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
	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
Subtask-1	**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
	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
Subtask-2	**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.

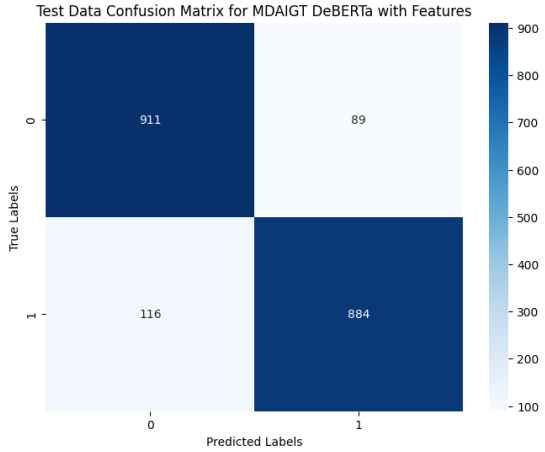

Figure 1: Confusion matrix of DeBERTa model with proposed approach on Subtask 1.

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.