Minimax Theory and its Applications

Volume 11 (2026), No. 4s, 145 to 164

Contrastive Transformer Framework for Explainable Detection of AI-Generated and Rephrased English Text

Deshpandey Pavan Kumar[1], Dr.E.S.Sharmila Sigamany[2], Mr Nalluri Babu Rao[3], Angalakuduru Aravind[4], Rendla Ashlesha[5], Elangovan Muniyandy[6], Kommalapati Prasanthi[7], Mrs N. Prasanna Lakshmi[8]

[1]Research Scholar, Department of English, koneru Lakshmaiah Education foundation , Vaddeswaram campus, Andhra Pradesh , India, Email: Pavandeshpandey437@gmail.com
[2]Associate professor , Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India, Email: sharmilajoe@kluniversity.in
[3]Asst Professor, Department of English , Anurag University, Hyderabad, Email: baburaohs@anurag.edu.in
[4]Assistant professor, Department of English, Anurag Engineering College, Kodad,Telangana,India, Email: aravindangalakuduru@gmail.com
[5]Department Of English Kalasalingam Academy Of Research And Education Krishnankoil Tamilnadu-626126 India, Email: Rendla9@Gmail.Com
[6]Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences,Chennai - 602 105, Email: muniyandy.e@gmail.com
[7]Assistant Professor Department of Computer Science of Engineering , Kallam Harinadhareddy Institute of Technology, Chowdavaram, Guntur, Andhra Pradesh, India, Email: prasanthik06@gmail.com
[8]Assistant Professor , Department of CSE(AI&ML), RVR &JC College Of Engineering, Chowdavaram, Guntur, Andhra Pradesh , India, Email: lakshmiprasannagali62@gmail.com

Abstract
Generative Artificial Intelligence (GAI) and LLMs have transformed English textgeneration, editing, summarization, and rephrasing, creating increasing concerns regarding academic integrity, authenticity, misinformation, and the credibility of digital communication. Distinguishing human-written, AI-generated, and AI-rephrased text remains challenging because AI rephrasing can preserve semantic meaning while substantially altering linguistic characteristics. Existing detection approaches often focus on either contextual semantic representations or handcrafted linguistic features and provide limited interpretability of classification decisions. To address these limitations, this study proposes a Contrastive Transformer Framework for explainable three-class detection of human-written, AI-generated, and AI-rephrased English text. The framework employs DeBERTaV3 to capture contextual semantic representations and multi-level contrastive learning to enhance separation among closely related text classes. Textagon-based lexical, syntactic, readability, and stylistic features are extracted and encoded through a Hierarchical Attention Network (HAN), after which semantic and linguistic representations are fused for classification. Layer Integrated Gradients provides interpretable evidence supporting each prediction, while cross-domain evaluation examines generalization between educational and news text. The proposed framework was evaluated using 12,000 texts comprising balanced human-written, AI-generated, and AI-rephrased samples from both domains. The complete framework achieved 94.6% accuracy, 94.4% precision, 94.2% recall, and 94.3% F1-score, outperforming the Neural Network, XGBoost, and PPL-based Random Forest baselines. Cross-domain accuracy reached 90.8% for educational-to-news and 91.4% for news-to-educational evaluation. The novelty lies in integrating contrastive DeBERTaV3 representations, Textagon-HAN linguistic modeling, semantic-linguistic fusion, and Layer Integrated Gradients for explainable three-class and cross-domain detection. The findings demonstrate accurate, interpretable, and robust identification of AI-generated and AI-rephrased English text.

Keywords: Automated text generation, English language, Human-written text, Language models, Text classification, Text detection

1. Introduction

Generative AI (GAI) and Large Language Models (LLMs) have revolutionized the process of creating, editing, summarizing, and rephrasing English text[1]. In education, the field of journalism and research, and digital

communication, systems like ChatGPT can produce coherent and grammatically correct text that is strikingly similar to human writing, opening up new possibilities[2]. AI's growing capabilities, however, have sparked worries about academic integrity, authenticity of writing, plagiarism, academic dishonesty, and the presence of misinformation[3]. The tasks become more difficult when the existing human written text needs to be rephrased with AI, as the resulting text can capture the meaning of the original text, but significantly alter the language's surface features[4]. Thus, the problem of separating human written, AI generated and AI rephrased English text is an important research challenge[5]. In teaching and reporting contexts, such as schools and media outlets, it is especially important to be able to trace back and validate the source and veracity of text-based information, such as affecting student evaluation, credibility and information trustworthiness[6]. AI-text detection is different from conventional plagiarism detection, which involves recognizing subtle semantic, syntactic, stylistic and probabilistic patterns that are associated with the generation and transformation of text[7]. Thus, detection systems must be able to pick up on the deeper meaning of text in addition to the linguistic features that distinguish between the writing of humans and AI-generated and AI-rephrased writing[8].

Previously, AI-generated text detection has been explored through various methods such as detection tools, handcrafted linguistic features, statistical features, pretrained language models, adversarial analysis, and cross-domain learning[9]. Previous studies have found that traditional detection methods have been less than effective in identifying high-quality AI-generated content and have been incorrect in identifying text written by humans[10]. Further analyses focused on semantic, perplexity, readability, error-based, document and text-vector features for the purpose of detecting AI-generated and/or AI-rephrased text in the context of educational and news texts[11]. Additionally, other studies investigated how to enhance the reliability of detection by using adversarial robustness, paraphrase effects, attention-based characteristics, dependency-based text reordering, pretrained language models, and prompt-tuning strategies[12]. These methods have proven to be effective for detecting AI-generated text, but there are some drawbacks[13]. Most current approaches focus either on semantic representations or on explicitly designed linguistic features, instead of learning both types of information simultaneously. Moreover, some conventional classification methods might not be adequate to increase the distinction between human written, AI-generated, and AI-rephrased classes when they are close to each other. As more and more documents are being rephrased using AI, detectors have to deal with this challenge as well, as changed documents can appear similar to naturally written ones even though they may have altered meaning. Furthermore, few works have focused their efforts on incorporating contrastive representation learning into linguistic feature analysis and presented explainable evidence for each prediction. These restrictions suggest a more comprehensive framework for detection which should be capable of learning discriminative representations and also provide an explanation of the rationale behind its classification choices.

To tackle these challenges, the present study suggests a Contrastive Transformer Framework for Explainable Detection of AI-Generated and Rephrased English Text. The proposed framework aims to provide more detailed classification compared to traditional binary detection methods, by specifically targeting three classes: human-written, AI-generated, and AI-rephrased English text. Firstly, DeBERTaV3 is used as the context-based semantic encoder to extract the relationship between words and wider textual context. Multi-level contrastive representation learning is then introduced to augment the distinction of text categories generated from different sources but having similar semantics. Simultaneously, linguistic characteristics are extracted from the text with Textagon and then linguistic representation is encoded by a Hierarchical Attention Network (HAN) to represent measurable lexical, syntactic, readability, and stylistic features. This is followed by the fusion of contextual semantic and linguistic representations to give the comprehensive representation for three class classification. To make the model more transparent, layer integrated gradients is used to surface the text evidence that is relevant to each prediction. The framework is further tested in other domains, namely in education and in news, to investigate the generalization ability of the framework beyond the scope of one textual context. In the proposed study, contextual Transformer representations, contrastive learning, linguistic feature analysis, feature fusion, explainability and cross domain evaluation are combined to offer a more powerful and interpretable method for separating human-written, AI-generated, and AI-rephrased English text.

1.1 Problem Statement

Reliable differentiation between human-written, AI-generated and AI-rephrased English text is becoming more challenging due to the growing use of generative AI[14]. Current detection methods may fail to yield reliable results in cases of AI-generated rephrasing, modification, or AI-generated text that looks like human writing[15]. A lot of methods are based primarily on hand-made language features or context representations and fail to effectively combine these two sources of information. In addition, the limited interpretability allows understanding of the reasons for assigning a specific text to a category, which is hard to achieve. Hence, it is necessary to have a powerful, discriminative, and explainable framework to classify AI-generated and AI-rephrased English text in various domains accurately.

1.2 Research Motivation

The growing sophistication of generative AI and AI-based rephrasing makes reliable text detection increasingly challenging. Current solutions are not always able to detect the nuances between human and AI-generated text and offer poor interpretability. Accordingly, this study is driven to build an accurate, explainable and robust framework for three-class English text detection in various domains.

1.3 Research Significance

This study is significant because it provides an integrated approach for distinguishing human-written, AI-generated, and AI-rephrased English text. The proposed framework integrates contextual semantic representations, contrastive learning, linguistic feature analysis and explainability to enhance the reliability of detection while enhancing transparency. In addition to this, its cross-domain assessment also confirms its usability in educational, academic and news fields.

1.4 Key Contributions

- Proposes a three-class framework for distinguishing human-written, AI-generated, and AI-rephrased English text.
- Uses DeBERTaV3 and multi-level contrastive learning to improve contextual representation and separation between text classes.
- Extracts Textagon-based linguistic features and uses HAN to capture important linguistic patterns in the text.
- Combines semantic and linguistic features and applies Layer Integrated Gradients (LIG) to provide understandable explanations for model predictions.
- Evaluates the framework across educational and news domains to examine its detection performance and cross-domain generalization.

1.5 Rest of the Section

The paper is organized as follows: Section 2 presents the literature review and research gap. Section 3 describes the proposed methodology. Section 4 discusses the experimental results and findings. Section 5 presents the conclusion and future research directions.

2. Literature Review

Elkhatat et al. [16]examined text detection for AI-generated texts in academic writing in the English language, specifically between AI generated content and human written responses, particularly from ChatGPT. Multiple AI detection tools, such as OpenAI, Writer, CrossPlag, Copyleaks, and GPTZero were examined with engineering texts on cooling towers. The experimental analysis focused on the effects on detection effectiveness as a function of sophistication of the generated content, and on the comparison of predictions with human-authored responses. The results indicated that detection capability varied and error rates in classification of human-written text occurred. The study also underscored the challenge of accurately detecting complex AI-generated text using traditional methods to detect plagiarism. The observations highlight the importance of developing strong detection systems with better discrimination between human and AI-generated English text and less false classification, which would motivate further development of more semantic and linguistic detection methods.

Weber-Wulff et al. [17]examined the reliability of AI-generated text detection tools in academic settings, focusing on their ability to distinguish ChatGPT-generated content from human-written text. This study assessed the accuracy and types of errors of an original document set and examined publicly available and commercial detection systems. It also explored the influence of machine translation and obfuscation of content on detection performance. The study identified significant challenges in current detection technologies, such as the inability to accurately identify AI-generated content and misidentify human-generated text as AI. The further down the page, the less they could be detected because of the obfuscation. The research underscores the importance of improved detection methods that can process AI-generated text in various forms and minimise classification mistakes in academic English-language contexts.

Schaaff et al. [18]examined how humans, AI, and AI rephrased text are detectable in the educational and news sectors. The analysis investigated various types of features such as perplexity, semantic, readability, error-based, document and text-vector features for the purpose of distinguishing naturally written material from AI-altered material. The researchers created a corpus of English and other European-language texts generated by AI and

human editors, which is called the Multilingual Human-AI-Generated Text Corpus (HAIG). To assess detection in more human-like AI generation and rephrasing scenarios, both basic and advanced prompting scenarios were taken into account. The results showed that feature types shared across various domains and languages could be used for detection and the proposed classifiers were superior to other detection systems. The results underscore the need to leverage multiple textual attributes to effectively identify AI-generated and AI-rephrased content.

Fishchuk and Braun [19]studied the robustness of neural text detectors to adversarial attacks for artificial English text. This study concentrated on the growing human-like nature of the generated content and the necessity of a reliable detection when users deliberately try to hide the use of AI. English texts were subjected to various adversarial modification strategies and several commercial and research-based neural detectors were evaluated. The analysis focused on how well these detectors could identify AI-generated text with intentional modifications.The analysis compared the performance of these detectors in detecting AI-generated text when it was intentionally modified. The results showed that there were differences in the robustness of detectors, some being more robust than others, depending on the attack method. The study also underscores the need for testing the adversarial performance of AI-text detection and building detection systems that are robust when faced with adversarial documents that are generated by AI or rephrased by AI and then manipulated to be presented as human documents.

Cooperman and Brandão [20]explored the ability of AI detection tools to identify AI and human-written scientific articles in foot and ankle surgery literature. The study examined a variety of publicly available detection systems based on AI-generated and human-written abstracts and also explored the impact of AI rewording on detection. The analysis showed there are significant differences between the detection tools, and it was not easy to accurately detect AI-generated content and prevent false positives. The rewording experiment also highlighted the significant impact that rewording an AI-generated abstract could have on their detectability. The study stresses the shortcomings of current detection systems regarding the detection of text-challenged AI-generated text and the importance of powerful detection systems that can detect sophisticated and rephrased AI-generated text in scientific and other English-language fields.

Lau and Zubiaga [21]explored paraphrasing effects on text detection by LLMs, especially regarding limitations of current detectors with paraphrased text. The study presented the Human & LLM Paraphrase Collection (HLPC) which included human-generated texts, human paraphrases, LLM-generated texts, and LLM-generated paraphrases. Experiments on classification compared existing text detectors trained on LLM output with watermark-based techniques, leveraging content from various LLM with both generated and paraphrased versions. The experiments analyzed the impact of paraphrases written by humans on detector performance and compared detector behavior for the original and modified texts. Results demonstrated that manually created paraphrases had a significant impact on the performance of the detectors, suggesting that paraphrasing can be a challenge for accurate AI-text detection. The study highlights the need for detection frameworks specifically designed to handle both AI-generated and paraphrased English text.

Z. Wang et al. [22]suggested BENATTEN as an approach to detect AI-generated text based on the differences in the attention paid to tokens by users vs. AI writers. The method is based on Benford's law to describe the distributions of attention and on differentiating probabilistic generation behaviour from human writing patterns. The study tested BENATTEN on general-domain and specialized-domain datasets, to assess its detection capability and cross-domain robustness. The proposed approach was tested extensively and compared to existing AI-text detection systems, and further explored in adversarial settings and in the presence of interfering AI systems. The results showed high performance in detection and high cross-domain/cross-platform transferability. The study shows the promise of attention-based language patterns in effective text detection using AI and encourages the construction of holistic semantic and linguistic feature-based patterns for text detection.

S. Wang et al.[23]proposed a few-shot black-box method for identifying LLM-generated text, and made suggestions for overcoming the drawbacks of existing zero-shot approaches and handcrafted feature-based approaches. The approach proposed by the authors recast text detection from a LLM as a cross domain classification problem by utilizing a small amount of labeled data from the source domain and a large number ofunlabeled data from the target domain. A language model was pre-trained and then a soft prompt-tuning model was trained in the source domain and iteratively finetuned to extract label information in the target domain. Voting, another method of combining predicted labels, was used to enhance the decision-making of detection. Experimental comparisons were made with the state of the art baselines in the literature under various data distributions. The results showed that cross-domain adaptation with prompt tuning significantly enhanced the detection accuracy, suggesting its promising potential in achieving effective cross-domain adaptation of LLM-generated text detection.

Sun and Lv[12] proposed a black-box, zero-shot approach to detect LLM-generated text based on the contrast between linguistic ordering of human and AI-generated text. The study creates a collection of medical paper introductions, consisting of human-written papers and papers created by LLM, and proposes a method for detecting using dependency-tree-based reordering of the texts. A similarity measure was used to compare the reordered text with the original text, and log-likelihood features were added as supporting evidence. These features were amalgamated to determine a threshold for discriminating between human and LLM generated content. The method was tested with generated text and LLM polished abstracts and benchmarked with current zero-shot detection methods. The results showed that it successfully detected the content and could identify the polished content as well. The study underscores the significance of structural linguistic features in achieving robust text detection with AI-generated text.

Qorich and El Ouazzani [24]proposed a deep learning framework for distinguishing human-written and AI-generated essays, motivated by limitations of conventional AI-content detection tools in academic settings. The authors fine-tuned GPT-2, RoBERTa, and BART on large data sets and optimized both the models and their hyperparameters with manual and automated optimization. The authors build a BART-CNN model using the optimal BART configuration and a convolutional classification head based on the comparative experiments. The framework was tested with data from various sources and explored further with cross domain validation to examine generalization. The results showed high classification performances and better than the baseline models. This study demonstrates the effectiveness of introducing pretrained language models with specialized classification layers for AI-generated text detection and points to the need for solid cross-domain evaluation.

The reviewed studies exhibit remarkable advances in the detection of AIGT with the use of detection tools, linguistic features, pretrained language models, adversarial analysis, and cross domain adaptation[18]. Most of the current methods, however, have been developed for binary classification tasks between human written and AI-generated texts, and have not been applied to identify AI-rephrased texts. Studies have also generally focused on either semantic representations or handcrafted linguistic features independently, with little research investigating their combined effect on detection. Moreover, strong explanations for the reasons behind the classification of a text as human-written, AI-generated, or AI-rephrased are still scarce. A further study of the cross domain generalization of the educational and news English is needed. Thus, the proposed research is the development of a CTF model that combines the DeBERTaV3 model with multi-level contrastive learning, linguistic features extracted from the text (textagon), HAN, semantic linguistics fusion, and LIG for explainable three-class detection and cross-domain testing of English text.

3. Proposed Methodology for Explainable Detection of AI-Generated and Rephrased English Text

The suggested approach offers a systematic mechanism to find human written, AI generated, and AI rephrased English text. The framework combines multi-level contrastive learning with contextual representation using DeBERTaV3 to enhance the discrimination among the three text categories. Linguistic features are also extracted and encoded from the written text based on Textagon to represent explicit writing characteristics in a Hierarchical Attention Network. The classifiers are trained in a three-class classification task, where semantic and linguistic representations are fused; the Layer Integrated Gradients evidence gives the interpretable evidence for model prediction. The framework is tested with within domain and cross domain experiments, to verify performance in detection and generalization abilities.

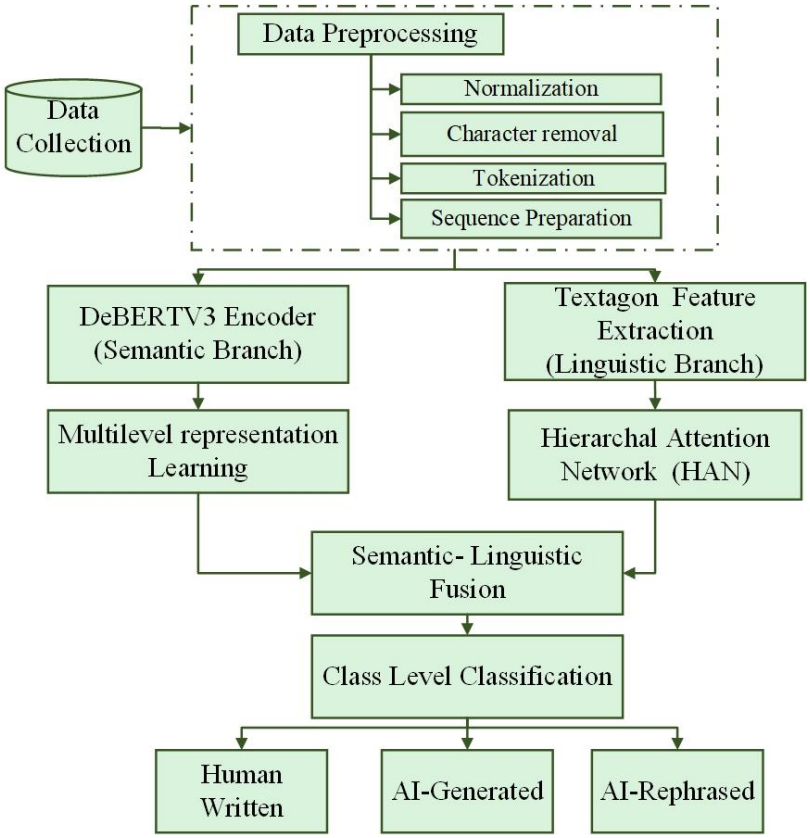


Fig. 1. Proposed Methodology Workflow

The workflow in Fig.1 starts with data collection and is followed by the preprocessing stage which includes normalization, character removal, tokenization, and sequence preparation. The preprocessed text is then sent to two parallel branches: a semantic branch in which the text is encoded by the DeBERTaV3 model, followed by multi-level contrastive representation learning; and a linguistic branch in which the features are extracted by Textagon and then encoded by the Hierarchical Attention Network (HAN). Semantic–linguistic fusion is performed on the outputs of the two branches, and the classification stage at the class level classifies each text into one of three classes: Human-Written, AI-Generated, or AI-Rephrased.

3.1 Data Collection

The study uses the publicly available Human-AI-Generated Text Corpus as the primary dataset for detecting AI-generated and AI-rephrased English text[25]. A corpus of human-written, AI-generated and AI-rephrased texts in the educational and news domains, available in English, French, German and Spanish. The data used for the development and evaluation of the proposed detection framework are organized into 3 target categories: Human-written, AI-generated and AI-rephrased.

3.2 Data Preprocessing

Preprocessing filters away irrelevant textual differences while retaining semantic, syntactic, lexical and stylistic information to aid the identification of AI-generated and AI-rephrased writing. The preprocessing stage involves normalization of text, the removal of unwanted characters and tokenization with sequence preparation. This representation is then passed to the DeBERTaV3 encoder and linguistic feature extraction module.

3.2.1 Text Normalization

Text normalization standardizes the collected English documents before model processing. Capitalisation, spacing and text formatting vary less and the meaning of words and sentence structure is preserved. This operation will make sure that the identical textual patterns are consistently represented in the three target classes. This normalized representation is used for further cleaning and tokenization. The Text normalization is given in eqn. (1).

$$X_{norm} = \text{Normalize}(X) \quad (1)$$

Where Xrepresents the original text and X_{norm}represents the normalized text.Normalize represents the standardization operation applied to the original textual input.The output is a standardized representation of the original English text.

3.2.2 Unwanted Character Removal

The cleaned texts are normalized, with irrelevant symbols and formatting artefacts excluded that doesn't play a role in authorship discrimination. Redundant formatting and characters are removed and meaningful text and required punctuation are preserved. All three classes are cleaned in the same way to ensure consistency. This results in a clean input stream of text for further language processing and Transformer-based processing. The Mathematical expression is given in eqn. (2).

$$X_{clean} = X_{norm} \setminus C_{noise} \quad (2)$$

Where $X_{norm}$is the normalized text, C_{noise}represents unwanted characters or formatting artifacts, and X_{clean}denotes the cleaned text.The output is clean text containing the information required for detection.

3.2.3 Tokenization and Sequence Preparation

The cleaned English text is then tokenized into subwords by the DeBERTaV3 tokenizer.Subword tokens are generated by the DeBERTaV3 tokenizer from the cleaned English text. Documents are encoded as a sequence of tokens, special tokens are used for the structure of the input sequence. Sequences are truncated or padded to the maximum input length selected. This results in a sequence of tokens that are passed to the contextual representation module and it is given in eqn. (3).

$$T = \text{Tokenize}(X_{clean}) = [t_1, t_2, \dots, t_n] \quad (3)$$

Where X_{clean}represents the cleaned text, $T$is the token sequence, $t_i$represents the i-th token, and $n$represents the sequence length. The processed text is now represented in a form suitable for Transformer encoding.

3.3 DeBERTaV3 Contextual Representation

DeBERTaV3 models the relationship of tokens in context and generates the deep semantic representation of the input documents. These representations allow for the detection of small differences between texts written by humans, AI-generated texts, and AI-rephrased texts. The document representation obtained is sent to the contrastive learning module and subsequently merged with the explicit linguistic features and it is given in eqn. (4).

$$H = \text{DeBERTaV3}(T) \quad (4)$$

Where Trepresents the prepared token sequence and Hrepresents the contextual token-level representations generated by DeBERTaV3.The Transformer converts the token sequence into contextual semantic representations and it is given in eqn. (5).

$$h_{cls} = H_{[CLS]} \quad (5)$$

Where $H_{[CLS]}$denotes the representation associated with the classification token and h_{cls}represents the document-level contextual representation.This representation summarizes the semantic characteristics of the input document.The contextual representations of all valid tokens are averaged to obtain a single representation that captures the overall semantic information of the English text and it is given in eqn. (6).

$$h_{doc} = \frac{1}{n}\sum_{i=1}^{n} H_i \quad (6)$$

Where $H_i$represents the contextual representation of the i-th token generated by DeBERTaV3, $n$denotes the number of valid tokens in the input sequence, and h_{doc}represents the overall document-level contextual representation.

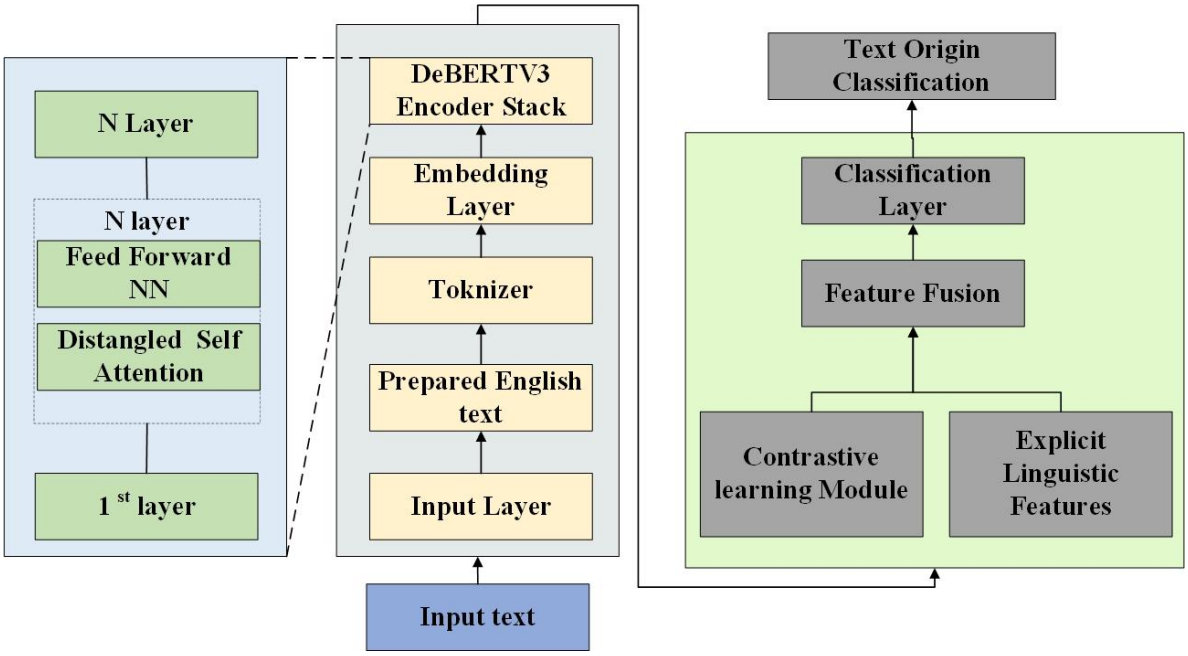


Fig 2. DeBERTaV3 Architecture

Fig.2 shows the input text passing through the input layer, the prepared English text stage, the tokenizer, the embedding layer, and the DeBERTaV3 encoder stack. The expanded view on the left illustrates the encoder stack consisting of N stacked layers, each consisting of a disentangled self-attention block and a feed-forward neural network. The output of the encoder is fed to the contrastive learning module which is then fused with the explicit linguistic features via feature fusion, and the classification layer generates the final text origin classification.

3.4 Multi-Level Contrastive Representation Learning

A multi-level contrastive representation learning strategy is used to enhance the discrimination between the three text categories. The contrastive component pushes similar representations towards being more similar, and different representations away from one another. This is especially crucial when differentiating AI-generated content from AI-rephrased content, as both types may share certain semantic and stylistic features. The learned representation is thus a more discriminative feature space for subsequent classification and it is given in eqn. (7).

$$z_i = g\left(h_{cls}^i\right) \tag{7}$$

Where h_{cls}^irepresents the DeBERTaV3 representation of sample i, $g(\cdot)$denotes the projection function, and $z_i$represents the projected contrastive representation.The projection creates a representation suitable for comparing samples and it is given in eqn. (8).

$$s_{ij} = \frac{z_i^T z_j}{\|z_i\|\|z_j\|} \tag{8}$$

Where $z_i$and $z_j$represent two sample embeddings and s_{ij}represents their cosine similarity.Higher similarity indicates greater representational closeness between two samples and it is given in eqn. (9).

$$\mathscr{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_{ip}/\tau)}{\sum_{j \neq i} \exp(s_{ij}/\tau)} \tag{9}$$

Where Ndenotes the number of samples, $p$identifies the positive sample associated with anchor i, s_{ip}denotes positive-pair similarity, s_{ij}denotes similarity with other samples, and τrepresents the temperature parameter.The contrastive loss improves separation among the three writing categories.

3.5 Textagon-Based Linguistic Feature Extraction

The framework is able to extract explicit linguistic features and supplement the contextual features obtained from DeBERTaV3. Lexical, syntactic, readability and stylistic features for the cleaned English texts are extracted using the textagon based linguistic feature extraction. The following features give explicit examples of how the text was written that can be a difference between human written, AI written and AI rephrased texts. The linguistic representation mined from the text is then fed into a Hierarchical Attention Network (HAN) to learn the relative weight of these characteristics and it is given in eqn. (10).

$$L = \text{Textagon}(X_{clean}) \tag{10}$$

Where X_{clean}denotes the cleaned text and Lrepresents the extracted linguistic feature vector.The linguistic feature vector captures measurable writing characteristics and it is given in eqn. (11).

$$L' = \text{HAN}(L) \tag{11}$$

Where Lrepresents the extracted linguistic features and L'represents the attention-based linguistic representation generated by HAN.The HAN representation emphasizes the most informative linguistic characteristics.

3.6 Semantic–Linguistic Feature Fusion

The contextual representation from DeBERTaV3 and the linguistic representation from HAN are fused together to take advantage of complementary information. While the linguistic branch offers explicit lexical, syntactic, readability and stylistic information, DeBERTaV3 offers detailed contextual and semantic information. The union of these representations allows the classifier to use a learned semantic pattern as well as measurable linguistic trait. The fused representation is used in the final three-class classification and it is given in eqn. (12).

$$F = [h_{cls}; L'] \tag{12}$$

Where h_{cls}represents the DeBERTaV3 contextual representation, L'represents the HAN-based linguistic representation, and Fdenotes the fused feature vector.The fusion produces a unified representation containing semantic and linguistic evidence.

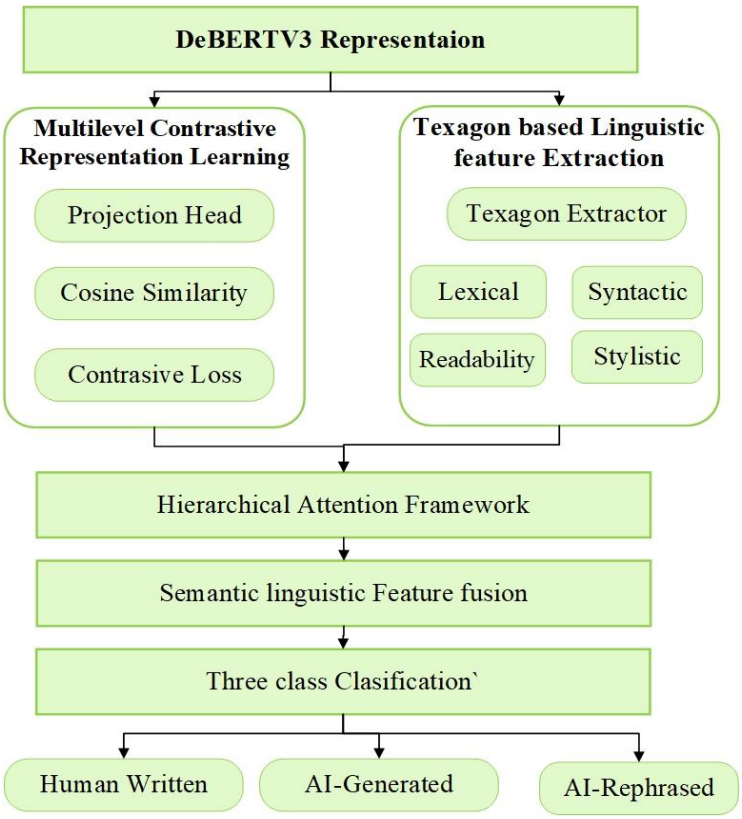


Fig 3. Linguistic Feature Extraction and Classification

The linguistic feature extraction and classification stages after the DeBERTaV3 representation is depicted in Fig.3. The representation is learned by two modules: multilevel contrastive representation learning (a projection head, cosine similarity, and contrastive loss) and Textagon-based linguistic feature extraction (lexical, syntactic, readability, and stylistic features generated by the Textagon extractor). The results of the hierarchical attention framework are fed into a semantic–linguistic representation that is categorized as Human-Written, AI-Generated, or AI-Rephrased text.

3.7 Three-Class Classification

The fused representation is passed to a fully connected classification head that classifies the three target categories. The classifier classifies each input as Human-written, AI-generated, or AI-rephrased text. The classification scores are passed to a Softmax function to get probabilities for the three classes. The class with the greatest likelihood is the last class that is chosen to be the ultimate prediction and it is given in eqn. (13).

$$o = WF + b \tag{13}$$

Where Frepresents the fused feature vector, Wdenotes the learnable classification weight matrix, $b$represents the bias vector, and $o$denotes the output logits.The classification layer converts the fused features into class-specific scores and it is given in eqn. (14).

$$P(y = k \mid F) = \frac{\exp(o_k)}{\sum_{c=1}^{3} \exp(o_c)} \tag{14}$$

Where $o_k$represents the score of class k, $c$indexes the three classes, and $P(y = k \mid F)$represents the probability of class k.The highest probability determines the predicted text category.

3.8 Explainable Prediction Using Layer Integrated Gradients

The Layer Integrated Gradient (LIG) is added to provide an understanding of the prediction produced by the proposed classifier. LIG uses gradients along a path from a baseline representation to the actual input to identify the contribution of the input representation to the selected class prediction. This offers feature-level attribution of the model decision. The values of the attribute relationships are used to identify linguistic and contextual evidence that affects the classification and it is given in eqn. (15).

$$IG_i(x) = \left(x_i - x_i^{'}\right) \int_0^1 \frac{\partial S_c\left(x^{'} + \alpha(x - x^{'})\right)}{\partial x_i} \, d\alpha \tag{15}$$

Where $x_i$represents the input feature, $x_i^{'}$represents the baseline feature, $S_c$represents the model score for class c, and αrepresents the integration coefficient.The attribution identifies how each input feature influences the selected class and it is given in eqn. (16).

$$A_i = \mid IG_i(x) \mid \tag{16}$$

Where $IG_i(x)$represents the integrated-gradient attribution and $A_i$denotes the absolute attribution magnitude.The highest attribution values identify the most influential evidence for the prediction.

3.9 Cross-Domain Evaluation

The cross-domain evaluation is conducted to identify the strength of the detector when faced with domain variation. The complete framework is trained by jointly optimizing the classification objective and the contrastive representation objective. The classification part enables better discrimination between the three target classes, and the contrastive part enhances the separation of their learned representations. The proposed framework is tested against the baseline models: PPL based Random Forest, XGBoost and Neural Network models.

The proposed approach is based on Human-AI-Generated Text Corpus, aiming to explore the analysis of human written text, AI-generated text, and AI-rephrased text in English. The English samples are created by normalization, cleaning and DeBERTaV3 compatible tokenization. DeBERTaV3 is able to produce contextual representations and separation is enhanced between the three target categories with the help of multi-level contrastive learning. The explicit linguistic features are extracted by Textagon and encoded by a Hierarchical Attention Network. A fully connected Softmax classifier fuses and classes the contextual and the linguistic

representations. Layer Integrated Gradients gives interpretable evidence for the class predicted. Finally, within and across domain experiments are run against the baseline models, based on PPL, to evaluate classification performance and generalization.

Algorithm: Proposed Contrastive Transformer-Based Detection Framework

Input: English text
Output: Predicted class, probability, and explanation

```
BEGIN

Read English text from the dataset

IF text is available THEN
   Normalize and clean the text
   Tokenize the text
   Generate contextual representation using DeBERTaV3
   Apply multi-level contrastive learning
   Extract Textagon-based linguistic features
   Encode linguistic features using HAN
   Fuse DeBERTaV3 and HAN representations
   Predict the text class using Softmax
   IF predicted class is Human-written THEN
      Assign Human-written label
   ELSE IF predicted class is AI-generated THEN
      Assign AI-generated label
   ELSE
      Assign AI-rephrased label
   END IF
   Apply Layer Integrated Gradients
   Generate explanation for the prediction
ELSE
   Return "No valid text available"
END IF
Evaluate the prediction using the selected performance metrics
END
```

The algorithm:1 describes the Steps involved in the selected corpus. These texts are then preprocessed and converted to contextual representations using DeBERTaV3. The results show that multi-level contrastive learning helps to distinguish the three writing categories, and Textagon and HAN offer complementary linguistic representations. In this way these semantic and linguistic representations are fused and classified by the Softmax classification head. Lastly, Layer Integrated Gradients predicts and cross域 evaluates the robustness of the proposed framework.

4. Results and Discussion

The experimental results of the proposed Contrastive Transformer Framework for detecting human-written, AI-generated and AI-rephrased English text are presented in this section. It first explains the composition of the dataset and the results of the pre-processing, and then shows the contextual representation obtained by DeBERTaV3 and the effect of multi-level contrastive learning. The extracted linguistic features from Textagon and encoded by the Hierarchical Attention Network are subsequently analyzed, along with semantic-linguistic fusion. Lastly, the section presents three-class classification, explainability, cross-domain evaluation, the ablation study, and comparing with baseline methods.

4.1 Dataset Composition and Class Distribution

Table 1. Class-Wise Dataset Distribution

Text Class	Educational	News	Total
Human-Written	2,000	2,000	4,000
AI-Generated	2,000	2,000	4,000
AI-Rephrased	2,000	2,000	4,000
Total	6,000	6,000	12,000

The class-wise distribution of the data set for both the educational and the news domain is shown in Table 1. There are the same number of human-written, AI-generated, and AI-rephrased texts in each domain, so that there is a balanced class representation. This dataset includes 12,000 texts, 6,000 texts in each domain and 4,000 texts per text class.

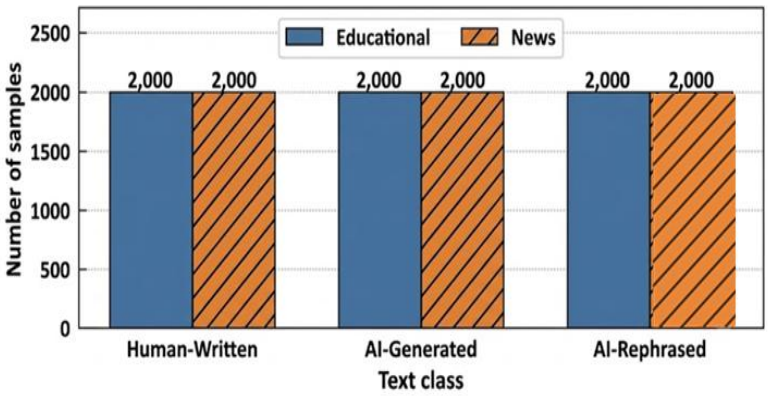


Fig 4. Class Distribution

The distribution of the each text class in educational and news domain is presented in Fig.1. There are 2000 samples in each domain for each of the three classes (Human-Written, AI-Generated, and AI-Rephrased). This is a balanced distribution so that there is no class bias in training and testing.

4.2 Text Preprocessing Results

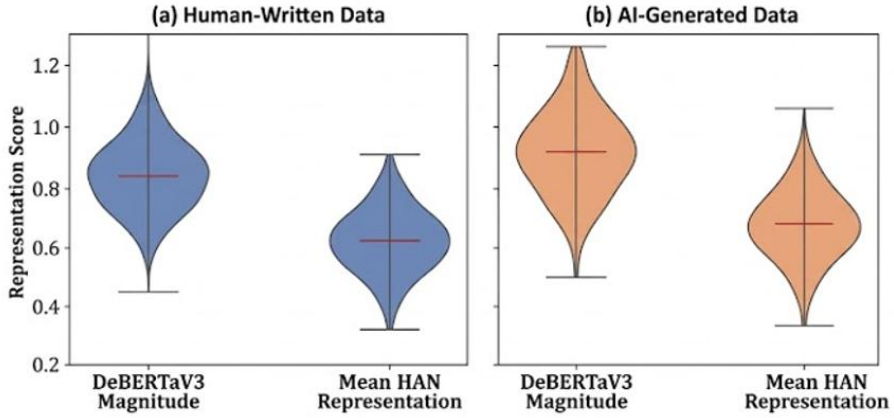


Fig 5. Token Length Distribution Across Text Classes

Fig.5 shows violin plots of the DeBERTaV3 magnitude of the data written by humans and the mean HAN representation scores of the data that were generated by AI for both (a) human-written data and (b) AI-generated data. The median DeBERTaV3 magnitude is around 0.84 for humans and 0.92 for AI-generated content, whereas the median mean HAN representation is 0.62 and 0.68 respectively. The larger scores for AI generated text in both representations suggest that they contain class dependent information.

4.3 DeBERTaV3 Contextual Representation

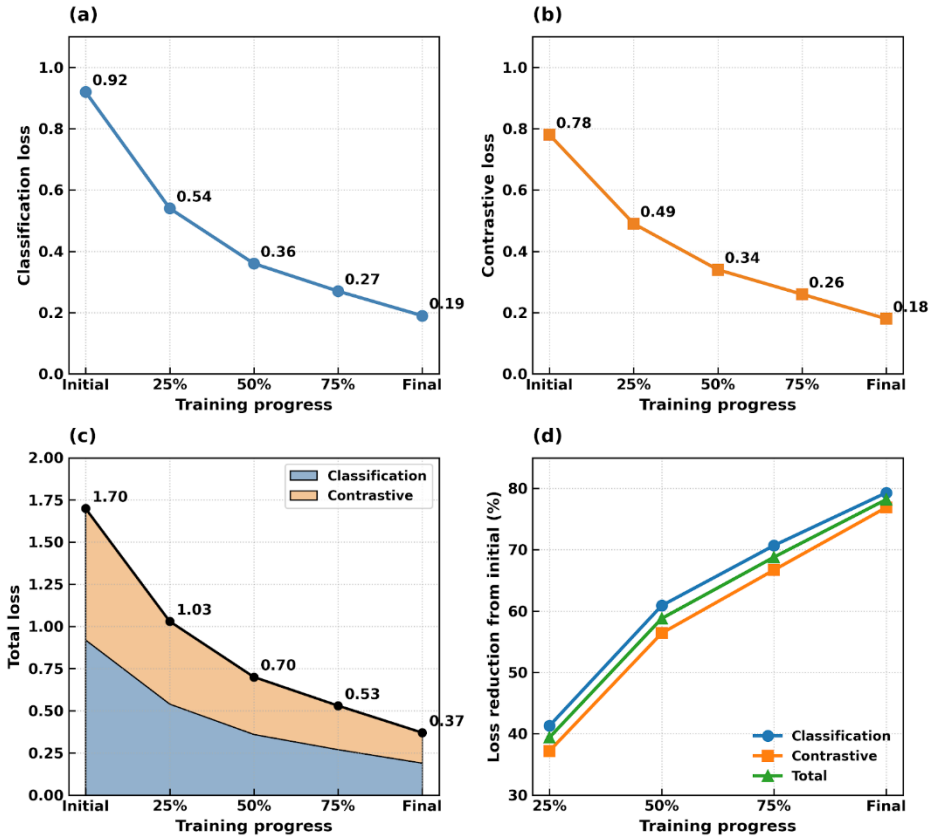


Fig 6. DeBERTaV3 Contextual Representation Distribution

Fig.6 illustrates training behavior of the framework as a function of training progress. The classification loss decreases from 0.92 to 0.19 as seen in (a) and the contrastive loss decreases from 0.78 to 0.18 as seen in (b).

Subfigure (c) reveals that the loss value is being reduced from 1.70 to 0.37, while subfigure (d) demonstrates that the loss value is being reduced to about 79% (classification), 77% (contrastive) and 78% (total loss) at the final stage.

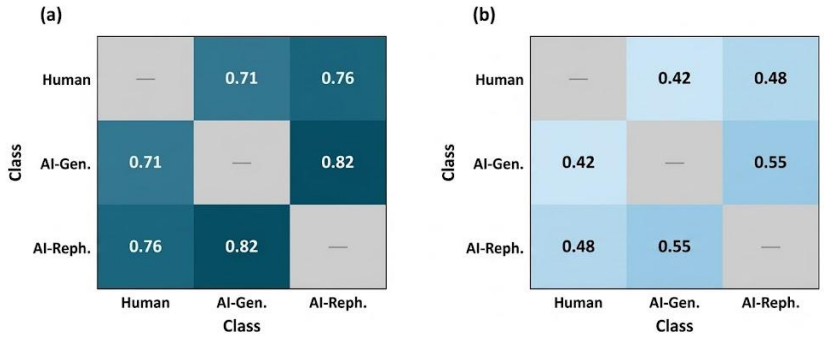


Fig 7. Class-Wise Semantic Representation Separation

Fig. 7 shows the inter-class representation similarity (a) before and (b) after multi-level contrastive learning. Before contrastive learning, the similarity is 0.71 for Human vs AI-Generated, 0.76 for Human vs AI-Rephrased, and 0.82 for AI-Generated vs AI-Rephrased. Following contrastive learning, these values drop to 0.42, 0.48, and 0.55, respectively, thereby showing a better separation among the three classes.

4.4 Multi-Level Contrastive Representation Learning

Table 2. Inter-Class Representation Similarity Before and After Contrastive Learning

Class Pair	Before Contrastive Learning	After Contrastive Learning
Human vs AI-Generated	0.71	0.42
Human vs AI-Rephrased	0.76	0.48
AI-Generated vs AI-Rephrased	0.82	0.55

As shown in Table 2, contrastive learning significantly decreases inter-class representation similarity in all class pairs. The similarities between AI-generated and AI-rephrased texts drop from 0.82 to 0.55 and Human vs AI-Rephrased and Human vs AI-Generated similarities also decrease. This means that the separation between the three text classes is better.

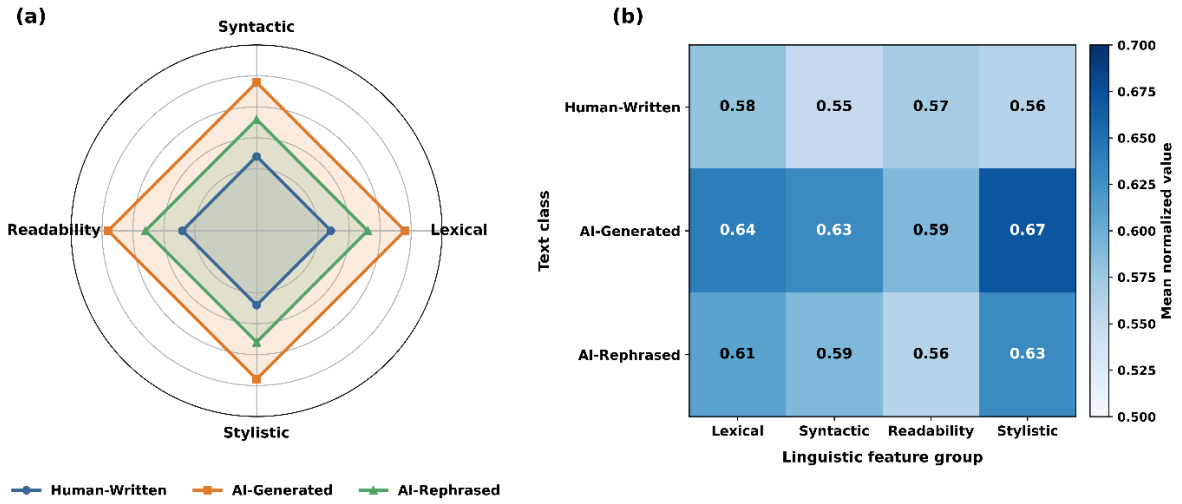


Fig 8. Contrastive and Classification Loss Convergence

The linguistic feature patterns for the three classes of text are compared in Fig. 8. Subfigure (a) shows a radar chart for the lexical, syntactic, readability, and stylistic feature groups, with AI-generated text having the highest values, AI-rephrased text in between, and human-written text having the lowest values. The mean normalized values are presented as a heatmap in subfigure (b): Human written text ranges from 0.55 to 0.58, AI-generated text ranges from 0.59 to 0.67 with the strongest value for stylistic features, and AI-rephrased text ranges from 0.56 to 0.63.

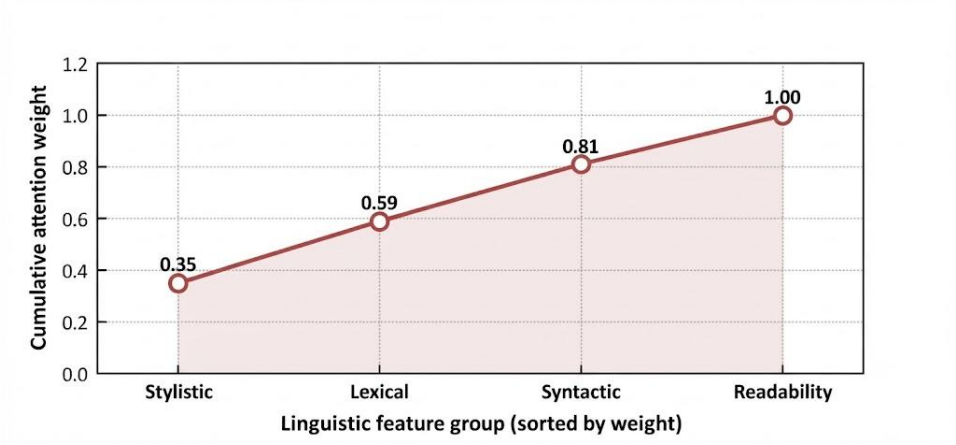


Fig 9. Representation Separation After Contrastive Learning

The HAN's cumulative attention weight for the linguistic feature groups ordered by weight is presented in Fig. 9. The cumulative weight increases from 0.35 for stylistic features to 0.59 for lexical features, 0.81 for syntactic features, and 1.00 for readability features. This indicates that the attention is given mostly to stylistic features, then to lexical and syntactic features, and then to readability features.

4.5 Textagon-Based Linguistic Feature Extraction

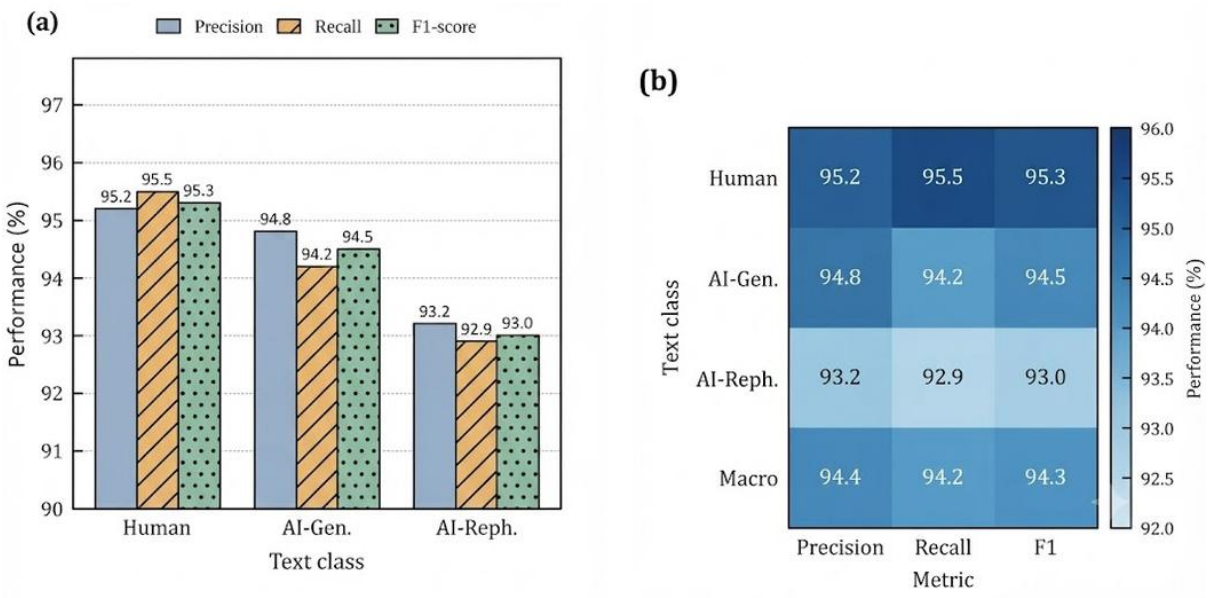


Fig 10. Class-Wise Linguistic Feature Patterns

Fig. 10 presents the class-wise precision, recall, and F1-score as (a) grouped bars and (b) a heatmap that also includes the macro average. Human-written text achieves the highest scores (95.2% precision, 95.5% recall, and 95.3% F1-score), followed by AI-generated text (94.8%, 94.2%, and 94.5%), while AI-rephrased text is the most difficult class (93.2%, 92.9%, and 93.0%). The macro averages are 94.4% precision, 94.2% recall, and 94.3% F1-score.

4.6 Semantic–Linguistic Feature Fusion

Table 3. Performance of Semantic, Linguistic, and Fused Representations

Representation	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DeBERTaV3 Representation	91.8	91.6	91.4	91.5
HAN Linguistic Representation	84.7	84.3	84.1	84.2
Fused Representation	94.6	94.4	94.2	94.3

Table3 shows that the fused representation achieves the best performance in all the evaluation metrics. The semantic representation of DeBERTaV3 is better than that of HAN-based linguistic features, and fusion of the two representations enhances the classification ability. With the fused representation, the accuracy is 94.6% and the F1-score is 94.3%, demonstrating the advantage of using semantic and linguistic information together.

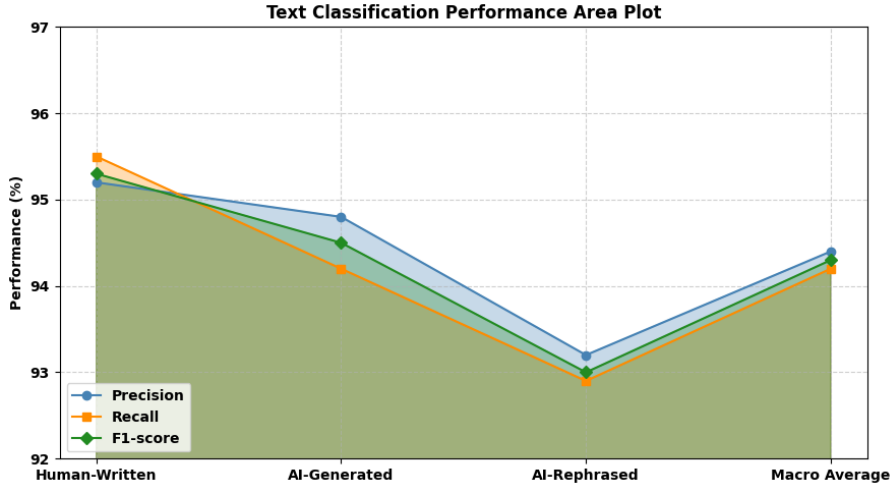


Fig 11. Class-Wise Performance of the Fused Representation

The precision, recall and F1 score of the fused representation are presented in Fig 11, per class and macro average. The best performance is achieved by human written text with a recall of 95.5%, and the lowest score is obtained by AI rephrased text with a score of about 93%. The macro-average values are not far from 94.3%, which shows that the system performs similarly across the three classes.

4.7 Three-Class Classification, Explainability and Cross-Domain Evaluation

Table 4. Cross-Domain Classification Performance

Training Domain	Testing Domain	Accuracy (%)	F1-Score (%)
Educational	Educational	95.1	94.9
News	News	95.0	94.8
Educational	News	90.8	90.3
News	Educational	91.4	91.0

The cross domain classification accuracy is shown in Table 4 for the educational and news domain. Within domain evaluation produces higher performance and cross domain testing produces a moderate drop. The framework obtains high accuracy on news (91.4%) and high F1-score on educational text (91.0%) for training on news and testing on educational text, showing that it has good generalizability.

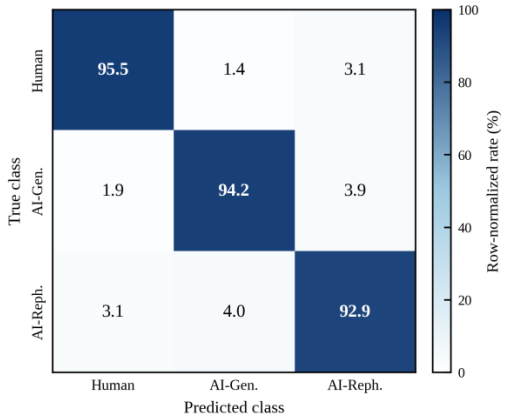


Fig 12. Confusion Matrix of the Three-Class Classification

The result is a row-normalised confusion matrix for the three-class classification as is depicted in Fig. 12. This proposed framework achieves a 95.5% accuracy with human-written texts, 94.2% with AI-generated texts, and 92.9% with AI-rephrased texts. The majority of the remaining errors are between AI-generated and AI-rephrased texts (3.9% and 4.0%) supporting the fact that these two classes are more similar.

4.8 Ablation Study

Table 5. Ablation Study

Framework Configuration	**Accuracy (%)**	**F1-Score (%)**
DeBERTaV3 Only	91.8	91.5
Without Contrastive Learning	92.7	92.4
Without Textagon + HAN	91.9	91.6
Without Semantic–Linguistic Fusion	92.1	91.8
Complete Framework	94.6	94.3

The contribution of individual framework components was found by ablation analysis as shown in Table 5. The whole framework obtains the highest accuracy, 94.6% and F1-score 94.3%. Removing contrastive learning (Textagon + HAN) or semantic–linguistic fusion (Textagon) decreases performance, thus establishing that each of them is necessary for the efficacy of the proposed framework.

4.9 Comparison with Baseline Methods

Table 6. Comparison with Existing Baselines

Method	**Dataset**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
PPL-Based Random Forest	Human-AI-Generated Text Corpus	82.4	81.9	81.5	81.7
XGBoost	Human-AI-Generated Text Corpus	85.7	85.2	84.9	85.0
Neural Network	Human-AI-Generated Text Corpus	88.6	88.1	87.8	87.9
Proposed Framework	Human-AI-Generated Text Corpus	94.6	94.4	94.2	94.3

Table 6 compares the proposed framework with the Random Forest, XGBoost, and Neural Network baselines. The proposed approach has the best performance for all the metrics with 94.6% accuracy and 94.3% F1-score. The results show the efficacy of using a combination of contextual, contrastive, and linguistic representation to detect three classes of text.

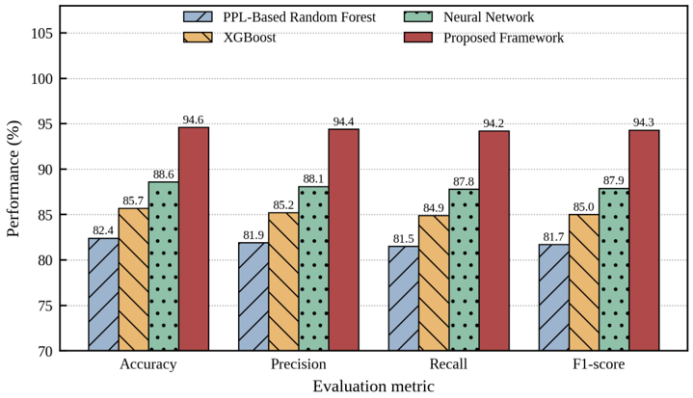


Fig 13. Performance Comparison with Baseline Methods

Fig. 13 presents the accuracy, precision, recall, and F1 score of the proposed framework against the three baseline models: PPL based Random Forest, XGBoost, and Neural Network. The proposed framework achieves 94.6% accuracy, 94.4% precision, 94.2% recall, and 94.3% F1-score, which is higher than the Neural Network (88.6%, 88.1%, 87.8%, and 87.9%), XGBoost (85.7%, 85.2%, 84.9%, and 85.0%), and PPL-based Random Forest (82.4%, 81.9%, 81.5%, and 81.7%) in every metric.

4.10 Discussion

The proposed framework performs well and consistently in separating human-written, AI-generated, and AI-rephrased English text. When combining DeBERTaV3 semantic features with HAN-encoded linguistic features, there is clearly an improvement over using either representation alone, proving that they complement each other. Multi level contrastive learning minimizes the similarity between classes, particularly between AI generated and AI rephrased text, which is the most similar. There are consistently reduced training losses with stable convergence. The most attention in the HAN is given to stylistic features, followed by lexical, syntactic, and readability features. The text classes that can be detected most accurately are human written texts, while the most difficult to detect are when the texts are rephrased by AI, and the largest number of errors falls between the two classes of AI rephrased text. The drop of cross domain testing is small when compared to within domain testing, suggesting good cross domain generalization between educational and news texts. The ablation study reveals that all the components are contributing and the framework is better than all the baseline methods on all metrics.

5. Conclusion and Future Works

This study proposed a Contrastive Transformer Framework for explainable detection of human-written, AI-generated, and AI-rephrased English text. DeBERTaV3 captured contextual semantics, while multi-level contrastive learning increased the separation between classes. Textagon features encoded by a Hierarchical Attention Network supplied explicit lexical, syntactic, readability, and stylistic evidence, and Layer Integrated Gradients explained each prediction. On the Human-AI-Generated Text Corpus, the fused representation outperformed the semantic-only and linguistic-only representations, reaching 94.6% accuracy and 94.3% F1-score. The ablation study confirmed that every component contributes, and the framework surpassed the PPL-based Random Forest, XGBoost, and Neural Network baselines on all metrics. Cross-domain testing between educational and news texts showed only a modest performance drop, indicating good generalization. AI-rephrased text remained the most difficult class, with most errors occurring between the two AI-related categories. Overall, the framework offers an accurate, interpretable, and robust solution for distinguishing AI-generated and AI-rephrased writing in educational and news settings, supporting academic integrity and information trust.

Future work will extend the framework to multilingual and cross-lingual detection using the French, German, and Spanish texts in the corpus. Additional LLMs and domains will be tested to improve generalization. Robustness against paraphrasing and human editing will be evaluated, and sentence-level detection will be

explored for mixed documents. Model compression will also support real-time use in classrooms and newsrooms.

Data Availability Statement

The dataset used in this study was obtained from Github and is publicly available at: https://github.com/iu-ai-research/human-AI-generatedTextCorpus

References

[1] J. Golec and T. Hachaj, "Ten Natural Language Processing Tasks with Generative Artificial Intelligence," *Applied Sciences*, vol. 15, no. 16, p. 9057, Aug. 2025, doi: 10.3390/app15169057.

[2] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023, doi: 10.1109/ACCESS.2023.3294090.

[3] J. Q. J. Liu *et al.*, "The great detectives: humans versus AI detectors in catching large language model-generated medical writing," *Int J Educ Integr*, vol. 20, no. 1, p. 8, May 2024, doi: 10.1007/s40979-024-00155-6.

[4] Y. Zhao, A. Borelli, F. Martinez, H. Xue, and G. M. Weiss, "Admissions in the age of AI: detecting AI-generated application materials in higher education," *Sci Rep*, vol. 14, no. 1, p. 26411, Nov. 2024, doi: 10.1038/s41598-024-77847-z.

[5] Y. Hui, "Using generative adversarial network to improve the accuracy of detecting AI-generated tweets," *Sci Rep*, vol. 14, no. 1, p. 29322, Nov. 2024, doi: 10.1038/s41598-024-78601-1.

[6] R. Ardeshirifar, "Comparing hand-crafted and deep learning approaches for detecting AI-generated text: performance, generalization, and linguistic insights," *AI Ethics*, vol. 5, no. 4, pp. 4197–4209, Aug. 2025, doi: 10.1007/s43681-025-00699-4.

[7] R. Alharthi *et al.*, "Responsible Detection and Mitigation of AI-Generated Text Using Hybrid Neural Networks and Feature Fusion: Toward Trustworthy Content Management in the Era of Large Language Models," *Int J Comput Intell Syst*, vol. 18, no. 1, p. 274, Nov. 2025, doi: 10.1007/s44196-025-01025-w.

[8] J. Milička, A. Marklová, O. Drobil, and E. Pospíšilová, "Learning to detect AI texts and learning the limits," *PLoS One*, vol. 20, no. 10, p. e0333007, Oct. 2025, doi: 10.1371/journal.pone.0333007.

[9] A. Cheng *et al.*, "Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content," *Adv Simul*, vol. 10, no. 1, p. 66, Nov. 2025, doi: 10.1186/s41077-025-00396-6.

[10] I. Muneer, A. Shehzadi, M. A. Ashraf, and R. M. A. Nawab, "Has machine paraphrasing skills approached humans? Detecting automatically and manually generated paraphrased cases," *Big Data Research*, vol. 39, p. 100507, Feb. 2025, doi: 10.1016/j.bdr.2025.100507.

[11] A. Kartelj, M. Mladenović, and S. Vujičić Stanković, "Comparison of algorithms for the recognition of ChatGPT paraphrased texts," *J Big Data*, vol. 12, no. 1, p. 28, Feb. 2025, doi: 10.1186/s40537-025-01082-0.

[12] J. Sun and Z. Lv, "Zero-shot detection of LLM-generated text via text reorder," *Neurocomputing*, vol. 631, p. 129829, May 2025, doi: 10.1016/j.neucom.2025.129829.

[13] X. Liu, S. Wang, Y. Li, and K. Li, "Robust AI generated text detection through multi-grained latent feature denoising and contrastive representation learning," *Intelligent Data Analysis: An International Journal*, p. 1088467X261441866, Apr. 2026, doi: 10.1177/1088467X261441866.

[14] G. A. Adam *et al.*, "GPTZero: Robust Detection of LLM-Generated Texts," 2026, *arXiv*. doi: 10.48550/ARXIV.2602.13042.

[15] Z. Ahmad, M. Torres-Ruiz, A. Mahmood, R. Quintero, I. Ameer, and N. Bölücü, "Human or Machine? A Survey on Machine-Generated Text Detection," *IEEE Access*, vol. 14, pp. 34113–34136, 2026, doi: 10.1109/ACCESS.2026.3666781.

[16] A. M. Elkhatat, K. Elsaid, and S. Almeer, “Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text,” *Int J Educ Integr*, vol. 19, no. 1, p. 17, Sep. 2023, doi: 10.1007/s40979-023-00140-5.

[17] D. Weber-Wulff *et al.*, “Testing of detection tools for AI-generated text,” *Int J Educ Integr*, vol. 19, no. 1, p. 26, Dec. 2023, doi: 10.1007/s40979-023-00146-z.

[18] K. Schaaff, T. Schlippe, and L. Mindner, “Classification of human- and AI-generated texts for different languages and domains,” *Int J Speech Technol*, vol. 27, no. 4, pp. 935–956, Dec. 2024, doi: 10.1007/s10772-024-10143-3.

[19] V. Fishchuk and D. Braun, “Robustness of generative AI detection: adversarial attacks on black-box neural text detectors,” *Int J Speech Technol*, vol. 27, no. 4, pp. 861–874, Dec. 2024, doi: 10.1007/s10772-024-10144-2.

[20] S. R. Cooperman and R. A. Brandão, “AI tools vs AI text: Detecting AI-generated writing in foot and ankle surgery,” *Foot & Ankle Surgery: Techniques, Reports & Cases*, vol. 4, no. 1, p. 100367, 2024, doi: 10.1016/j.fastrc.2024.100367.

[21] H. T. Lau and A. Zubiaga, “Understanding the effects of human-written paraphrases in LLM-generated text detection,” *Natural Language Processing Journal*, vol. 11, p. 100151, Jun. 2025, doi: 10.1016/j.nlp.2025.100151.

[22] Z. Wang, G. Xu, and M. Ren, “Can attention detect AI-generated text? A novel Benford’s law-based approach,” *Information Processing & Management*, vol. 62, no. 4, p. 104139, Jul. 2025, doi: 10.1016/j.ipm.2025.104139.

[23] S. Wang, Y. Zhu, and P. Li, “A smaller model can be better: Domain adaptation for LLM-generated text detection via soft prompt-tuning,” *Information Processing & Management*, vol. 63, no. 3, p. 104566, Apr. 2026, doi: 10.1016/j.ipm.2025.104566.

[24] M. Qorich and R. El Ouazzani, “Detection of artificial intelligence-generated essays for academic assessment integrity using large language models,” *Expert Systems with Applications*, vol. 291, p. 128405, Oct. 2025, doi: 10.1016/j.eswa.2025.128405.

[25] Schaaff, “GitHub - iu-ai-research/human-AI-generatedTextCorpus: Human- and AI-generated texts from two domains as well as the four languages: EN, FR, DE, and ES,” GitHub. Accessed: Oct. 05, 2026. [Online]. Available: https://github.com/iu-ai-research/human-AI-generatedTextCorpus