Journal of Computer Science

Research Article

Detecting Novelty in Indonesian Research Grant Proposals Using Machine Learning and Natural Language Processing

Walanstar Alimcan Sitorus and Evaristus Didik Madyatmadja

School of Information Systems, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 20-02-2026
Revised: 09-07-2026
Accepted: 15-07-2026

Corresponding Author:
Walanstar Alimcan Sitorus
School of Information Systems,
Bina Nusantara University,
Jakarta, Indonesia
Email: walanstar@gmail.com

Abstract: Assessing the novelty of research proposals represents a critical challenge in large-scale research grant management, particularly when the number of submitted proposals is substantial and topic diversity continues to increase. This study proposes an analytical approach to detect the novelty of Indonesian academic research proposal titles by leveraging semantic similarity based on Natural Language Processing (NLP). The CRISP-DM framework is adopted as the research methodology, with a comparative evaluation of three text representation approaches: TF-IDF, Latent Semantic Analysis (LSA), and IndoBERT. The dataset comprises 52,084 research proposal titles submitted between 2020 and 2024, which, after administrative deduplication, resulted in 44,429 unique titles. Inter-title similarity is computed using cosine similarity, followed by similarity distribution analysis and the application of a Top-K nearest neighbors strategy. In this study, novelty is conceptualized as novelty risk, defined as an analytical indicator of relatively low novelty derived from Top-1 similarity values and patterns of semantic similarity distribution. The results demonstrate that IndoBERT exhibits higher semantic sensitivity and a more stable similarity distribution compared to TF-IDF and LSA. Novelty risk estimation is conducted in an aggregated manner using rule-based analytical interpretation, without manual labeling or automated decision-making. The proposed approach functions as a decision support system to assist in the pre-screening of research proposals, while preserving the central role of expert reviewers in qualitative evaluation.

Keywords: Novelty Detection, Research Proposal, Latent Semantic Analysis, IndoBERT, Semantic Similarity

Introduction

In recent years, research activities conducted by university lecturers in Indonesia have grown substantially, reflecting national efforts to strengthen the research and innovation ecosystem. This growth has been supported by various competitive research grant schemes administered by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Information System (BIMA). As a centralized platform, BIMA manages a large volume of research proposals submitted annually by academic institutions across the country.

Between 2020 and 2024, more than 52.000 research proposal titles were funded under multiple grant schemes, including basic research, applied research, early-career researcher grants, and doctoral dissertation research. While this expansion indicates increasing research participation, it also introduces practical challenges in the proposal evaluation process. In particular, assessing the novelty of proposed research ideas becomes increasingly difficult when reviewers are required to evaluate thousands of proposals within limited timeframes.

In current practice, novelty assessment relies primarily on manual review. Although existing systems are able to detect identical titles and verify administrative completeness, they are not designed to identify semantic similarity between proposal titles. As a result, proposals with closely related research themes may be funded under different schemes or across different years, potentially leading to thematic overlap and reduced efficiency in research funding allocation.

SCIENCE Publications

© 2026 Walanstar Alimcan Sitorus and Evaristus Didik Madyatmadja. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Advances in artificial intelligence, especially in Natural Language Processing (NLP), offer opportunities to support large-scale text analysis in research management systems. Recent studies have shown that semantic similarity analysis can be used to capture meaningful relationships between texts beyond surface-level lexical overlap. However, most existing approaches focus on benchmark datasets or small-scale experiments and are rarely examined in the context of real-world research proposal management at a national scale.

This study addresses this gap by investigating the use of semantic similarity analysis as an analytical indicator to support early-stage evaluation of research proposal titles. Rather than treating novelty as a binary outcome, this work introduces the concept of novelty risk, defined as the relative likelihood that a proposal exhibits low novelty based on its semantic similarity to previously funded proposals. The proposed approach operates in an unsupervised manner and is intended to assist preliminary screening without replacing substantive evaluation by expert reviewers.

To examine this approach, three text representation models with different levels of semantic expressiveness are evaluated: TF-IDF, Latent Semantic Analysis (LSA), and IndoBERT. All models are assessed using a consistent similarity measurement framework to ensure comparability. The analysis is conducted on a large-scale dataset of funded research proposal titles, allowing empirical observation of model behavior, computational trade-offs, and their suitability for deployment in a decision support context.

The main contributions of this study are twofold. First, it provides a comparative analysis of TF-IDF, LSA, and IndoBERT for measuring semantic similarity among research proposal titles in a large-scale Indonesian funding dataset. Second, it demonstrates how similarity-based analysis can be used as an interpretable screening mechanism to identify titles that merit closer review for potential thematic overlap. These contributions are relevant for improving the efficiency and consistency of proposal evaluation in national research funding systems.

Literature Review

The digital transformation of research grant governance has significantly reshaped national research funding ecosystems. In Indonesia, the Basis Informasi Penelitian dan Pengabdian kepada Masyarakat (BIMA), managed by the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek), serves as the primary platform for administering thousands of lecturer research proposals across higher education institutions. This system supports transparency and efficiency in proposal submission, evaluation, and research outcome reporting (Kemristekdikti, 2019). Despite its critical role in digitising national research management, the BIMA system remains limited in its analytical capabilities, particularly in detecting semantic similarity among research proposal titles. Currently, BIMA is able to identify exact textual duplicates and perform administrative verification; however, it is not designed to recognise semantic similarities between titles that differ lexically but convey conceptually similar research ideas. Consequently, research proposals with overlapping themes may be funded under different schemes or across different funding periods, potentially leading to inefficiencies in national research fund allocation.

In the literature on text similarity detection, there is no universally accepted standard or absolute threshold for determining whether two texts should be considered similar. Previous studies emphasise that similarity thresholds are inherently contextual and must be defined empirically according to analytical objectives and dataset characteristics (Ghosal et al., 2022). Similarity scores are therefore better interpreted as indicators of semantic proximity rather than rigid decision boundaries. In this context, similarity measures function as analytical tools to support decision-making, rather than as automatic indicators of duplication or plagiarism.

This study should also be positioned in relation to three adjacent research streams: Plagiarism detection, semantic duplicate detection, and novelty detection. Plagiarism detection is primarily concerned with identifying unauthorized textual reuse or close reproduction across documents, often with emphasis on lexical, structural, or semantic evidence of copying (Foltýnek et al., 2020). Semantic duplicate or near-duplicate detection, by contrast, focuses on identifying texts that are highly similar in meaning, even when they differ in wording or surface form. Novelty detection addresses a different question, namely whether a text introduces information that is relatively new with respect to previously observed content or existing knowledge (Ghosal et al., 2022). The present study does not aim to detect plagiarism, nor does it classify proposal titles as exact duplicates in a strict retrieval sense. Instead, semantic similarity is used as an analytical basis for identifying novelty risk, namely the possibility that a proposal title may be thematically too close to previously funded proposals to warrant closer review. In this sense, the study is positioned closer to similarity-assisted novelty screening than to plagiarism checking or duplicate-document detection (Foltýnek et al., 2020; Ghosal et al., 2022; Ihnaini et al., 2024). Although prior studies have examined semantic similarity, novelty-related analysis, and duplicate detection in various textual settings, fewer studies have addressed title-level semantic screening in large-scale national research proposal evaluation, particularly in the Indonesian-language context.

Most prior studies on text similarity analysis have focused on longer textual documents, such as scientific

articles, policy reports, or research abstracts (Ghosal et al., 2022). Classical approaches, including Term Frequency–Inverse Document Frequency (TF-IDF) and Latent Semantic Analysis (LSA), have been widely applied to measure document similarity and semantic relationships between terms (Deerwester et al., 1990; Salton and Buckley, 1988). While these methods have demonstrated effectiveness in large-scale document analysis, they rely primarily on lexical overlap or linear semantic representations, which limits their ability to capture contextual meaning, particularly in short-text scenarios such as research proposal titles.

To address these limitations, more recent studies have adopted transformer-based models, such as BERT, which significantly improve semantic representation by capturing contextual relationships between words (Devlin et al., 2019). For the Indonesian language, IndoBERT, a language-specific variant pretrained on large-scale Indonesian corpora, has demonstrated strong performance in understanding semantic relationships within natural linguistic contexts (Wilie et al., 2020). However, large-scale empirical studies that systematically apply and compare TF-IDF, LSA, and IndoBERT for semantic similarity detection in Indonesian research proposal titles remain limited. Although semantic similarity and novelty-related text analysis have been widely studied in NLP (Ghosal et al., 2022), the present study does not claim methodological novelty in the sense of introducing a new detection model or algorithm. Instead, its contribution lies in the application and comparative evaluation of TF-IDF, LSA, and IndoBERT for large-scale Indonesian research proposal titles. In this respect, the study is positioned as an applied and context-specific contribution, with particular relevance for research management systems and decision-support for novelty risk screening.

Scalability is another key challenge in large-scale similarity analysis. Exhaustive pairwise similarity computation becomes infeasible as dataset size increases. To mitigate this issue, several studies employ nearest neighbour techniques, particularly the Top-K nearest neighbour approach, which focuses analysis on the most relevant similarity pairs while reducing computational cost (Hanafi and Saadatfar, 2022). Efficient similarity search libraries, such as FAISS, further support large-scale semantic similarity analysis by enabling rapid vector-based search in high-dimensional spaces (Johnson et al., 2021).

Recent studies in novelty detection increasingly conceptualise novelty as a relative and context-dependent construct rather than a binary attribute. In the context of higher education research governance, novelty and originality are fundamental principles embedded within national academic regulations and research quality frameworks (Kemdikbud, 2020; Pusat, 2014). From an analytical perspective, novelty can be approximated through semantic similarity or semantic distance relative to existing knowledge, particularly in settings where labelled novelty data are unavailable (Reimers and Gurevych, 2019). Transformer-based semantic representations have been shown to effectively capture contextual semantic differences, making them suitable as analytical indicators of novelty in short-text scenarios (Ghosal et al., 2022).

It is important to distinguish semantic similarity from novelty. Semantic similarity refers to the degree of closeness in meaning between text representations, whereas novelty concerns the extent to which a text introduces information, combinations of ideas, or thematic content that is new relative to prior knowledge or previously observed documents. In this sense, high similarity may indicate thematic overlap, but it does not automatically imply the absence of novelty. Likewise, a proposal may still be novel in its contribution even when it shares conceptual vocabulary with prior work. As noted by Ghosal et al. (2022), novelty is fundamentally a relational concept defined against what has already been seen or known. At the same time, (Ihnaini et al., 2024) describe semantic similarity as a measure of meaning-related proximity rather than originality itself. In the context of scientific texts, (Jeon et al., 2023) also show that novelty can be operationalized in relation to prior textual patterns, but should not be reduced to similarity alone. Therefore, the present study does not treat similarity as equivalent to novelty, but rather uses similarity patterns as an analytical basis for identifying potential novelty risk in research proposal titles.

Based on this body of work, the present study positions novelty as an analytical decision support indicator rather than as an automated evaluative judgement. By comparatively analysing lexical (TF-IDF), latent semantic (LSA), and contextual semantic (IndoBERT) approaches for detecting semantic similarity among Indonesian lecturer research proposal titles, this study aims to provide empirical insights that support scalable and objective novelty assessment. The findings are expected to contribute both academically, by enriching NLP research for the Indonesian language, and practically, by informing the development of analytical decision support features within the BIMA system.

Materials and Methods

This study employs an analytical, data-driven, and unsupervised research design to examine semantic similarity among Indonesian research proposal titles as an indicator of novelty. The methodological workflow follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which provides a systematic structure for large-scale text analysis, from problem formulation to analytical deployment recommendations, as shown in Fig. 1.

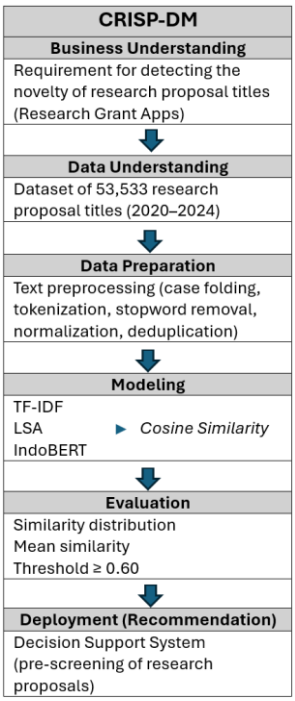

Fig. 1: CRISP-DM framework for research proposal title novelty detection

Research Design and Framework

The research is designed as an analytical study focusing on large-scale textual data without supervised labels or predefined ground truth. Novelty is not treated as a binary outcome but as a relative analytical indicator derived from semantic similarity patterns. The CRISP-DM framework is adopted to ensure methodological rigor and transparency across six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This framework is suitable for text-based analytics involving large datasets and supports reproducibility of results.

Research Procedures

The research procedure was conducted sequentially, beginning with the collection of funded research proposal title data for the 2020–2024 period. This stage was followed by data understanding and exploratory analysis to examine the distribution of proposals across funding schemes and time periods. Subsequently, text preprocessing and administrative deduplication were performed to ensure consistency and quality of the textual data prior to analysis. The preprocessed data were then used to construct text representations and to perform semantic similarity modelling using TF-IDF, Latent Semantic Analysis (LSA), and IndoBERT. The similarity analysis results were evaluated through descriptive distribution analysis and Top-K nearest neighbor strategies to estimate novelty risk analytically. Finally, based on the evaluation outcomes, recommendations were formulated for the deployment of an analytical novelty detection module as a decision support component to support the pre-screening stage of research proposal evaluation.

Dataset and Data Understanding

The dataset consists of 52,084 funded research proposal titles submitted by Indonesian university lecturers between 2020 and 2024, covering 17 national research funding schemes. Each record includes the proposal title, funding year, funding scheme, and research focus area. An initial data exploration was conducted to examine proposal distribution across schemes and funding periods. The analysis revealed substantial variation in submission volume across schemes, indicating differences in thematic concentration and potential overlap of research ideas. All records contain complete title information. Therefore, no missing textual data were identified at the title level, and no imputation procedures were required.

Data Preparation and Deduplication

Data preparation was performed to ensure textual consistency prior to similarity analysis. Preprocessing steps included case folding, tokenisation, Indonesian stopword removal, and basic word normalisation. These steps were applied uniformly to all titles to reduce irrelevant linguistic variation while preserving semantic content. Administrative deduplication was conducted using a multi_year indicator to identify titles submitted by the same researcher across multiple funding years. Such entries were treated as valid administrative repetitions rather than research duplication and were excluded from the main similarity analysis. Exact string-based deduplication after normalisation reduced the dataset from 52,084 to 44,429 unique proposal titles.

Text Representation and Similarity Modeling

Text representation and similarity modeling in this study were conducted using three Natural Language Processing approaches applied in parallel to capture different levels of semantic information in research proposal titles. The selected methods Term Frequency–Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), and IndoBERT represent lexical, mathematical semantic, and contextual semantic modeling paradigms, respectively. TF-IDF was employed as a lexical baseline model due to its computational efficiency and widespread use in large-scale text analysis. This approach represents each proposal title as a weighted vector based on term frequency relative to the entire corpus, enabling rapid detection of lexical overlap, although it remains limited in capturing implicit semantic

relationships (Saadatfar et al., 2020; Xiao et al., 2025). To extend similarity analysis beyond direct word matching, Latent Semantic Analysis (LSA) was applied by performing dimensionality reduction on the TF-IDF matrix using Singular Value Decomposition (SVD). This process allows latent semantic structures to be represented in a lower-dimensional vector space, enabling titles with different lexical expressions but similar thematic content to exhibit higher similarity scores (Saadatfar et al., 2020). In addition to these classical approaches, IndoBERT was utilised as a transformer-based contextual semantic model pretrained on large-scale Indonesian language corpora. Each proposal title was encoded as a fixed-length sentence embedding that captures contextual meaning based on the relationships among words within the title. This representation enables the detection of semantic similarity that cannot be adequately captured by lexical or classical semantic models, particularly for short texts such as research proposal titles. The use of IndoBERT aligns with recent findings demonstrating the effectiveness of BERT-based models in semantic textual similarity tasks for the Indonesian language (Krisnawati et al., 2024). By applying these three representation approaches in parallel, the study enables a comprehensive comparison of semantic sensitivity and computational characteristics across different modeling paradigms, providing an empirical basis for novelty risk estimation and system-level recommendation.

Similarity Measurement and Data Analysis Techniques

Cosine similarity was used consistently across all representation models to measure inter-title semantic similarity. This metric is independent of text length and well-suited for vector-based representations. A similarity threshold of cosine similarity ≥0.60 was adopted as an analytical indicator to flag potential semantic overlap. This threshold was defined empirically and contextually, following established practices in text similarity and near-duplicate detection research, and is used as an early warning indicator rather than a strict decision boundary (Indasari and Tjahyanto, 2024).

Threshold Justification

The cosine similarity threshold of 0.60 was used in this study as a screening threshold to detect proposal titles with potentially meaningful semantic overlap, rather than as a strict decision boundary for determining novelty. In semantic similarity research, threshold selection is generally influenced by the purpose of the analysis, the characteristics of the text, the language, and the representation model used (Scalvini and Mashaghi, 2025; Brglez, 2023; Fellah, 2021). This consideration is particularly relevant in the present study, where the input consists of short and domain-specific proposal titles. In such cases, a moderate threshold is more appropriate for detecting thematic proximity than for making definitive judgments about originality.

This interpretation is consistent with previous studies. Scalvini and Mashaghi, (2025) used a cosine similarity threshold of 0.60 to capture a moderate degree of thematic or contextual overlap. Brglez (2023) showed that optimal cosine thresholds may vary substantially across languages, highlighting the importance of context-sensitive threshold selection. (Fellah, 2021) also emphasized that threshold selection is an important factor in similarity-based detection performance. Taken together, these findings support the use of 0.60 as a practical analytical threshold for identifying titles that merit closer inspection.

Accordingly, proposal titles with cosine similarity scores of 0.60 or higher were not treated as automatically non-novel, but rather as flagged cases indicating possible thematic closeness to previously funded proposals. In this sense, the threshold functions only as an early detection mechanism for novelty risk and serves as a decision-support aid for further human review.

In this study, semantic similarity was used as a proxy for novelty risk, not as a direct measure of novelty. The purpose of the similarity analysis was to detect titles that are semantically close to previously funded proposals and therefore may warrant further review for possible thematic overlap. This approach follows the view of Ghosal et al. (2022) that novelty should be understood relative to prior knowledge, rather than inferred from similarity scores alone. It is also consistent with the broader understanding of semantic similarity as a measure of closeness in meaning (Ihnaini et al., 2024) and with prior work showing that textual novelty in scientific publications must be interpreted in relation to earlier textual patterns and scholarly context (Jeon et al., 2023). Accordingly, the similarity scores reported in this study should be interpreted as analytical indicators for screening and decision support, rather than as definitive judgments of whether a proposal is novel or non-novel.

Descriptive Similarity Distribution Analysis

Descriptive statistical analysis was conducted to examine the distribution of similarity scores generated by each model. The analysis focused on measures such as mean similarity, median similarity, and the proportion of title pairs exceeding the similarity threshold. This distribution-based analysis provides insight into the overall semantic overlap patterns across the dataset and supports the interpretation of novelty risk at an aggregate level. No inferential statistical testing was performed, as the study is exploratory and unsupervised in nature.

Top-K Nearest Neighbor Strategy

To manage computational complexity in large-scale similarity analysis, a Top-K nearest neighbor strategy was employed. For each proposal title, only the K most similar

previously funded titles were retained for further analysis. In this study, K was fixed at 10 for all three representation methods (TF-IDF, LSA, and IndoBERT) to ensure methodological consistency and comparability across models. Because each title naturally appears as its own nearest match when retrieval is performed on the same corpus, K + 1 neighbors were initially retrieved and the self-match was subsequently removed, leaving the 10 nearest neighboring titles for analysis. A value of K = 10 was considered sufficient to retain the most relevant neighbouring titles for each proposal while avoiding unnecessary expansion of less informative similarity pairs, thereby providing a practical balance between analytical coverage and computational efficiency. This strategy reduces computational burden while preserving the most semantically relevant similarity relationships.

The use of Top-K retrieval is consistent with prior research on efficient nearest-neighbor search and large-scale similarity analysis. (Halder et al., 2024) emphasize that the choice of K is an important parameter in k-nearest neighbor methods because it influences sensitivity to local patterns, robustness to noise, and computational efficiency. In addition, recent studies on large-scale similarity search systems have highlighted the importance of efficient nearest-neighbor retrieval for high-dimensional data, particularly when exhaustive pairwise comparisons become increasingly expensive as dataset size grows (Huici et al., 2025). For the IndoBERT representation, FAISS was used to perform efficient similarity search in the dense embedding space, thereby avoiding the need to construct a full pairwise similarity matrix across all proposal titles (Douze et al., 2026).

Therefore, the study adopts an unsupervised approach to novelty risk identification based on semantic similarity scores. Because no manually annotated novelty labels were available, the study was designed as an unsupervised similarity-based screening framework rather than as a supervised novelty classification task.

Results and Discussion

This section presents the results of the semantic similarity analysis of Indonesian research proposal titles and discusses their implications for estimating novelty risk as a decision-support indicator. The findings are described at an aggregate level, reflecting the unsupervised design of the study. Rather than generating automated decisions, the analysis aims to provide analytical insights that can assist evaluators in assessing proposal originality.

Dataset Characteristics

The dataset analyzed in this study consists of 52,084 funded research proposal titles submitted between 2020 and 2024. After administrative deduplication and data cleaning, a total of 44,429 unique proposal titles were retained for analysis, as shown in Table 1.

The dataset spans 17 national research funding schemes with varying thematic scopes and submission volumes, thereby representing real-world conditions of national-scale research grant management. The distribution of proposal titles across funding schemes shows substantial variation in submission intensity. Schemes targeting early-career researchers and fundamental research account for a large proportion of proposals, while applied and strategic schemes exhibit relatively lower volumes. This variation provides essential contextual information for interpreting semantic similarity patterns and novelty risk across schemes.

Results of Text Similarity Analysis

Semantic similarity analysis was performed using a Top-K nearest neighbor strategy with cosine similarity across three text representation approaches: TF-IDF, Latent Semantic Analysis (LSA), and IndoBERT. As the study adopts an unsupervised design, the evaluation focuses on descriptive comparison of similarity characteristics rather than accuracy against ground truth labels, as shown in Table 2.

The results demonstrate clear differences among the three approaches, as shown in Table 2 TF-IDF produces low average similarity values, reflecting its reliance on lexical overlap and its limited ability to capture semantic relationships beyond shared vocabulary LSA yields higher similarity values than TF-IDF, indicating improved sensitivity to latent thematic relationships.

Table 1: Dataset Summary Before and After Administrative Deduplication

Description	Number of Records
Initial dataset (funded proposal titles, 2020–2024)	52,084
Duplicate titles identified (administrative repetition/Multi Years)	7,655
Final dataset after deduplication	44,429

Table 2: Average cosine similarity scores across text representation models

Model	Mean Top-K Similarity	Median Top-K Similarity	% Top-K ≥ 0.60	Mean Top-1 Similarity	% Top-1 ≥ 0.60
TF-IDF	0.318	0.302	0.88	0.421	6.18
LSA	0.666	0.657	71.57	0.747	92.97
Indo BERT	0.806	0.809	99.92	0.833	99.98

IndoBERT produces the highest and most stable similarity values, highlighting its effectiveness in capturing contextual semantic similarity in short texts such as research proposal titles. These findings suggest that the choice of text representation model substantially influences the observed similarity patterns and, consequently, the interpretation of novelty risk.

Similarity Distribution Results

The high concentration of IndoBERT similarity scores should not be interpreted solely as evidence of superior semantic sensitivity. It may also reflect a compression effect in the embedding space, where many short proposal titles are mapped into a relatively narrow similarity range. As a result, IndoBERT appears highly effective for capturing broad thematic relatedness, but it may provide less dispersion for distinguishing finer-grained novelty differences among titles. This interpretation is consistent with the heatmap analysis, in which IndoBERT shows denser off-diagonal similarity patterns than TF-IDF and LSA. Therefore, the high-score concentration of IndoBERT should be understood both as a strength for detecting semantic proximity and as a limitation when similarity values are used as a proxy for novelty risk.

An analysis of cosine similarity score distributions was conducted to examine overall semantic overlap patterns across the dataset, as shown in Fig. 2.

TF-IDF similarity scores are concentrated at lower ranges, indicating sparse lexical overlap among proposal titles. LSA similarity scores are distributed around mid-range values, reflecting its ability to capture latent semantic relationships while remaining constrained by linear vector representations. In contrast, IndoBERT similarity scores are predominantly concentrated at higher ranges with relatively low variance. The very high similarity concentration observed in IndoBERT should not be interpreted solely as stronger semantic precision. It may also indicate a compression effect in the embedding space, where many short proposal titles are mapped into a relatively narrow similarity range.

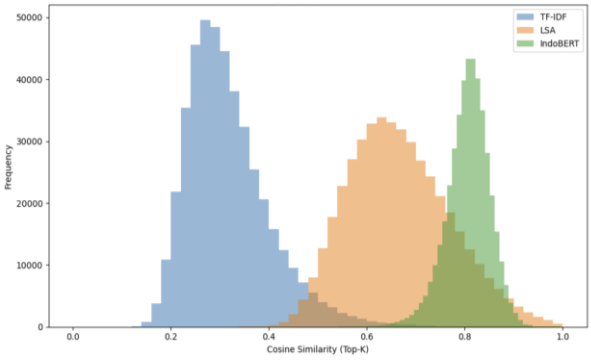

Fig. 2: Distribution of cosine similarity scores for TF-IDF, LSA, and IndoBERT models

In practical terms, this means that IndoBERT is highly sensitive to broad thematic relatedness, but may provide less dispersion for distinguishing finer-grained novelty differences among titles. This pattern is consistent with the heatmap results, where IndoBERT shows denser off-diagonal similarity values than TF-IDF and LSA. Therefore, while IndoBERT appears effective for detecting general semantic proximity, its high-score concentration should be interpreted cautiously when used for novelty-risk screening. This distribution suggests that transformer-based contextual embeddings are more effective in identifying semantic proximity among research proposal titles, even when lexical forms differ significantly. The distribution analysis provides an empirical basis for defining analytical thresholds and supports novelty risk interpretation without relying on binary novelty classification.

Fig. 3 shows the distribution of Top-1 similarity scores across TF-IDF, LSA, and IndoBERT. The boxplot indicates a clear shift in the central tendency and spread of similarity values across representation methods. TF-IDF produces the lowest median similarity scores, reflecting its stronger dependence on explicit lexical overlap. LSA yields substantially higher scores, suggesting that latent semantic structure captures broader thematic relatedness beyond surface-level wording. IndoBERT exhibits the highest median similarity values and a comparatively concentrated distribution in the upper range, indicating that contextual embeddings generate denser semantic proximity among proposal titles. These patterns are consistent with the histogram-based results and further confirm that representation choice strongly affects how thematic similarity is quantified.

Heatmap-Based Comparison of Similarity Patterns

To provide a more intuitive comparison of local similarity structure across representation methods, heatmap-based visualization was applied to a subset of 10 sampled proposal titles. Fig. 4 presents pairwise cosine similarity heatmaps generated using TF-IDF, LSA, and IndoBERT representations for the same sampled titles, allowing direct visual comparison across models.

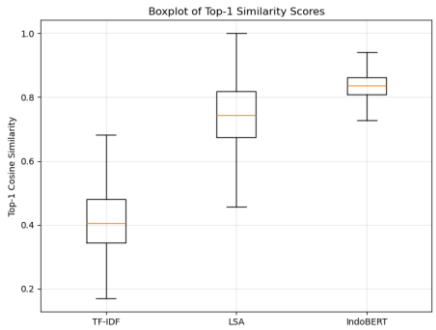

Fig. 3: Distributional comparison of Top-1 cosine similarity scores across TF-IDF, LSA, and IndoBERT representations

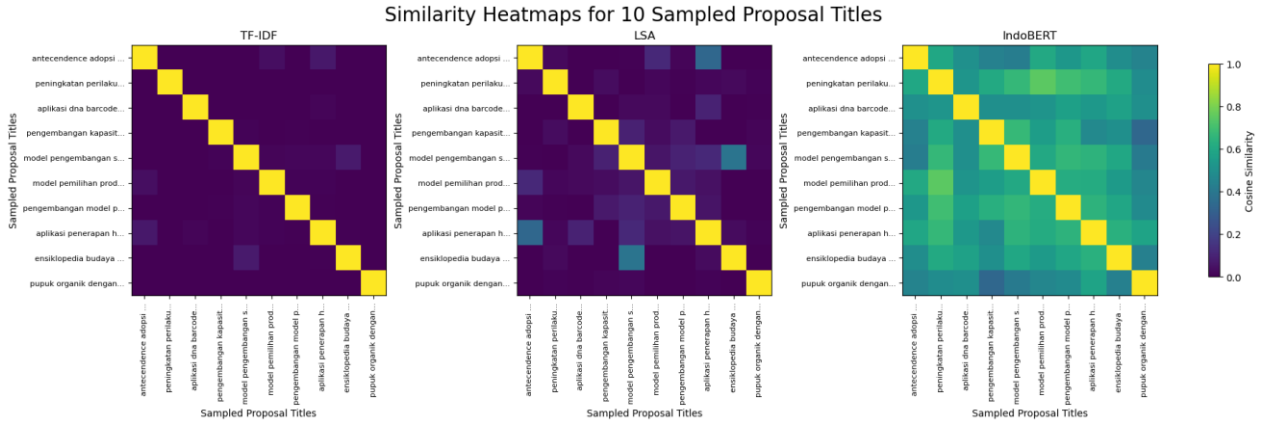

Fig. 4: Similarity heatmaps of 10 sampled proposal titles generated using TF-IDF, LSA, and IndoBERT representations

The TF-IDF heatmap shows that most off-diagonal values remain close to zero, indicating that lexical-based matching captures only limited overlap among titles. This pattern suggests that TF-IDF is highly selective and primarily sensitive to direct word-level similarity. As a result, titles that do not share many surface-level terms are less likely to be identified as strongly related under this representation. The LSA heatmap reveals a moderate increase in off-diagonal similarity values, suggesting that latent semantic relationships begin to emerge after dimensionality reduction. Compared with TF-IDF, LSA produces a richer local similarity structure, indicating that semantically related titles can still be identified even when explicit lexical overlap is limited. This intermediate pattern is consistent with the role of LSA in capturing broader thematic associations beyond exact term matching. Among the three methods, IndoBERT displays the densest similarity pattern, with consistently higher off-diagonal values across many title pairs. This finding indicates that IndoBERT is more effective in capturing broad semantic relatedness between proposal titles. At the same time, the relatively uniform high-similarity pattern suggests a narrower discriminative range compared with TF-IDF and LSA. In this sense, IndoBERT appears to cluster titles more strongly in the semantic space, but the dense visual structure also indicates that similarity values are more compressed for this corpus.

Overall, the heatmap comparison complements the quantitative findings reported earlier. TF-IDF produces a sparse similarity structure, LSA provides an intermediate representation with moderate semantic grouping, and IndoBERT yields the strongest and most consistent semantic clustering. These visual differences confirm that the choice of text representation substantially affects how semantic proximity is captured in large-scale proposal title analysis.

Illustrative Examples of Similarity Scores

To complement the aggregate similarity distributions, several representative title pairs were examined qualitatively in order to clarify how similarity scores should be interpreted across different representation methods. Table 3 presents three illustrative examples corresponding to lower lexical similarity, moderate thematic similarity, and high semantic similarity. These examples help show that TF-IDF, LSA, and IndoBERT do not respond identically to the same title pair, but instead capture different aspects of textual proximity. Accordingly, the table is intended to support interpretation of the numerical results by linking similarity scores to concrete differences in topic, terminology, and semantic focus.

Table 3 provides qualitative examples to clarify how similarity scores should be interpreted across representation methods. In the first example, both titles concern damage detection, but they differ in application context, namely vehicle diagnostics and road monitoring. This results in relatively low TF-IDF similarity, while LSA and IndoBERT still capture broader thematic relatedness. The second example reflects moderate thematic similarity, as both titles are situated in the domain of advanced material engineering and share partially overlapping terminology related to carbon nanotubes, although their specific chemical systems and end applications differ. The third example shows strong semantic proximity across all three methods, especially in LSA and IndoBERT, because both titles address legal protection in the context of fintech peer-to-peer lending in Indonesia and share substantial thematic as well as lexical overlap. These examples reinforce that similarity scores in this study represent degrees of thematic closeness rather than definitive judgments of duplication or scientific originality.

Novelty Risk Interpretation

A cosine similarity threshold of ≥ 0.60 was applied as an analytical indicator to flag potential semantic overlap.

Table 3: Illustrative examples of title-pair similarity across TF-IDF, LSA, and IndoBERT

No	Example Category	Proposal Title A	Closest Title B	TF-IDF	LSA	IndoBERT	Brief Interpretation
1	Lower lexical similarity	sistem deteksi kerusakan melalui on board diagnostic kendaraan menggunakan microcontroller berbasis nirkabel untuk mendukung sdgs 9	deteksi kerusakan jalan raya secara real time berbasis perangkat bergerak dengan convolutional neural network	0.3366	0.6194	0.7840	Related in broad topic, but with limited lexical and thematic overlap
2	Moderate thematic similarity	pengembangan katalis heterogen nano seng oksida berpenyangga multi-walled carbon nanotubes mwcntsz untuk transesterifikasi minyak kesambi	sintesis komposit fe2o3 mxene multiwall carbon nanotubes mwcnt dengan prekursor alami etilen-gelatin untuk produksi hidrogen	0.4666	0.7715	0.8619	Shared thematic area with partially overlapping terminology
3	High semantic similarity	kepastian hukum sebagai upaya perlindungan hukum terhadap praktek fintech peer to peer lending di indonesia	penguatan pengaturan fintech peer to peer lending dalam upaya memberikan perlindungan terhadap pengguna layanan yang berkeadilan	0.6864	0.9047	0.8690	Strong thematic and semantic proximity with substantial shared terminology

This threshold is not treated as a universal or deterministic boundary, but rather as an early warning signal to support further qualitative evaluation. In this study, novelty risk is defined as the relative potential for low novelty indicated by high semantic similarity between a proposal title and previously funded titles. Novelty risk is derived from aggregated similarity indicators, including Top-1 similarity values, similarity score distributions, and the proportion of titles exceeding the similarity threshold. By positioning novelty risk as a relative and context-dependent construct, the analysis avoids binary novelty decisions and aligns with the intended role of the proposed approach as a decision support mechanism.

Results Across Research Schemes

Similarity analysis conducted at the research-scheme level revealed broadly consistent Top-1 similarity patterns across funding programmes, with only limited variation between categories. While some schemes showed slightly higher or lower average similarity values, the overall results indicate that semantically related proposal titles are distributed across multiple schemes rather than concentrated in a single programme type. These scheme-level patterns suggest that semantic proximity is a general feature of the dataset and should be interpreted in relation to the thematic context of each scheme.

Although the average Top-1 similarity values reported in Table 4 are numerically close, this pattern is still informative. The concentration of scores within a similar range suggests that semantic proximity among proposal titles is consistently observed across multiple research schemes, rather than being limited to a single category. Therefore, Table 4 should be interpreted not as evidence of large numerical contrasts between schemes, but as an indication of relatively stable thematic similarity patterns throughout the dataset. This consistency further suggests that the proposed screening framework may be applied across schemes in a broadly uniform manner.

The scheme-level results also indicate that semantically related proposal titles are not confined to one specific programme category. Instead, thematic overlap appears across a range of schemes, including academic, collaborative, and applied research programmes. This finding is important because it suggests that similarity-based screening may support proposal evaluation across diverse funding contexts, while still requiring contextual interpretation by reviewers. Accordingly, the reported scheme-level patterns should be understood as comparative analytical signals rather than definitive classifications of proposal originality.

Discussion and System-Level Implications

The results indicate a clear progression in semantic sensitivity across the three representation methods. TF-IDF remains highly dependent on direct lexical overlap, LSA captures broader latent relationships, and IndoBERT provides the strongest semantic grouping among titles. These differences suggest that model selection has important implications for how thematic proximity is identified in proposal screening systems, particularly when the objective is to support consistent, scalable, and interpretable early-stage evaluation.

Summary of Results and Discussion

Overall, the findings show that semantic similarity analysis can provide a practical basis for identifying proposal titles that are closely related to previously funded work. The comparative results also confirm that

representation choice strongly influences the range and structure of similarity values, from sparse lexical matching in TF-IDF to denser semantic clustering in IndoBERT.This makes similarity-based analysis particularly useful in large-scale and resource-constrained evaluation environments where rapid preliminary screening is needed.

These findings should be interpreted cautiously, particularly because semantic similarity is used in this study as an indicator of novelty risk rather than as a direct measure of scientific novelty. A more detailed discussion of these limitations is presented in the following subsection.

Limitations of Similarity-Based Novelty Screening

This study has several limitations that should be acknowledged.

First, the analysis is conducted in an unsupervised setting without manual annotation or labelled novelty ground truthTherefore, the reported similarity patterns cannot be interpreted as definitive evidence that the model directly measures scientific novelty. Instead, the framework should be understood as identifying semantic proximity that may indicate novelty risk.

Table 4: Similarity Results across Research Schemes

No	Research Scheme	Number of Proposals	Mean Top-1 Similarity	Median Top-1 Similarity	% Top-1 ≥ 0.60	Novelty Risk	Analytical Interpretation
1	Strategic Research Collaboration	337	0.8690	0.8621	99.70%	High (low novelty)	High thematic concentration
2	Inter-University Research Collaboration	314	0.8440	0.8474	100.00%	Medium	Similarity patterns in collaborative schemes
3	Early-Career Lecturer Research	17,660	0.8360	0.8386	99.99%	Medium	Recurrent themes in early-stage proposals
4	Early-Career Lecturer Research (Affirmation)	722	0.8356	0.8387	100.00%	Medium	Recurrent themes in early-stage proposals
5	Master's Thesis Research	5,493	0.8353	0.8375	100.00%	Medium	Academic research focus similarity
6	Domestic Research Collaboration	558	0.8349	0.8365	100.00%	Medium	Similarity patterns in collaborative schemes
7	Fundamental Research – Regular	5,159	0.8330	0.8360	99.98%	Medium	Limited thematic variation
8	Doctoral Dissertation Research	3,570	0.8327	0.8353	99.97%	Medium	Academic research focus similarity
9	Applied Research – National Competitive	176	0.8324	0.8304	99.97%	Medium	Limited thematic variation
10	Basic Research – National Competitive	552	0.8322	0.8319	99.97%	Medium	Limited thematic variation
11	Applied Research – Institutional Excellence	2,267	0.8296	0.8321	99.97%	Medium	Limited thematic variation
12	Master-to-Doctoral Education Research	936	0.8296	0.8328	99.89%	Medium	Academic research focus similarity
13	Applied Research	1,358	0.8288	0.8308	99.97%	Medium	Limited thematic variation
14	Applied Research Downstreaming Track	238	0.8278	0.8328	99.97%	Medium	Limited thematic variation
15	Basic Research Institutional Excellence	3,350	0.8264	0.8299	99.97%	Medium	Limited thematic variation
16	Basic Research	1,388	0.8236	0.8279	99.78%	Medium	Limited thematic variation
17	Prototype Research	351	0.7883	0.7951	100.00%	Medium	Relatively higher thematic variation

Particularly when a proposal title appears thematically close to previously funded titles. Second, the analysis is based only on proposal titles rather than full proposal documents, which means that the semantic evidence available for assessing thematic overlap is necessarily limited. Short titles often compress complex research ideas into narrow textual expressions, making similarity estimates sensitive to shared terminology, writing conventions, and domain-specific phrasing. Third, semantic similarity should not be interpreted as equivalent to novelty. A title may appear highly similar to previously funded work while still offering a distinct contribution in terms of method, perspective, context, data, or application. Conversely, a title with lower similarity does not automatically guarantee substantive originality. As emphasized by Ghosal et al. (2022), novelty depends on what is new relative to previously known information, while (Ihnaini et al., 2024) describe semantic similarity primarily as a measure of semantic closeness. Jeon et al., (2023) further show that novelty in scientific texts may be examined through textual patterns, but such measures should still be interpreted cautiously within the broader scholarly context. Fourth, the threshold-based screening strategy proposed in this study is intended for early detection and decision support, not for automatic judgment or exclusion. The framework is therefore designed to support reviewer attention and prioritization, rather than to replace expert evaluation.

Finally, the findings are derived from a specific national proposal dataset and may reflect the thematic structure, language patterns, and administrative characteristics of that corpus. For this reason, the generalizability of the results to other institutional, linguistic, or national research funding contexts should be considered with caution.

Conclusion

This study proposes an analytical framework for examining semantic proximity among research proposal titles using TF-IDF, LSA, and IndoBERT representations. The results show that each method offers distinct advantages, with TF-IDF providing stricter lexical discrimination, LSA capturing intermediate latent structure, and IndoBERT producing the broadest semantic clustering. These findings suggest that similarity-based analysis can serve as a useful decision-support component in proposal evaluation, particularly for early-stage screening.

From a system perspective, the findings support the integration of similarity-based text analysis into research funding platforms to assist reviewers in identifying potentially overlapping proposal titles more efficiently. Such integration may improve the consistency, scalability, and transparency of preliminary proposal assessment, while still preserving the central role of expert judgment in interpreting thematic relevance and research contribution.

Acknowledgment

The authors acknowledge the use of artificial intelligence tools solely for improving the clarity and grammar of the English language in this manuscript. All research design, data processing, analysis, and interpretation of results were conducted independently by the authors. The findings and conclusions presented in this paper reflect the authors' original work. The dataset and related materials are openly accessible at: https://doi.org/10.5281/zenodo.18502698.

Funding Information

This study was conducted as part of the author's Master's degree program, which was fully funded through the Beasiswa Unggulan scholarship granted by the Kementerian Pendidikan Dasar dan Menengah, Republic of Indonesia. The scholarship supported the author's academic training during the research period. The funding institution did not influence the research design, data analysis, interpretation of findings, or the preparation of this manuscript.

Author's Contributions

Walanstar Alimcan Sitorus: Conceptualized the research idea, designed the methodology, conducted data processing and analysis, and prepared the original manuscript. He is a Master's degree student and served as the primary author of this study.

Evaristus Didik Madyatmadja: Provided academic supervision, research guidance, and critical feedback throughout the research process, including methodological direction and manuscript refinement.

Ethics

The authors confirm that this manuscript has not been published previously and is not under consideration for publication elsewhere.

References

Brglez, M. (2023). Dispersing the clouds of doubt: can cosine similarity of word embeddings help identify relation-level metaphors in Slovene? *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, 61–69. https://doi.org/10.18653/v1/2023.bsnlp-1.8

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2026). The Faiss Library. *IEEE Transactions on Big Data*, *12*(2), 346–361. https://doi.org/10.1109/tbdata.2025.3618474

Fellah, A. (2021). All-Three: Near-optimal and domain-independent algorithms for near-duplicate detection. *Array*, *11*, 100070. https://doi.org/10.1016/j.array.2021.100070

Foltýnek, T., Meuschke, N., & Gipp, B. (2020). Academic Plagiarism Detection. *ACM Computing Surveys*, *52*(6), 1–42. https://doi.org/10.1145/3345317

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, *48*(1), 77–117. https://doi.org/10.1162/coli_a_00429

Halder, R. K., Uddin, M. N., Uddin, Md. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113. https://doi.org/10.1186/s40537-024-00973-y

Hanafi, N., & Saadatfar, H. (2022). A fast DBSCAN algorithm for big data based on efficient density calculation. *Expert Systems with Applications*, *203*, 117501.https://doi.org/10.1016/j.eswa.2022.117501

Huici, D., Rodríguez, R. J., & Mena, E. (2025). APOTHEOSIS: An efficient approximate similarity search system. *SoftwareX*, *29*, 102016. https://doi.org/10.1016/j.softx.2024.102016

Ihnaini, B., Abuhaija, B., Mills, E. A., & Mahmuddin, M. (2024). Semantic similarity on multimodal data: A comprehensive survey with applications. *Journal of King Saud University - Computer and Information Sciences*, *36*(10), 102263. https://doi.org/10.1016/j.jksuci.2024.102263

Indasari, S. S., & Tjahyanto, A. (2024). Decision Support Model in Compiling Owner Estimate for FMCGs Products from Various Marketplaces with TF-IDF and LSA-based Clustering. *Procedia Computer Science*, *234*, 455–462. https://doi.org/10.1016/j.procs.2024.03.027

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). Measuring the novelty of scientific publications: A fastText and local outlier factor approach. *Journal of Informetrics*, *17*(4), 101450. https://doi.org/10.1016/j.joi.2023.101450

Johnson, J., Douze, M., & Jegou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/tbdata.2019.2921572

Kemdikbud. (2020). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 tentang Standar Nasional Pendidikan Tinggi*. https://doi.org/10.52431/murobbi.v3i1.172

Kemristekdikti. (2019). *Peraturan menteri riset, teknologi*.

Krisnawati, L. D., Mahastama, A. W., Haw, S.-C., Ng, K.-W., & Naveen, P. (2024). Indonesian-English Textual Similarity Detection Using Universal Sentence Encoder (USE) and Facebook AI Similarity Search (FAISS). *CommIT (Communication and Information Technology) Journal*, *18*(2), 183–195. https://doi.org/10.21512/commit.v18i2.11274

Pusat, P. (2014). Peraturan Pemerintah (PP) Nomor 4 Tahun 2014 tentang Penyelenggaraan Pendidikan Tinggi Dan Pengelolaan Perguruan Tinggi. *CommIT (Communication and Information Technology) Journal*, *18*(2), 183–195.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., & Shamshirband, S. (2020). A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics*, *8*(2), 286. https://doi.org/10.3390/math8020286

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Scalvini, B., & Mashaghi, A. (2025). Semantic Topology: a New Perspective for Communication Style Characterization. *Findings of the Association for Computational Linguistics: ACL 2025*, 9223–9233. https://doi.org/10.18653/v1/2025.findings-acl.479

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857. https://doi.org/10.18653/v1/2020.aacl-main.85

Xiao, Z., Ning, X., & Duritan, M. J. M. (2025). BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching. *Alexandria Engineering Journal*, *126*, 231–246. https://doi.org/10.1016/j.aej.2025.04.061