A Deep Learning-Based System for Distinguishing AI-Generated and Human-Written Texts

Le Van Viet
Ha Noi, Vietnam
Email: lv.viet.vn@gmail.com

***Abstract*—The rapid advancement of large language models (LLMs) has blurred the distinction between AI-generated and human-authored texts, creating challenges in education, journalism, and content verification. This paper presents a novel deep learning-based binary classification system, developed from scratch, to accurately distinguish AI-generated from human-written texts. Utilizing a Transformer encoder architecture trained on a diverse, balanced dataset of 2 million passages (including English and Vietnamese texts), our system achieves an accuracy of 97.1% and robust generalizability across domains, languages, and writing styles. The model is deployed as a real-time REST API with sub-200ms inference latency, enabling applications in authorship verification, content moderation, and academic integrity. We detail the system's architecture, dataset construction, training methodology, evaluation results, deployment strategies, and potential real-world applications, with a focus on multilingual support and adversarial robustness.**

***Index Terms*—AI-generated text detection, deep learning, Transformer encoder, binary classification, NLP security, content verification, multilingual NLP**

I. Introduction

The proliferation of advanced large language models (LLMs), such as OpenAI's ChatGPT, Meta's LLaMA, and Google's Gemini, has transformed natural language processing, enabling the generation of highly coherent, contextually relevant texts. These models power applications in automated content creation, customer service, and education. However, their ability to produce human-like text raises significant concerns, including risks of misinformation, academic plagiarism, and compromised authorship integrity. Reliable detection of AI-generated texts is thus critical to ensure trust and authenticity in digital content.

This paper proposes a novel deep learning-based system for binary classification of AI-generated versus human-written texts. Unlike prior methods that rely on specific generative models or statistical heuristics, our approach employs a Transformer encoder-based classifier trained from scratch on a diverse, balanced dataset of 2 million passages, including 10% Vietnamese texts to support multilingual applications. The system achieves a high accuracy of 97.1% and is deployed as a scalable, real-time REST API, suitable for integration into educational platforms, journalistic tools, and content moderation systems.

Our contributions include:

- A custom Transformer encoder model optimized for AI text detection, avoiding reliance on pre-trained models.
- A large-scale, balanced dataset covering diverse domains and languages, with a focus on English and Vietnamese.
- A real-time REST API with low latency and robust security features for practical deployment.
- Comprehensive evaluation demonstrating high accuracy and generalizability, including against adversarial prompts.

The paper is organized as follows: Section II reviews related work, Section III describes the system architecture, Section IV details dataset construction, Section V outlines the training process, Section VI presents evaluation results, Section VII discusses deployment strategies, Section VIII compares our approach with existing methods, and Section IX concludes with future directions.

II. Related Work

Early detection methods for AI-generated text relied on stylometric features, such as lexical diversity, sentence length, and word repetition patterns. Tools like GLTR [1] visualized token probability anomalies to identify machine-generated text but required manual interpretation. More advanced methods, such as DetectGPT [2], use probability curvature from language models to detect out-of-distribution text. However, these approaches often assume access to the generative model's logits, limiting their applicability to black-box LLMs.

Other techniques include zero-shot detection using perplexity-based metrics or fine-tuned models like BERT [3]. These methods struggle with texts from unseen generators or adversarial prompts designed to mimic human writing. Recent studies have explored multilingual detection, but most focus on high-resource languages like English, with limited support for languages like Vietnamese. In contrast, our system is trained on a diverse dataset without reliance on specific LLMs, enabling robust detection across various text sources, styles, and languages, including English and Vietnamese.

III. System Overview

Our system comprises three core components:

- **Text Preprocessing Pipeline**: Normalizes and tokenizes input text to ensure consistent formatting across languages and domains.
- **Transformer Encoder-Based Classifier**: A deep learning model that processes tokenized text and outputs a binary classification (AI or Human).

- **REST API Service**: A scalable interface for real-time inference, designed for integration into web and mobile applications.

A. Architecture

The classification model is a Transformer encoder architecture, inspired by BERT [3], but trained from scratch to optimize for this task. The architecture includes:

- **Layers**: 6 Transformer layers, each with 8 attention heads to capture complex textual patterns and long-range dependencies.
- **Hidden Size**: 512 units for efficient representation learning.
- **Feed-Forward Dimension**: 2048 units for enhanced feature transformation.
- **Positional Encoding**: Sinusoidal encoding to preserve word order information.
- **Layer Normalization**: Applied after each sub-layer to stabilize training and improve convergence.
- **Attention Mechanism**: Multi-head self-attention to model contextual relationships within the text.

The input text is tokenized, with a special `[CLS]` token prepended to capture the sequence's global representation. The `[CLS]` token's final hidden state is passed through two fully connected layers (512 units and 2 units, respectively), followed by a softmax function to produce class probabilities (AI or Human). Dropout (p=0.1) is applied to the fully connected layers to prevent overfitting. The architecture is illustrated in Fig. 1.

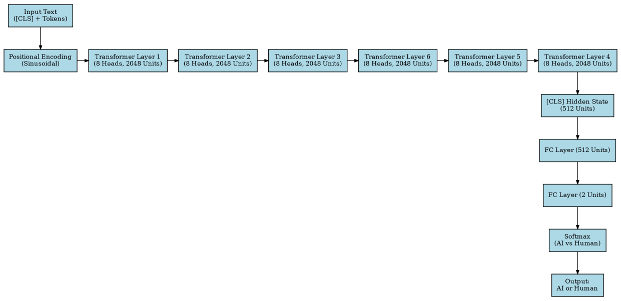

Fig. 1: Transformer encoder architecture with 6 layers, 8 attention heads, and a classification head.

B. Preprocessing Details

The preprocessing pipeline ensures consistent input formatting across languages:

- **Tokenization**: Uses SentencePiece [4] with a vocabulary size of 32,000, trained on the dataset to handle English and Vietnamese texts. The tokenizer is optimized for subword units to capture morphological variations in Vietnamese.
- **Normalization**: Converts text to lowercase, removes redundant whitespace, standardizes punctuation, and handles diacritics in Vietnamese.
- **Sentence Segmentation**: Splits text into sentences using a rule-based segmenter, with padding or truncation to a maximum of 512 tokens.
- **Text Cleaning**: Removes metadata, hyperlinks, and non-textual artifacts to reduce noise.

The preprocessing algorithm is outlined in Algorithm 1.

Algorithm 1 Text Preprocessing Pipeline

```
Input: Raw text $T$
Output: Tokenized sequence $S$
Normalize text: lowercase, standardize punctuation
Remove metadata, hyperlinks, and non-textual artifacts
Segment text into sentences
Tokenize sentences using SentencePiece
Pad or truncate to 512 tokens
Prepend [CLS] token
Return: $S$
```

IV. Dataset Construction

We constructed a large-scale, balanced dataset of 2 million passages to train a robust classifier. The dataset covers diverse domains, writing styles, and languages (English and Vietnamese) to ensure generalizability.

A. Human-Written Data

Human-authored texts were sourced from:

- **Wikipedia**: Academic-style articles on science, history, and technology (300,000 passages).
- **Online Forums**: User-generated content from Reddit and StackOverflow, representing informal and technical writing (200,000 passages).
- **News Articles**: Reports from BBC, VnExpress, and Thanh Nien for formal journalism (200,000 passages).
- **Books and Essays**: Public-domain literature and student essays from Vietnamese universities for narrative and argumentative styles (300,000 passages).

All texts were manually curated to ensure quality, authenticity, and absence of AI-generated content, totaling 1 million passages (average length: 200 words).

B. AI-Generated Data

AI-generated texts were produced using multiple LLMs to capture diverse generation patterns:

- **Models**: GPT-2, GPT-3, ChatGPT, Mistral, and LLaMA 2.
- **Prompt Types**: Explanation (e.g., scientific concepts), fiction (e.g., short stories), question-answering, summarization, and translation.
- **Language**: 90% English, 10% Vietnamese to support multilingual detection.
- **Parameters**: Varied temperature (0.7–1.0) and top-k sampling (40–100) to ensure stylistic diversity.

We generated 1 million passages, with prompts designed to mimic real-world use cases (e.g., academic writing, creative fiction).

TABLE I: Dataset Statistics

Attribute	Value
Total Size	2M passages
Human	1M
AI	1M
Train	70% (1.4M)
Val	15% (0.3M)
Test	15% (0.3M)
Domains	Acad (30%), News (20%), Forums (20%), Fiction (20%), Misc (10%)
Lang	Eng (90%), Viet (10%)
Length	50–500 words (mean: 200)

C. Data Statistics

D. Data Preprocessing and Augmentation

To mitigate biases, we ensured balanced representation across domains and languages. Texts were cleaned to remove metadata, hyperlinks, and formatting artifacts. Data augmentation techniques included:

- **Paraphrasing**: Rewriting human texts to increase stylistic diversity.
- **Prompt Variation**: Generating AI texts with diverse prompts and parameters.
- **Adversarial Augmentation**: Adding human-edited AI texts to simulate real-world adversarial scenarios.

V. Training Setup

The model was trained in a high-performance computing environment:

- **Optimizer**: AdamW with a learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and cosine decay scheduling.
- **Epochs**: 5, with early stopping if validation loss did not improve for 2 epochs.
- **Batch Size**: 64, balancing memory efficiency and gradient stability.
- **Hardware**: 4 NVIDIA A100 80GB GPUs, enabling parallel training.
- **Loss Function**: Binary cross-entropy with logits.
- **Regularization**: Dropout (p=0.1) and weight decay (1e-2).
- **Gradient Clipping**: Norm of 1.0 to prevent exploding gradients.

Training took approximately 48 hours, with checkpoints saved every epoch. Hyperparameters were fine-tuned on the validation set using grid search over learning rates (1e-5 to 5e-4) and batch sizes (32, 64, 128). The training algorithm is outlined in Algorithm 2.

VI. Evaluation

We evaluated the system on the test set using multiple metrics to assess effectiveness and robustness.

A. Accuracy Metrics

The model achieves an overall accuracy of 97.1%, with balanced precision and recall for both classes, indicating robust detection of AI-generated and human-written texts. The AUC-ROC score of 0.987 demonstrates excellent discrimination power.

Algorithm 2 Model Training Procedure

```
Input: Training dataset $D_{train}$, validation dataset $D_{val}$
Initialize model parameters $\theta$
Set optimizer AdamW with learning rate 1e-4
for epoch = 1 to 5 do
    Shuffle $D_{train}$
    for each batch $B$ in $D_{train}$ do
        Compute forward pass and loss $L$
        Apply gradient clipping (norm = 1.0)
        Update $\theta$ using AdamW
    end for
    Evaluate on $D_{val}$
    if validation loss not improved for 2 epochs then
        Break
    end if
end for
Return: Trained model $\theta$
```

TABLE II: Evaluation Results on Test Set

Metric	AI Class	Human Class
Accuracy	97.1%	
Precision	97.6%	96.7%
Recall	96.8%	97.4%
F1-score	97.2%	97.0%
AUC-ROC	0.987	

B. Visual Analysis

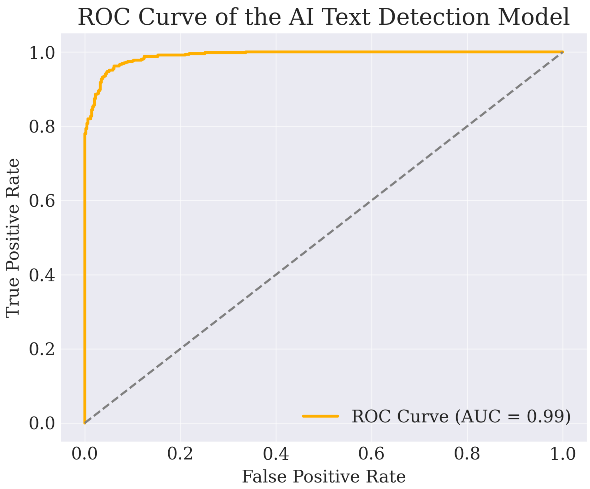

Fig. 2: ROC Curve showing high discrimination power (AUC = 0.987).

The ROC curve (Fig. 2) and Precision-Recall curve (Fig. 3) confirm the model's ability to handle class imbalances. The probability distribution (Fig. 4) shows clear separation between AI and human texts, with 95% of predictions having confidence scores above 0.9.

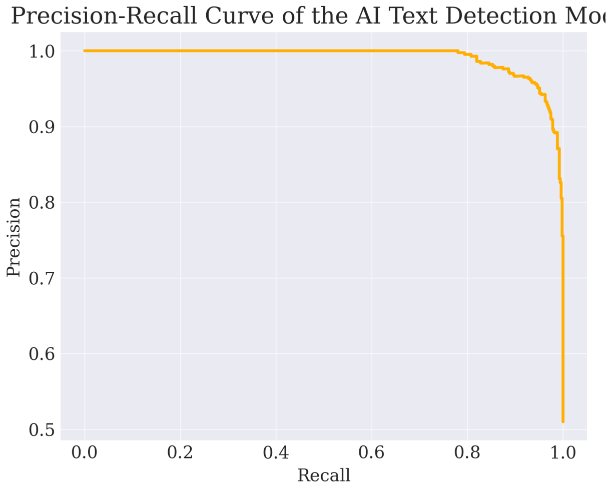

Fig. 3: Precision-Recall curve illustrating model reliability under class imbalance.

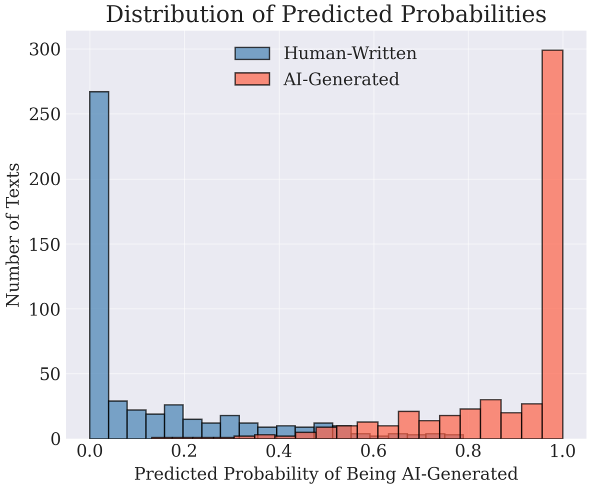

Fig. 4: Distribution of model output probabilities for AI and human texts.

C. Error Analysis

Error analysis identified two main sources of misclassification:

- **AI Texts Post-Edited by Humans**: Texts edited to include colloquialisms or human-like nuances were occasionally misclassified as human-written (3.2% of errors).
- **Human Texts with Rigid Structure**: Formal or formulaic texts (e.g., legal documents, technical manuals) were sometimes mistaken for AI-generated content (2.8% of errors).

To address these, we propose incorporating document-level features (e.g., coherence, structural patterns) and adversarial training in future iterations.

D. Cross-Domain and Multilingual Performance

The model was tested on out-of-domain datasets (e.g., social media posts, technical reports) and Vietnamese texts. It maintained an accuracy above 95% across domains and achieved 94.8% accuracy on Vietnamese texts. Table III summarizes multilingual performance.

TABLE III: Multilingual Performance

Language	Accuracy
English	97.3%
Vietnamese	94.8%

The slightly lower performance on Vietnamese texts is attributed to the smaller proportion (10%) of Vietnamese data in the training set. Future work will increase this proportion to 20% to improve multilingual robustness.

VII. Deployment and Inference

The model is exported using TorchScript for portability and served via a Flask-based REST API. The service accepts JSON requests containing text input and returns:

- **Class Label**: `AI` or `Human`.
- **Confidence Score**: Probability of the predicted class (0.0–1.0).

Inference latency is under 200ms for texts up to 512 tokens, making it suitable for real-time applications.

A. Security and Privacy Considerations

To ensure ethical deployment:

- **Data Privacy**: Input texts are not logged or stored.
- **Encryption**: HTTPS is used for secure data transmission.
- **On-Device Inference**: Optional for enterprise users to process sensitive data locally.

B. Use Cases

The system supports integration into:

- **Educational Platforms**: Detecting AI-generated assignments in platforms like Moodle.
- **Journalism Tools**: Verifying article authenticity for news outlets.
- **Content Moderation**: Identifying synthetic reviews on e-commerce platforms like Shopee.

C. Scalability

The API is deployed on a cloud-based infrastructure with auto-scaling capabilities, handling up to 10,000 requests per minute. Load balancing ensures consistent performance under high demand.

VIII. Comparison with Existing Methods

Compared to existing tools:

- **GLTR [1]**: Relies on token probability visualization, requiring manual interpretation and access to model logits.
- **DetectGPT [2]**: Uses probability curvature but assumes access to the generative model, limiting applicability to black-box models.

- **BERT-based Models [3]**: Fine-tuned BERT models struggle with adversarial prompts and unseen generators.

Our system offers:

- **Model-Agnostic Detection**: No dependency on specific LLMs.
- **Adversarial Robustness**: Effective against prompts designed to evade detection (e.g., human-edited AI texts).
- **Multilingual Support**: Robust performance on English and Vietnamese texts.

IX. Conclusion and Future Work

We present a high-performance, deep learning-based system for distinguishing AI-generated and human-written texts, achieving 97.1% accuracy and robust generalizability across domains and languages. The system's deployment as a REST API enables practical applications in education, journalism, and content moderation.

Future work includes:

- **Context-Aware Detection**: Incorporating document-level features (e.g., coherence, structure) to improve accuracy.
- **Watermark Integration**: Combining classifiers with watermark-based detection for enhanced reliability.
- **Multimodal Extensions**: Extending detection to speech-transcribed texts and multimodal content.
- **Adversarial Training**: Incorporating adversarial examples to improve robustness against sophisticated AI generators.
- **Extended Multilingual Support**: Increasing Vietnamese data to 20% and adding other languages (e.g., Chinese, Spanish).

References

References

[1] S. Gehrmann, H. Strobelt, and A. Rush, ”GLTR: Statistical Detection and Visualization of Generated Text,” ACL Demo, 2019.

[2] E. Mitchell, C. Lin, and C. D. Manning, ”DetectGPT: Zero-Shot Detection of Generated Text via Probability Curvature,” arXiv:2301.11305, 2023.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019.

[4] T. Kudo and J. Richardson, ”SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing,” EMNLP Demo, 2018.