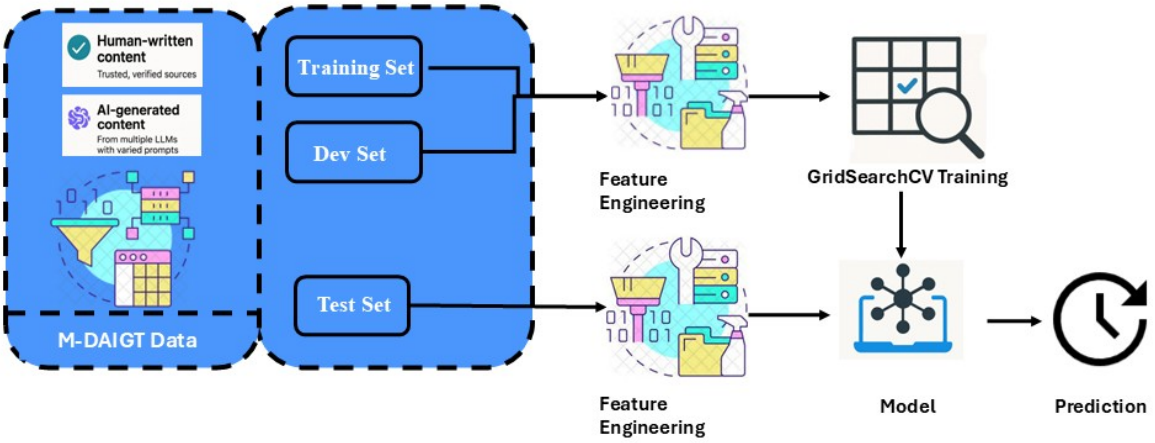

Figure 1: Overall Architecture of the Proposed Model

3.1 Dataset

The News Article Detection (NAD) dataset applied in this study was developed for a M-DAIGT shared task (Lamsiyah et al., 2025) that aims at binary classification of textual content as human-authored or machine-generated. It comprises two principal categories:

- Human-written texts: These samples were gathered from reliable and authenticated reports, such as established news outlets and academic journals.
- AI-generated texts: These examples were created using many cutting-edge large language models (LLMs), including GPT-3.5, GPT-4, and Claude.

The dataset was divided into three subsets as shown in Table 1.

Dataset	Samples	Notes
Train	10,000	Labeled samples
Dev	2,000	Labeled validation samples
Test	3,000	Unlabeled test samples

Table 1: NAD Dataset Statistics.

3.2 Experimental Setup

To examine the effectiveness of classical machine learning models in detecting AI-generated content, we designed a comprehensive experimental pipeline consisting of data preprocessing, feature extraction, model training, hyperparameter tuning, and evaluation.

The dataset was already cleaned before use to ensure high-quality input for training and evaluation. Nevertheless, we applied a two-step text normalization procedure as an additional safeguard. Textual data was transformed into numerical representations using TF-IDF vectorization. The vectorizer was configured with varying max-df thresholds, a range of maximum features, and n-gram ranges to capture both unigram and bigram. In machine learning research, selecting the best-performing model often requires more than just choosing a suitable algorithm. A model's effectiveness significantly depends on its hyperparameters, which are parameters not learned from the data but set before the learning process begins. In our research, we utilized `GridSearchCV`. A grid search with 5-fold cross-validation was conducted to identify optimal hyperparameters for each model. Although we developed two additional models based on deep learning and transformer architectures to conduct a comparative study. The first model is a BiLSTM-based deep learning model that utilizes GloVe pre-trained word embedding, to capture semantic relationships in Arabic text. The second model is a transformer-based architecture built by fine-tuning the XLM-RoBERTa model on our dataset. This setup allows us to compare the effectiveness of static word embeddings with contextualized language representations in text classification. The macro-averaged f1-score was used as the primary evaluation metric during tuning to ensure balanced performance across both classes. All the parame-

ters used in our model are shown in Table 2

Component	Hyperparameter	Values Tested
Classical ML Models		
TfidfVectorizer	max_df	0.85, 0.95
TfidfVectorizer	max_features	5000, 10000
TfidfVectorizer	n-gram_range	(1,1), (1,2)
LinearSVC	dual	False
LR	solver	`liblinear`
LR	max_iter	1000
DT	random_state	42
GridSearchCV	C	0.1, 1, 5
GridSearchCV	CV	5
Deep Learning (BiLSTM) Model		
Embedding Layer	input_dim	vocab_size (based on tokenizer)
Embedding Layer	output_dim	100, 200, 300
Embedding Layer	trainable	True
BiLSTM Layer	units	128
BiLSTM Layer	return_sequences	True
Dropout Layer	rate	0.3
LSTM Layer	units	64
Dense Layer	activation	sigmoid
Model Compile	loss	binary_crossentropy
Model Compile	optimizer	adam
Transformer-Based (XLM-RoBERTa) Model		
TrainingArguments	per_device_train_batch_size	8
TrainingArguments	per_device_eval_batch_size	16
TrainingArguments	warmup_steps	500
TrainingArguments	weight_decay	0.01
TrainingArguments	logging_steps	10

Table 2: Hyperparameters and model settings for NAD

4 Results and discussion

The News Article Detection (NAD) dataset utilized in this study was developed as part of the M-DAIGT shared task, which focused on the binary classification of textual content into a human-written or AI-generated text. Our team participated in this shared task and achieved a high ranking, securing second place on the official leaderboard. The best-performing models are SVM and LR achieving accuracy and f1-score of 0.99 and 0.99 respectively, demonstrating exceptional performance in distinguishing between human- and machine-generated news content.

In addition to classical machine learning models, we also developed two deep learning-based models to conduct a comprehensive comparative study. The first was a BiLSTM model using pretrained word embeddings (GloVe and AraVec), which achieved an f1-score of 0.96. The second model was based on the XLM-RoBERTa transformer architecture, fine-tuned on our dataset, achieving an f1-score of 0.98. While deep learning and transformer models showed competitive performance, the classical machine learning models, particularly SVM and Logistic Regression, offered nearly equivalent results with significantly less computational cost and faster training times. These findings highlight a key trade-off: while deep and transformer-based models can capture more complex patterns. In some cases, traditional linear classifiers remain highly effective and efficient for high-dimensional text classification tasks such as AI-generated content detection.

Model	Accuracy	Precision	Recall	f1-score	Macro Avg
SVM	0.99	0.98	0.99	0.99	0.99
LR	0.98	0.98	0.99	0.99	0.99
DT	0.90	0.90	0.90	0.90	0.90
BiLSTM	0.96	0.95	0.96	0.96	0.96
XLM-RoBERTa	0.98	0.98	0.97	0.98	0.98

Table 3: NAD dataset results

5 Conclusion and Future work

In this study, we explored the task of distinguishing between human-written and AI-generated news articles using the News Article Detection (NAD) dataset from the M-DAIGT shared task. The proposed SVM-based model achieved competitive results and ranked second on the official leaderboard, demonstrating strong performance across all evaluation metrics. Comparisons with Logistic Regression and Decision Tree classifiers further validated the robustness of linear models for this binary classification task. The dataset was well-balanced and carefully cleaned, which contributed to the reliability of our results. Importantly, our findings suggest that not all models or classification tasks require complex transformer architectures or deep multi-layer training to achieve strong results. Simpler, well-tuned models like SVMs can perform competitively, especially when the classification boundaries are clear. Future research will investigate deeper neural networks and transformer-based models to better capture subtle distinctions in AI-generated text. We also plan to incorporate richer semantic and syntactic features to enhance model understanding. Exploring ensemble methods could further boost detection accuracy. Furthermore, expanding the dataset with a wider variety of sources and outputs from different large language models will help improve the generalization and robustness of the system across diverse domains and writing styles.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilker Cingillioglu. 2023. Detecting ai-generated essays: the chatgpt challenge. *The International Journal of Information and Learning Technology*, 40(3):259–268.

Ahmed M. Fetouh and Hamada Nayel. 2023. BFCAI at coli-tunglish@fire 2023: Machine learning based model for word-level language identification in code-mixed tulu texts. In *Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023*, volume 3681 of *CEUR Workshop Proceedings*, pages 205–212. CEUR-WS.org.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (M-DAIGT). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Hamada Nayel. 2020. NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2086–2089, Barcelona (online). International Committee for Computational Linguistics.

Hamada Nayel and Ghada Amer. 2021. A simple n-gram model for urdu fake news detection. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 1150–1155. CEUR-WS.org.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Eman Shehab, Hamada Nayel, and Mohamed Taha. 2024. Character n-gram model for toxicity prediction. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(4):4380–4387.

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.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

Author Index