CNLP-NITS-PP at GenAI Detection Task 1: AI-Generated Text Using Transformer-Based Approaches

Annepaka Yadagiri, Reddi Mohana Krishna, L D M S Sai Teja,
M Srikar Vardhan and Partha Pakray
Department of Computer Science & Engineering
National Institute of Technology Silchar, Assam, India, 788010
{annepaka22_rs, redd_pg_23, lekkalad_ug,
mangadoddis_ug, partha}@cse.nits.ac.in

Abstract

In the current digital landscape, distinguishing between text generated by humans and that created by large language models has become increasingly complex. This challenge is exacerbated by advanced LLMs such as the Gemini, ChatGPT, GPT-4, and LLaMa, which can produce highly sophisticated, human-like text. This indistinguishability introduces a range of challenges across different sectors. Cybersecurity increases the risk of social engineering and misinformation, while social media aids the spread of biased or false content. The educational sector faces issues of academic integrity, and within large, multi-team environments, these models add complexity to managing interactions between human and AI agents. To address these challenges, we approached the problem as a binary classification task using an English-language benchmark COLING dataset. We employed transformer-based neural network models, including *BERT*, *DistilBERT*, and *RoBERTa*, fine-tuning each model with optimized hyperparameters to maximize classification accuracy. Our team, *CNLP-NITS-PP* has achieved the 23rd rank in subtask 1 at Coling-2025 for machine-generated text detection in English with a Main Score (*F1 Macro*) of 0.6502 and *micro-F1* score of 0.6876.

1 Introduction

Large Language Models (*LLMs*) represent a significant advancement in Natural Language Processing (*NLP*), advancing development in applications such as machine translation, text analysis, text generation, and question answering (Bommasani et al., 2021; Chowdhery et al., 2023). In academic, industrial, and everyday contexts, the increasing deployment of LLM-powered applications, such as ChatGPT[1], highlights their transformative potential. However, this rapid integration also underscores the importance of understanding their capabilities and limitations to manage expectations and address ethical, societal, and technical challenges effectively (Bender et al., 2021).

Detecting AI-Generated Text (*AGT*) focuses on leveraging Artificial Intelligence (*AI*) to identify and distinguish content produced by AI from Human-Written Text (*HWT*). This area has gained significant importance due to the rapid advancements in Deep Learning (*DL*), which have enabled widespread applications of AGT in content creation, virtual assistants, and more. However, these developments also introduce challenges, including the propagation of misinformation, potential privacy violations, and ethical risks (Al-kfairy et al., 2024). Consequently, AGT detection has emerged as a critical domain in AI research, aimed at mitigating these challenges and ensuring accountability in using generative AI technologies (Bender et al., 2021).

AGT detection research has emerged as a critical area within NLP, driven by advancements in DL. The introduction of robust models, such as Recurrent Neural Networks (*RNN*) (Lipton, 2015), Long Short-Term Memory Networks (*LSTM*) (Hochreiter, 1997), and Transformers (Vaswani, 2017), has significantly enhanced AI capabilities in text generation. These models now produce high-quality content, including articles, dialogues, and news reports. However, their misuse poses substantial risks, such as disseminating misinformation, deception of readers, and propagation of harmful content.

DL plays a pivotal role in generating AGT. DL models can produce realistic and coherent text by analyzing linguistic patterns and structures through training on extensive datasets. Pre-Trained Language Models (PLMs) based on the Transformer architecture (*e.g., GPT-3, BERT*) have demonstrated exceptional performance across various NLP tasks, contributing significantly to AGT development (Vaswani, 2017; Brown, 2020).

However, the widespread adoption of DL tech-

[1] https://chatgpt.com/

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 197–202
January 19, 2025.

niques for AGT generation has raised several challenges. These include the potential for spreading misinformation, such as AI-generated fake news and deceptive advertisements that could influence public opinion (Al-kfairy et al., 2024). Additionally, personal data may be exploited to generate misleading or targeted fraudulent content (Bender et al., 2021). Furthermore, DL-powered AI can be misused to create inappropriate material, including violent, pornographic, or hate speech content, which may be widely disseminated (Zellers et al., 2019).

Numerous researchers are developing strategies to detect and identify problematic content to address the challenges associated with AGT. These strategies include rule-based and statistical approaches, as well as ML techniques like Support Vector Machines (*SVM*) and Random Forests (*RF*), which are commonly used for building detection models (Aristantia et al., 2024). Additionally, combining these techniques with DL models, such as those based on the Transformer architecture, is being explored to improve detection accuracy. This paper introduces a tool designed to detect LLM-generated AI text using Transformer-based models to improve detection accuracy and provide insights for future research.

2 Related Work

Various commercial and open-source tools, such as GPTZero, ZeroGPT[2], AI Content Detector, and GPT-2 Output Detector (Mitchell et al., 2023), have emerged to detect AI-generated content effectively. Additionally, active research focuses on curating specialized datasets and determining which features and classifiers can enhance classification performance. For example, (Yu et al., 2023) compiled a dataset of human and AI-generated abstracts to assess commercial and non-commercial detection systems, though the dataset is currently limited to English.

Recent studies have experimented with different detection methodologies, such as using *XGBoost classifiers* (Shijaku and Canhasi, 2023), decision tree algorithms (Zaitsu and Jin, 2023), and transformer-based models (Guo et al., 2023). Notably, analyzed text from English customer reviews, developing a transformer-based classifier that achieved a classification accuracy of 79%. These efforts demonstrate a trend toward optimizing detection systems, enhancing reliability, and expanding detection capabilities across various languages and text types.

[2] https://www.zerogpt.com/

3 Proposed Methodology

3.1 Problem Statement

Given the rapid advancements and adoption of LLMs, it is increasingly challenging to differentiate between HWT and AGT. Identifying AGT can be defined as a classification problem: determining whether a given sequence of words $S = \{w_1, w_2, \ldots, w_n\}$ was generated by an AI model or by a human.

Formally, let S represent a text sample of n words. The problem can then be framed as:

- **Input:** A text sample S where $S = \{w_1, w_2, \ldots, w_n\}$.
- **Output:** A binary label $y \in \{0, 1\}$, where:
  - $y = 0$ *denotes HWT,*
  - $y = 1$ *denotes AGT.*

3.2 Dataset Description

In the COLING Workshop on MGT Detection Task 1, a binary classification approach is employed to distinguish whether a given text is generated by an AI or authored by a human (Wang et al., 2025). A diverse dataset is compiled by initially gathering English-language datasets, including HC3 (Guo et al., 2023), MAGE (Li et al., 2024), and M4GT (Wang et al., 2024). These datasets are subsequently merged and refined into a final consolidated dataset for further analysis. The statistical properties of the refined dataset are presented in Table 1. In contrast, Figure 1 visually compares the training and development datasets.

Label	English	
	Train Count	Dev Count
Human	228,922	98,328
AI	381,845	163,430
Total	610,767	261,758

Table 1: Dataset Label Counts for English Train and English Dev

3.3 System Description

This paper presents our methodology and results for the MGT Detection Task 1, which focuses on identifying AGT. The primary objective of this task

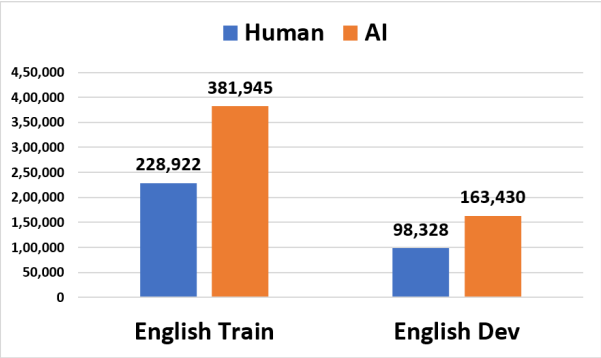

Figure 1: Visually compared to Train and Dev dataset

Parameter	Value
Activation Function	Sigmoid
Optimizer	AdamW
Loss Function	binary_crossentropy
Learning Rate	2×10^{-5}
Batch Size	16
Number of Epochs	03
Dropout	0.2
ModelCheckpoint	Yes
EarlyStopping	Yes
Patience	5

Table 2: Hyperparameters utilized across all experiments

is to classify whether a given text segment has been produced by a machine or authored by a human. Our participation was specifically in Subtask A, which deals exclusively with English texts.

We employed a two-pronged approach combining the fine-tuned **DistilBERT** model (Sanh, 2019), optimized for capturing semantic nuances, and a rule-based feature extraction strategy. Key **hyperparameters** were adjusted to enhance performance for the task as shown in Table 2.

In parallel, we implemented a rule-based approach to extract a set of **linguistic and statistical features** that could complement the semantic insights of the model. These features included measures such as **Average Line Length, Vocabulary richness, Word Density, and Part-Of-Speech (POS) tag distributions**, computed for each text sample. Such features were chosen based on their potential to highlight subtle stylistic and structural differences between HWT and AGT. The following sample text is from the COLING dataset.

Sample text = *"Hitler's plans for the succession and power structure after his death are shrouded in mystery, as he never explicitly wrote down his intentions. However, it is known that he designated several potential successors, including Heinrich Himmler, Hermann Göring, and Joseph Goebbels, each with their own strengths and weaknesses."*

Average Line Lenght: In NLP, the average line length refers to the mean number of words per line in a given text dataset (Guo et al., 2023).

$$\text{Average line length} = \frac{\text{Total word count}}{\text{Total line count}} \quad (1)$$

For example, consider the above sample text: Total word count = 63, Total line count = 2. Thus, the average line length is 31.5 words per line.

Vocabulary Richness: It quantifies the uniqueness of words within a given text (Guo et al., 2023).

$$\text{Vocabulary Richness} = \frac{\text{Total Number of Words}}{\text{Number of Unique Words}} \quad (2)$$

For the above sample text, the number of unique words is 43, and the total number of words is 63, so we got the vocabulary richness as 0.746.

Word Density: In NLP, word density measures the concentration of unique words in a given text (Guo et al., 2023).

$$\text{Word Density} = \frac{100 \cdot \text{Vocabulary Size}}{\text{No of Lines} \cdot \text{Average Line Len}} \quad (3)$$

For the above sample text, the unique word count is 43, the average line length is 31.5, and the number of lines is 2, so we got the word density as 74.6.

Part-Of-Speech tag: POS tags are labels assigned to each word in a text to indicate its grammatical category, such as noun, verb, adjective, etc (Guo et al., 2023).

The above sample text contains various POS distributed as follows: Nouns (*NN, NNS, NNP, NNPS*) appear 13 times, with words like "plans", "succession", "power", and proper nouns like "Himmler", "Goebbels", and "Göring." Verbs (*VB, VBD, VBG, VBN, VBP, VBZ*) are used 7 times, including "are," "shrouded," "wrote," and "designated." Punctuation marks (*., „ :, (,), ", ”, ‘‘‘, !, ?, ;,* -) occur 7 times, such as in "mystery", "death", and "intentions". Determiners (*DT, PDT, WDT*) appear 6 times, including words like "the" and "his". Pronouns (*PRP*) are used 6 times, such as "his," "he," and "it". Proper nouns (*NNP, NNPS*) also occur 7 times, such as "Hitler's", "Heinrich", and "Himmler". Adjectives (*JJ, JJR, JJS*) appear 2 times with words like "potential" and "own". Auxiliary verbs (*MD*) do not appear in the text. Adverbs (*RB, RBR, RBS*) are used 3 times, including "never" and "explicitly". Particles (*RP*) appear once with "down".

Subordinating conjunctions (*IN*) are used 6 times, including “for," “after," and “that". Numbers (*CD*) are absent in this text. Foreign words (*FW*) and interjections (*UH*) do not appear. Prepositions (*IN*) like “for" and “with" are used 6 times. Symbols (*SYM*) and spaces (*SP*) are not present, and coordinating conjunctions (*CC*) such as “and" appear 3 times.

The extracted linguistic features were subsequently integrated with the DistilBERT-based embeddings, creating a hybrid feature set that combines text-based and numerical characteristics. This integration aimed to enhance the system’s ability to distinguish between the two types of content by leveraging deep semantic understanding and surface-level textual patterns. Below is the architecture description.

Architecture Description: The custom model architecture described in Figure 2 combines a pre-trained DistilBERT model with additional feature processing for sequence classification. The model uses the **DistilBERTForSequenceClassification** module, which includes the **DistilBERT-Model** for generating contextual embeddings. The model leverages a transformer-based architecture, comprising six layers of TransformerBlocks, each consisting of MultiHeadSelfAttention and a **Feed-Forward Network (FFN)** with *GELU* activation. These blocks enable the model to capture complex relationships within input sequences. The model further incorporates a pre-classifier layer that refines the BERT output by projecting it to a 768-dimensional space, followed by a final classifier layer that reduces the dimensionality to two output nodes for classification. A dropout layer with a 0.2 rate is also used to prevent overfitting. Beyond the BERT layers, the model also integrates a fully connected layer **(feature_fc)** that processes additional input features, followed by a **ReLU** activation and another dropout layer (*0.3 rate*). It is important to note that the additional features are not separately normalized, and the activation function (*ReLU*) is applied directly within the neural network layers, which ensures that normalization is not exclusively used on the features. Finally, the outputs from the BERT model (*768*) and the additional features from the neural network layer with ReLU activation (*64*) are concatenated and passed through a final classifier layer (*832 input features*) and subjected to a **sigmoid** activation function, which outputs probabilities for each of the two classes. This design combines the robust contextual understanding of DistilBERT with additional feature-based inputs for enhanced predictive performance in sequence classification tasks.

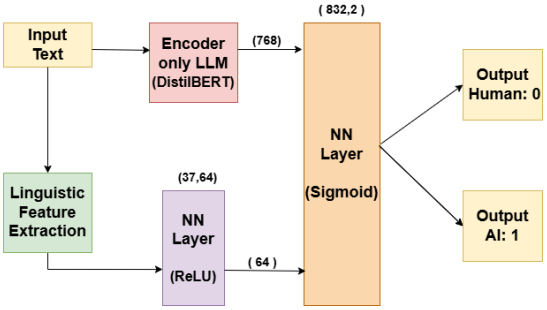

Figure 2: Training Architecture for AGT Detection

Our final system demonstrated strong performance, achieving an **F1 score** of **0.6513** on the test set for Subtask A. This placed us **23rd** out of 36 participating teams, reflecting the competitiveness of our approach. While there is room for further improvement, our results underscore the **effectiveness of combining transformer-based embeddings with handcrafted linguistic features**, showcasing the potential of hybrid models in AI text detection tasks.

3.4 Results Analysis

The comparison between the baseline results in Table 3 and the results obtained using linguistic features in Table 4 underscores the significant benefits of incorporating linguistic features. Training accuracy increased notably from **84.28** to **92.32**, accompanied by a substantial reduction in training loss from **0.355** to **0.193**. Validation metrics also showed measurable improvements, with accuracy rising from **79.25** to **81.62** and the F1-score increasing from **0.773** to **0.818**. These results highlight linguistic feature’s efficacy in enhancing training and validation performance, indicating their value in improving model robustness and generalization.

The results from the leaderboard are displayed in Table 5, showcasing our team’s achievements. These results were achieved using the DistilBERT model enhanced with linguistic features, with hyperparameters tuned as illustrated in Table 2.

4 Conclusion

Our approach to the MGT Detection Task 1 effectively combined DistilBERT’s semantic embeddings with rule-based linguistic features and hyperparameters. This hybrid strategy enhanced

Epoch	Dataset	Accuracy	Loss	F1 Score
1	Train	0.842	0.355	0.835
	Val	0.792	0.631	0.773
2	Train	0.878	0.286	0.872
	Val	0.799	0.343	0.801
3	Train	0.890	0.225	0.895
	Val	0.801	0.372	0.804

Table 3: Baseline Results for Training and Validation Metrics

Epoch	Dataset	Accuracy	Loss	F1 Score
1	Train	0.858	0.323	0.863
	Val	0.808	0.428	0.812
2	Train	0.894	0.253	0.897
	Val	0.815	0.433	0.816
3	Train	0.923	0.193	0.925
	Val	0.816	0.486	0.818

Table 4: Training and Validation Metrics Using Linguistic Features

the model's ability to distinguish between human-written and AI-generated text. The model showed steady improvement during training, with training accuracy rising from 85.83 to 92.32 and the F1 score increasing from 0.863 to 0.925. On the test set, our system achieved a **Main F1 Macro score of 0.6502** and an **Auxiliary F1 Micro score of 0.6876**, ranking 23rd out of 36 teams. These results demonstrate the effectiveness of our feature integration approach, though future work should focus on improving generalization through better feature selection and regularization.

Task	Main (F1 Macro)	Auxiliary (F1 Micro)
English	0.6502	0.6876

Table 5: Test Results by Leaderboard

References

Mousa Al-kfairy, Dheya Mustafa, Nir Kshetri, Mazen Insiew, and Omar Alfandi. 2024. Ethical challenges and solutions of generative ai: an interdisciplinary perspective. In *Informatics*, volume 11, page 58. MDPI.

Dita Wahyuni Aristantia, Muhammad Baharuddin, Nur Mazidah, and Zaenab Tri Lestari. 2024. Learning model of arabic in indonesia?: A study of the curriculum system at bahrul ulum tambakberas islamic boarding school, jombang and an-nuqayah, madura. *Edumaspul: Jurnal Pendidikan*, 8(1):485–499.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Zachary Chase Lipton. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv Preprint, CoRR, abs/1506.00019*.

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.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rexhep Shijaku and Ercan Canhasi. 2023. Chatgpt generated text detection. *Publisher: Unpublished*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *arXiv preprint arXiv:2304.12008*.

Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing chatgpt (-3.5,-4)-generated and human-written papers through japanese stylometric analysis. *PLoS One*, 18(8):e0288453.

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.