Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 1, November 2025, pp. 113∼126
ISSN: 2476-9843, accredited by Kementerian Pendidikan Tinggi, Sains dan Teknologi, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i1.5046

Assessing the Semantic Alignment in Multilingual Student-Teacher Concept Maps Using mBERT

Nadindra Dwi Ariyanta[1], Didik Dwi Prasetya[1], Ilham Ari Elbaith Zaeni[1], Reo Wicaksono[1], Tsukasa Hirashima[2]
[1]Universitas Negeri Malang, Malang, Indonesia
[2]Hiroshima University, Hiroshima, Japan

Article Info

Article history:

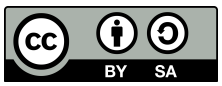
Received March 24, 2025
Revised July 20, 2025
Accepted September 30, 2025

Keywords:

Concept Map;
mBert;
Multilingual;
Open-ended;
Semantic Similarity;
TF-IDF.

ABSTRACT

This study examines the effectiveness of mBERT (Multilingual Bidirectional Encoder Representations from Transformers) in assessing semantic alignment between student and teacher concept maps in multilingual educational contexts, comparing its performance with TF-IDF. Using datasets in both Indonesian and English, the study demonstrates that mBERT outperforms TF-IDF in capturing complex semantic relationships, achieving 96% accuracy, 96% precision, 100% recall, and a 98% F1 score in the Indonesian dataset. In contrast, TF-IDF achieved higher precision (73%) and accuracy (79%) in the English dataset, where mBERT recorded 54% accuracy, 47% precision, but 90% recall. Semantic alignment was measured using cosine similarity to calculate the cosine of the angle between vectors representing textual embeddings generated by both models. This method facilitates cross-linguistic semantic comparison, overcoming challenges related to word frequency and syntactic variations. While mBERT's computational demands and the study's limited linguistic scope suggest room for improvement, the findings highlight the potential for hybrid models and emphasize the transformative impact of AI-driven tools, such as mBERT, in fostering inclusive and effective multilingual education.



Corresponding Author:

Didik Dwi Prasetya, +628129700011,
Department of Electrical and Informatics Engineering,
Universitas Negeri Malang, Malang, Indonesia,
Email: didikdwi@um.ac.id.

How to Cite:
N. D. Ariyanta, D. D. Prasetya, I. A. Elbaith Zaeni, T. Hirashima, and R. Wicaksono, "Assessing the Semantic Alignment in Multilingual Student-Teacher Concept Maps Using mBERT", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 1, pp. 113-126, November, 2025

1. INTRODUCTION

The increasing complexity of educational environments requires a comprehensive understanding of how multilingualism affects learning outcomes, particularly through tools such as concept maps. Concept maps serve as powerful visual aids for organizing knowledge and illustrating relationships among concepts [1]. Recent studies have shown that concept mapping significantly enhances comprehension and retention, especially in multilingual settings where language barriers can impede learning [2]. Consequently, integrating multilingual assessments is crucial to identify discrepancies in understanding and pinpoint areas requiring instructional support [3]. Open-ended concept maps, which enable students to create and connect concepts freely, are particularly beneficial in diverse linguistic environments. They encourage creativity and deeper engagement, enabling students to express understanding beyond language barriers [4]. In multilingual classrooms, assessing the semantic alignment between student and teacher concept maps is essential for identifying comprehension gaps. This alignment provides critical insights into how well students grasp material presented in different languages, particularly in contexts where instruction is not in their mother tongue [5]. The importance of this assessment lies in its ability to foster a more inclusive learning environment where students' diverse linguistic backgrounds are recognized and valued [6]. By employing concept maps, educators can better understand students' conceptual frameworks and adapt teaching strategies, ultimately enhancing educational equity [7].

The relevance of multilingual assessments is underscored by increasing diversity in global educational settings. Developing assessment tools that reflect and leverage students' linguistic resources is therefore crucial [8]. Research indicates that multilingual education can lead to improved cognitive flexibility and problem-solving skills, essential for success in today's globalized world [9]. Furthermore, multilingual assessments can help preserve and revitalize indigenous and minority languages, contributing to the preservation of cultural heritage and identity [10]. This study aims to explore the effectiveness of the multilingual BERT (mBERT) model in assessing semantic alignment between student and teacher concept maps in both Indonesian and English contexts. mBERT was chosen due to its proven ability to process and understand multiple languages simultaneously, a key aspect of its theoretical foundation as a transformer-based model pre-trained on extensive multilingual corpora [11]. Previous studies have highlighted mBERT's effectiveness in capturing complex semantic relationships across languages, making it an ideal tool for evaluating concept maps in diverse linguistic contexts [12]. Its architecture leverages shared linguistic features, which is particularly beneficial when students express similar concepts in different languages [13].

Moreover, the application of mBERT aligns with contemporary trends in educational technology, emphasizing the integration of artificial intelligence in learning environments. Research demonstrates the potential of AI-driven tools to enhance formative assessments by providing real-time feedback and insights into student understanding [14]. By utilizing mBERT, this study investigates the model's capacity to facilitate semantic alignment assessments between student and teacher concept maps, contributing to the ongoing discourse on the role of technology in education [15]. In addition to mBERT, the evaluation methods employed in this study, including TF-IDF and cosine similarity, are well-established techniques for measuring semantic similarity. These methods allow for a quantitative analysis of concept maps, providing a clear framework for comparing the conceptual understandings of students and teachers [16]. Comparative studies highlight that while deep learning models like LSTM achieve high accuracy in sequential interaction analysis, lightweight methods such as TF-IDF and cosine similarity offer computational efficiency a critical factor for scalable multilingual assessments [17]. The use of cosine similarity enables quantitative analysis of concept maps, providing a clear framework for comparing conceptual understandings and informing instructional strategies [18]. Furthermore, incorporating statistical measures such as the Intraclass Correlation Coefficient (ICC) and Pearson correlation will enhance the robustness of the findings, offering a comprehensive view of alignment [19].

The significance of this research extends beyond concept mapping, as it contributes to the broader field of multilingual education by providing insights into effective assessment practices. As educators increasingly recognize the value of students' linguistic diversity, studies like this can inform the development of pedagogical strategies that embrace multilingualism as a resource rather than a barrier [20]. By accurately assessing semantic alignment in multilingual settings, this research aims to highlight the potential for improved educational outcomes through tailored instruction that respects and incorporates students' linguistic backgrounds [21]. In conclusion, assessing semantic alignment between student and teacher concept maps in multilingual settings is a critical area of research that addresses the complexities of language diversity in education. This study is particularly important as it directly tackles the challenge of identifying discrepancies in understanding and comprehension gaps among multilingual learners, especially when they are acquiring knowledge in a non-native language, thereby fostering a more inclusive and equitable learning environment. By employing the mBERT model and established evaluation methods, this study aims to provide valuable insights into the effectiveness of concept mapping as a pedagogical tool in both Indonesian and English contexts, thereby contributing to the assessment of AI-based semantic alignment and its application within multilingual education. The findings will not only contribute significantly to the understanding of multilingual assessments but also support educators in creating truly inclusive and effective learning environments for all students [22].

2. RESEARCH METHOD

This study aims to evaluate the semantic alignment between student and teacher concept maps in a multilingual educational context. The research process consists of four main stages: Data Collection, Data Preparation, Modeling, and Evaluation. Each phase is carried out systematically to ensure the validity and reliability of the results. This approach enables an in-depth analysis of how students' concept maps reflect their understanding and how semantic alignment can be utilized to enhance learning in multilingual classrooms.

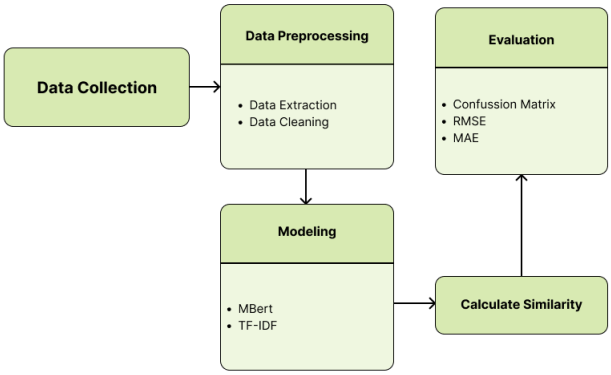


Figure 1. Flowchart research

Figure 1 illustrates the research process flow, which is divided into four main stages: Data Collection, Data Preprocessing, Modeling, and Evaluation. The first stage, Data Collection, involves gathering raw data, followed by Data Preprocessing, which includes data extraction and cleaning to prepare the data for further analysis. The Modeling stage utilizes two models, mBERT and TF-IDF, to process the data. Afterward, the Evaluation stage assesses the models' performance through metrics such as the confusion matrix, RMSE, and MAE. Finally, the Calculate Similarity step measures the semantic alignment between the student's and teacher's concept maps, thereby linking all components of the research workflow.

2.1. Data Collection

Two primary and distinct data sources were utilized in the data collection phase to ensure a comprehensive evaluation across different linguistic and disciplinary contexts. The Indonesian dataset was collected from a database course that focuses on relational database topics. These materials, including concept maps that cover basic concepts such as entities, relationships between tables, and normalization processes, were derived from actual course content designed by university instructors to help students understand the structure and fundamental principles of relational databases. This dataset was selected to represent a real-world academic scenario where students learn foundational concepts in their native language, providing a clear context for assessing semantic alignment in a structured educational setting. The dataset comprised concept maps from 27 individual students (as derived from Table 1, "Teacher - Student 1" to "Teacher - Student 27"), with each student's concept map formatted as an individual .txt file, allowing for effective textual analysis to capture their conceptual understanding [23].

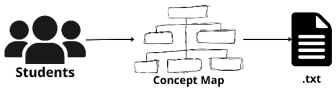


Figure 2. Data collection

Figure 2 demonstrates the Data Collection process, which begins with students creating concept maps that represent their understanding of a particular topic. These concept maps are then converted into text files (.txt format) for further analysis. The conversion from graphical concept maps to text-based formats enables easier processing and semantic analysis in subsequent stages of the research, facilitating comparison between student and teacher concept maps.

The English dataset was obtained from open-source materials on GitHub, with a specific focus on user authentication in cybersecurity. This dataset comprises concept maps that illustrate various aspects of user authentication, including methods, password management, and encryption techniques. This dataset was chosen to represent a distinct domain (technical/cybersecurity) and a different language, allowing for the examination of model robustness across varied lexical and conceptual complexities outside of a direct classroom collection. It provides industry-relevant content and facilitates the comparison of conceptual representations across academic and technical contexts. Unlike the Indonesian dataset, the English dataset required additional processing to extract textual representations for analysis. The English dataset comprised 24 student concept maps (as derived from the results discussion in Section 3.1 and Table 2's mention in relation to English dataset accuracy), alongside the corresponding teacher's reference map [24].

2.2. Data Preprocessing

During the data pre-processing phase, the datasets are cleaned and structured to ensure consistency and usability. For the Indonesian dataset, individual .txt files from students are consolidated into a single .csv file, with the teacher's file placed at the top as a reference. Similarly, for the English data set, textual data is extracted from student concept maps and then consolidated into a .csv file, following the same format. Placing the teacher's data at the top facilitates direct comparison and alignment during semantic analysis.

Preprocessing steps include removing irrelevant punctuation such as periods, commas, question marks, and exclamation points to avoid interference during modeling. All text is converted to lowercase to maintain consistency and simplify analysis. Additionally, excess white space is removed to ensure the data is clean and well-structured. These steps prepare the data sets for advanced semantic analysis using models such as mBERT, ensuring accurate and reliable results [25].

2.3. Modeling

In the modeling phase, the multilingual BERT (mBERT) model was utilized to assess the semantic alignment between student and teacher concept maps. Specifically, the 'bert-base-multilingual-cased' pre-trained model was utilized from the Hugging Face Transformers library, chosen for its proven capability in cross-lingual understanding and its broad coverage of 104 languages, including Indonesian and English. mBERT generates high-dimensional numerical embeddings from the input text, which encapsulate the contextual semantic meaning of words and phrases. These embeddings were then evaluated using cosine similarity to compute semantic alignment scores. Thresholds are applied to classify alignment levels, with values above 0.7 in the Indonesian dataset and 0.6 in the English dataset considered high alignment (1) [26, 27]. These specific thresholds were determined empirically through preliminary experiments and cross-validation on a subset of the respective datasets, aiming to optimize the balance between precision and recall for each language, given the characteristics of the concept map data. The lower threshold for English reflected a recognition of potentially greater lexical variability or semantic complexity in the cybersecurity domain compared to the more structured academic content in Indonesian, allowing mBERT to classify more instances as aligned without excessive false negatives. As a baseline for comparison, the TF-IDF model is also used to compute semantic alignment scores. TF-IDF represents text based on word frequency and significance, using cosine similarity for evaluation. A uniform threshold of 0.4 is set for TF-IDF in both datasets to reflect its statistical approach. This threshold was selected to provide a consistent and conservative baseline for TF-IDF, reflecting its inherent focus on direct term overlap. While TF-IDF provides a useful baseline, mBERT's deep learning capabilities allow for a more nuanced understanding of semantic relationships across multiple languages, which this study seeks to validate [28].

2.4. Evaluation

In the evaluation phase, the alignment scores are compared to the ground truth data to assess the model's performance. For the Indonesian dataset, the ground truth is binary, with alignment values classified as 0 or 1. For the English dataset, the ground truth initially represents the alignment as percentages or decimals (0 to 0.6), which are normalized to a range of 0 to 1 by dividing by 0.6. After normalization, a threshold of 0.6 is applied to classify high (1) and low (0) semantic alignment [29].

Several metrics are used to evaluate performance, including accuracy, precision, recall, F1 score, root mean square error (RMSE), and mean absolute error (MAE). Accuracy provides an overall measure of correct classifications, while Precision and Recall provide insight into specific types of errors. F1 Score balances Precision and Recall, highlighting the model's reliability in different scenarios. RMSE and MAE quantify prediction error, with RMSE penalizing larger errors more severely. These metrics provide a comprehensive view of the effectiveness of the models in assessing semantic alignment [30].

$$Accuracy = \frac{TP + TN}{N} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

The performance of mBERT is evaluated using several metrics to assess its ability to capture semantic alignment between student and teacher concept maps. Accuracy is calculated as the ratio of correctly predicted positive and negative instances to the total number of samples, as described in Formula 1. Precision measures the proportion of true positives (TP) to the total predicted positives, as explained in formula 2. Recall is the ratio of true positives (TP) to the total number of actual positives, as shown in Formula 3. The F1-Score, which combines both precision and recall, is calculated using Formula 4. Additionally, Root Mean Square Error (RMSE) is used to evaluate prediction error by computing the square root of the average squared differences between the ground truth values y_i and the predicted values $(\hat{y}_i)$, as explained in formula 5. Mean Absolute Error (MAE), which calculates the average absolute difference between predicted and actual values, is described in formula 6. These evaluation metrics ensure a comprehensive assessment of mBERT's performance, providing insights that can inform educational strategies and support multilingual learning environments.

The evaluation of mBERT's performance in capturing semantic alignment is carried out using several metrics, including accuracy, precision, recall, and F1-score. These metrics are calculated using the following parameters: True Positives (TP) represent instances that are correctly predicted as positive. At the same time, True Negatives (TN) refer to instances that are correctly predicted as negative. False Positives (FP) are instances that are incorrectly classified as positive, and False Negatives (FN) are those that are incorrectly classified as negative. The ground truth value is denoted as y_i, and the predicted value is represented as $(\hat{y}_i)$. The number of samples in the dataset is denoted as n. By employing these evaluation methods, the study ensures a robust assessment of mBERT's ability to align student and teacher concept maps semantically, offering valuable insights that can guide instructional strategies and enhance multilingual education.

3. RESULT AND ANALYSIS

In text-based data processing, accuracy and efficiency in document similarity analysis are crucial aspects. The system developed in this study uses text representation methods, specifically mBERT and TF-IDF, to generate comparable numerical vectors. Cosine similarity is used to measure the degree of similarity between documents based on the angle between their vectors. This approach aims to identify documents with certain levels of similarity according to predetermined thresholds.

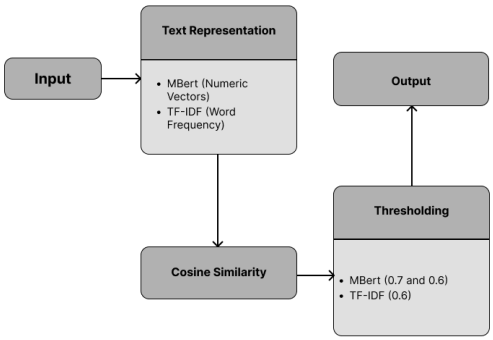


Figure 3. Flowchart system

3.1. Results

The similarity scores between teachers' and students' concept maps, calculated using TF-IDF and mBERT models, are presented in Table 1. For the Indonesian dataset, the results reveal a notable disparity: the average TF-IDF similarity score was 0.4629, while mBERT achieved a significantly higher average of 0.8711. The significant difference observed in the Indonesian dataset can be attributed to mBERT's deep contextual understanding capabilities. Indonesian, as an agglutinative language, often conveys meaning through prefixes, suffixes, and infixes, which can significantly alter the meaning and relationships of words. mBERT's pre-training on extensive multilingual corpora enables it to effectively process complex morphological structures and capture nuanced semantic connections that simpler bag-of-words models, such as TF-IDF, might miss, as TF-IDF primarily relies on individual term frequencies. For example, consider a student concept map in Indonesian discussing 'pengelolaan data' (data management) and a teacher's map using 'manajemen basis data' (database management). TF-IDF might assign a lower similarity due to the lexical differences. However, mBERT, leveraging its contextual understanding of Indonesian, would likely recognize the strong semantic equivalence between 'pengelolaan' and 'manajemen' in this context, resulting in a high similarity score, as evidenced in pairs like Guru-Siswa 2, which showed a TF-IDF score of 0.2156 but an mBERT score of 0.7826.

A similar trend in average similarity scores was observed in the English dataset, where TF-IDF yielded an average score of 0.3918, compared to mBERT's 0.6434. However, as further detailed in Table 2, TF-IDF ultimately outperformed mBERT in terms of accuracy (79.17% vs. 54%) and precision (0.7273 vs. 0.47) for the English dataset. This outcome in the English dataset suggests that the linguistic characteristics of the cybersecurity content, often characterized by precise terminology and less contextual ambiguity compared to general language, might align better with TF-IDF's frequency-based approach. While mBERT's contextual embeddings are powerful, its broader semantic understanding, when coupled with the specific thresholding applied, appears to lead to a higher rate of false positives in this particular English domain, impacting its precision and overall accuracy compared to TF-IDF's more conservative classification. For instance, in a cybersecurity concept map, if a student uses the exact technical term "multi-factor authentication" and the teacher also uses it, TF-IDF's direct term frequency match would contribute strongly to a high and precise similarity. mBERT might struggle more when distinguishing between highly specific, yet semantically related, technical jargon, sometimes over-generalizing and leading to false positives where maps are deemed similar but lack the precise alignment required for accurate classification in a highly technical domain.

Table 1 displays the detailed similarity scores for each student-teacher pair in the Indonesian dataset. For instance, while TF-IDF scores ranged moderately (e.g., 0.2156 for teacher-student 2), mBERT consistently produced higher values (e.g., 0.7826 for the same pair), with multiple instances nearing the maximum score (e.g., 0.9554 for teacher-student 5). This pattern underscores mBERT's robustness in aligning conceptual representations across linguistic and pedagogical contexts.

Table 1. Text Representation Results

Perbandingan	TF-IDF	M-Bert
Teacher - Student 1	0.5184	0.7826358
Teacher - Student 2	0.2156	0.7826358
Teacher - Student 3	0.5048	0.6952449
Teacher - Student 4	0.5069	0.7826358
Teacher - Student 5	0.5697	0.95541185
Teacher - Student 6	0.5588	0.8980879
Teacher - Student 7	0.5601	0.77951205
Teacher - Student 8	0.4692	0.9143783
Teacher - Student 9	0.5048	0.95541185
Teacher - Student 10	0.2949	0.9030227
Teacher - Student 11	0.4669	0.7144241
Teacher - Student 12	0.4309	0.7572096
Teacher - Student 13	0.4695	0.95541185
Teacher - Student 14	0.4028	0.886523
Teacher - Student 15	0.4778	0.95541185
Teacher - Student 16	0.4979	0.95541185
Teacher - Student 17	0.4453	0.8202536
Teacher - Student 18	0.513	0.7826358
Teacher - Student 19	0.5349	0.95541185
Teacher - Student 20	0.5349	0.95541185

(continued on next page)

Table 2 (continued)

Perbandingan	TF-IDF	M-Bert
Teacher - Student 21	0.5565	0.95541185
Teacher - Student 22	0.2778	0.9077907
Teacher - Student 23	0.2454	0.76943403
Teacher - Student 24	0.5423	0.9030227
Teacher - Student 25	0.3515	0.8984235
Teacher - Student 26	0.5487	0.95541185
Teacher - Student 27	0.4978	0.95541185

In the first analysis phase, similarity scores were calculated for both models, TF-IDF and mBERT, to compare students' concept maps with those of the teachers. Figure 4(a) and Figure 4(c) display the scatter plot similarity scores for TF-IDF on the Indonesian and English datasets, respectively. For the Indonesian dataset, the average TF-IDF similarity score was 0.4629, which is represented in Figure 4(a), showing a moderate alignment between the student and teacher concept maps. Similarly, Figure 4(c) shows the TF-IDF results for the English dataset, with an average score of 0.3918, indicating a lower degree of alignment.

In contrast, Figures 4(b) and 4(d) show the scatter plot similarity scores for mBERT on the Indonesian and English datasets. For the Indonesian dataset, mBERT scored significantly higher, with an average similarity score of 0.8711, as shown in Figure 4(b), demonstrating a much better alignment between the concept maps. Likewise, Figure 4(d) for the English dataset shows mBERT achieving an average similarity score of 0.6434, which is notably higher than the TF-IDF score in Figure 4(c). These results underscore mBERT's superior ability to capture semantic relations, as reflected in the higher similarity scores and better alignment with the teacher's concept map. The visual comparison in the scatter plots clearly demonstrates the robustness of mBERT in multilingual contexts, capturing more nuanced semantic alignments than TF-IDF.

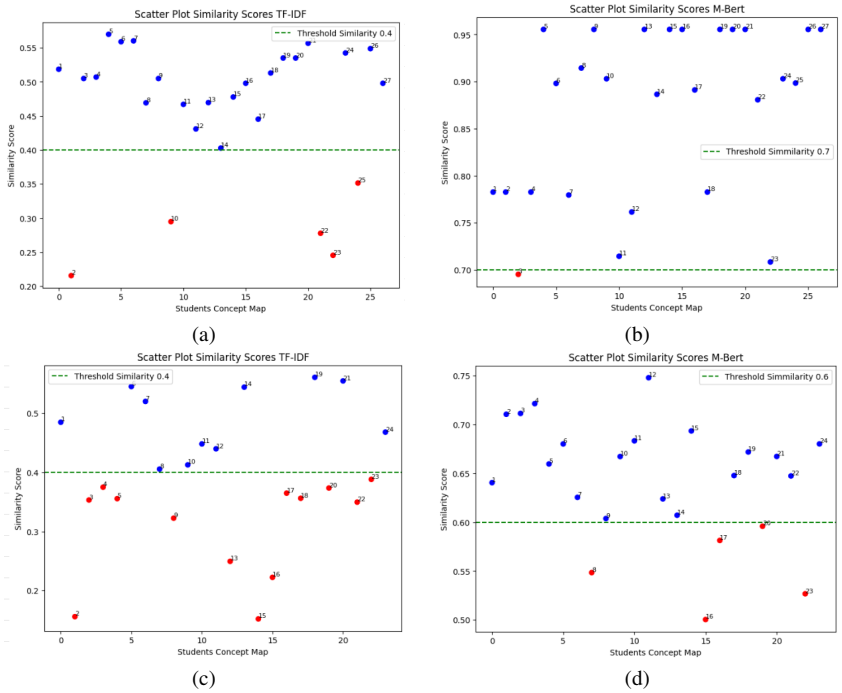


Figure 4. Thresholding

The thresholding process is crucial for converting similarity scores into binary classifications, enabling the structured analysis of semantic relationships. For TF-IDF, a threshold of 0.4 was applied uniformly across all datasets, resulting in moderate sensitivity. In contrast, mBERT thresholds were set at 0.7 for the Indonesian and 0.6 for the English datasets to account for dataset-specific variability. These thresholds improved the recall of mBERT without significantly compromising precision, especially for the Indonesian dataset.

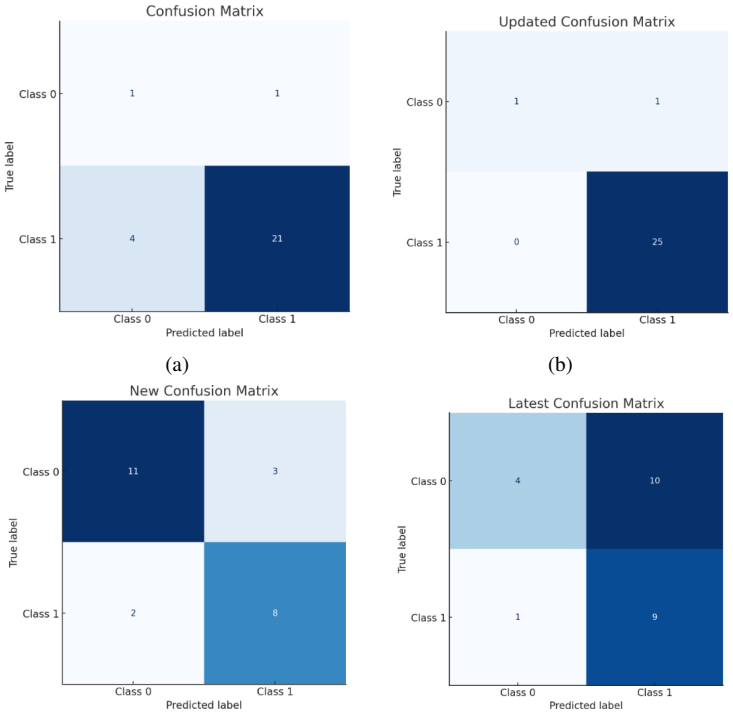


Figure 5. Confussion matrix

Figures 5(a) and 5(b) show the confusion matrices for the Indonesian dataset, with mBERT performing in the initial and updated phases. In Figure 5(a), the model exhibits a higher number of false positives, as shown by the misclassification of Class 1 instances as Class 0, leading to an imbalance in precision. This suggests that initially, mBERT might have identified some semantic connections that, while plausible in a broad sense, did not meet the strict ”aligned” criteria of the ground truth, leading to an over-prediction of similarity for some unaligned pairs. However, Figure 5(b) demonstrates an improvement, with mBERT achieving minimal false negatives and better overall classification of Class 1 instances, reflecting high recall. This indicates that mBERT effectively learned to identify most truly aligned concept maps in Indonesian, even if the student's expression of concepts differed lexically from the teacher's, validating its strength in handling semantic nuances in agglutinative languages.

For the English dataset, Figures 5(c) and 5(d) show the confusion matrices for the new and latest phases, respectively. In Figure 5(c), mBERT captures more true positives but also misclassifies some Class 0 instances as Class 1, as indicated by the higher false positives. Specifically, these false positives in the English dataset often arose when student maps contained general cybersecurity terms that were broadly related to the teacher's map but lacked the precise, domain-specific conceptual links required for true alignment. mBERT's contextual breadth, while powerful, sometimes led to over-generalization in this highly technical domain. By the latest phase in Figure 5(d), the model shows better balance between precision and recall, with fewer false positives and improved identification of both classes. This suggests that through refinement, mBERT became slightly more discerning, reducing instances where it incorrectly predicted high alignment. These results underscore mBERT's ability to improve as it adapts to linguistic nuances in different languages, with the confusion matrices providing a more detailed insight into prediction distributions and errors.

Table 2. Experiments Results

Language	**MODEL**	**Experiment Results**					
		Accuracy	***Precision***	***Recall***	**F1 - Score**	**RMSE**	**MAE**
Indonesia	MBERT	**0.96**	**0.96**	**1**	**0.98**	**0.19**	**0.04**
	TF-IDF	0.81	0.95	0.84	0.89	0.43	0.18
English	MBERT	0.54	0.47	**0.9**	0.62	0.68	0.46
	TF-IDF	**0.79**	**0.73**	0.8	**0.76**	**0.46**	**0.21**

Table 2 presents the results of the binary classification experiments using various metrics, including accuracy, precision, recall, F1 score, RMSE, and MAE. In the Indonesian dataset, mBERT outperformed TF-IDF across nearly all metrics. mBERT achieved an accuracy of 96.30%, with a precision of 0.9615, recall of 1.0000, and F1 score of 0.9804, as shown in Table 2. In contrast, TF-IDF demonstrated an accuracy of 81.48%, precision of 0.9545, and recall of 0.8400, reflecting a more cautious approach but with a higher risk of missing relevant relationships, as seen in the lower recall. The lower recall for TF-IDF here indicates a higher number of false negatives, as it failed to identify many truly aligned concept maps. This is likely because it was unable to capture the semantic equivalence between lexically different but conceptually similar terms, which mBERT successfully handled.

In the English dataset, TF-IDF outperformed mBERT overall, achieving an accuracy of 79.17%, a precision of 0.7273, and a recall of 0.8000, as detailed in Table 2. These variations highlight the influence of threshold selection on each model's sensitivity and specificity, demonstrating how mBERT excels in capturing semantic relationships in the Indonesian dataset. At the same time, TF-IDF is more balanced but slightly more conservative in the English dataset. The lower precision of mBERT in English, coupled with its higher recall, confirms its tendency towards more false positives (over-predicting alignment). In contrast, TF-IDF's higher precision means it was more accurate when it did predict alignment, but it missed more true positives (hence lower recall), indicating more false negatives. This suggests TF-IDF's strength in this context lies in its conservative, term-matching approach, which avoids many of the false positives generated by mBERT's broader semantic interpretations.

3.2. Discussion

The results of this study clearly highlight the advantages of mBERT in capturing complex semantic relationships, particularly within the Indonesian dataset. Its superior recall in this context demonstrates its effectiveness in identifying a wider range of semantic similarities, which is crucial in educational settings where a comprehensive understanding of student knowledge is paramount. By providing a more nuanced analysis, mBERT offers deeper insights into students' conceptual understanding, enabling educators to better gauge alignment with teacher-provided concept maps. This strong performance on the Indonesian dataset, where mBERT significantly outperformed TF-IDF (0.8711 vs. 0.4629 average similarity), can be attributed to mBERT's ability to capture contextual semantic relationships through its deep learning architecture. Being pre-trained on extensive multilingual corpora, mBERT effectively understands the meaning of words and phrases in context, even across languages. This pre-training enables mBERT to effectively utilize contextual embeddings for tasks that require nuanced understanding, thereby enhancing its performance on simple mappings where the relationships between concepts are more straightforward and the text is less dense. Studies have shown that transformer-based models such as mBERT excel at tasks requiring contextual understanding, such as semantic similarity and text classification, especially when the input text is concise and well structured [31–33]. The representation of context within mBERT's architecture allows it to adapt to different linguistic nuances in different languages, making it particularly strong at interpreting short texts effectively [34–36].

However, the English dataset revealed notable trade-offs, with mBERT exhibiting lower accuracy (54% vs. TF-IDF's 79%) and precision (47% vs. TF-IDF's 73%) despite achieving high recall (90%). This suggests that mBERT's performance may vary depending on dataset characteristics, such as linguistic complexity, domain specificity (e.g., cybersecurity technical terms), and variability in ground truth values. For instance, mBERT may struggle with long or highly complex concept maps, where the relationships between concepts are more intricate and the text is more verbose. This complexity can exacerbate problems such as overfitting and misinterpretation of semantic relationships, leading to lower accuracy and an increased rate of false positives. Recent studies suggest that while deep learning models can capture broad semantic relationships, their reliance on contextual embeddings may misrepresent more subtle relationships found in more complex text, ultimately reducing their effectiveness in tasks that require precision [37–39].

Conversely, TF-IDF's more conservative, frequency-based approach yielded higher accuracy and precision in the English dataset, making it advantageous in applications where minimizing false positives is critical. Its reliance on word frequency patterns allows for a more cautious classification process. TF-IDF performs effectively with long or complex concept maps because it focuses on the statistical significance of individual words rather than their contextual relationships. Research suggests that by focusing on term frequency, TF-IDF is inherently less susceptible to noise and overfitting, especially in longer texts, where the frequency of specific terms can serve as a reliable indicator of semantic alignment [40–42]. Despite these advantages, this precision-focused mechanism comes at the expense of recall, as TF-IDF may fail to capture subtle semantic relationships, particularly in shorter or simpler texts where contextual understanding is more important than mere word frequency. In essence, while TF-IDF successfully curates a statistically sound representation for longer texts, its limitations stem from an inability to discern the deeper meanings often required for nuanced interpretations found in less verbose data [43–45].

This study's findings have significant implications for educational technology and multilingual pedagogy. The demonstrated potential of mBERT in assessing semantic alignment provides a powerful tool for formative assessment, enabling educators to gain real-time insights into student comprehension across diverse linguistic contexts. This can inform adaptive instructional strategies,

leading to more tailored and effective learning experiences. Theoretically, the study contributes to the understanding of how advanced NLP models can bridge linguistic barriers in knowledge representation, pushing the boundaries of automated assessment beyond traditional word-matching techniques.

Despite these contributions, the study has certain limitations. The primary focus on only Indonesian and English datasets restricts the generalizability of the results to other language contexts. Expanding future research to include languages with different linguistic structures, such as tonal or highly agglutinative languages beyond Indonesian, would provide broader and more robust insights. Furthermore, the empirical thresholds used in this study (0.7 for Indonesian mBERT, 0.6 for English mBERT, and 0.4 for TF-IDF across both) may not generalize universally to other datasets, suggesting a need for systematic optimization or adaptive thresholding methods in future work. The computational requirements of mBERT also present practical challenges for implementation in resource-constrained educational environments, which may limit its immediate scalability.

Future research should explore the integration of mBERT with traditional models such as TF-IDF to develop hybrid models that combine their respective strengths. Such a hybrid approach could address the observed trade-offs, potentially improving overall performance across diverse linguistic and textual complexities. Additionally, exploring real-time classroom applications of mBERT could revolutionize formative assessment by providing teachers with immediate, actionable feedback, enabling dynamic and adaptive instructional strategies tailored to students' evolving needs. Ultimately, leveraging mBERT's capabilities to support and assess learning in indigenous and minority languages holds significant potential for promoting educational equity and preserving cultural heritage. These findings collectively underscore the transformative role of AI tools, such as mBERT, in creating inclusive and effective multilingual learning environments.

4. CONCLUSION

This study evaluated the effectiveness of TF-IDF and mBERT in assessing semantic alignment between student and teacher concept maps in multilingual educational contexts, with a specific focus on datasets in Indonesian and English. The results show that mBERT significantly outperformed TF-IDF in capturing complex semantic relationships, especially in the Indonesian dataset, where it achieved 96% accuracy, 96% precision, 100% recall, and a 98% F1 score. In contrast, TF-IDF in the Indonesian dataset showed 81% accuracy, 95% precision, 84% recall, and an 89% F1 score. However, mBERT's performance in the English dataset showed limitations, achieving 54% accuracy and 47% precision, despite a 90% recall rate. Conversely, TF-IDF demonstrated stronger performance in the English dataset, achieving 79% accuracy and 73% precision (with 80% recall), making it more suitable for applications where minimizing false positives is critical.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AI USAGE STATEMENT

The authors acknowledge that Artificial Intelligence tools, including ChatGPT developed by OpenAI, were utilized to support language refinement, grammar correction, and paraphrasing in the manuscript preparation process. The authors confirm that all ideas, data interpretations, and conclusions are their own and not generated by the AI tool.

AUTHOR CONTIBUTION

Reo Wicaksono, the first author, conceived and designed the research, conducted data collection and analysis, and drafted the initial manuscript. Didik Dwi Prasetya, the second author, contributed to the experimental design, performed statistical analysis, and provided critical revisions to improve the manuscript. Author 3, Ilham Ari Elbaith, provides technical expertise, develops models or tools, and reviews the statistical or computational methods. Author 4, Nadindra Dwi Ariyanta, contributed to dataset extraction and research review. Author 5, Senior Advisor, Tsukasa Hirashima, offers guidance on framing the research, reviews the manuscript critically, and ensures it meets academic standards for publication.

FUNDING STATEMENT

This research was funded independently by the researchers, without any financial support or assistance from government institutions, private organizations, or other external sources.

COMPETING INTEREST

The authors confirm that there are no conflicts of interest, either financial or non-financial, that could influence the research results

and interpretation of the data in this article.

REFERENCES

[1] D. D. Prasetya, T. Widiyaningtyas, and T. Hirashima, "Interrelatedness patterns of knowledge representation in extension concept mapping," vol. 20, p. 009, May,2024, https://doi.org/10.58459/rptel.2025.20009.

[2] M. Konu KadiRhanogullari, "The Effect of Teaching with Concept Maps on Academic Success in Biology Teaching: A Meta-Analysis Study," no. 58, pp. 2781–2796, December, 2023, https://doi.org/10.53444/deubefd.1324169.

[3] C. G. M. Fine and E. M. Furtak, "A framework for science classroom assessment task design for emergent bilingual learners," vol. 104, no. 3, pp. 393–420, May, 2020, https://doi.org/10.1002/sce.21565.

[4] D. D. Prasetya, A. Pinandito, Y. Hayashi, and T. Hirashima, "Analysis of quality of knowledge structure and students' perceptions in extension concept mapping," vol. 17, no. 1, p. 14, December, 2022, https://doi.org/10.1186/s41039-022-00189-9.

[5] J. M. Goodrich, L. Thayer, and S. Leiva, "Evaluating Achievement Gaps Between Monolingual and Multilingual Students," vol. 50, no. 7, pp. 429–441, October, 2021, https://doi.org/10.3102/0013189X21999043.

[6] S. Bhatia, S. Bhatia, and I. Ahmed, "Automated Waterloo Rubric for Concept Map Grading," vol. 9, pp. 148 590–148 598, 2021, https://doi.org/10.1109/ACCESS.2021.3124672.

[7] J. Nordmeyer, "From Testing to Teaching: Equity for Multilingual Learners in International Schools," vol. 125, no. 7–8, pp. 247–275, August, 2023, https://doi.org/10.1177/01614681231194413.

[8] J. Mancilla-Martinez, J. K. Hwang, and M. H. Oh, "Assessment Selection for Multilingual Learners' Reading Development," vol. 75, no. 3, pp. 351–362, November, 2021, https://doi.org/10.1002/trtr.2053.

[9] M. E. Flognfeldt, D. Tsagari, D. Šurkalović, and T. Tishakov, "The practice of assessing Norwegian and English language proficiency in multilingual elementary school classrooms in Norway," vol. 17, no. 5, pp. 519–540, October, 2020, https://doi.org/10.1080/15434303.2020.1827409.

[10] M. Aparici, E. Rosado, and L. Tolchinsky, "Multilingual use assessment questionnaire: A proposal for assessing language and literacy experience," vol. 9, p. 1394727, May, 2024, https://doi.org/10.3389/fcomm.2024.1394727.

[11] N. Donmez Usta, E. ultay, and N. ultay, "Reading the Concept Map of Physics Teacher Candidates: A Case of Light," vol. 31, no. 1, pp. 14–21, March, 2020, https://doi.org/10.33828/sei.v31.i1.2.

[12] N. Fauziah, N. Izzati, and H. Handoko, "Development of Cooperative Integrated Reading and Composition Learning Model with Mind Mapping Method to Improve Students' Understanding of Mathematical Concepts," vol. 1, no. 3, pp. 117–130, November, 2022, https://doi.org/10.58421/gehu.v1i3.27.

[13] H. L. Blake, "Intelligibility Enhancement via Telepractice During COVID-19 Restrictions," vol. 5, no. 6, pp. 1797–1800, December, 2020, https://doi.org/10.1044/2020_PERSP-20-00133.

[14] A. L. Ferrell, L. Soltero-González, and S. Kamioka, "Beyond English centrality: Integrating expansive conceptions of language for literacy programming into IEPs," vol. 9, p. 1347503, May, 2024, https://doi.org/10.3389/feduc.2024.1347503.

[15] D. Colla, E. Mensa, and D. P. Radicioni, "LessLex: Linking Multilingual Embeddings to SenSe Representations of LEXical Items," vol. 46, no. 2, pp. 289–333, June, 2020, https://doi.org/10.1162/coli_a_00375.

[16] J. Heuzeroth and A. Budke, "The Effects of Multilinguality on the Development of Causal Speech Acts in the Geography Classroom," vol. 10, no. 11, p. 299, October, 2020, https://doi.org/10.3390/educsci10110299.

[17] F. A. S. Laily, D. D. Prasetya, A. N. Handayani, and T. Hirashima, "Revealing Interaction Patterns in Concept Map Construction Using Deep Learning and Machine Learning Models," vol. 24, no. 2, pp. 207–218, Februari, 2025, https://doi.org/10.30812/matrik.v24i2.4641.

[18] S. Qin, L. Orchakova, Z.-Y. Liu, Y. Smirnova, and E. Tokareva, "Using the Learning Management System "Modular Object-Oriented Dynamic Learning Environment" in Multilingual Education," vol. 17, no. 03, pp. 173–191, Februari, 2022, https://doi.org/10.3991/ijet.v17i03.25851.

[19] J. O. Uguru, "A Lexico-phonetic Comparison of Olukumi and Lukumi: A Procedure for Developing a Multilingual Dictionary," vol. 31, no. 1, May, 2021, https://doi.org/10.5788/31-1-1643.

[20] R. A. J. R. Peixoto, "Political Boundaries in Language Policies: A Discussion on Institutional Settings," vol. 24, pp. e–1982–4017–24–17, 2024, https://doi.org/10.1590/1982-4017-24-17.

[21] A. I. Anisimova, N. A. Safonova, M. Y. Dobrushyna, N. O. Lysenko, and I. H. Bezrodnykh, "Verbalization of the concept language policy: Online research," vol. 5, no. S4, pp. 1301–1311, November, 2021, https://doi.org/10.21744/lingcure.v5nS4.1779.

[22] M. Perquin, S. Viswanathan, M. Vaillant, O. Risius, L. Huiart, J.-C. Schmit, N. J. Diederich, G. R. Fink, and J. Kukolja, "An individualized functional magnetic resonance imaging protocol to assess semantic congruency effects on episodic memory in an aging multilingual population," vol. 14, p. 873376, July, 2022, https://doi.org/10.3389/fnagi.2022.873376.

[23] J. M. Giesinger, F. L. Loth, N. K. Aaronson, J. I. Arraras, G. Caocci, F. Efficace, M. Groenvold, M. Van Leeuwen, M. A. Petersen, J. Ramage, K. A. Tomaszewski, T. Young, and B. Holzner, "Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research," vol. 118, pp. 1–8, Februari, 2020, https://doi.org/10.1016/j.jclinepi.2019.10.003.

[24] K. Ro, J. Y. Kim, H. Park, B. H. Cho, I. Y. Kim, S. B. Shim, I. Y. Choi, and J. C. Yoo, "Deep-learning framework and computer assisted fatty infiltration analysis for the supraspinatus muscle in MRI," vol. 11, no. 1, p. 15065, July, 2021, https://doi.org/10.1038/s41598-021-93026-w.

[25] B. A. Polascik, J. Peck, N. Cepeda, S. Lyman, and D. Ling, "Reporting Clinical Significance in Hip Arthroscopy: Where Are We Now?" vol. 16, pp. 527–533, December, 2020, https://doi.org/10.1007/s11420-020-09759-3.

[26] R. A. Binder, G. F. Fujimori, C. S. Forconi, G. W. Reed, L. S. Silva, P. S. Lakshmi, A. Higgins, L. Cincotta, P. Dutta, M.-C. Salive, V. Mangolds, O. Anya, J. M. Calvo Calle, T. Nixon, Q. Tang, M. Wessolossky, Y. Wang, D. A. Ritacco, C. S. Bly, S. Fischinger, C. Atyeo, P. O. Oluoch, B. Odwar, J. A. Bailey, A. Maldonado-Contreras, J. P. Haran, A. G. Schmidt, L. Cavacini, G. Alter, and A. M. Moormann, "SARS-CoV-2 Serosurveys: How Antigen, Isotype and Threshold Choices Affect the Outcome," vol. 227, no. 3, pp. 371–380, Februari, 2023, https://doi.org/10.1093/infdis/jiac431.

[27] M. Franceschini, A. Boffa, E. Pignotti, L. Andriolo, S. Zaffagnini, and G. Filardo, "The Minimal Clinically Important Difference Changes Greatly Based on the Different Calculation Methods," vol. 51, no. 4, pp. 1067–1073, March, 2023, https://doi.org/10.1177/03635465231152484.

[28] J. Bordon, "The Importance of Cycle Threshold Values in the Evaluation of Patients with Persistent Positive PCR for SARS-CoV-2: Case Study and Brief Review," vol. 4, no. 1, pp. 1–5, 2020, https://doi.org/10.18297/jri/vol4/iss1/54.

[29] J. Bullard, K. Dust, D. Funk, J. E. Strong, D. Alexander, L. Garnett, C. Boodman, A. Bello, A. Hedley, Z. Schiffman, K. Doan, N. Bastien, Y. Li, P. G. Van Caeseele, and G. Poliquin, "Predicting Infectious Severe Acute Respiratory Syndrome Coronavirus 2 From Diagnostic Samples," vol. 71, no. 10, pp. 2663–2666, December, 2020, https://doi.org/10.1093/cid/ciaa638.

[30] S. Jayatilake, J. M. Bunker, A. Bhaskar, and M. Miska, "Time–space analysis to evaluate cell-based quality of service in bus rapid transit station platforms through passenger-specific area," vol. 13, no. 2, pp. 395–427, June, 2021, https://doi.org/10.1007/s12469-021-00267-z.

[31] M. F. Bonner and R. A. Epstein, "Object representations in the human brain reflect the co-occurrence statistics of vision and language," vol. 12, no. 1, p. 4081, Februari, 2021, https://doi.org/10.1038/s41467-021-24368-2.

[32] C. Qu, M. F. Bonner, N. K. DeWind, and E. M. Brannon, "Contextual coherence increases perceived numerosity independent of semantic content." vol. 153, no. 8, pp. 2028–2042, August, 2024, https://doi.org/10.1037/xge0001595.

[33] M. C. Iordan, T. Giallanza, C. T. Ellis, N. M. Beckage, and J. D. Cohen, "Context Matters: Recovering Human Semantic Structure from Machine Learning Analysis of Large-Scale Text Corpora," vol. 46, no. 2, p. e13085, Februari, 2022, https://doi.org/10.1111/cogs.13085.

[34] B. Cao and J. Liu, "Combining bidirectional long short-term memory and self-attention mechanism for code search," vol. 35, no. 10, p. e7662, May, 2023, https://doi.org/10.1002/cpe.7662.

[35] R. Richie and S. Bhatia, "Similarity Judgment Within and Across Categories: A Comprehensive Model Comparison," vol. 45, no. 8, p. e13030, August, 2021, https://doi.org/10.1111/cogs.13030.

[36] D. Rose and P. Bex, "The Linguistic Analysis of Scene Semantics: LASS," vol. 52, no. 6, pp. 2349–2371, December, 2020, https://doi.org/10.3758/s13428-020-01390-8.

[37] J. C. Yang, "The prediction and analysis of heart disease using XGBoost algorithm," vol. 41, no. 1, pp. 61–68, Februari, 2024, https://doi.org/10.54254/2755-2721/41/20230711.

[38] L. Xu, S. Liu, S. Wang, D. Sun, and N. Li, "Word's Predictability Can Modulate Semantic Preview Effect in High-Constraint Sentences," vol. 13, p. 849351, March, 2022, https://doi.org/10.3389/fpsyg.2022.849351.

[39] A. Onan and H. Alhumyani, "Contextual Hypergraph Networks for Enhanced Extractive Summarization: Introducing Multi-Element Contextual Hypergraph Extractive Summarizer (MCHES)," vol. 14, no. 11, p. 4671, May, 2024, https://doi.org/10.3390/app14114671.

[40] Y. Zhu, W. Zheng, and H. Tang, "Interactive Dual Attention Network for Text Sentiment Classification," vol. 2020, pp. 1–11, March, 2020, https://doi.org/10.1155/2020/8858717.

[41] B. Tang, J. Wang, H. Qiu, J. Yu, Z. Yu, and S. Liu, "Attack Behavior Extraction Based on Heterogeneous Cyberthreat Intelligence and Graph Convolutional Networks," vol. 74, no. 1, pp. 235–252, 2023, https://doi.org/10.32604/cmc.2023.029135.

[42] N. H. Hameed, A. M. Alimi, and A. T. Sadiq, "Short Text Semantic Similarity Measurement Approach Based on Semantic Network," vol. 19, p. 1581, May, 2022, https://doi.org/10.21123/bsj.2022.7255.

[43] W. Pasisingi, A. Mariana, and D. Husain, "A Semantic Analysis on Maroon 5 Songs," vol. 2, no. 1, August, 2022, https://doi.org/10.30984/jeltis.v2i1.1948.

[44] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images," vol. 59, no. 1, pp. 426–435, Januari, 2021, https://doi.org/10.1109/TGRS.2020.2994150.

[45] W. Ma, Y. Wu, F. Cen, and G. Wang, "MDFN: Multi-scale deep feature learning network for object detection," vol. 100, p. 107149, April, 2020, https://doi.org/10.1016/j.patcog.2019.107149.

[This page intentionally left blank.]