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.

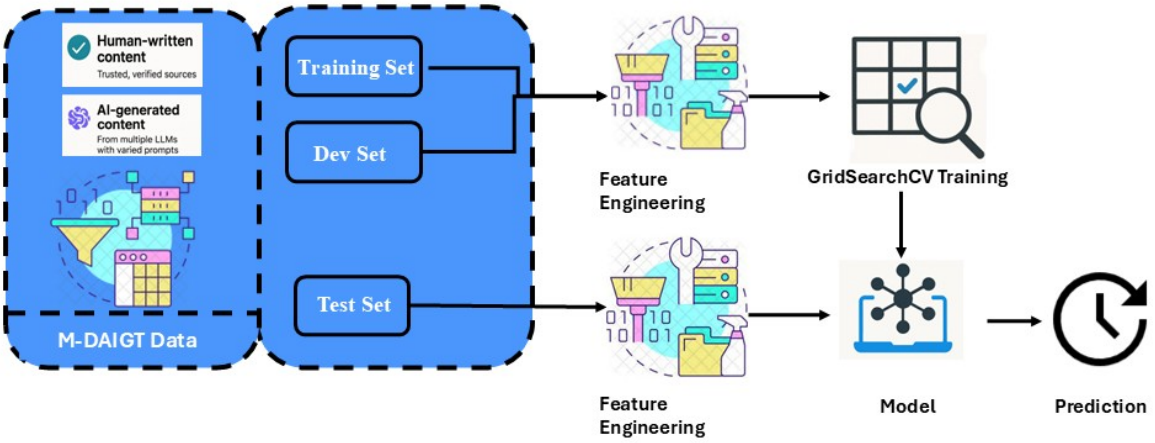

Figure 1: Overall Architecture of the Proposed Model

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.