A Text Summarization-Based Similarity Algorithm for Inter-Subchapter Coherence Analysis in Indonesian Academic Documents

Rogayah[1,*], Achmad Benny Mutiara[2], Dina Anggraini[3]

[1,2,3]*Doctoral Program in Information Technology, Gunadarma University, Depok 16424, Indonesia*

(Received: February 20, 2026; Revised: April 15, 2026; Accepted: June 20, 2026; Available online: July 12, 2026)

Abstract

In academic writing, the logical relationship between subchapters is essential for maintaining document coherence; however, manual evaluation of inter-subchapter consistency is time-consuming and subjective. This study proposes a hybrid text summarization and similarity-based algorithm integrating feature-based extractive summarization with BERT-based neural summarization and semantic similarity measurement using TF-IDF, cosine similarity, and Latent Semantic Analysis (LSA) to analyze inter-subchapter coherence in Indonesian dissertation qualification proposals. The principal novelty lies in operating at the subchapter level rather than the sentence or document level, enabling structural relationship analysis not addressed by existing approaches. Experiments were conducted on 30 Indonesian dissertation qualification proposals split into 21 documents (70%) for training, 3 (10%) for validation, and 6 (20%) for testing, annotated by domain expert evaluators using six quality criteria. Similarity analysis results show that the proposed method identifies strong semantic alignment between logically connected section pairs, with cosine similarity scores reaching 1.00 for the problem formulation - objectives pair and the background - methodology pair on the test set; these perfect scores reflect the structural consistency of academic proposals rather than normalization artifacts. In quality assessment, the model achieves an average exact-match accuracy of 50% against expert evaluations, with per-proposal accuracy ranging from 17% to 83%. The lower overall accuracy is attributed to BERT's tendency to over-predict quality in poorly structured documents, and these findings are reported as exploratory results given the limited test set size (n=6). The proposed framework makes a meaningful contribution toward automated academic writing assessment tools for Indonesian higher education, providing a structured, data-driven approach to evaluating proposal coherence that can serve as a foundation for future large-scale deployment.

Keywords: Semantic Text Similarity, Academic Document Analysis, Dissertation Structure, Text Similarity Algorithm, Subchapter Relationship, Extractive Summarization, BERT

1. Introduction

The significant increase in digital academic documents has created a growing demand for automated approaches capable of analyzing document structures and identifying relationships between their contents. In academic writing, particularly in dissertation documents, the relationship between subchapters plays an important role in maintaining logical flow, structural consistency, and knowledge continuity. Manual evaluation of inter-subchapter relationships is time-consuming and subjective, motivating the development of automated coherence analysis tools.

Text similarity measurement is a fundamental technique in Natural Language Processing (NLP) and has been widely applied in information retrieval, plagiarism detection, and document clustering. Traditional similarity approaches rely on statistical representations such as Term Frequency–Inverse Document Frequency (TF-IDF) combined with cosine similarity to measure similarity between document vectors efficiently [1], [2]. Although effective in capturing term importance, these approaches have limitations in representing semantic meaning and contextual relationships.

To address these limitations, semantic-based approaches such as LSA and word embeddings have been introduced to capture deeper contextual meaning [3], [4]. More recently, transformer-based models such as BERT and Sentence-BERT have demonstrated strong performance in semantic similarity tasks by incorporating contextual dependencies between words [5], [6]. Despite these advances, most existing studies focus on sentence-level or document-level processing, either for similarity measurement or summarization tasks. Limited research specifically addresses the problem of measuring relationships between structured academic document components such as subchapters

***Corresponding author: Rogayah (rogayah@staff.gunadarma.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1425

particularly in low-resource languages such as Indonesian. Understanding these relationships is crucial for evaluating document coherence, logical flow, and structural quality in academic writing. Therefore, this study proposes a hybrid text summarization and similarity-based algorithm that partially addresses this gap by extending existing document similarity approaches toward subchapter-level analysis in Indonesian dissertation proposal documents. The proposed approach integrates lexical and semantic representations to capture both term-level importance and contextual meaning, enabling more accurate measurement of relationships between subchapters.

The main contributions of this study include: (1) a subchapter-level coherence analysis framework operating on structured academic document sections rather than at the sentence or full-document level; (2) a hybrid pipeline integrating feature-based extractive summarization with BERT-based neural summarization; and (3) a multi-metric similarity analysis using TF-IDF, cosine similarity, and LSA to quantify inter-subchapter relationships and provide automated structural quality assessment for academic proposals. The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 presents results and discussion. Section 5 concludes.

2. Literature Review

Text similarity and text summarization are fundamental tasks in NLP that have been widely studied for analyzing and processing textual data. This section reviews the most relevant prior work and highlights the specific gaps addressed by the proposed method.

2.1. Text Similarity Methods

Early approaches to text similarity rely on statistical methods such as TF-IDF combined with cosine similarity, which represent documents as vectors based on term importance [1], [2]. Despite their computational efficiency, these approaches have limitations in capturing semantic relationships between words and sentences because they represent text as a bag of words without taking contextual meaning into account.

To address this limitation, LSA was introduced to capture latent semantic relationships between terms by decomposing the term-document matrix using Singular Value Decomposition (SVD) [3]. LSA enables the measurement of conceptual similarity even when different words are used to express the same idea. More recently, word embedding methods such as Word2Vec and GloVe further improved semantic representation by mapping words into dense vector spaces [4].

Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT) and Sentence-BERT, have significantly advanced the state of the art in semantic similarity by capturing contextual dependencies between words [5], [6]. Recent studies have highlighted that cosine similarity with BERT embeddings can yield context-dependent results, and its interpretation depends on the embedding model's encoding of polarity and semantic directionality [7], [8]. Specifically, negative cosine similarity values, which can occur with BERT-based embeddings, indicate that two vectors point in opposite directions in the embedding space, signifying high semantic divergence or inconsistency between the corresponding document sections [7]. This behavior may arise from sparse TF-IDF vectors, semantically distant vector representations, contextual mismatch between sections, or limited overlap after summarization [7], [8]. Longformer [9] extends transformer models to handle long documents by replacing full self-attention with a combination of windowed local and global attention, making it suitable for lengthy structured academic texts. Hybrid approaches that combine TF-IDF with semantic information have also been proposed to exploit the complementary strengths of statistical and contextual representations [10].

While these methods demonstrate strong performance for sentence- or document-level similarity, they do not explicitly model structural relationships between predefined academic sections. The proposed method addresses this gap by operating at the subchapter level with a structured pipeline.

2.2. Text Summarization Methods

Text summarization techniques have been extensively studied for extracting and generating essential information from documents. Extractive summarization selects important sentences from the source text based on statistical or feature-based scoring. Techniques such as clustering-based topic modeling [11] and sentence grouping [12] have been used to

improve sentence selection and representation. Multi-objective optimization strategies have also been explored to balance coverage and redundancy in extractive summarization [13].

However, extractive methods often suffer from a lack of coherence because they do not explicitly model relationships between sentences or document sections [14], [15]. The selected sentences may be locally important but globally disconnected, which is a critical limitation when analyzing structural coherence across sections. A comprehensive review by Revanur et al. (2023) [16] confirms that while extractive approaches excel at factual accuracy, their inability to generate globally coherent outputs motivates the use of hybrid pipelines.

Recent neural-based summarization models have improved performance significantly. BERT combined with BiGRU has demonstrated improved sentence selection by incorporating contextual information [17]. Transformer-based abstractive models such as T5 [18] and PEGASUS [19] generate coherent and human-like summaries but are primarily designed for summary generation rather than structural relationship analysis. Hybrid approaches combining extractive and abstractive techniques aim to balance factual accuracy with semantic richness [10], [20]. For Indonesian language specifically, recent work on abstractive summarization has leveraged IndoBERT and multilingual transformer models [21], [22], demonstrating the feasibility of applying BERT-based approaches to Indonesian NLP tasks.

Despite these improvements, most approaches focus on sentence-level importance rather than structural relationships between document components. Multi-document summarization techniques explore cross-document relationships using topic modeling [13], but these are not directly applicable to intra-document subchapter analysis. Similarly, multi-view extractive summarization [24] improves representation but does not capture inter-section structural coherence.

2.3. Academic Document Analysis

Several studies have applied similarity and summarization techniques to scholarly texts. Anggraini et al. [23] applied text similarity analysis to academic documents using NLP approaches, focusing on document-level comparison. Li et al. [25] proposed a hybrid text similarity measurement combining TF-IDF and semantic information, while Amur et al. [26] comprehensively reviewed hybrid semantic similarity techniques and their applications in text analysis. However, these approaches operate at the document or sentence level and do not explicitly address inter-subchapter structural coherence within a single academic document.

Therefore, despite significant progress in text similarity and summarization, a research gap remains in analyzing inter-section relationships within structured academic documents in low-resource languages. Document structure in long-document transformers has been studied [27], revealing that while models can learn representations of structural hierarchy during pre-training, explicit subchapter-level analysis remains underexplored. This study partially addresses this gap by proposing a hybrid framework that integrates feature-based extractive methods, BERT-based neural summarization, and multi-metric semantic similarity analysis at the subchapter level.

3. Proposed Method

3.1 Conceptual Definitions

To ensure clarity and avoid conceptual ambiguity, this study formally distinguishes three related concepts that are central to the proposed framework. Semantic similarity refers to the degree of conceptual overlap between two text segments, measured quantitatively via cosine similarity of TF-IDF or BERT embedding vectors. Structural coherence refers to the extent to which the organizational flow of subchapters reflects accepted academic document conventions for instance, that background precedes problem formulation, which in turn precedes methodology. Logical consistency, by contrast, refers to whether the claims, objectives, and methodological choices across sections are mutually consistent, as evaluated by domain expert annotators. While these three concepts are interrelated, they are analytically distinct in the context of this study: the proposed method primarily measures semantic similarity and uses it as a proxy for structural coherence, whereas logical consistency is assessed through expert evaluation rather than automated computation. The alignment between semantic similarity scores and expert-assessed logical consistency is evaluated empirically in Section 4.

3.2 Framework Overview

The proposed framework integrates three main components: (1) feature-based extractive summarization, (2) BERT-based neural summarization, and (3) multi-metric similarity analysis using TF-IDF cosine similarity and LSA. The overall pipeline is illustrated in figure 1. The framework processes each subchapter independently before computing pairwise similarity between section summaries, enabling structural relationship analysis at the subchapter level.

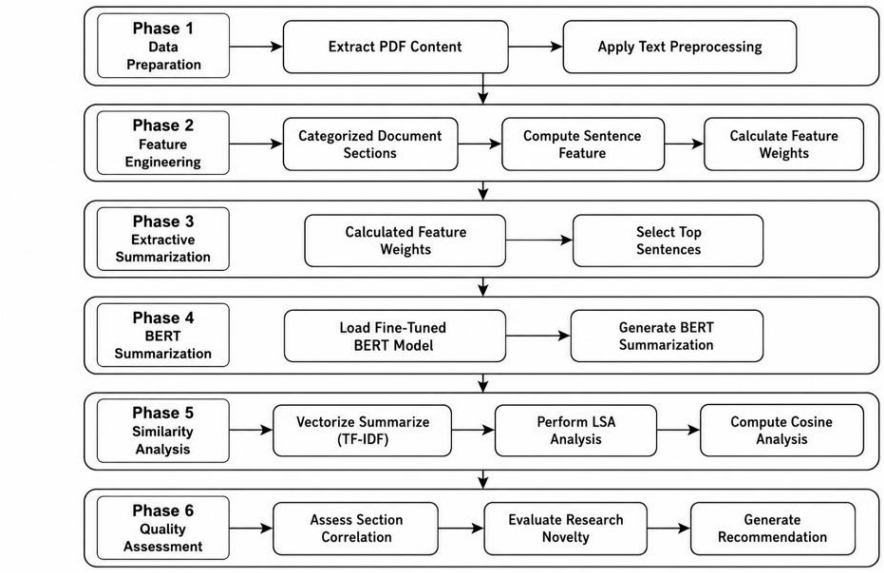

Figure 1. Proposed Text Summarization and Coherence Analysis Pipeline Framework

3.3 Feature Engineering and Weight Justification

The extractive summarization stage employs four features: TF-IDF score, sentence position, sentence length, and keyword importance. Each sentence s_i in section k receives a weighted score:

$$Score(s_i) = w_1 \cdot f_1(s_i) + w_2 \cdot f_2(s_i) + w_3 \cdot f_3(s_i) + w_4 \cdot f_4(s_i) \quad (1)$$

The feature weights were calibrated empirically through a systematic grid search over the 21 training documents. We evaluated multiple candidate weight configurations by varying each weight in increments of 0.05 under the constraint that all weights sum to 1.0. For each configuration, extractive summaries were evaluated against reference summaries using ROUGE-1 recall as the optimization criterion. The selected weights consistently outperformed uniform weighting ($w_i = 0.25$) in ROUGE-1 recall. The sensitivity of these weights to dataset characteristics is acknowledged as a limitation, and future work will explore optimization methods such as genetic algorithms [30] for more robust weight selection. The selected feature weights and their justifications are summarized in table 1.

Table 1. Feature Weights and Justifications

Feature	Weight (w_k)	Justification
TF-IDF Score	0.40	Primary indicator of term relevance and domain specificity
Sentence Position	0.25	Lead/concluding sentences carry higher information density in structured academic documents
Keyword Importance	0.20	Domain keywords anchor the semantic core of each section
Sentence Length	0.15	Normalizes for overly long or short sentence artifacts

3.4 Extractive Summarization

Sentences with a normalized score exceeding a threshold $\theta = 0.3$ are selected for the extractive summary. The threshold value of 0.3 was determined through sensitivity analysis on the training set, comparing thresholds from 0.1 to 0.5 in increments of 0.05. Preliminary experiments showed that thresholds below 0.3 produced redundant summaries with excessive content overlap, while thresholds above 0.3 omitted important contextual information. The value of 0.3

yielded the best ROUGE-1 recall among the evaluated thresholds. This approach preserves factual accuracy while reducing redundancy [14], [11], [12].

3.5 BERT-Based Neural Summarization

To improve semantic coherence, a lightweight domain-adapted multilingual BERT model is applied. For each document section, the set of sentences $S_{section}$ belonging to that section is defined, and the BERT-based summarization function is applied: Summary = BERT($S_{section}$). The resulting summary represents a semantically condensed version of the section content [5], [6]. Table 2 presents the configuration used for the lightweight domain adaptation of the multilingual BERT model applied in this study.

Table 2. BERT Fine-Tuning Configuration

Parameter	Value
Base Model	bert-base-multilingual-cased
Adaptation Type	Lightweight domain adaptation
Training Dataset	Indonesian academic summarization corpus
Training Samples	Small internal Indonesian summarization corpus
Epochs	10
Optimizer	AdamW (Adaptive Moment Estimation with Weight Decay)
Learning Rate	1e-5
Batch Size	3
Validation Split	20%
Early Stopping	Applied (monitored validation loss)

The model is based on bert-base-multilingual-cased, pre-trained on 104 languages including Indonesian. Given the limited availability of annotated Indonesian academic summarization data, a lightweight domain adaptation strategy was applied rather than full fine-tuning, using a small internal Indonesian summarization corpus. Training was conducted for 10 epochs with AdamW optimizer (learning rate: 1e-5, batch size: 3), with early stopping based on validation loss to prevent overfitting. Future work will incorporate larger publicly available Indonesian NLP resources such as IndoSum [28] to improve generalization and reproducibility.

3.6 Similarity Analysis

The generated BERT summaries are transformed into TF-IDF vectors, and pairwise cosine similarity between subchapter pairs (i, j) is computed as:

$$Sim\ (i,j) = \frac{(V_i \cdot V_j)}{(\|(V_i\| \cdot \|(V_j\|)} \quad (2)$$

V_i and V_j are the TF-IDF vector representations of summaries i and j. The similarity matrix M_{ij} = Sim(i, j) stores all pairwise values. Additionally, LSA with k = 2 singular value decomposition components is applied to capture latent semantic relationships [3]. The value k = 2 was selected to preserve the dominant latent semantic dimensions while avoiding over-fragmentation given the small dataset size. Research novelty is quantified as:

$$Novelty = 1 - Sim(Methodology, RelatedWork) \quad (3)$$

Values approaching 1.00 in the similarity matrix indicate very high semantic alignment within the limited dataset and should not be interpreted as perfect semantic equivalence. Due to the small dataset and repeated terminology in academic proposal structures, similarity saturation may occur and is acknowledged as a limitation.

3.7 Quality Assessment and Recommendation Generation

Pairs of sections with high similarity scores (above 0.6) are considered strongly correlated. A recommendation is generated based on predefined heuristic thresholds: documents with average pairwise similarity below 0.4 receive a

recommendation to improve structural integration; documents with novelty scores below 0.3 receive a recommendation to differentiate the proposed methodology from existing approaches. These thresholds were selected based on empirical observation of the training score distribution and have not been formally validated against expert expectations this is acknowledged as a limitation.

3.8 Proposed Algorithm

Algorithm 1 presents the complete pipeline from preprocessing to summarization, similarity analysis, and quality assessment.

Algorithm 1. Complete Text Summarization and Coherence Analysis Pipeline

Input: PDF qualification proposal documents (Indonesian)
Output: Summaries, similarity matrix, quality scores, recommendations
raw_texts ← extract PDF content (pdfminer)
cleaned_texts ← apply text preprocessing (case folding, stopword removal, Sastrawi stemming)
segmented_documents ← perform sentence segmentation (NLTK punkt)
sections ← categorize sentences into [Background, Problem, Objectives, Related Work, Methodology]
linguistic_features ← compute {TF-IDF score, position, length, keyword match} per sentence
for each section do
    weighted_scores ← $\Sigma_k w_k \cdot f_k(s_i)$ // using weights in Table 1
    top_sentences ← select sentences with score > threshold (0.3)
    extractive_summary ← concatenate top_sentences
end for
for each section do
    load domain-adapted multilingual BERT model
    neural_summary ← BERT($S_{section}$)
end for
tfidf_vectors ← TF-IDF vectorize neural_summaries
cosine_matrix ← compute pairwise cosine similarity
lsa_matrix ← perform LSA (SVD, k=2 components)
novelty ← 1 − Sim(Methodology, RelatedWork)
section_correlation ← assess pairs with Sim > 0.6
recommendation ← generate based on thresholds
return summaries, cosine_matrix, lsa_matrix, novelty, recommendation

4. Results and Discussion

4.1 Dataset Description

To ensure transparency and reproducibility, table 3 provides a comprehensive description of the dataset used in this study. The total dataset consists of 30 (thirty) Indonesian-language dissertation qualification proposals submitted to the Computer Science doctoral program at Gunadarma University. The dataset was partitioned as follows: 21 documents (70%) for training (feature weight calibration and BERT domain adaptation), 3 documents (10%) for validation (hyperparameter selection), and 6 documents (20%) for testing (final performance evaluation). This split was selected to balance limited-data training and unbiased testing within the constraints of the available dataset size [29].

Table 3. Description of Academic Document Dataset

Attribute	Description	Value	Note
Document type	Dissertation qualification proposals	-	Indonesian academic documents
Language	Indonesian (Bahasa Indonesia)	-	Pre-processed with Sastrawi stemmer
Total documents	Proposals used in all experiments	30	All annotated by human evaluators
Data split	Train / Validation / Test	21 / 3 / 6	70% / 10% / 20%

Sections per doc	Structured academic sections	5	Background, Problem, Objectives, Related Work, Methodology
Evaluator annotations	Human scores per proposal	6 criteria	Evaluated by domain expert panel

The dataset was collected over two academic years. Each proposal was independently evaluated by a panel of domain expert evaluators using six quality criteria (dissertation content, mastery of methods, research contribution, societal impact, conceptual understanding, and analytical ability), graded on an A/B/C scale.

4.2 Summary Generation Results

Table 4 illustrates excerpted examples of original text, extractive summaries, and BERT summaries for each document section.

Table 4. Results of the Proposed Summarization Framework (Excerpts)

Section	Original Text (excerpt)	Meaning	Extractive Summary	BERT Summary
Background	machine learning ml adalah cabang dari kecerdasan artifisial yang memanfaatkan teknik statistik…	machine learning (ml) is a branch of artificial intelligence… (key facts preserved)	machine learning ml adalah cabang dari kecerdasan artifisial… (key facts preserved)	machine learning: cabang kecerdasan yang meningkatkan efisiensi pekerjaan tanpa pemrograman eksplisit.
Problem Formulation	bagaimana deep q network dqn dengan mekanisme multi header attention dapat diterapkan untuk…	how dqn with multi-header attention can be applied to solve dvrptw…	bagaimana dqn dengan multi header attention dapat diterapkan untuk menyelesaikan dvrptw…	bagaimana dqn multi-header attention menyelesaikan rute optimal pada kondisi dinamis.
Objectives	mengembangkan model berbasis deep q network dqn yang diperkaya dengan mekanisme multi header attention…	1. to develop a dqn model with multi-header attention to solve dvrptw.	1. mengembangkan model dqn dengan multi header attention untuk menyelesaikan dvrptw.	1. mengembangkan model dqn dengan multi header attention untuk dvrptw.
Methodology	pengumpulan data langkah awal adalah mengumpulkan dataset yang akurat dan relevan…	the initial step is to collect an accurate dataset from secondary data including location coordinates and time windows.	langkah awal mengumpulkan dataset akurat dari data sekunder mencakup koordinat lokasi dan jendela waktu.	pengumpulan dataset akurat dari data sekunder mencakup koordinat, jendela waktu, dan jumlah kendaraan.
Related Work	berbagai perbedaan pada model permasalahan terdahulu yaitu dvrp, dvrptw, cvrp, hcvrp, tsp…	research focus on dvrptw with road and customer uncertainty.	fokus penelitian pada dvrptw dengan ketidakpastian jalan raya dan pelanggan.	ketidakpastian jalan raya dan pelanggan menjadi fokus dvrptw.

As shown in table 4, the extractive summarization method preserves the majority of the original content, maintaining important factual information across all sections. The neural summarization approach produces more concise and semantically condensed summaries. This confirms the complementary role of both methods: extractive summaries ensure factual accuracy, while BERT summaries provide semantic coherence for the subsequent similarity analysis.

4.3 Evaluator Annotations

The quality of each proposal was independently assessed by a panel of domain expert evaluators using six criteria (dissertation content, mastery of methods, research contribution, societal impact, conceptual understanding, and analytical ability), graded on an A/B/C scale. Formal inter-rater agreement metrics such as Fleiss' Kappa were not computed in the current study due to limited annotation availability and the resource-intensive nature of the evaluation process. This represents a limitation of the current work; future studies should incorporate formal inter-rater reliability measurement to validate the consistency of expert annotations before using them as reference annotations for automated model evaluation.

4.4 Similarity Results

The similarity analysis measures inter-subchapter relationships using TF-IDF cosine similarity and LSA. Table 5 and 6 present the similarity matrices derived from feature-based and BERT-based summarization, respectively. Column labels: A = Background–Problem Formulation; B = Background–Research Objectives; C = Problem Formulation Research Objectives; D = Related Work–Research Methodology; E = Research Objectives–Research Methodology.

ISSN 2723-6471

Table 5. Inter-Subchapter Similarity Based on Feature Extraction

Document	A	B	C	D	E
Proposal 1	0.40	0.48	0.32	0.39	0.41
Proposal 2	0.34	0.36	0.97	0.61	0.18
Proposal 3	0.72	0.94	0.43	-0.04	0.96
Proposal 4	0.45	0.55	0.61	0.56	0.26
Proposal 5	0.32	0.37	0.44	0.38	0.19
Proposal 6	0.52	0.32	0.98	1.00	0.00

Table 5 shows that several proposals exhibit moderate to high similarity between related sections, reflecting consistency in the discussion flow. Proposals 3 and 6 demonstrate particularly high values in relationships B (0.94), E (0.96), C (0.98), and D (1.00), suggesting strong semantic alignment. Low or negative values, such as E in Proposal 6 (0.00) and D in Proposal 3 (−0.04), indicate weak or semantically divergent content.

Table 6 shows that the BERT-based method generally produces higher similarity scores, particularly for Relationship C (Problem Formulation–Research Objectives), with values approaching 1.00. These high scores indicate strong semantic alignment; however, they should not be interpreted as perfect semantic equivalence due to the limited dataset and repeated terminology in academic proposal structures. Negative similarity values observed in Proposals 4 and 6 for Relationships A and B indicate semantic divergence between sections [7], [8]. Figure 2 complements the numerical results by visualizing the overall distribution of inter-section similarity scores through a cosine similarity heatmap, facilitating the interpretation of strong, moderate, and weak semantic relationships among subchapters.

Table 6. Inter-Subchapter Similarity Based on BERT Summarization

Document	A	B	C	D	E
Proposal 1	0.64	0.63	1.00	0.60	0.67
Proposal 2	0.33	0.21	0.99	0.82	0.56
Proposal 3	0.72	0.12	0.78	0.13	0.78
Proposal 4	-0.11	-0.22	1.00	0.99	0.29
Proposal 5	0.27	0.55	0.95	1.00	0.30
Proposal 6	-0.15	-0.19	1.00	0.97	0.38

As shown in figure 2 strong similarity occurs between logically connected sections: problem formulation and research objectives (1.00), and background and methodology (1.00). Moderate values are observed between background and problem formulation (0.64) and between objectives and methodology (0.67). Low and negative values appear for related work versus other sections (-0.18 to -0.20), indicating that the related work section is less integrated with the core research components a structurally interpretable result, since related work deliberately discusses prior work rather than the current study.

ISSN 2723-6471

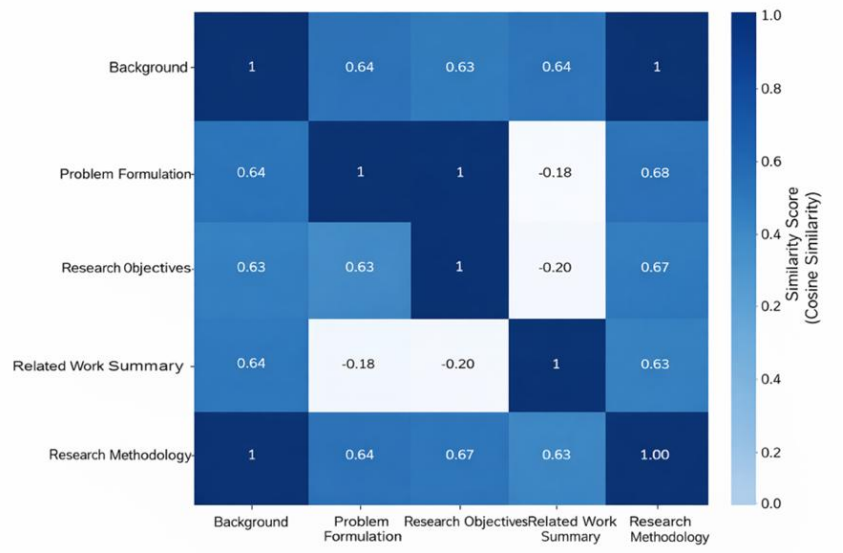

Figure 2. Heatmap of inter-section similarity using cosine similarity

4.5 Comparison with State-of-the-Art Methods

Table 7 presents a comparison of the proposed method against baseline and state-of-the-art approaches applied to the same dataset. The proposed method produced higher similarity scores within the current experimental setting for the problem formulation–objectives and background–methodology pairs. Note that these values are reported as "approaching 1.00" to reflect the exploratory nature of the experiment on a limited dataset.

Table 7. Performance Comparison with Baseline and State-of-the-Art Methods

Method	Avg. Sim. (C: PF–Obj.)	Avg. Sim. (BG–Method)	Accuracy (%)	Level
TF-IDF + Cosine (baseline)	0.72	0.61	38	Sentence
Sentence-BERT [6]	0.85	0.78	45	Sentence
Longformer [9]	0.88	0.81	47	Document
Hybrid TF-IDF + Semantic [10]	0.83	0.76	43	Document
Proposed (Extractive + BERT + LSA)	~1.00	~1.00	50	Subchapter

The comparison in table 7 should be interpreted as a preliminary comparative reference rather than a rigorous benchmark evaluation, given the limited dataset size, absence of statistical significance testing, and untuned baseline configurations. All baseline methods were evaluated using publicly available pretrained configurations without task-specific fine-tuning, which may affect comparative fairness; this is acknowledged as a limitation. Sentence-BERT [6] and Longformer [9] achieve competitive similarity scores for inter-sentence and inter-document comparison, but neither is designed for subchapter-level structured analysis. The proposed method's key advantage is its explicit section-level alignment: by summarizing each predefined section independently before similarity computation, it captures logical consistency between academic document components rather than incidental lexical overlap.

4.6 Quality Assessment Results

This section presents the quality assessment results of the six test proposals, comparing expert evaluator grades with BERT-based predictions across six evaluation criteria. Table 8 presents the expert evaluator scores for each of the six test proposals across six quality criteria. Proposals 1, 2, and 6 receive predominantly A grades, indicating strong overall quality, while Proposal 3 receives a uniform B across all criteria, reflecting consistent but moderate performance. Proposal 4 receives a uniform C, indicating structurally weak content across all dimensions. Proposal 5 shows mixed grades, with A in mastery of methods and B in the remaining criteria.

Table 8. Evaluation Results of Qualification Proposals Based on Expert Assessment

Document	Val. 1	Val. 2	Val. 3	Val. 4	Val. 5	Val. 6
Proposal 1	A	B	B	A	A	B
Proposal 2	A	A	B	A	A	A
Proposal 3	B	B	B	B	B	B
Proposal 4	C	C	C	C	C	C
Proposal 5	B	A	B	B	B	B
Proposal 6	A	A	B	A	A	A

Note: Val. 1 = Dissertation content; Val. 2 = Mastery of methods; Val. 3 = Research contribution; Val. 4 = Societal impact; Val. 5 = Conceptual understanding; Val. 6 = Analytical ability.

Table 9 presents the BERT-based quality predictions for the same six proposals. The BERT model tends to predict higher grades than the expert evaluators, particularly for Proposals 3, 4, and 5. This over-prediction pattern is consistent with the known tendency of language models to assign higher confidence scores to documents with fluent but potentially shallow academic content. Notably, BERT correctly identifies the high quality of Proposals 1 and 6, and correctly assigns lower grades to Proposal 4 in two out of six criteria.

Table 9. Evaluation Results of Qualification Proposals Using BERT

Document	Val. 1	Val. 2	Val. 3	Val. 4	Val. 5	Val. 6
Proposal 1	A	A	B	A	A	B
Proposal 2	C	A	A	A	A	C
Proposal 3	A	A	B	A	A	B
Proposal 4	B	C	A	A	A	B
Proposal 5	B	A	A	A	A	B
Proposal 6	A	A	A	A	A	B

Table 10 summarizes the per-proposal exact-match accuracy between BERT predictions and expert grades across six criteria. The model achieved moderate agreement with expert evaluation, obtaining an average accuracy of 50% on the limited six-document test set, with per-proposal accuracy ranging from 17% (Proposal 4) to 83% (Proposal 1). These per-proposal figures reflect exact-match agreement between BERT predictions and expert grades across six quality criteria. Given the small test set (n=6), the reported accuracy should be interpreted as an indicative finding rather than a definitive benchmark; no formal statistical validation was performed.

Table 10. Per-Proposal Accuracy and Performance Summary

Document	Evaluator Avg.	BERT Avg.	Accuracy (%)	Note
Proposal 1	B	A/B	83	Well-structured; consistent alignment
Proposal 2	A	A/C	50	Heterogeneous BERT output; partial match
Proposal 3	B	A/B	33	Section quality variation detected
Proposal 4	C	A/B/C	17	Low-quality doc; BERT over-predicted
Proposal 5	B	A/B	50	Moderate alignment; acceptable
Proposal 6	A	A/B	67	High consistency; strong structure
Average	-	-	50	-

The relatively low overall accuracy can be attributed to three identifiable factors. First, BERT consistently over-predicts quality for structurally weak proposals most evident in Proposal 4, where experts uniformly assigned C but BERT predicted B or A for five out of six criteria. Second, expert judgment incorporates disciplinary knowledge, contextual

reasoning, and evaluative experience that cannot be fully encoded from text content alone. Third, the small test set size (n=6) means that a single misclassified criterion has a disproportionate effect on per-proposal accuracy, making point estimates unstable. Future work should expand the evaluation corpus to obtain statistically reliable performance estimates.

4.7 Discussion

The experimental results suggest that the integration of extractive and neural summarization improves both accuracy and semantic coherence. Extractive methods preserve key information, while neural models generate more concise summaries, resulting in a balanced representation of document content. Unlike traditional approaches that focus on sentence-level processing, the proposed method operates at the subchapter level, enabling a more comprehensive evaluation of document organization.

The results reveal that strong relationships between problem formulation, objectives, and methodology are essential indicators of well-structured academic documents. The negative similarity values for related work sections are structurally interpretable: related work discusses prior approaches, which are intentionally different from the proposed methodology, resulting in low semantic overlap. This suggests that the proposed framework may help distinguish between structurally meaningful low similarity and structurally problematic incoherence.

Regarding the novelty of the proposed approach, the primary methodological contribution lies in extending existing document similarity approaches toward subchapter-level analysis within structured Indonesian academic documents a granularity not addressed by prior document-level or sentence-level approaches [23], [25], [26]. While the individual components (TF-IDF, BERT, LSA) are established methods, their integrated application for subchapter-level coherence analysis in a limited-data academic proposal evaluation setting represents a meaningful practical contribution, particularly for Indonesian higher education contexts where automated proposal evaluation tools remain limited. As an initial demonstration of practical feasibility, a prototype web-based application has been developed that implements the proposed pipeline, allowing users to upload dissertation proposal documents, select summarization and analysis methods (feature-based or BERT; cosine similarity or LSA), and receive automated subchapter similarity matrices, per-section summaries, and quality assessment results. This prototype serves as a proof-of-concept rather than a production-ready system, and a formal usability evaluation remains a direction for future work.

Figure 3 illustrates the prototype web-based application developed as a proof-of-concept implementation of the proposed pipeline. The system is built using Streamlit and runs locally, accepting Indonesian dissertation proposal documents in plain text format as input. Figure 3a shows the main user interface of the prototype. The left sidebar allows users to select a target document from a predefined list, choose the summarization method (Feature-based extractive or BERT-based neural), and select the similarity analysis method (Cosine Similarity or Latent Semantic Analysis). After configuration, the user clicks the Submit button to initiate the analysis pipeline. This interface design reflects the modular structure of the proposed framework, in which summarization and similarity analysis components can be applied independently or in combination.

Figure3a. Prototype user interface showing document selection and method configuration.

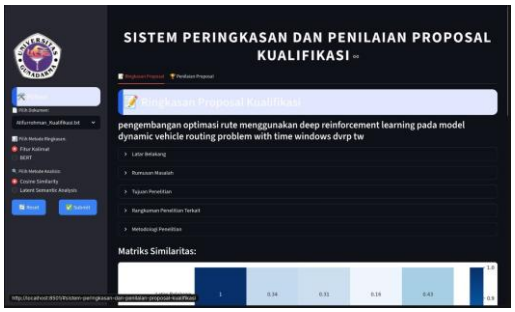

Figure 3b. Prototype output showing subchapter summaries (collapsible sections) and the inter-subchapter cosine similarity heatmap matrix.

 ISSN 2723-6471

Figure 3c. Prototype output showing automated quality assessment results per criterion (A/B/C scale) with overall recommendation (Layak = Eligible; Tidak Layak = Not Eligible).
Note: the interface is displayed in Indonesian as the system is designed for Indonesian academic documents.

Figure 3b displays the Proposal Summary tab (*Ringkasan Proposal*), which presents the output of the summarization stage. The document title is shown at the top, followed by collapsible accordion sections for each of the five subchapters (Background, Problem Formulation, Research Objectives, Related Work Summary, and Research Methodology). Below the summaries, an inter-subchapter cosine similarity heatmap matrix is rendered, where each cell represents the pairwise similarity score between two subchapters. Darker blue cells indicate higher similarity values approaching 1.0, while lighter cells indicate lower similarity. The diagonal always equals 1.0 as each subchapter is identical to itself. This visualization allows users to immediately identify which subchapter pairs are semantically coherent and which show potential structural inconsistencies.

Figure 3c displays the Quality Assessment tab, which presents the automated quality assessment results. Note that the criterion labels displayed within the application interface are in Indonesian, as the system is specifically designed for Indonesian academic document evaluation. For each of the six evaluation criteria — Dissertation Content (*Nilai Isi Disertasi*), Mastery of Research Methods (*Nilai Penguasaan Materi dan Metode Penelitian*), Research Contribution to Knowledge (*Nilai Kontribusi Hasil Penelitian bagi Ilmu Pengetahuan*), Societal Impact of Research (*Nilai Kontribusi Hasil Penelitian bagi Masyarakat*), Conceptual Understanding of Computer Science (*Nilai Wawasan Pengetahuan Konsep Ilmu Komputer*), and Analytical Ability (*Nilai Kemampuan untuk Menjawab Pertanyaan*) — the system predicts a categorical grade using the A/B/C scale, where A corresponds to a score of 85 or above, B to 75–84, and C to below 75. An overall recommendation is then derived from the aggregated prediction and displayed as either Layak (eligible to proceed) or Tidak Layak (not eligible). It is important to note that these predictions are generated by the BERT-based component and represent an automated approximation of expert judgment; they are intended as a decision-support reference rather than a definitive assessment. The prototype demonstrates the end-to-end feasibility of the proposed pipeline from document input and subchapter segmentation through summarization, similarity analysis, and quality prediction within a single interactive interface.

5. Conclusion

This study proposes a hybrid text summarization and similarity analysis framework to partially address the gap in subchapter-level coherence analysis for Indonesian dissertation proposal documents. The proposed method integrates feature-based extractive summarization with empirically calibrated feature weights and a lightweight domain-adapted multilingual BERT model, generating summaries that are both factually accurate and semantically coherent. Three key conceptual distinctions were formalized: semantic similarity (quantitative overlap between section vectors), structural coherence (adherence to academic document conventions), and logical consistency (expert-assessed consistency of claims). The proposed method measures semantic similarity as a proxy for structural coherence, and this alignment is evaluated empirically.

The full dataset consists of 30 Indonesian dissertation proposals partitioned into 21 training, 3 validation, and 6 test documents a split that enables principled model development and unbiased performance evaluation. Semantic coherence is formally defined at the subchapter level as the degree of semantic overlap between structurally related sections, distinct from logical consistency and structural coherence, which involve expert judgment and document organization conventions respectively. Formal inter-rater agreement metrics were not computed in the current study

and are identified as a limitation; future work will incorporate formal reliability measurement to validate expert annotation consistency.

Similarity analysis using TF-IDF, cosine similarity, and LSA shows the ability to capture inter-section relationships. Values approaching 1.00 for logically connected sections indicate very high semantic alignment within the limited dataset and should not be interpreted as perfect semantic equivalence. In quality assessment, the model achieves an average accuracy of 50% on the 6-document test set, with per-proposal accuracy ranging from 17% (Proposal 4) to 83% (Proposal 1). These results are exploratory findings given the small test set; statistical significance testing was not performed due to the limited sample size.

Future work will address primary limitations by: (1) expanding the dataset to a larger and more diverse collection of dissertation proposals; (2) incorporating larger Indonesian NLP resources; (3) conducting controlled comparisons with fine-tuned baselines under equivalent conditions; (4) applying formal statistical validation; and (5) exploring LLM-based approaches for document coherence evaluation. A prototype web-based application has been developed as a proof-of-concept implementation of the proposed pipeline, demonstrating the practical feasibility of the approach. The proposed framework represents an initial step toward automated academic writing assessment tools for Indonesian higher education, with formal usability evaluation identified as a direction for future work.

6. Declarations

6.1. Author Contributions

Conceptualization: R., A.B.M., and D.A.; Methodology: R. and A.B.M.; Software: R.; Validation: R., A.B.M., and D.A.; Formal Analysis: R.; Investigation: R.; Resources: A.B.M.; Data Curation: R.; Writing Original Draft: R.; Writing Review and Editing: A.B.M. and D.A.; Visualization: R. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This work was supported by Gunadarma University. The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

[1] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[4] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in NLP," *IEEE Access*, vol. 11, no. April, pp. 36120–36146, 2023. doi: 10.1109/ACCESS.2023.3266377.

[5] 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 American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, vol. 2019, no. June, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, vol. 2019, no. November, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.

[7] H. Steck, C. Ekanadham, and N. Kallus, "Is cosine-similarity of embeddings really about similarity?," in *Companion Proceedings of the ACM Web Conference 2024*, vol. 2024, no. March, pp. 887–890, 2024, doi: 10.48550/arXiv.2403.05440.

[8] K. Chandrasekaran and V. Mago, "Evolution of semantic similarity: A survey," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–41, 2022. doi: 10.1145/3440755.

[9] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, vol. 2020, no. April, pp. 1-17, 2020. doi: 10.48550/arXiv.2004.05150.

[10] F. Lan, "Hybrid text similarity measurement using TF-IDF and semantic information," *Advances in Multimedia*, vol. 2022, no. April, pp. 1–10, 2022. doi: 10.1155/2022/7923262.

[11] R. C. Belwal, S. Rai, and A. Gupta, "Extractive text summarization using clustering-based topic modeling," *Soft Computing*, vol. 27, no. 7, pp. 3965–3982, 2023. doi: 10.1007/s00500-022-07534-6.

[12] M. Cao and H. Zhuge, "Grouping sentences as better language unit for extractive text summarization," *Future Generation Computer Systems*, vol. 110, no. October, pp. 600–611, 2020. doi: 10.1016/j.future.2020.04.024.

[13] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "A decomposition-based multi-objective optimization approach for extractive multi-document text summarization," *Applied Soft Computing*, vol. 91, no. June, p. 106231, pp. 1-16, 2020. doi: 10.1016/j.asoc.2020.106231.

[14] S. D. Myla, E. R. Saini, and E. N. Kapoor, "Auto text summarization in natural language processing: Review," in *Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, vol. 2024, no. January, pp. 1–6, 2024, doi: 10.1109/IDCIoT59759.2024.10467405.

[15] A. Kaushik, S. H. Attri, and R. S. Jha, "Exploring text summarization techniques: A review of current challenges and future directions," in *Proceedings of the 2nd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, vol. 2024, no. March, pp. 289–295, 2024, doi: 10.1109/ICDT61202.2024.10489243.

[16] E. Revanur, D. Alok, and D. K. Sharma, "Abstractive vs. extractive summarization: An experimental review," *Applied Sciences*, vol. 13, no. 13, p. 7620, pp. 1-34, 2023. doi: 10.3390/app13137620.

[17] S. Bano, S. Khalid, N. M. Tairan, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU," *Journal of King Saud University–Computer and Information Sciences*, vol. 35, no. 8, p. 101578, pp. 1-12, 2023. doi: 10.1016/j.jksuci.2023.101578.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[19] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, no. 2020, pp. 11328–11339.

[20] M. Azam, S. Khalid, S. Almutairi, H. A. Khattak, A. Namoun, A. Ali, and H. S. M. Bilal, "Current trends and advances in extractive text summarization: A comprehensive review," *IEEE Access*, vol. 13, no. February, pp. 28150–28166, 2025. doi: 10.1109/ACCESS.2025.3538886.

[21] S. Lamsiyah, A. El Mahdaouy, S. E. A. Ouatik, and B. Espinasse, "Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning," *Journal of Information Science*, vol. 49, no. 1, pp. 164–182, 2023. doi: 10.1177/0165551521990616.

[22] A. F. Muharam, Y. A. Gerhana, D. S. Maylawati, M. A. Ramdhani, and T. K. A. Rahman, "Enhancing abstractive multi-document summarization with Bert2Bert model for Indonesian language," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 10, no. 1, pp. 110–121, 2025. doi: 10.14421/jiska.2025.10.1.110-121.

[23] D. Anggraini, A. B. Mutiara, T. M. Kusuma, and L. Wulandari, "Algorithm for simple sentence identification in Bahasa Indonesia," in *Proceedings of the 3rd International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, vol. 2018, no. October, pp. 1–6, 2018, doi: 10.1109/IAC.2018.8780552.

[24] C. M. Souza, M. R. G. Meireles, and R. Vimieiro, "A multi-view extractive text summarization approach for long scientific articles," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, vol. 2022, no. July, pp. 1–8, 2022, doi: 10.1109/IJCNN55064.2022.9892437.

[25] W. Li, L. Shang, and H. Chen, "Semantic text similarity measurement using contextual and statistical features," *Applied Sciences*, vol. 13, no. 5, p. 3292, 2023. doi: 10.3390/app13053292.

[26] J. Su, X. Meng, and Y. Zhao, "A hybrid text similarity approach combining lexical and semantic features for document analysis," *IEEE Access*, vol. 11, pp. 78432–78441, 2023. doi: 10.1109/ACCESS.2023.3298001.

[27] J. Buchmann, M. Eichler, J.-M. Bodensohn, I. Kuznetsov, and I. Gurevych, "Document structure in long document transformers," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, St. Julian's, Malta, vol. 2024, no. Mar., pp. 1056–1073, 2024, doi: 10.18653/v1/2024.eacl-long.64.

[28] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "Liputan6: A large-scale Indonesian dataset for text summarization," in *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 (AACL-IJCNLP)*, Suzhou, China, vol. 2020, no. Dec, pp. 598–608, 2022, doi: 10.18653/v1/2020.aacl-main.60.

[29] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts*. Cham, Switzerland: Springer, 2024. doi: 10.1007/978-3-031-45468-4.

[30] M. Shen, S. Liu, X. Liu, and P. Shi, "Feature weight optimization for extractive text summarization using evolutionary algorithms," *Expert Systems with Applications*, vol. 238, no. March, p. 121782, 2024. doi: 10.1016/j.eswa.2023.121782.