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

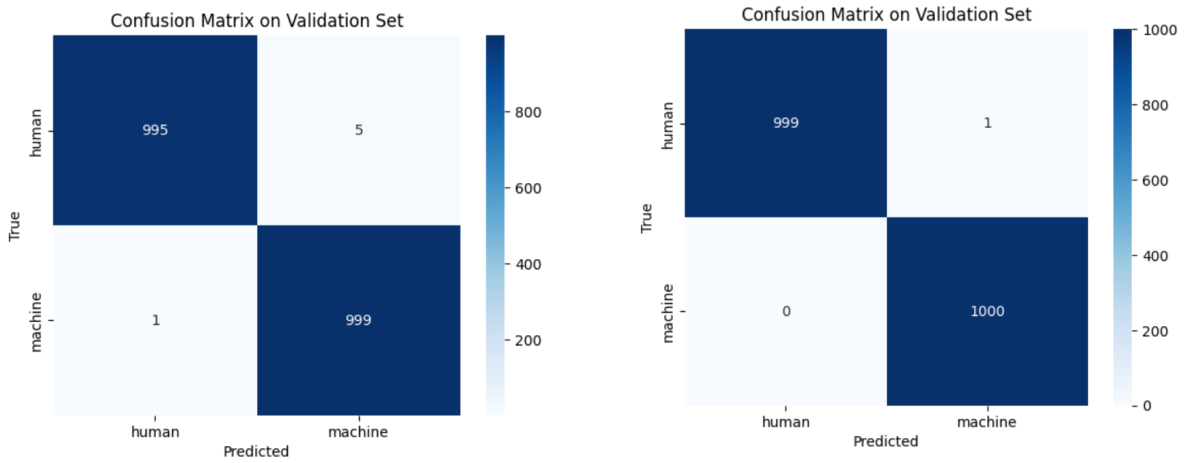

Figure 1: Confusion matrices on the validation sets (News subtask on the left).

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.