Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v7i2.457

Research Article

Zero-Shot Detection of IndoT5-Synthesized Indonesian Scientific Abstracts Using mDeBERTa v3

Aldo Syahputra [1]; Aris Wahyu Murdiyanto [2]; Ulfi Saidata Aesyi [3]
[1] Universitas Jenderal Achmad Yani, Yogyakarta, Indonesia, aldosyahputra80@gmail.com
[2] Universitas Jenderal Achmad Yani, Yogyakarta, Indonesia, ariswahyumurdiyanto@gmail.com
[3] Universitas Jenderal Achmad Yani, Yogyakarta, Indonesia, ulfiaesyi@gmail.com

Correspondence should be addressed to Aldo Syahputra; aldosyahputra80@gmail.com

Received 10 March 2026; Accepted 15 June 2026; Published 31 July 2026

© Authors 2026. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:
Introduction: Distinguishing human-written scientific abstracts from AI-synthesized text remains challenging, particularly when machine-generated language appears fluent and formally structured. This study evaluates mDeBERTa v3 in a zero-shot Natural Language Inference (NLI) setting for detecting Indonesian scientific abstracts specifically synthesized using IndoT5-base-paraphrase. **Method**: A balanced dataset of 2,274 abstracts comprising 1,137 human-written abstracts from SINTA 3 journals and 1,137 IndoT5-synthesized counterparts was analyzed. Seven linguistic features were examined using the Mann–Whitney U test, followed by zero-shot mDeBERTa v3 classification using one-, three-, and five-aspect NLI instruction scenarios. A Random Forest classifier using the same linguistic features was included as a supervised baseline. **Results and Discussion**: All seven linguistic features differed significantly between classes ($p < 0.001$), with AI texts showing substantially higher sentence-length variation than human texts. The targeted one-aspect NLI scenario achieved the highest recall of 76.52% but only 53.52% accuracy because 790 human abstracts were misclassified as AI. Increasing instruction complexity further reduced recall. In contrast, Random Forest achieved 91.21% accuracy and an F1-score of 0.9130, confirming that the identified linguistic anomalies are strong learnable signals. **Conclusion**: Zero-shot mDeBERTa v3 can detect generator-specific structural artifacts but remains insufficiently precise for standalone academic-integrity screening and should be supplemented by supervised methods and human review.

Keywords: AI Text Detection; IndoT5; mDeBERTa v3; Scientific Abstract; Zero-Shot Classification.

1. Introduction

The rapid advancement of Large Language Models (LLMs) has fundamentally transformed text generation, simultaneously introducing critical challenges to academic integrity [1]. As reliance on AI tools increases, educational institutions face a growing crisis in distinguishing authentic human writing from machine-generated text [2]. Scientific abstracts, serving as the primary gateway to research findings, are particularly vulnerable to AI-driven manipulation [3], [4].

Previous studies demonstrate that human evaluators correctly identify only 68% of AI-generated abstracts, with a concerning 14% of authentic human texts mistakenly classified as machine-generated [5]. This high false-positive rate underscores the urgent need for objective, automated computational detection systems [6]. While earlier automated detection efforts have made significant strides [7], their application remains largely unoptimized for specific linguistic domains such as Indonesian academic text [8].

To address this vulnerability, this study proposes the implementation of the Multilingual Decoding-enhanced BERT with Disentangled Attention (mDeBERTa v3) model [9]. Unlike standard BERT architectures [10], mDeBERTa v3 utilizes a disentangled attention mechanism that separates content and relative word position, making it highly sensitive to the structural anomalies often present in AI-generated text. Rather than employing resource-

intensive model retraining, this research utilizes a zero-shot classification approach based on Natural Language Inference (NLI) [11]. This method evaluates the model's innate pre-trained capability to identify statistical "model signatures" such as unnatural sentence length fluctuations and semantic hallucinations left behind by generative AI [12].

This study specifically focuses on Indonesian scientific abstracts. To ensure a pure baseline, authentic human abstracts were scraped from SINTA 3 accredited journals published before 2018, predating the widespread use of advanced generative LLMs [13]. The AI-generated counterparts were synthesized using the IndoT5-base-paraphrase model, which adapts the unified text-to-text transformer architecture [14]. This transformation is essential because language models fundamentally operate by modeling the probabilities of text sequences [13]. IndoT5-base-paraphrase was selected as the AI-text generator because it is, to date, one of the few openly available paraphrase-oriented sequence-to-sequence models specifically fine-tuned for the Indonesian language, making it a realistic proxy for AI-assisted academic writing scenarios, such as a student using a local paraphrasing tool to rephrase existing text, rather than fully AI-authored content, which is more commonly associated with English-centric decoder-only LLMs. These probabilistic word selection patterns and sentence structures form unique statistical signatures, enabling Natural Language Processing (NLP) systems to distinguish between human-written and machine-generated content [15]. Furthermore, this study evaluates the model's performance across varying levels of instructional complexity, specifically utilizing 1, 3, and 5 NLI evaluation aspects. The primary objective is to compare how minimal, intermediate, and complex prompt instructions affect classification accuracy. This comparison is theoretically linked to the concept of attention dilution in Transformer-based architectures [16], wherein increasing the number of concurrent evaluation constraints forces the model's self-attention mechanism to distribute its probability weights across a larger set of tokens, potentially reducing the model's sensitivity to subtle anomaly markers and altering its zero-shot detection performance.

It is important to delineate the scope of this study relative to the broader AI-generated-text detection literature. AI-generated academic text detection can be broadly divided into three distinct problems, detecting fully generative text produced end-to-end by large language models such as ChatGPT, detecting paraphrased text derived from an existing human source, and detecting degraded or corrupted paraphrase output where semantic fidelity has been compromised during transformation. The present study falls primarily within the second and third categories, the AI class was not independently authored by a generative model, but produced by paraphrasing authentic human abstracts through IndoT5-base-paraphrase. Consequently, the two classes share substantial lexical and topical content, differing mainly in structural artifacts introduced during the paraphrasing process. This study should therefore be understood as an investigation of generator-specific AI-synthesized text detection, and its findings should not be generalized to the detection of independently written, fully generative AI abstracts (those produced by ChatGPT-like decoder-only LLMs) without further validation. Within this narrower but still practically relevant scope, the primary novelty of this study lies in demonstrating how a single, precisely targeted NLI instruction, one that mirrors a generator's specific mechanical error signature, can outperform broader, multi-aspect instructions in a zero-shot setting, and in identifying the structural anomaly that makes such targeting possible.

2. Method

This study employs a quantitative research design utilizing text mining and machine learning approaches to detect AI-generated Indonesian scientific abstracts. The overall research workflow is structured sequentially, starting from data collection and preprocessing to statistical testing and final performance evaluation. The complete research flow is illustrated in the flowchart presented in **Figure 1**.

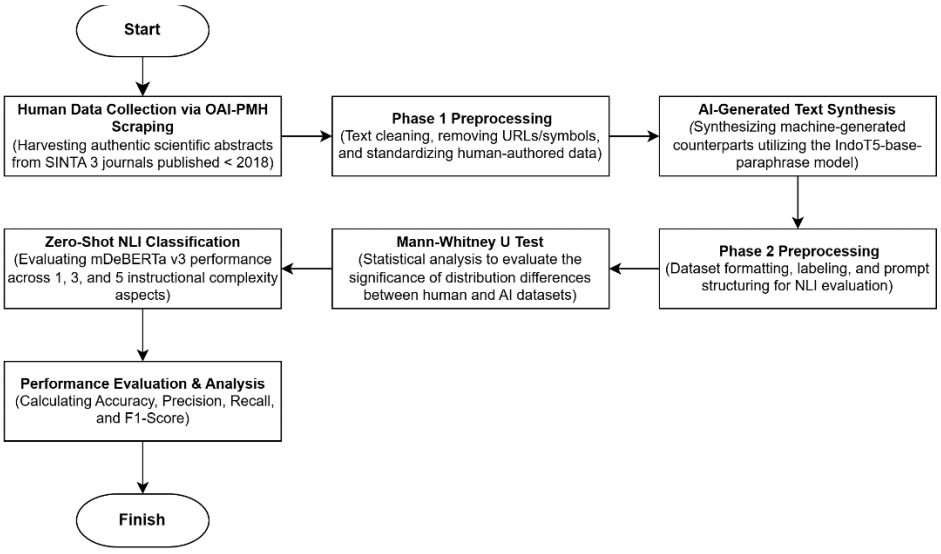


Figure 1. Research Flow

The collected raw data were initially subjected to a primary preprocessing stage to remove irrelevant elements such as HTML tags and extreme length anomalies. Subsequently, the authentic human abstracts were synthesized using the IndoT5-base-paraphrase model to generate the AI text class. A secondary preprocessing phase, including truncation and undersampling, was then applied to balance the dataset. Prior to the classification process, the dataset's linguistic features were statistically validated using the Mann-Whitney U test. The text classification was performed using the mDeBERTa v3 architecture under a zero-shot Natural Language Inference (NLI) framework. The performance of the model was evaluated using a confusion matrix, from which accuracy, precision, recall, and F1-score were calculated.

Data Selection

The data used in this study consist of Indonesian scientific abstracts collected to establish a pure baseline of human-written text, as the quality of raw data significantly determines the success of subsequent Natural Language Processing (NLP) modeling phases [17]. Data collection was conducted through an automated harvesting process using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) across 18 national scientific journal portals accredited as SINTA 3. The selection of SINTA 3 journals ensures that the textual data meets national accreditation standards for formal linguistic structure [18], while OAI-PMH facilitates structured and legal metadata extraction directly from Open Journal Systems [19]. To guarantee a baseline dataset free from advanced generative AI contamination, data retrieval was strictly limited to publications issued prior to 2018, aligning with the evolutionary timeline of Large Language Models (LLMs) which achieved massive coherent text generation capabilities only post-2018 [13], [19]. Through this harvesting process, a total of 1,139 raw human abstracts were successfully obtained. **Table 1** details the distribution of raw abstracts harvested across these 18 sources.

Table 1. Distribution of Raw Data by SINTA 3 Journal Source

No	Journal Name	Subject Area	Portal URL (OAI-PMH)	Abstract Count
1	SIMETRIS, UMK	Mech./Electrical Eng., CS	jurnal.umk.ac.id/index.php/simet/oai	312
2	Jurnal INFOTEL, IT Telkom Purwokerto	IT & Telecom	ejournal.ittelkom-pwt.ac.id/index.php/infotel/oai	171
3	JIP, Polinema	Informatics	jurnal.polinema.ac.id/index.php/jip/oai	131
4	Sistemasi, UNISI	Information Systems	sistemasi.ftik.unisi.ac.id/index.php/stmsi/oai	82
5	Jurnal Sisfokom, ISB Atma Luhur	IS & Computing	jurnal.atmaluhur.ac.id/index.php/sisfokom/oai	78

No	Journal Name	Subject Area	Portal URL (OAI-PMH)	Abstract Count
6	EXPLORE, UBL	IS & Telematics	jurnal.ubl.ac.id/index.php/explore/oai	62
7	ILKOM, UMI Makassar	Computer Science	jurnal.fikom.umi.ac.id/index.php/ILKOM/oai	61
8	Jurnal Komputer Terapan, PCR	Applied Computing	jurnal.pcr.ac.id/index.php/jkt/oai	52
9	JUSIFO, UIN Raden Fatah	Information Systems	jurnal.radenfatah.ac.id/index.php/jusifo/oai	35
10	CESS	Computer Eng. & Systems	jurnal.unimed.ac.id/2012/index.php/cess/oai	33
11	JTIP, UNP	IT & Education	tip.ppj.unp.ac.id/index.php/tip/oai	27
12	JISKA, UIN Sunan Kalijaga	Informatics	ejournal.uin-suka.ac.id/saintek/JISKA/oai	24
13	Jurnal COREIT, UIN SUSKA	Information Technology	ejournal.uin-suska.ac.id/index.php/coreit/oai	19
14	Jurnal TEKNOINFO, Teknokrat	Information Technology	ejurnal.teknokrat.ac.id/index.php/teknoinfo/oai	19
15	Jurnal Tekno Kompak, Teknokrat	Computing	ejurnal.teknokrat.ac.id/index.php/teknokompak/oai	13
16	INTI Nusa Mandiri	Informatics	ejournal.nusamandiri.ac.id/index.php/inti/oai	10
17	ELTIKOM, POLIBAN	Electrical, Telecom & CE	eltikom.poliban.ac.id/index.php/eltikom/oai	5
18	AITI, UKSW	Information Technology	ejournal.uksw.edu/aiti/oai	5
Total				**1139**

Of 20 initially targeted SINTA 3 portals, 18 returned valid OAI-PMH responses. The remaining two, JSiI UNSERA and Telematika Jurnal Informatika UPNYK, repeatedly failed due to server request timeouts and were excluded from the harvesting process.

Preprocessing

To ensure data quality, reduce noise, and maintain reproducibility prior to the artificial intelligence synthesis and classification stages, a structured and sequential data-cleaning procedure was applied. The preprocessing workflow was initiated through a primary cleaning phase on the harvested human corpus, which involved duplicate removal to eliminate repeated abstracts, alongside relevance filtering to exclude HTML tags, formatting artifacts, and text anomalies. Furthermore, a length filtering criterion was applied to eliminate abstracts containing fewer than 50 words, ensuring that every retained document possessed sufficient linguistic density for deep learning analysis.

Following this initial cleaning phase, the dataset volume was refined. A comparison of the data volume before and after each data-cleaning stage is presented in **Table 2**.

Table 2. Data Volume After Harvesting and Cleaning Stages

Category	Volume
Harvesting Result	1,139
After Duplicate Removal	1,139
After Length Filtering (≥ 50 words)	1,138

AI Synthesis

To construct the comparative machine-generated text class, the validated corpus of authentic human abstracts was processed using the generative *IndoT5-base-paraphrase* model. This transformation process is theoretically grounded in the mechanics of Large Language Models (LLMs) and generative architectures, where text is produced through probability distribution over token sequences [13]. According to foundational generative theory, models learn hidden

data distributions to synthesize outputs that mimic human data, yet this mathematical process inherently maintains structural patterns distinct from natural human data distribution [13].

The application of the IndoT5-base-paraphrase model systematically alters natural human sentence structures into machine-reconstructed equivalents. Because the model generates text by computing the probabilistic distribution of words and token sequences rather than relying on human cognitive intent, this algorithmic transformation leaves distinct statistical traces known as "model signatures" [15]. Consequently, the resulting text retains semantic similarity with the original human abstract while embedding the algorithmic and statistical anomalies characteristic of machine generation [13], [14]. Based on this principle, any text transformed through this sequence-to-sequence transformer model is systematically designated and labeled as AI-generated text, establishing a reliable comparative dataset for subsequent evaluation.

Text synthesis was performed using the Wikidepia/IndoT5-base-paraphrase checkpoint via the Hugging Face AutoModelForSeq2SeqLM interface. Each human abstract was tokenized with a maximum input length of 512 tokens. The specific decoding parameters were selected based on standard practice for controlling beam-search quality in sequence-to-sequence generation. A beam size of 4 was used as a balanced configuration between generation diversity and computational cost, consistent with findings that excessively small beam widths produce overly literal candidates while widths beyond 5–6 yield diminishing quality gains at substantially higher inference cost [21]. A repetition penalty of 1.3, combined with a no-repeat n-gram size of 3, was applied to suppress a well-documented failure mode of beam-search decoding, repetitive phrase looping, following the repetition-penalty mechanism introduced for controllable Transformer generation [22] and the n-gram repeat-blocking strategy established for sequence-to-sequence summarization models [23]. A length penalty of 1.0 (neutral, neither favoring longer nor shorter outputs) was applied using the length-normalization procedure for beam-search scoring introduced by Wu et al. [24], chosen specifically to avoid systematically biasing the AI class toward artificially short or long outputs relative to the source text, which could otherwise introduce a length-based classification shortcut independent of genuine stylistic differences. Early stopping was enabled to terminate beam search once all beam hypotheses reach the end-of-sequence token, preventing unnecessary computation without affecting output quality. As beam search is a deterministic decoding strategy, no stochastic sampling parameters (top-k, top-p, or temperature) were applicable, and no random seed was required for output reproducibility at the generation stage.

Post-Synthesis Preprocessing and Dataset Balancing

Following the AI synthesis stage, a secondary preprocessing phase was executed on the combined dataset to eliminate generation noise and prevent length-based classification biases, as rigorous data cleansing and normalization are fundamental in data mining and text processing pipelines [16], [20]. This advanced phase commenced with case folding to standardize all alphabetic characters into lowercase, reducing feature space dimensionality and lexical redundancy [17]. This was followed by a targeted artifact removal process specifically eliminating generative formatting anomalies produced during the IndoT5 paraphrasing, such as empty brackets and redundant transition clauses. However, structural punctuation marks, including commas and periods, were strictly preserved as vital syntactic features for detecting structural anomalies in AI-generated text [12].

To prevent the classification algorithm from developing shortcut learning based on document length disparities, a common vulnerability in text classification where models exploit length variations rather than genuine semantic content, a truncation technique was applied to limit all abstracts to a maximum of 100 words, aligning with optimal sequence length management in transformer architectures [13]. This specific threshold was chosen to standardize the input window, forcing the zero-shot classifier to evaluate dense linguistic structures and mechanical anomalies rather than simply classifying texts based on output-length discrepancies generated by IndoT5. While this rigid truncation introduces minor boundary artifacts, it ensures that the NLI model's decision is driven purely by internal sentence characteristics rather than macroscopic document length. Finally, an undersampling technique was executed to address minor data shrinkage caused by artifact filtering in the AI class, mitigating class imbalance which can heavily degrade classifier generalizability [16], [20]. By randomly removing a single instance from the human abstract class, the dataset achieved a perfectly balanced distribution.

Through this comprehensive preprocessing and balancing procedure, the final corpus was established with a total of 2,274 documents, comprising 1,137 authentic human abstracts and 1,137 AI-synthesized texts. A detailed summary of the data reduction and balancing stages across the entire pipeline is presented in **Table 3**.

Table 3. Summary of Data Reduction and Balancing Stages

Stage	Human Class	AI Class	Total Corpus
Initial Harvesting / Synthesis	1.138	1.138	2.276
Artifact & Invalid Filtering	0	-1	-1
Undersampling	-1	0	-1
Final Dataset Corpus	1.137	1.137	2.274

Mann-Whitney U Test

To prove that the linguistic feature differences are valid and not merely random variations, an objective statistical test is required. Because the numeric metrics extracted from natural language text documents generally do not follow a normal distribution, conventional parametric statistical tests become invalid [27]. Therefore, the approach used to examine the differences within this text dataset is the Mann-Whitney U test. According to Field [27], the Mann-Whitney U test is a nonparametric statistical method that can be used to analyze textual or numerical data, designed to find differences between two independent groups or conditions. This statistical method does not rely on calculating or comparing the mean values of the two data groups [27]. Instead, the underlying concept of this test operates by sorting all data points and assigning global ranks from the smallest to the largest value, followed by evaluating the total sum of ranks for each respective data group to determine if they are distributed in roughly equal proportions [27]. The core output of this computation is the U-statistic (U), which measures the degree of separation between the distributions of the two groups by quantifying how many times a score from one group precedes a score from the other group [27].

To determine the statistical significance of differences between groups, the test evaluates a significance probability value, or p-value, using a significance threshold of $\alpha = 0.05$ [27]. If the test results show a p-value < 0.05, the null hypothesis is rejected, which mathematically proves that there is a real difference in linguistic characteristics between human-written text and artificial intelligence [27]. However, because the p-value only confirms the presence or absence of a difference without explaining its magnitude, an effect size metric symbolized by r is utilized [27].Based on statistical testing standards put forward by Cohen in Field [27], the interpretation of the separation gap between classes using the r-value is categorized into a small effect when $|r| \geq 0.1$, a medium effect when $|r| \geq 0.3$, and a large effect when $|r| \geq 0.5$. Ultimately, the larger the r-value obtained from a linguistic feature, the sharper and more separated the boundary of stylistic characteristics will be between pure human text and machine-generated text in the dataset.

Zero-Shot NLI Classification

One of the crucial aspects in the development of modern Natural Language Processing (NLP) is the paradigm shift toward utilizing pre-trained language models to perform zero-shot inference [28]. Leveraging the intrinsic capabilities of these models enables the evaluation of detection effectiveness without requiring task-specific fine-tuning [13]. The zero-shot classification approach is implemented through the Natural Language Inference (NLI) framework. Yin et al. [27] proposed utilizing this capability where the text to be classified serves as the premise, while each category option is transformed into a hypothesis or statement to be evaluated for truth. This approach is reinforced by Laurer et al. [27], who demonstrated that NLI is universal, meaning that nearly all classification tasks can be solved simply by converting category names into meaningful hypothesis sentences without needing to retrain the model from scratch.

In the application of NLI-based zero-shot classification, model effectiveness depends on the complexity level of the evaluated hypotheses, where instructing too many criteria simultaneously makes the model susceptible to attention dilution. Based on research by Hahn [16], the self-attention mechanism in the Transformer architecture has a theoretical limitation where the distribution of probability weights becomes increasingly fragmented when processing longer and more complex inputs. This is further reinforced by the findings of Liu et al. [30], who proved that adding

overly dense information and instructions actually reduces the language model's sensitivity in capturing primary signals within the text.

To operationalize the three instructional-complexity scenarios, five candidate hypothesis pairs were defined, each targeting a distinct textual dimension, sentence-structure stability, empirical grounding of data, technical-term usage, conclusion validity, and punctuation mechanics. For every abstract, the classifier independently evaluated all five hypothesis pairs using the template "{}.", and the top-ranked hypothesis per pair was recorded as a binary vote (human/AI). The three scenarios were then constructed via a voting scheme over these five independent evaluations, as detailed in **Table 4**.

Table 4. NLI Hypothesis Pairs and Scenario Voting Configuration

No	Target Dimension	Human Hypothesis	AI Hypothesis
1	Sentence structure stability	Sentence structure in this abstract is consistent, orderly, and naturally well-bounded in length	Sentence structure in this abstract is highly unstable, fluctuates unnaturally, and excessively stacks clauses
2	Empirical grounding	Data and figures in this abstract originate from real experiments conducted directly by the researcher	Data and figures in this abstract are automatically generated and do not always originate from actual experiments
3	Technical-term usage	Technical terms in this abstract are used precisely, reflecting the author's deep understanding of the research field	Technical terms in this abstract are used mechanically, without deep contextual understanding by the author
4	Conclusion validity	The conclusion of this abstract is limited to findings genuinely substantiated by the research	The conclusion of this abstract overreaches beyond the actual findings, claiming results larger than what was proven
5	Punctuation mechanics	Punctuation in this abstract is used organically and fluidly, without a rigid structural comma pattern	Commas in this abstract are used rigidly and mechanically to initiate transition words and link long clauses

Scenario 1-Aspect used only hypothesis pair no 1 (sentence-structure stability), the dimension directly targeting the anomaly identified. Scenario 3-Aspect evaluated three criteria by utilizing hypothesis pairs no 1 through no 3. Scenario 5-Aspect incorporated all five hypothesis pairs. This design allowed direct comparison between a narrowly targeted single-criterion instruction and progressively broader multi-criterion ensembles under an identical zero-shot backbone.

Supervised Baselined Comparison

To contextualize the zero-shot classification results, a supervised Random Forest classifier was implemented as a comparison baseline, using the same seven linguistic features validated through the Mann-Whitney U test. Random Forest was selected because, as an ensemble of decision trees, it does not assume a normal feature distribution or a linear decision boundary, consistent with the nonparametric framing already established for the linguistic feature analysis, and because it natively handles features of differing scale without requiring normalization. This choice is further supported by recent literature demonstrating that Random Forest classifiers trained on stylometric or TF-IDF-derived linguistic features achieve strong, computationally lightweight performance for AI-generated text detection, including academic and scientific writing contexts [7], [30], [31].The classifier was configured with 200 estimators and default scikit-learn hyperparameters otherwise, trained on an 80/20 stratified train-test split (random_state = 42) to preserve class balance, and further validated using 5-fold cross-validation to assess result stability. This baseline serves purely as a reference point for evaluating the relative discriminative power of the underlying linguistic features, independent of the zero-shot NLI framework.

mDeBERTa v3

mDeBERTa v3 (*Multilingual Decoding-enhanced BERT with Disentangled Attention*) is a multilingual variant of the Transformer-based language model introduced by He et al. [9]. This model retains the core architectural advantages of previous DeBERTa versions [33] while being specifically pre-trained on the CC100 corpus, which encompasses 100 different languages, including Indonesian. The CC100 corpus is a large-scale text repository

constructed by Conneau et al. [34] through rigorous filtering of raw data extracted from Common Crawl, providing over 2TB of clean text data representing a hundred languages worldwide. In this study, the specific pre-trained checkpoint MoritzLaurer/mDeBERTa-v3-base-mnli-xnli is utilized to perform zero-shot classification without additional fine-tuning.

A primary architectural enhancement in mDeBERTa v3 is the incorporation of the disentangled attention mechanism [9]. Unlike standard BERT models that combine word content and position embeddings directly, mDeBERTa v3 decouples content and position into two distinct representations, a content vector (H_i) representing word meaning and a relative position vector ($P_{i|j}$) representing the relative distance between tokens. Within the self-attention mechanism, attention scores are computed by evaluating interactions among content-to-content, content-to-position, and position-to-content components. This decoupling is particularly crucial for AI-generated text detection, as machine-generated texts often carry specific structural rigidities and positional patterns (model signatures), enabling the model to capture subtle structural anomalies more effectively than standard architectures.

Furthermore, mDeBERTa v3 adopts an ELECTRA-style pre-training strategy known as Replaced Token Detection (RTD), which trains the model as a discriminator to identify whether a token is original or replaced across 100 languages simultaneously [9]. It also incorporates Gradient-Disentangled Embedding Sharing (GDES) to manage embedding sharing between the generator and discriminator with isolated gradient flows, thereby ensuring stable training and high computational efficiency during processing [9].

During the inference process, input texts are tokenized using the multilingual SentencePiece tokenizer, where special tokens such as [CLS] are introduced to capture the global contextual representation of the entire abstract [9]. After passing through the Transformer encoder blocks enhanced by disentangled attention and the Enhanced Mask Decoder (EMD), the hidden state representation of the [CLS] token (h_{CLS}) is projected through a linear classification layer [9]. Finally, the Softmax function computes the normalized probability distribution over the target categories, enabling the zero-shot classification framework to effectively distinguish between human-written and AI-synthesized texts [9].

Classification Evaluation Metrics

Evaluating model performance in text classification tasks requires objective statistical parameters to determine the extent to which a model can accurately distinguish between human-written texts and AI-generated texts. Since this study implements a zero-shot classification approach, the model evaluation purely aims to measure the reliability of the model's intrinsic knowledge in generalizing classification across unseen abstract datasets [35].

Based on studies by Sujon et al. [36] and Sathyanarayanan [37], the primary performance metrics utilized to evaluate the classification model comprise Accuracy, Precision, Recall, and F1-Score. Accuracy measures the ratio of total correct predictions to the overall number of data samples. Precision evaluates the model's exactness in predicting the positive class (AI-generated text) to minimize false positives, whereas Recall measures the model's capability to retrieve all actual AI-generated text instances to minimize false negatives. Furthermore, the F1-Score serves as a crucial harmonic balance metric between Precision and Recall, ensuring a comprehensive assessment of the model's classification performance across balanced datasets.

Additionally, the confusion matrix serves as a fundamental analytical tool used to map the model's decision distribution in detail. This matrix enables researchers to examine the model's error patterns through four primary components.

Table 5. Confusion Matrix for Classification AI and Human

Actual \ Predicted	AI	Human
AI	TP	FN
Human	FP	TN

True Positive (TP) represents the condition where the actual data is AI-generated text and the model successfully predicts it as such. True Negative (TN) denotes when the actual data is human-written text and the model accurately predicts it as human text. Conversely, False Positive (FP) occurs when the actual data is human text but the model incorrectly predicts it as AI-generated, while False Negative (FN) refers to when the actual data is AI text but the model mistakenly predicts it as human text.

3. Result and Discussion

Linguistic Feature Validation

Prior to classification, the balanced corpus of 2,274 abstracts (1,137 human and 1,137 AI-synthesized) was statistically validated using the Mann-Whitney U test across seven linguistic features. **Table 6** summarizes the results.

Table 6. Mann-Whitney U Test Results on Linguistic Features

Feature	Human Mean	AI Mean	U-Statistic	p-value	r-value	Effect Size
Word Count	98.62	84.89	1,047,230.5	< 0.001	0.620	Large
Average Sentence Length	20.40	29.44	316,916	< 0.001	0.510	Large
Sentence Length Variation	7.99	15.44	261,758.5	< 0.001	0.595	Large
Sentence Count	5.31	3.35	1,079,522	< 0.001	0.670	Large
Comma Ratio	0.036	0.082	209,865.5	< 0.001	0.675	Large
Complex Punctuation Ratio	0.027	0.059	389,161.5	< 0.001	0.398	Medium
Conjunction Word Ratio	0.23	0.25	458,482	< 0.001	0.291	Small

All seven features returned $p < 0.001$ (with substantial U-statistics ranging from 209,865.5 for Comma Ratio to 1,079,522 for Sentence Count), rejecting the null hypothesis and confirming that the stylistic gap between human and AI-synthesized text is statistically genuine rather than incidental. Five features, Comma Ratio, Sentence Count, Word Count, Sentence Length Variation, and Average Sentence Length, exhibited large effect sizes ($r \geq 0.5$). The Comma Ratio showed the strongest separation ($r = 0.675$), with AI text using commas roughly 2.3 times more frequently than human text (0.082 vs. 0.036), suggesting a mechanical over-reliance on comma-linked clauses to compensate for broken sentence flow. This pattern is consistent with AI text also producing longer average sentences (29.44 vs. 20.40 words) while containing fewer sentences overall per abstract (3.35 vs. 5.31), ideas are compressed into fewer, longer, comma-heavy sentences rather than being distributed across a natural sentence structure.

The most theoretically significant result concerns Sentence Length Variation. Under the conventional assumption that generative models produce uniform, homogeneous phrasing, AI text would be expected to show *lower* variation than human text. Instead, the opposite was observed, AI-synthesized text exhibited a mean standard deviation of 15.44, nearly double that of human text (7.99), with a large effect size (r = 0.595). This anomaly directly motivated the design of the subsequent zero-shot NLI instructions.

Anomaly Context-Window Exhaustion in IndoT5-base-paraphrase

To explain this counterintuitive variation, paired samples of original and paraphrased abstracts were manually inspected. To systematically document this failure mode, **Table 7** presents representative paired examples of original and paraphrased text alongside the corresponding error type.

Table 7. Representative Semantic Hallucination Sample in IndoT5-base-paraphrase Output

Original (Human)	Paraphrased (AI)	Error Type
"...*objek penelitian cabai*..."	"...*objek penelitian acar*..."	Lexical substitution (research object)
"...*studi kasus di Tembilahan*..."	"...*studi kasus di Tembakan*..."	Place-name distortion
"...*sistem informasi percetakan*..."	"...*metode pengembangan waterfall*..."	Method–domain misattribution

These errors concentrate toward the middle and end of longer abstracts, consistent with context-window exhaustion as the model's attention capacity is consumed by earlier tokens, severely degrading its ability to preserve long-range semantic dependencies [30]. In Transformer-based sequence-to-sequence architectures, the self-attention mechanism distributes probability weights across token sequences, but as input length and lexical density increase, the internal representations, particularly within intermediate and final decoding layers, become susceptible to attention dilution [16].

As established in prior literature evaluating transformer failure modes in long-text generation tasks, when a model exceeds its effective attention span or encounters out-of-domain specialized terminology (such as complex scientific nomenclature in Indonesian academic abstracts), the decoder loses tracking of the global syntactic structure [30]. This operational breakdown manifests not only as localized lexical substitutions, but also as catastrophic repetition and fragmented clause boundaries. Consequently, the abnormal sentence-length fluctuations observed in IndoT5-synthesized texts do not reflect sophisticated human-like stylistic variation; rather, they serve as a mechanical symptom of generative instability driven by attention degradation over dense token sequences [15, 29].

Zero-Shot Classification Performance Across NLI Scenarios

Table 8 presents the comparative performance of the mDeBERTa v3 zero-shot NLI classifier across three instruction scenarios of increasing evaluative complexity (1, 3, and 5 aspects).

Table 8. Evaluation Metrics per NLI Scenario

Scenario	Accuracy	Precision	Recall	F1-Score
1-Aspect	0.5352	0.5241	0.7652	0.6221
3-Aspect	0.5101	0.5112	0.4626	0.4857
5-Aspect	0.5075	0.5089	0.4666	0.4641

The single-aspect instruction, which specifically targeted the sentence-fluctuation anomaly identified, achieved the highest overall performance (Recall = 76.52%, F1 = 0.6221), substantially outperforming both the 3-aspect and 5-aspect configurations. This result directly addresses the study's core research question, a narrowly targeted zero-shot hypothesis that mirrors the model's *actual* error signature outperforms broader, seemingly more "thorough" instruction sets.

Contrary to the expectation that additional evaluative criteria (factual accuracy, terminology precision) would improve discrimination, performance instead degraded monotonically as complexity increased — Recall dropped from 76.52% to 46.26% (3-Aspect) and further to 46.66% (5-Aspect, essentially near-random). This pattern is consistent with the theoretical concept of attention dilution [16], as the number of concurrent evaluation constraints grows, the model's self-attention distributes probability mass across a wider set of hypothesis tokens, weakening its sensitivity to the specific mechanical anomaly (sentence-length fluctuation) that most reliably distinguishes IndoT5-synthesized text. In effect, additional instructional aspects introduce competing decision signals that are less discriminative than the single fluctuation-based cue, diluting rather than reinforcing the classifier's judgment. **Table 9** corroborates this at the raw prediction level.

Table 9. Correct, Incorrect, and Error Predictions per Scenario

Scenario	Correct	Incorrect	Error
1-Aspect	1,217 (53.52%)	1,057 (46.48%)	0 (0.00%)
3-Aspect	1,160 (51.01%)	1,114 (48.99%)	0 (0.00%)
5-Aspect	1,154 (50.75%)	1,120 (49.25%)	0 (0.00%)

Confusion Matrix Analysis

The per-class distribution of predictions, disaggregated by scenario, is visualized in Figures 2–4 and summarized in **Table 10**.

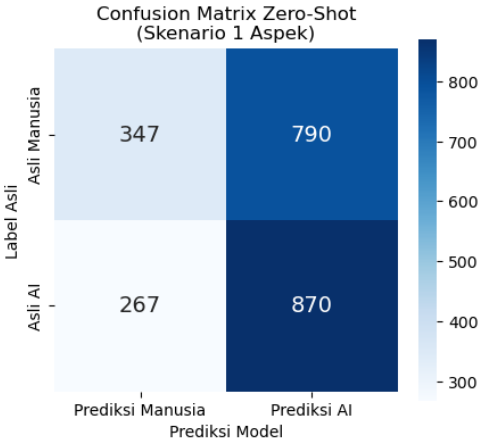


Figure 2. Confusion Matrix, Scenario 1-Aspect

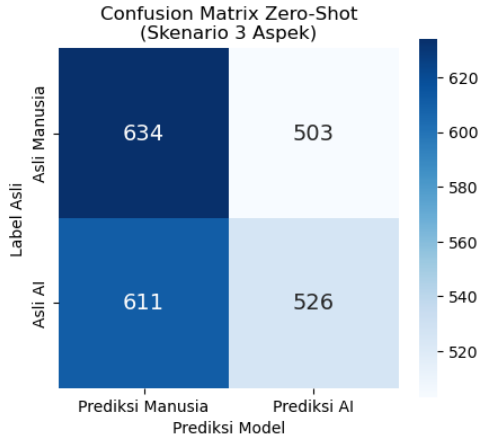


Figure 3. Confusion Matrix, Scenario 3-Aspect

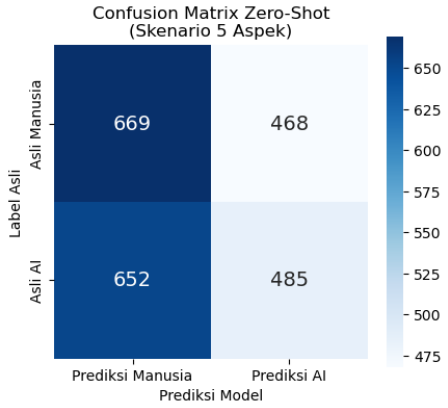


Figure 4. Confusion Matrix, Scenario 5-Aspect

Table 10. Confusion Matrix Summary per Scenario

Scenario	TP (AI correct)	FN (AI misclassified)	TN (Human correct)	FP (Human misclassified)
1-Aspect	870	267	347	790
3-Aspect	526	611	634	503
5-Aspect	485	652	669	468

1-Aspect scenario produced the darkest concentration of predictions in the AI-predicted column, 870 of 1,137 AI-synthesized abstracts were correctly identified (TP), but this same column also captured 790 human abstracts misclassified as AI (FP), visibly the largest single cell in the matrix aside from the true-positive cell itself. This confirms that the model's high sensitivity to sentence-fluctuation cues extended indiscriminately to any text exhibiting similar surface variation, regardless of authorship.

Inversion of this pattern as instructional complexity increased. In Scenario 3-Aspect, the darkest cell shifts to the human-correct quadrant (TN = 634), while AI recall collapses to 526 correctly identified documents, with 611 AI abstracts now misread as human (FN). This shift becomes more pronounced in Scenario 5-Aspect, where TN rises

further to 669 while TP falls to its lowest point at 485, with 652 AI abstracts misclassified as human. A clear reallocation of the model's decision boundary, from an aggressive, AI-leaning classifier under a single targeted instruction, toward an increasingly conservative, human-leaning classifier as more evaluative aspects are added, reinforcing the attention-dilution interpretation rather than the expected improvement in discriminative precision.

Sensitivity and False Positive Analysis

To characterize the false-positive and false-negative errors beyond the aggregate confusion matrix (Section D), 20 randomly sampled false positives and 20 false negatives from Scenario 1-Aspect were manually reviewed, with corresponding linguistic feature values recalculated per document.

The dominant driver of false positives was found to be a truncation artifact rather than genuine stylistic ambiguity, 17 of the 20 sampled human abstracts (85%) were cut off mid-sentence at the 100-word truncation boundary without terminal punctuation ("...penelitian dilakukan menggunakan data beras dari bulog kota"). Because the sentence-extraction function requires terminal punctuation to count a sentence, this truncated final clause is excluded from the sentence count, artificially distorting the computed sentence-length statistics independent of the author's actual writing style. This finding directly addresses the methodological trade-off of the 100-word limit, while truncation successfully prevented length-based shortcut learning, it inherently penalized human texts that possess long contextual introductions, artificially inflating their structural instability at the document boundary. This validates the necessity of exploring untruncated processing capabilities in future iterations. One human abstract exhibited a more genuine ambiguity, a single long compound sentence connected entirely by commas before the first period (std = 0, num_sentences = 1), a legitimate rhetorical structure in formal Indonesian academic writing that happens to be structurally indistinguishable from the instability pattern targeted by the Scenario 1 hypothesis.

False negatives showed a complementary but distinct pattern, 8 of 20 sampled AI-synthesized abstracts (40%) contained two or fewer detected sentences, and 3 exhibited zero measurable sentence-length variation (std = 0), while 7 (35%) showed visible word-level repetition degeneration ("...speed speed, 255 speed adalah speed sangat mempengaruhi hasil speed..."). These represent IndoT5's most severe generation failures, yet they were paradoxically the hardest for the classifier to detect, when paraphrase degeneration prevents proper sentence segmentation, the resulting text collapses into a single run-on unit for which sentence-length variance is mathematically undefined (or trivially zero), depriving the Scenario 1 hypothesis, which specifically targets variation between sentences of any signal to act on.

Taken together, this error analysis reveals that classification failures at both ends of the confusion matrix converge on the same underlying structural condition, collapse of the multi-sentence structure into a single sentence-like unit, whether caused by natural human compound-sentence style, dataset truncation, or catastrophic AI generation failure. This indicates that the sentence-length-fluctuation heuristic, while effective for moderately unstable AI text, is not merely imprecise at the margins, it is structurally blind to the most severe AI generation failures, and structurally vulnerable to an artifact introduced by the study's own preprocessing pipeline (truncation) rather than the AI-generation process itself. This is empirically evidenced by the detailed *False Positive* and *False Negative* sample analysis in **Table 11**, which demonstrates that classification errors are driven by three distinct structural conditions, natural long compound sentences, pipeline truncation artifacts, and catastrophic word-repetition degeneration in AI-synthesized texts.

Table 11. Representative FP/FN Error Examples

Type	Source	Excerpt	Structural Cause
FP	ILKOM UMI Makassar	"*...sistem kendali adalah suatu alat...sistem yang digunakan untuk membuat suatu perangkat menjadi terkendali...*"	Natural long compound sentence (no truncation involved)
FP	Jurnal INFOTEL	"*...dengan memanfaatkan teknologi pemrosesan citra sebagai alat bantu instrumen pengukuran yang dapat menampilkan nilai ukur yang jelas*	Truncated mid-sentence at 100-word limit

Type	Source	Excerpt	Structural Cause
		terbaca. penelitian dilakukan menggunakan data beras dari bulog kota"	
FN	Jurnal INFOTEL (IndoT5)	"...speed speed, 255 speed *adalah* speed *sangat mempengaruhi hasil speed yang relatif stabil speed diperlukan*..."	Catastrophic word-repetition degeneration
FN	JIP (IndoT5)	"*...untuk mendapatkan keakuratan atau ketepatan perhitungan dalam mendiagnosis gejala penyakit pada burung kenia*."	Premature truncation of generation, low sentence count

Lexical Similarity and Its Contribution to False Positives

To quantitatively explain the persistently high false-positive rate and provide a more robust lexical analysis than visual word clouds alone, word-frequency analysis was conducted alongside the word clouds. **Figures 5** and **6** visualize the text corpus, while **Table 12** presents the absolute frequency of the top 10 dominant terms across both classes.

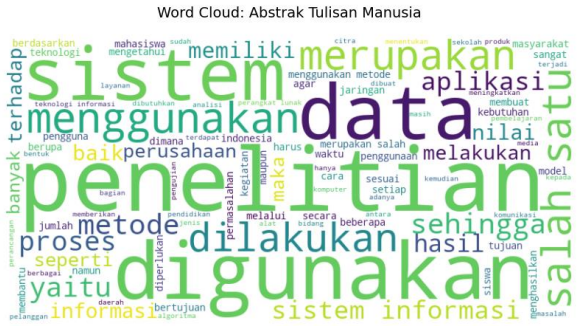


Figure 5. Word Cloud of Human-Written Abstracts



Figure 6. Word Cloud of AI-Synthesized Abstracts

Table 12. Top 10 Most Frequent Words in Human vs. AI-Synthesized Abstracts

Rank	Human Text (Word)	Frequency	AI-Synthesized Text (Word)	Frequency
1	*sistem*	1,311	*sistem*	1,264
2	*menggunakan*	956	*data*	836
3	*data*	902	*menggunakan*	834
4	*informasi*	848	*aplikasi*	714
5	*penelitian*	826	*digunakan*	706
6	*metode*	773	*informasi*	699
7	*digunakan*	608	*metode*	628
8	*merupakan*	595	*penelitian*	510
9	*dilakukan*	558	*hasil*	497
10	*salah*	510	*sehingga*	407

The frequency analysis and word clouds reveal a massive lexical overlap between the two classes. Both datasets are dominated by identical core technical terms characteristic of computer science and information technology abstracts, "*sistem*," "*data*," "*menggunakan*," "*metode*," and "*informasi*". This strong overlap confirms that the IndoT5-base-paraphrase model successfully preserves the formal domain terminology of the source texts, generating paraphrases that are lexically indistinguishable from human writing at a glance.

However, a closer inspection of the term rankings reveals a structural divergence beneath this surface-level similarity. The human abstracts contain a higher frequency of narrative, context-building, and descriptive connective

words within their top ranks, such as "*merupakan*" (rank 8), "*dilakukan*" (rank 9), and "*salah*" (rank 10, typically part of the phrase "*salah satu*"). In contrast, the AI-synthesized abstracts prioritize outcome and artifact-oriented terms much earlier in the frequency hierarchy. For instance, the word "*aplikasi*" surges to rank 4 (714 occurrences) in the AI class, whereas it only ranks 18th in the human class. Conversely, foundational research terms like "penelitian" drop significantly in the AI output.

This quantitative shift indicates that AI-synthesized text tends to compress human background elaboration, abandoning narrative connectives to jump more directly toward function-stating phrasing. This aligns seamlessly with the sentence-fragmentation and context-loss phenomenon inherent to the model's generation process.

Most importantly, this near-identical core vocabulary provides a direct lexical explanation for the high false-positive classification rate. Because mDeBERTa v3 functions under a zero-shot NLI approach lacking task-specific fine-tuning, the model becomes confused by the heavy vocabulary overlap. Lacking clear lexical boundaries, the classifier is forced to rely excessively on structural anomalies (such as sentence-length fluctuation) to separate the classes. While structurally diagnostic, this reliance causes the model to misfire on genuine human texts that happen to share similar fluctuations, proving that lexical-level features alone are fundamentally insufficient for reliable AI-text detection in paraphrased academic documents.

Comparison with Prior Work and Theoretical Implications

These findings complicate the framing in Gao et al. [5], where human reviewers struggled with AI-generated abstracts partly because such text *appeared* fluent and stylistically uniform. In the present Indonesian-language, paraphrase-based synthesis setting, AI text is in fact *less* uniform at the sentence-structure level than human writing, but this instability is not perceptible through casual human reading, since surface fluency and formal vocabulary remain intact. This suggests that AI-generated-text detectability is highly dependent on both the generation method (paraphrase-based vs. fully generative LLMs) and the target-language/domain match of the generator model, IndoT5-base-paraphrase, trained primarily on general-domain paraphrase corpora (PAWS/ParaBank-style data) rather than dense scientific text, appears especially prone to the context-exhaustion failure mode described here. This domain-mismatch hypothesis distinguishes the present anomaly from the "model signature" concept as originally framed for large decoder-only LLMs [12, 14], and suggests that detection strategies calibrated for ChatGPT-style generative text may not transfer directly to paraphrase-model-generated text.

Comparison with Prior Work and Theoretical Implications

To contextualize the zero-shot classifier's performance, the Random Forest baseline was evaluated on the same 2,274 document corpus. **Table 13** compares its performance against the best performing zero-shot configuration.

Table 13. Comparison Between Zero-Shot NLI and Supervised Baseline (Random Forest)

Method	Accuracy	Precision	Recall	F1-Score
Zero-Shot mDeBERTa v3 (Scenario 1-Aspect)	53.52%	52.41%	76.52%	0.6221
Random Forest (7 linguistic features, 80/20 split)	91.21%	90.13%	92.51%	0.9130

The supervised baseline substantially outperformed the best zero-shot configuration, achieving 91.21% test accuracy (F1 = 0.9130). Feature importance analysis further corroborates the Mann-Whitney findings, demonstrating that the features proven to be significantly different through nonparametric testing are precisely the ones that dominate the Random Forest classifier's decision-making. This convergence validates the practical utility of the Mann-Whitney analysis, showing that its statistical insights effectively translate into highly discriminative features for supervised learning. This large performance gap indicates that the seven linguistic features, when combined through a supervised, feature-weighted decision boundary, are considerably more discriminative than a single natural-language hypothesis evaluated in isolation under a zero-shot NLI framework. This result reinforces two conclusions: first, that the linguistic anomaly identified in this study is a genuinely strong and learnable signal, not a marginal statistical artifact, and second, that the current zero-shot approach substantially underutilizes this signal relative to what a lightweight

supervised model can achieve using the very same features directly motivating the hybrid mDeBERTa v3 + Random Forest architecture proposed in the Conclusion as a concrete, evidence-backed next step rather than a speculative suggestion.

4. Conclusion

This study evaluated the effectiveness of a zero-shot Natural Language Inference (NLI) approach using the mDeBERTa v3 model to detect Indonesian scientific abstracts specifically synthesized via IndoT5-base-paraphrase, not AI-generated text in general. Based on a balanced corpus of 2,274 abstracts (1,137 human-written and 1,137 IndoT5-synthesized), the study's central contribution is the identification of a structural anomaly in IndoT5-synthesized text, abnormally high sentence-length fluctuation, traced to context-window exhaustion during paraphrase generation, and the demonstration that zero-shot mDeBERTa v3 can partially track this anomaly through a single, precisely targeted instruction. However, the model cannot yet reliably separate this machine-induced instability from natural human sentence-length variation, and the resulting false-positive rate (790 of 1,137 human abstracts under the best-performing configuration) makes the current approach unsuitable for standalone deployment in academic-integrity screening. Three specific conclusions follow.

First, zero-shot classification with mDeBERTa v3 is able to distinguish human-written from AI-synthesized Indonesian scientific abstracts without any task-specific fine-tuning, but its overall effectiveness remains limited, reaching a peak accuracy of only 53.52% under the best-performing instruction configuration.

Second, the disentangled attention mechanism of mDeBERTa v3 is genuinely effective at capturing the statistical "model signature" left by IndoT5-base-paraphrase, provided the NLI instruction is precisely targeted. When the classifier was directed at the specific anomaly identified through Mann-Whitney U testing, abnormally high sentence-length fluctuation, a by-product of context-window exhaustion during paraphrase generation, Recall surged to reach 76.52% (Scenario 1-Aspect), confirming that a single, well-targeted hypothesis outperforms broader multi-aspect instructions, which instead suffered from attention dilution as complexity increased (Recall falling to 46.26% and 46.66% for the 3-Aspect and 5-Aspect scenarios, respectively).

Third, the model's reliability in minimizing classification errors remains low. The same sensitivity that enabled high Recall in Scenario 1-Aspect also caused the classifier to conflate mechanical AI instability with the natural sentence-length variation deliberately used by human writers for readability, resulting in 790 human abstracts being misclassified as AI-generated (False Positive). This limitation, combined with the near-identical core vocabulary shared by both classes, indicates that sentence-length fluctuation alone, while statistically significant at the corpus level is not yet a sufficiently precise standalone signal for reliable document-level detection.

Overall, these findings demonstrate that zero-shot NLI can meaningfully track a specific AI mechanical anomaly with a single, carefully targeted instruction, but currently lacks the precision required for standalone deployment in academic-integrity screening. Given the high risk of false accusation demonstrated in this study, the current model must not be used as a sole basis for academic-integrity decisions, any deployment should position it strictly as a supplementary screening signal requiring human review, not an automated verdict. Future work should explore fine-tuning mDeBERTa v3 on large-scale labeled Indonesian abstract corpora to build task-specific computational memory, extend testing to untruncated abstracts and more diverse scientific domains to validate real-world robustness, and adopt a hybrid detection architecture that pairs mDeBERTa v3's deep contextual analysis with numerical classifiers such as Random Forest or Support Vector Machine (SVM), using linguistic-feature statistics (comma ratio, sentence-length standard deviation) as a complementary second-stage discriminator to reduce false positives and produce a more balanced and trustworthy AI-text detection system.

References:

[1] J. G. Meyer *et al.*, “ChatGPT and large language models in academia: opportunities and challenges,” *BioData Min.*, vol. 16, no. 1, pp. 1–11, 2023, doi: https://doi.org/10.1186/s13040-023-00339-9.

[2] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,” *Innov. Educ. Teach. Int.*, vol. 61, no. 2, pp. 228–239, 2024, doi:

https://doi.org/10.1080/14703297.2023.2190148.

[3] F. M. Howard, A. Li, M. F. Riffon, and E. Garrett-mayer, "Characterizing the Increase in Artificial Intelligence Content Detection in Oncology Scienti fi c Abstracts From 2021 to 2023," 2023, doi: https://doi.org/10.1200/CCI.24.00077.

[4] P. C. Theocharopoulos, P. Anagnostou, A. Tsoukala, S. V. Georgakopoulos, S. K. Tasoulis, and V. P. Plagianakos, "Detection of Fake Generated Scientific Abstracts," Apr. 2023, doi: https://doi.org/10.1109/BigDataService58306.2023.00011.

[5] C. A. Gao *et al.*, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *npj Digit. Med.*, vol. 6, p. 75, 2023, doi: https://doi.org/10.1038/s41746-023-00819-6.

[6] O. Marsh, M. Yates, A. Rowe, and M. Song, "Adversarial Attacks and Robustness in LLM- Generated Text Detection Systems," 2025.

[7] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, and D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99\% accuracy using off-the-shelf machine learning tools," *Cell Reports Phys. Sci.*, vol. 4, no. 6, 2023.

[8] A. Alikhanov *et al.*, "AI Generated Text Detection," Jan. 2026

[9] P. He, J. Gao, and W. Chen, "Debertav3: Improving Deberta Using Electra-Style Pre-Training With Gradient-Disentangled Embedding Sharing," *11th Int. Conf. Learn. Represent. ICLR 2023*, no. Mlm, pp. 1–16, 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[11] W. Yin, J. Hay, and D. Roth, "Benchmarking Zero-shot Text Classification : Datasets , Evaluation and Entailment Approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 3914–3923.

[12] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," *Proc. Mach. Learn. Res.*, vol. 202, pp. 24950–24962, 2023.

[13] W. X. Zhao *et al.*, "A Survey of Large Language Models," Mar. 2026, [Online]. Available: http://arxiv.org/abs/2303.18223

[14] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, Sep. 2023.

[15] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-Generated Text be Reliably Detected?," pp. 1–37, Jan. 2025.

[16] M. Hahn, "Theoretical Limitations of Self-Attention in Neural Sequence Models," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 156–171, Dec. 2020, doi: https://doi.org/10.1162/tacl_a_00306.

[17] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[18] L. Lukman *et al.*, "Proposal of the S-score for measuring the performance of researchers, institutions, and journals in Indonesia," *Sci. Ed.*, vol. 5, no. 2, pp. 135–141, 2018, doi: https://doi.org/10.6087/KCSE.138.

[19] H. de Sompel and M. L. Nelson, "Reminiscing about 15 years of interoperability efforts," *D-Lib Mag.*, vol. 21, no. 11/12, 2015, doi: https://doi.org/10.1045/november2015-vandesompel.

[20] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[21] M. Freitag and Y. Al-Onaizan, "Beam Search Strategies for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*, T. Luong, A. Birch, G. Neubig, and A. Finch, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2017, pp. 56–60. doi:

https://doi.org/10.18653/v1/W17-3207.

[22] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," pp. 1–18, Sep. 2019.

[23] R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," *CoRR*, vol. abs/1705.0, Nov. 2017.

[24] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, vol. abs/1609.0, Oct. 2016.

[25] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014, doi: https://doi.org/10.1016/j.ipm.2013.08.006.

[26] B. Krawczyk, "Learning from imbalanced data : open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016, doi: https://doi.org/10.1007/s13748-016-0094-0.

[27] A. Field, *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

[28] M. Mars, "From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough," *Appl. Sci.*, vol. 12, no. 17, 2022, doi: https://doi.org/10.3390/app12178805.

[29] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI," *Polit. Anal.*, vol. 32, no. 1, pp. 84–100, Jan. 2024, doi: https://doi.org/10.1017/pan.2023.20.

[30] N. F. Liu, K. Lin, J. Hewitt, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle : How Language Models Use Long Contexts," 2023.

[31] I. V Pantic and S. Mugosa, "Artificial intelligence strategies based on random forests for detection of AI-generated content in public health," *Public Health*, vol. 242, pp. 382–387, May 2025, doi: https://doi.org/10.1016/j.puhe.2025.03.029.

[32] S. K. Aityan, W. Claster, K. S. Emani, S. Rais, and T. Tran, "A Lightweight Approach to Detection of AI-Generated Texts Using Stylometric Features," Jan. 2026.

[33] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-Enhanced Bert With Disentangled Attention," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[34] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440–8451.

[35] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge: Cambridge University Press, 2023.

[36] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, 2025, doi: https://doi.org/10.1186/s40537-025-01313-4.

[37] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, vol. 27, no. 4, pp. 4023–4031, 2024, doi: https://doi.org/10.53555/ajbr.v27i4s.4345.