Machine Learning for Open-ended Concept Map Proposition Assessment: Impact of Length on Accuracy

Reo Wicaksono[1], **Didik Dwi Prasetya**[1], **Ilham Ari Elbaith Zaeni**[1], **Nadindra Dwi Ariyanta**[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 31, 2025
Accepted September 24, 2025

Keywords:

Concept Map;
Classification;
Machine Learning;
Open-ended;
Text Representation.

ABSTRACT

Open-ended concept maps allow learners to freely connect concepts, enriching understanding by linking new and prior knowledge. However, manually assessing proposition quality is time-consuming and subjective. This study proposes an automatic classification model for proposition quality assessment using term frequency–inverse document frequency (TF-IDF), a text representation method based on word frequency, and several machine learning algorithms. Two datasets were used: a Relational Database with an average of 5 words per proposition and a Cybersecurity Authentication with an average of 10 words per proposition. Comparative experiments with Support Vector Machine (SVM), a supervised classification algorithm, K-Nearest Neighbor, Random Forest, and Long Short-Term Memory (LSTM), a recurrent neural network for sequence data, revealed that SVM with RBF kernel achieved the highest performance on shorter propositions 87% accuracy, Cohen's Kappa 0.76, while LSTM showed greater strength in handling longer propositions 85% accuracy, Cohen's Kappa 0.69. These findings suggest that proposition length influences model effectiveness. The proposed approach can reduce the burden of manual assessment, increase the objectivity of evaluation, and support more efficient implementation of concept maps in 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:
R. Wicaksono, D. Dwi Prasetya, I. A. Elbaith Zaeni, N. D. Ariyanta, and T. Hirashima, ”Machine Learning for Open-ended Concept Map Proposition Assessment: Impact of Length on Accuracy”, *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 1, pp. 1-14, November, 2025.

1. INTRODUCTION

Concept maps are visual representations that map relationships between concepts, often used as educational aids to improve students' conceptual understanding. The technique was first introduced by Novak in 1972 and has become a popular method in education due to its ability to simplify the complexity of concepts [1, 2]. In education, concept maps provide a framework for systematically designing, organising, and assessing knowledge [3]. In this digital age, its use has expanded with the help of digital tools that enable further integration of analysis and visualisation [**?**]. The benefit of concept maps lies in their flexibility in supporting constructivist-based learning. It helps students connect new knowledge with existing knowledge, thus deepening their understanding of the material learnt [4, 5]. In addition, concept maps are often used in various disciplines to improve students' analytical and critical thinking skills, including at the primary to university level [6]. However, a major challenge in implementing concept maps, especially open-ended ones, is assessing the quality of the resulting propositions. Manual assessment requires significant time and has a high subjectivity bias [1, 7]. Therefore, the development of automatic classification models to assess the propositional quality of concept maps is an important need in modern education. Several previous studies have explored machine learning algorithms for analysing concept maps, but most of them focus on closed types and lack consideration of open-ended concept maps [6].

Research on concept maps and their classification has shown the great potential of machine learning to accelerate and improve assessment accuracy. For example, the use of Support Vector Machine (SVM) and Random Forest algorithms has been applied for text classification in educational contexts, showing promising results [8, 9]. Other studies have explored the benefits of deep learning-based models such as BERT in understanding more complex text contexts, although their implementation on concept maps is still very limited [10]. However, most of these studies focus only on applying algorithms to general text classification or to closed concept maps and do not address the specific challenges of open-ended concept maps. In particular, they rarely compare how classical machine learning models and deep learning approaches perform when proposition characteristics, such as length, vary. In this research, the main novelty lies in combining classical approaches such as SVM, KNN, and Random Forest with TF-IDF and BERT-based representations to assess the quality of propositions in open-ended concept maps.

This study aims to fill the gap in the literature by analyzing the potential differences between short and long propositions in open-ended concept maps. The length of propositions is a critical factor: shorter propositions tend to be simpler and easier to classify, while longer propositions may involve more complex relationships and sequential patterns that challenge algorithms. Neglecting this variable can result in biased or less reliable automatic assessments. Therefore, investigating proposition length provides stronger justification for developing more robust and fair evaluation models. In this research, two concept map datasets are used: Relational Databases and Cyber Security Authentication, with average proposition lengths of 5 words and 10 words, respectively. We seek to examine the influence of TF-IDF text representation on the model's performance, as well as its performance on long and short propositions. This is expected to provide new insights into the automatic quality assessment of propositions in open-ended concept maps [11, 12]. Thus, this research not only focuses on the technical development of classification models but also on understanding the impact of proposition length on the model's effectiveness in the educational context. This issue is important because manual assessment of concept maps is not only time-consuming but also highly subjective, which limits their large-scale use in education. Moreover, ignoring the effect of proposition length can lead to biased or inaccurate automatic assessments. If this challenge is not addressed, educators will continue to face heavy workloads and unreliable evaluation outcomes, which may hinder the effective integration of concept maps and artificial intelligence in digital learning environments. It is hoped that the results of this study will significantly improve the efficiency and accuracy of assessing the quality of concept maps, as well as encourage the adoption of artificial intelligence technology in education. The implementation of this model is expected to help educators reduce the burden of manual assessment and improve the quality of learning evaluation, especially in concept-based education systems in the digital era [13, 14].

2. RESEARCH METHOD

This study employs a systematic approach to assess the propositional quality of open-ended concept maps using machine learning. Two datasets were used: a Relational Database with an average proposition length of five words and a Cybersecurity Authentication with an average proposition length of ten words. These datasets were selected to enable a comparison between short and long propositions. Data preprocessing included case folding, tokenization, and lemmatization, followed by numerical representation using TF-IDF with up to ten thousand features and a bigram range. Three classical machine learning models were implemented. The SVM applied a radial basis function kernel with C set to one, gamma set to scale, and class balancing enabled. The KNN algorithm was trained with five neighbors, while the Random Forest consisted of 100 trees with balanced weights. A Bidirectional LSTM was also employed, consisting of 128 hidden units, a dropout rate of 0.5, and a dense layer of 128 neurons, trained using the Adam optimizer for 50 epochs with a batch size of 32. Validation was conducted using stratified train-test splits, with ratios of 70:30 for SVM, KNN, and Random Forest, and 80:20 for LSTM. Model performance was evaluated using accuracy, precision, recall, F1-score, and Cohen's Kappa. The overall workflow is illustrated in Figure 1.

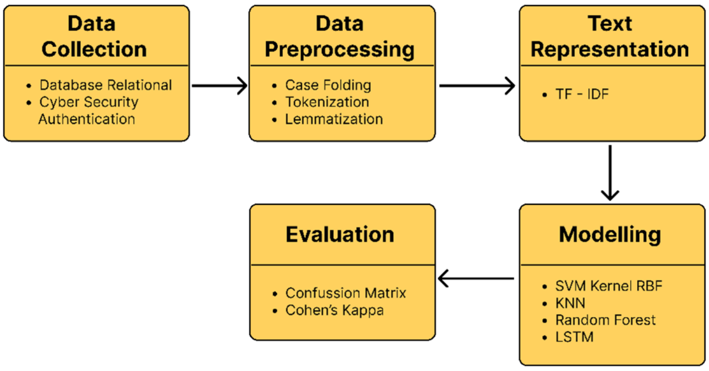

Figure 1. Flowchart Research

2.1. Data Collection

The dataset used in this research consists of 3,759 words from 30 concept maps, totaling 691 propositions. The average length of propositions in the Relational Database dataset is five words. Each proposition represents a sentence or claim made regarding a specific concept, such as ”Advantages of Relational Databases make data operations easy” or ”A relation must have a primary key”. Additionally, the Cybersecurity Authentication dataset consists of 1,465 propositions with a total of 13,840 words, with an average proposition length of 10 words [1, 15]. Each proposition is labelled with a score between 0 and 3 based on its quality by the expert [16]. A score of 0 indicates an irrelevant or incomprehensible proposition. A score of 1 indicates a proposition that is less precise or has an ambiguous sentence structure [17]. A score of 2 indicates a proposition that is fairly clear but has limitations in conceptual depth. Whereas a score of 3 represents a proposition that is fully relevant, clear, and shows a deep understanding of the concept [18]. This label reflects a manual judgement of the proposition's quality, which then serves as the target for the machine learning model in this study. Examples of propositions from the relational database concept map and the Cyber Security Authentication concept map are shown in Tables 1 and 2, respectively.

Table 1. Data Collection Database Relational

No	Propotition	Score
1	Relational Database Advantages easy to perform data operations	2
2	Relational Database simple advantage	1
. . .	. . .	. . .
690	Relation must have a primary key	1
691	Primary key consists of unique columns	2

Table 2. Data Collection Cyber Security Authentication

No	Propotition	Score
1	id attributes determine the user's privileges such as root or guest	3
2	Electronic monitoring examples Specific account attack	2
. . .	. . .	. . .
1465	the process of verifying an identity claimed by or for a system entity it has two steps verification step, which is presenting or generating authentication information that corroborates the binding between the entity and the identifier	3
1466	something the individual posseses like electronic key, or card the types of cards are 1. embossed 2. magnetic 3. memory 4. smart	3

2.2. Data Preprocessing

Data preprocessing is an important step in preparing data before it is used in the analysis or training of machine learning models [19]. In this research, data preprocessing is done through three main stages: case folding, tokenisation, and lemmatisation. These preprocessing steps aim to produce an optimal representation of the data and improve the accuracy of the machine learning model [20].

2.2.1. Case Folding

Case folding is the first step in converting all letters to lowercase, ensuring consistent text formatting. For example, 'Relational Database Advantages easy to perform data operations' is converted to 'relational database advantages easy to perform data operations' [21]. This step reduces variation caused by uppercase and lowercase letters, thereby optimising text-based analysis. Research shows that case folding greatly contributes to improving the accuracy and efficiency of machine learning models, particularly in text classification and in sentiment analysis [20].

2.2.2. Tokenization

Tokenisation breaks the text into small units called tokens. For example, the sentence 'Relational Database simple advantage' is converted into ['relational', 'database', 'simple', 'advantage']. This process is important to identify keywords in the text and helps the model understand the data structure. Research by Mashtalir and Nikolenko (2023) shows that tokenisation helps improve accuracy in domain-specific text classification [22]. Moreover, Kozhevnikov and Pankratova (2020) confirmed that tokenisation is a key step in text processing to increase the relevance of features in text classification tasks [23].

2.2.3. Lemmatization

Lemmatisation is the process of converting a word into its base form or lemma by considering the linguistic context. For example, the word 'consists' in the proposition 'Primary key consists of unique columns' is converted to 'consist'. Lemmatisation reduces data redundancy by unifying different forms of words that have the same meaning, thereby increasing the relevance of the analysis [24]. The study by Saputro and Hermawan (2021) shows that lemmatisation significantly improves the accuracy of the model by simplifying the text data [25].

2.3. Text Representation

Text representation is the process of converting raw text into a numerical form that machine learning algorithms can understand. This representation is very important in natural language processing (NLP) as it helps models capture patterns, relationships, and context in text data. In this research, two main methods for text representation are used: TF-IDF and BERT embedding. These two techniques offer different approaches to capture term-frequency information and semantic context, which contribute to more in-depth analysis. TF-IDF weights words based on their frequency in documents and a corpus [26].

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical-based method that measures the importance of a word in a document relative to the corpus [27]. In this research dataset, TF-IDF is used to measure the weight of words such as 'relational', 'database', and 'advantages' based on their frequency in specific documents and the whole corpus. This technique is effective in highlighting relevant words for text classification tasks. Research shows that TF-IDF excels in capturing discriminative features, especially in datasets with short to medium text structure, and provides consistent performance for various machine learning algorithms such as SVM and Random Forest [28]. Tf-idf is calculated based on Formulas 1, 2, and 3.

$$tf(w) = \frac{n(w)}{\sum_j n(w)^2} \tag{1}$$

The Term Frequency equation 1, represented as $tf(w) = \frac{n(w)}{\sum_j n(w)^2}$, calculates how frequently a word appears in a document relative to the total number of words. In this formula, $n(w)$ represents the number of times a specific word w appears in the document, while the denominator $\sum_j n(w)$ represents the total count of all words in that document. This normalization by total word count helps account for differences in document length, ensuring that longer documents don't automatically get higher term frequencies solely because of their length.

$$idf(w) = log\left(\frac{\Sigma_D}{D_w}\right) \tag{2}$$

The Inverse Document Frequency Equation 2, expressed as $idf(w) = log\left(\frac{\Sigma_D}{D_w}\right)$, measures how important or unique word is across all documents in the collection. In this formula, Σ_D represents the total number of documents in the corpus, while D_w

represents the number of documents that contain the word w. When a word appears in many documents, its IDF value becomes lower, indicating that it's a common word; conversely, if a word appears in few documents, its IDF value becomes higher, indicating that it's a more distinctive or specialized term.

$$tfidf(w) = tf(w) * idf(w) \tag{3}$$

The TF-IDF equation 3, written as $tfidf(w) = tf(w) * idf(w)$, combines the previous two measures to create a composite score that reflects both the word's importance within a single document and its distinctiveness across the entire document collection. This multiplication ensures that words that are both frequent in a specific document (high TF) and rare across the corpus (high IDF) receive the highest scores. The resulting TF-IDF score helps identify terms that are particularly characteristic or important for specific documents while downplaying terms that are either too rare to be significant or too common to be distinctive.

2.4. Modelling

2.4.1. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a margin-based machine learning algorithm used for classification and regression. It works by finding the optimal hyperplane that maximises the margin between the data classes, thus providing good generalisation to new data. In this study, a radial basis function (RBF) kernel SVM is used, which is known to capture non-linear patterns in data. The RBF kernel maps the data into higher dimensions, making non-linearly separable data more separable. Previous studies have shown that the use of RBF kernels in SVM provides superior performance in various classification tasks, especially in high-dimensional datasets with complex patterns [29]. RBF kernels have also proven efficient for text classification tasks, including network intrusion detection. Research by Yalsavar et al. (2022) showed that optimisation of RBF kernel parameters using the Sliding Mode Control approach significantly improved the convergence speed and accuracy of the model on various large datasets [30]. In addition, research in the health field shows that RBF kernel SVM is effective at handling unbalanced datasets, producing accurate predictions despite differences in class imbalance. With its advantages, RBF kernel SVM was chosen in this study to process datasets with non-linear distributions and provide optimal classification results. The following is the formula for the RBF kernel SVM from equations 4, 5, and 6.

$$K(x_i, x_j) = exp(-\gamma||x_i - x_i||^2) \tag{4}$$

Equation 4 represents the Gaussian Radial Basis Function (RBF) kernel, a fundamental component of SVM for handling non-linear classification tasks. and γ (gamma) is a parameter that controls the kernel's spread or influence. In this formula, x_i, x_j are input feature vectors $K(x_i, x_j) = exp(-\gamma||x_i - x_i||^2)$ represents the squared Euclidean distance between these vectors, and γ is a parameter that controls the kernel's spread or influence. This kernel function is particularly powerful as it can map input data into an infinite-dimensional space, enabling the SVM to find linear separating boundaries in transformed space, even when the original data is not linearly separable

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i - x) + b \tag{5}$$

Equation 5 represents the SVM decision function that classifies new data points. In this formula, α_i represents the Lagrange multipliers obtained during training. y_i represents the class labels (+1 or -1 for binary classification). $K(x_i, x)$ is the kernel function that measures similarity between the input point x and training points x_i, and b is the bias term that shifts the decision boundary.

$$min\frac{1}{2}||w||^2 \tag{6}$$

Equation 6 represents the primary optimization objective of SVM, which aims to find the optimal hyperplane by minimizing the norm of the weight vector w. This minimization problem is crucial because it directly relates to maximizing the margin between different classes in the feature space. The objective ensures that the resulting hyperplane is as far as possible from the nearest data points of any class, helping create a more robust classifier with better generalization capabilities. This optimization approach is fundamental to SVM's effectiveness across a range of real-world applications, from image classification to text analysis.

2.4.2. K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that determines the class of data based on the distance to its neighboring data [31]. In this research, KNN is used to classify data based on Euclidean distance, with an optimized K value [32]. KNN is very useful in simple classification tasks with small to medium datasets. Research shows that KNN can produce high accuracy in network intrusion detection. In medical applications, such as epileptic seizure prediction, KNN optimized using genetic algorithms achieved 92% accuracy [33]. Moreover, the KNN algorithm is simple and effective for multidimensional data classification in pattern recognition applications.

2.4.3. Random Forest

Random Forest, or RF, is an ensemble-based algorithm that uses multiple decision trees to make predictions. Each tree contributes to the final result via majority voting, thereby improving the model's accuracy and stability. RF is very effective in handling data with many features. Research shows that RF excels in classification tasks, such as SQL injection attack detection, with high accuracy on various datasets [34]. In diabetes prediction, RF shows the best performance compared to other algorithms with high accuracy, precision, and sensitivity [35]. RF is also used extensively in other medical diagnostic applications, such as lung cancer prediction, where the algorithm exhibits high stability and low error [36].

2.4.4. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of artificial neural network, a variant of the Recurrent Neural Network (RNN), specifically designed to capture long-term dependencies in sequential data. LSTMs are particularly effective for tasks that require context over long time spans, such as natural language processing and text classification. LSTM units use a regulatory mechanism called 'gates' to organise the flow of information, allowing the model to remember important patterns and forget irrelevant data over time [37]. This mechanism consists of three main gates-the input gate, the forgetting gate, and the output gate —working together to retain and modify information in memory cells. Thus, LSTM can overcome the vanishing gradient problem that often occurs in traditional RNNs, enabling it to learn and model long-term relationships in data. This capability makes LSTMs a powerful tool for text classification, especially for understanding complex context and sequential information.

2.5. Evaluation

Machine learning model evaluation assesses the performance of algorithms in classification tasks. In this study, two main metrics are used: the Confusion Matrix and Cohen's Kappa, which provide insight into the accuracy, fit, and level of agreement between the model's predictions and the actual data. A confusion matrix is a table that shows the comparison between the model's predicted results and the actual data. This matrix consists of four main elements [38, 39]. The accuracy is calculated using Equation 7.

$$Accuracy = \frac{TP + TN}{N} \tag{7}$$

The accuracy metric provides a fundamental measure of a model's overall performance by comparing the number of correct predictions to the total number of instances. In this formula, TP represents true positives and TN represents true negatives, while N denotes the total number of instances in the dataset. This metric serves as a baseline evaluation tool, though it may not always provide a complete picture of model performance, especially with an imbalanced dataset—the precision, as shown in Equation 8.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Precision and recall metrics offer deeper insights into model performance from different perspectives. Precision measures the accuracy of positive predictions by determining what proportion of identifications were actually correct, where FP represents false positives. The recall is calculated according to equation 9.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Recall, expressed as TP/(TP + FN), evaluates the model's ability to find all relevant instances in the dataset, where FN represents false negatives. These metrics work together to provide a more comprehensive understanding of a model's performance characteristics. The F1-Score, representing the harmonic mean of precision and recall, is computed using Equation 10.

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{10}$$

Cohen's Kappa (K) is a metric for measuring the level of agreement between two observers or prediction systems, taking into account the possibility of agreement by chance [40, 41]. Cohen's Kappa is very effective in evaluating classification models on unbalanced datasets. As shown in Equation 11, Cohen's Kappa is calculated using the formula.

$$K = \frac{P_o - P_e}{1 - P_e} \tag{11}$$

Where p_o represents the proportion of observations where the observer agrees and p_e represents the expected proportion of agreement under random sampling.

3. RESULT AND ANALYSIS

This research aims to compare the performance of several machine learning and deep learning algorithms, namely SVM RBF, KNN, Random Forest, and LSTM, in classifying propositions from Relational Database material. The expected benefit of this research is to gain a deeper understanding of the best algorithm for handling data with short and long proposition characteristics, thereby helping system developers and other researchers choose the most suitable approach. The following are the results of the model evaluation on the relational database dataset (Table 3) and the cybersecurity authentication dataset (Table 4).

Table 3. Experiments Results Database Relational

Model	***Accuracy***	***Precision***	***Recall***	**F1 - Score**	**Cohen's Kappa**
SVM RBF	**0.87**	**0.87**	**0.87**	**0.87**	**0.76**
KNN	0.85	0.83	0.85	0.83	0.69
Random Forest	0.85	0.84	0.85	0.84	0.70
LSTM	0.83	0.82	0.83	0.82	0.67

Experimental results on the relational database dataset show that the SVM RBF algorithm achieves the best performance, with accuracy, precision, recall, and F1-score all at 0.87, and Cohen's Kappa at 0.76. This demonstrates SVM's ability to handle data with non-linear distributions, thanks to the RBF kernel, which effectively separates classes with maximum margin. The KNN and Random Forest algorithms performed relatively equally, each achieving an accuracy of 0.85. However, Random Forest was slightly superior in Cohen's Kappa (0.70) compared to KNN (0.69), reflecting its stability and ability to handle more complex datasets through an ensemble approach. The LSTM algorithm achieved an accuracy of 0.83 and a Cohen's Kappa of 0.67, which, while slightly lower than the other models, indicates competitive performance. This brevity limits the LSTM's ability to leverage its strength in learning sequential patterns and long-term dependencies. LSTMs typically require larger datasets to train effectively due to their complex architectures and large number of parameters. With a smaller dataset, the LSTM may not have sufficient data to learn meaningful patterns without overfitting.

Table 4. Experiments Results Cyber Security Authentication

Model	***Accuracy***	***Precision***	***Recall***	**F1 - Score**	**Cohen's Kappa**
SVM RBF	0.85	0.84	0.85	0.84	0.67
KNN	0.70	0.76	0.70	0.72	0.46
Random Forest	0.82	0.81	0.82	0.81	0.61
LSTM	**0.85**	**0.85**	**0.85**	**0.85**	**0.69**

Experimental results on the Cyber Security Authentication dataset showed that both SVM RBF and LSTM algorithms achieved the best performance, each with an accuracy of 0.85. The SVM RBF model achieved a precision of 0.84, a recall of 0.85, and an F1-score of 0.84, with a Cohen's Kappa of 0.67. The LSTM algorithm exhibited slightly higher precision and F1-score (both 0.85) and achieved a Cohen's Kappa of 0.69, indicating slightly better agreement with the true labels than SVM RBF. Random Forest showed competitive performance, with an accuracy of 0.82 and a Cohen's Kappa of 0.61, though slightly below that of SVM RBF and LSTM. KNN had the lowest performance, with an accuracy of 0.70 and a Cohen's Kappa of 0.46, indicating limitations in capturing complex patterns using the TF-IDF representation. The inclusion of LSTM in the experiments highlights its ability to model sequential data and capture contextual information in textual data, which is particularly advantageous for the Cyber Security Authentication dataset, where propositions are longer, averaging 10 words. This suggests that LSTM networks can effectively handle longer text sequences because of their architecture, which is designed for sequential data processing.

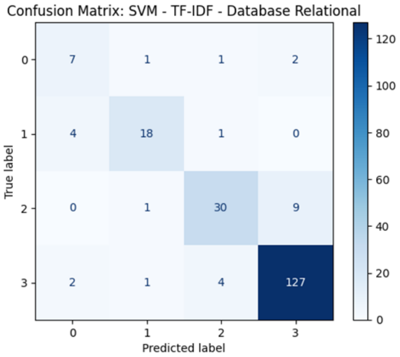

Figure 2. SVM database relational

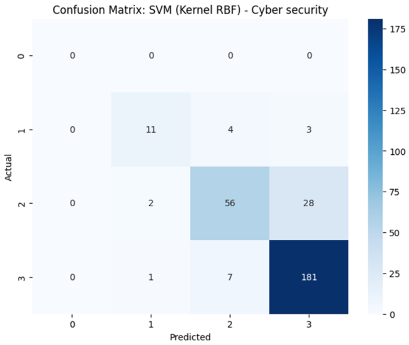

Figure 3. SVM cybersecurity

Based on the Confusion Matrix in Figure 2 for the relational database dataset, SVM with the TF-IDF representation demonstrates fairly stable performance, particularly in class 3, where the model's predictions often match the true labels. However, there are still some misclassifications in classes 1 and 2, which tend to be adjacent to each other. Meanwhile, as shown in Figure 3, the Confusion Matrix for the Cybersecurity Authentication dataset also shows high accuracy in class 3, but there are more errors, particularly in classes 0 and 1. This suggests that feature separation for those two classes is more challenging compared to the clearer class. Overall, TF-IDF representations in both datasets consistently achieve good performance with SVM, even though variations in class complexity are more pronounced in the Cybersecurity Authentication data.

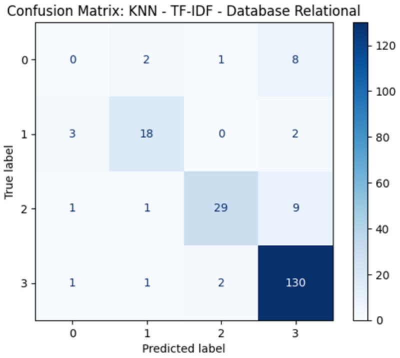

Figure 4. KNN database relational

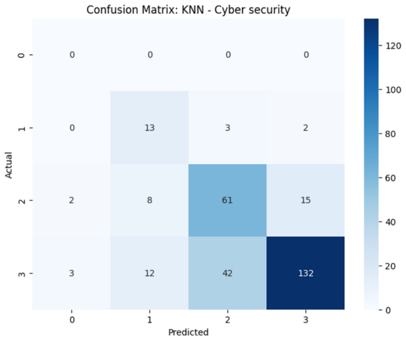

Figure 5. KNN cybersecurity

The results of applying the KNN algorithm to the cybersecurity authentication dataset show that most data in class 3 were correctly predicted, as indicated by the relatively large diagonal value in the lower right of the confusion matrix in Figure 5. However, there are still some misclassifications in other classes, such as 0, 1, and 2, suggesting that the KNN model sometimes struggles to distinguish classes with similar characteristics. Meanwhile, on the relational database dataset, KNN with TF-IDF representation also achieved fairly accurate predictions for class 3, as evidenced by the high number of data points correctly classified as 3 in Figure 4. Although some errors still occur in classes 0, 1, and 2, TF-IDF overall helps the model distinguish between classes, enabling most data to be correctly classified. This performance shows that adding TF-IDF word weights can improve KNN's ability to distinguish similar propositions in the Relational Database dataset.

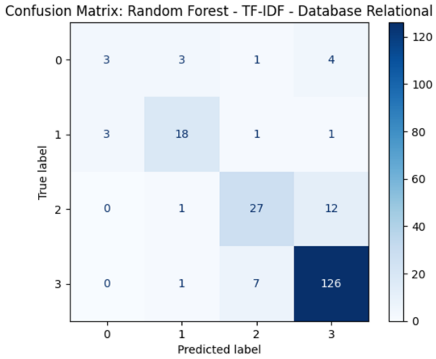

Figure 6. Random forest database relational

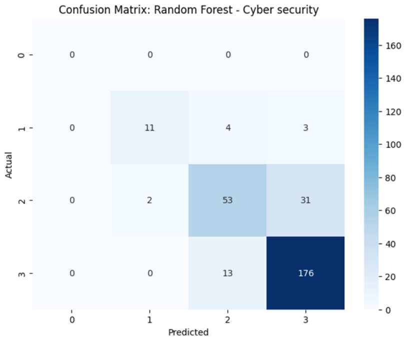

Figure 7. Random forest cybersecurity

In the Confusion Matrix for the cybersecurity dataset using Random Forest, shown in Figure 7, it is evident that most data in class 3 were correctly classified, as indicated by the dominant value of 176 in the diagonal cell for class 3. However, there are still some misclassifications in other classes; for instance, certain data in class 2 were predicted as class 3, and some in class 1 were predicted as class 2. This suggests that, even though Random Forest can accommodate many patterns in the cybersecurity dataset, separating features across classes other than class 3 remains challenging when their characteristics are similar. Meanwhile, Figure 6 displays the Confusion Matrix for the relational database dataset using Random Forest; applying TF-IDF improves class separation. While there are still a few shifts in classifications for classes 0, 1, and 2, their numbers are relatively smaller. Class 3 stands out with 126 correctly classified data points, indicating that the TF-IDF representation can highlight important keyword differences and thereby assist the Random Forest algorithm in recognizing distinct proposition characteristics in the relational database dataset.

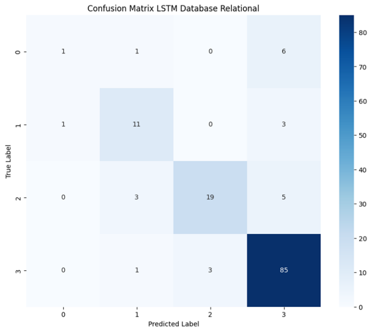

Figure 8. LSTM database relational

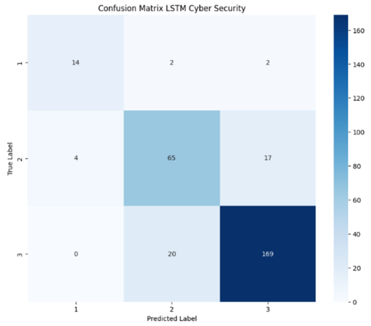

Figure 9. LSTM cybersecurity

Overall, the results from the Relational Database and Cybersecurity Authentication datasets highlight the consistently strong performance of SVM RBF with TF-IDF, especially on the Relational Database dataset, which contains shorter propositions, where SVM RBF achieves the highest accuracy of 0.87 and Cohen's Kappa of 0.76. On the more textually complex Cybersecurity dataset, LSTM rivaled SVM RBF, achieving an accuracy of 0.85 and a higher Cohen's Kappa of 0.69 compared to 0.67. According to the Landis and Koch interpretation, these values indicate substantial agreement for SVM on short propositions and moderate to substantial agreement for LSTM on longer propositions. This suggests that LSTM is better at handling sequential complexity, while SVM remains robust across different data distributions.

Although this study focused primarily on comparative performance, the observed differences highlight important practical implications. In educational assessment, even small increases in Cohen's Kappa can significantly reduce inconsistencies in grading and improve the fairness of evaluation. For instance, using LSTM for longer and more complex propositions may provide teachers with more reliable automatic scoring, while SVM offers stability across a variety of contexts. Future studies could strengthen this analysis by applying statistical significance tests to confirm whether the observed performance differences are systematic.

4. CONCLUSION

This study reveals important insights into the influence of proposition length on algorithm performance in open-ended concept map scoring. SVM RBF showed consistent results across datasets, achieving 87% accuracy with a Cohen's Kappa of 0.76 on shorter propositions. In comparison, LSTM matched SVM's accuracy of 85% on longer propositions and achieved a higher Kappa of 0.69, indicating better handling of complex sequential patterns. These findings extend existing literature by showing that proposition length is a critical factor in automated assessment and that different algorithms exhibit distinct strengths depending on text complexity. Beyond these contributions, several limitations must be acknowledged. The datasets were relatively small and limited to two domains, which constrained the generalizability of the findings and may have introduced domain-specific biases. The models also relied on TF-IDF and a single deep learning architecture, leaving room for further exploration of alternative feature representations and more diverse model families.

Despite these limitations, the results carry important practical implications. More reliable automated scoring can reduce educators' workload, enhance objectivity, and encourage wider adoption of concept maps in digital learning environments. To build on this work, future research should employ larger and more diverse datasets, explore hybrid models capable of dynamically adapting to varying proposition lengths, and develop more sophisticated preprocessing techniques to handle complex or cross-domain propositions. Extending automated scoring to multilingual contexts and designing more nuanced evaluation metrics will also be essential to enhance fairness and applicability in real-world educational settings.

5. ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to all those who have contributed to the completion of this research. We want to express our deepest appreciation to our supervisors, Didik Dwi Prasetya and Ilham Ari Elbaith Zaeni, for their unwavering support, insightful guidance, and invaluable feedback throughout the research. Our thanks also go to the Department of Electrical Engineering and Informatics at the State University of Malang for providing resources and facilities that greatly assisted in conducting this research. We thank the author's family and our colleagues for their assistance in reviewing this article.

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 study was conducted without any financial support.

COMPETING INTEREST

The authors declare no conflict of interest regarding the publication of 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] D. D. Prasetya and T. Hirashima, "Associated Patterns in Open-Ended Concept Maps within E-Learning," vol. 5, no. 2, p. 179, December,2022, https://doi.org/10.17977/um018v5i22022p179-187.

[3] Y. Cooper and E. Zimmerman, "Concept Mapping: A Practical Process for Understanding and Conducting Art Education Research and Practice," vol. 73, no. 2, pp. 24–32, March,2020, https://doi.org/10.1080/00043125.2019.1695478.

[4] A. Guiral Herrera and M. Pifarre Turmo, "Digital Cognitive Maps for Scientific Knowledge Construction in Initial Teacher Education:," vol. 26, no. 2, pp. 89–109, April,2023, https://doi.org/10.5944/ried.26.2.36067.

[5] K. E. De Ries, H. Schaap, A.-M. M. J. A. P. Van Loon, M. M. H. Kral, and P. C. Meijer, “A literature review of open-ended concept maps as a research instrument to study knowledge and learning,” vol. 56, no. 1, pp. 73–107, February,2022, https://doi.org/10.1007/s11135-021-01113-x.

[6] D. D. Prasetya, T. Hirashima, and Y. Hayashi, “Study on Extended Scratch-Build Concept Map to Enhance Students’ Understanding and Promote Quality of Knowledge Structure,” vol. 11, no. 4, pp. 144–153, 2020, https://doi.org/10.14569/IJACSA.2020.0110420.

[7] K. A. Kroeze, S. M. Van Den Berg, B. P. Veldkamp, and T. De Jong, “Automated Assessment of and Feedback on Concept Maps During Inquiry Learning,” vol. 14, no. 4, pp. 460–473, August,2021, https://doi.org/10.1109/TLT.2021.3103331.

[8] T. A. Assegie, A. O. Salau, G. Chhabra, K. Kaushik, and S. L. Braide, “Evaluation of Random Forest and Support Vector Machine Models in Educational Data Mining,” in *2024 2nd International Conference on Advancement in Computation &amp; Computer Technologies (InCACCT)*. IEEE, May,2024, pp. 131–135, https://doi.org/10.1109/InCACCT61598.2024.10551110.

[9] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, “Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review,” vol. 13, no. 1, pp. 6308–6325, 2020, https://doi.org/10.1109/JSTARS.2020.3026724.

[10] H. Choi, H. Lee, and M. Lee, “Optimal Knowledge Component Extracting Model for Knowledge-Concept Graph Completion in Education,” vol. 11, pp. 15 002–15 013, 2023, https://doi.org/10.1109/ACCESS.2023.3244614.

[11] S. V. Tytenko, “Concept Maps, Their Application Types and Methods in Information and Learning Systems,” vol. 10, no. 4, pp. 70–78, March,2021, https://doi.org/10.20535/kpisn.2020.4.227090.

[12] F. Sciarrone and M. Temperini, “A Sentence-Embedding-Based Dashboard to Support Teacher Analysis of Learner Concept Maps,” vol. 13, no. 9, p. 1756, May,2024, https://doi.org/10.3390/electronics13091756.

[13] P. Huang, S. Lin, J. Yuan, and H. Chen, “Course Achievement Evaluation Using Concept Map in Traditional Learning,” vol. 1624, no. 5, p. 052014, October,2020, https://doi.org/10.1088/1742-6596/1624/5/052014.

[14] S. Loizou, N. Nicolaou, B. A. Pincus, A. Papageorgiou, and P. McCrorie, “Concept maps as a novel assessment tool in medical education,” vol. 12, p. 21, https://doi.org/10.12688/mep.19036.1.

[15] 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.

[16] 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.

[17] S. Xia, P. Zhan, K. K. H. Chan, and L. Wang, “Assessing concept mapping competence using item expansion-based diagnostic classification analysis,” vol. 61, no. 7, pp. 1516–1542, September,2024, https://doi.org/10.1002/tea.21897.

[18] A. Caputo, D. Monterosso, and E. Sorrentino, “The use of concept maps as an assessment tool in students’ risk education about occupational safety and health,” vol. 58, no. 3, pp. 172–176, 2022, https://doi.org/10.4415/ANN_22_03_05.

[19] S. S. Bisarya, A. Shukla, and S. Kumar, “Data Preparation in Context of Social Sciences Research,” vol. 5, no. 3, pp. 1–10, June,2023, https://doi.org/10.36948/ijfmr.2023.v05i03.3937.

[20] Yuyun, A. D. Latief, T. Sampurno, Hazriani, A. O. Arisha, and Mushaf, “Next Sentence Prediction: The Impact of Preprocessing Techniques in Deep Learning,” in *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*. IEEE, October,2023, pp. 274–278, https://doi.org/10.1109/IC3INA60834.2023.10285805.

[21] T. Mustaqim, K. Umam, and M. A. Muslim, “Twitter text mining for sentiment analysis on government’s response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm,” vol. 1567, no. 3, p. 032024, June,2020, https://doi.org/10.1088/1742-6596/1567/3/032024.

[22] S. V. Mashtalir and O. V. Nikolenko, “Data preprocessing and tokenization techniques for technical Ukrainian texts,” vol. 6, no. 3, pp. 318–326, October,2023, https://doi.org/10.15276/aait.06.2023.22.

[23] V. A. Kozhevnikov and E. S. Pankratova, "Research of Text Pre-Processing Methods for Preparing Data in Russian for Machine Learning," vol. 84, no. 04, pp. 313–320, April,2020, https://doi.org/10.15863/TAS.2020.04.84.55.

[24] Z. Abidin, A. Junaidi, and Wamiliana, "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," vol. 10, no. 2, pp. 217–231, June, 2024, https://doi.org/10.20473/jisebi.10.2.217-231.

[25] T. H. Saputro and A. Hermawan, "The Accuracy Improvement of Text Mining Classification on Hospital Review through The Alteration in The Preprocessing Stage," vol. 10, no. 4, pp. 140–146, July,2021, https://doi.org/10.24203/ijcit.v10i4.138.

[26] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," vol. 4, pp. 27–33, August,2022, https://doi.org/10.56471/slujst.v4i.266.

[27] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," vol. 2171, no. 1, p. 012021, January,2022, https://doi.org/10.1088/1742-6596/2171/1/012021.

[28] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," vol. 10, no. 16, p. 35945, August,2024, https://doi.org/10.1016/j.heliyon.2024.e35945.

[29] V. Chaurasia and S. Pal, "Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer," vol. 1, no. 5, pp. 270–280, September,2020, https://doi.org/10.1007/s42979-020-00296-8.

[30] M. Yalsavar, P. Karimaghaee, A. Sheikh-Akbari, M.-H. Khooban, J. Dehmeshki, and S. Al-Majeed, "Kernel Parameter Optimization for Support Vector Machine Based on Sliding Mode Control," vol. 10, no. 1, pp. 17 003–17 017, 2022, https://doi.org/10.1109/ACCESS.2022.3150001.

[31] U. G. Inyang, F. F. Ijebu, F. B. Osang, A. A. Afoluronsho, S. S. Udoh, and I. J. Eyoh, "A Dataset-Driven Parameter Tuning Approach for Enhanced K-Nearest Neighbour Algorithm Performance," vol. 13, no. 1, pp. 380–391, January,2023, https://doi.org/10.18517/ijaseit.13.1.16706.

[32] Aqib Fawwaz Mohd Amidon, Z. M. Yusoff, N. Ismail, and M. N. Taib, "KNN Euclidean Distance Model Performance on Aquilaria Malaccensis Oil Qualities," vol. 48, no. 2, pp. 16–28, July,2024, https://doi.org/10.37934/araset.48.2.1628.

[33] H. Jagath Prasad and R. Marjorie S., "Optimized k-nearest neighbours classifier based prediction of epileptic seizures," vol. 13, no. 4, pp. 2442–2455, August,2024, https://doi.org/10.11591/eei.v13i4.6598.

[34] T. J. Angula and V. Hashiyana, "Detection of Structured Query Language Injection Attacks Using Machine Learning Techniques," vol. 15, no. 4, pp. 13–26, August,2023, https://doi.org/10.5121/ijcsit.2023.15402.

[35] A. Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," vol. 144, p. 102110, March,2023, https://doi.org/10.1016/j.datak.2022.102110.

[36] D. Li, G. Li, S. Li, and A. Bang, "Classification prediction of lung cancer based on machine learning method:," vol. 19, no. 1, pp. 1–12, November,2023, https://doi.org/10.4018/IJHISI.333631.

[37] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, and A. Muneer, "LSTM Inefficiency in Long-Term Dependencies Regression Problems," vol. 30, no. 3, pp. 16–31, May,2023, https://doi.org/10.37934/araset.30.3.1631.

[38] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," vol. 8, pp. 90 847–90 861, 2020, https://doi.org/10.1109/ACCESS.2020.2994222.

[39] G. Canbek, T. Taskaya Temizel, and S. Sagiroglu, "BenchMetrics: A systematic benchmarking method for binary classification performance metrics," vol. 33, no. 21, pp. 14 623–14 650, November,2021, https://doi.org/10.1007/s00521-021-06103-6.

[40] A. Casagrande, F. Fabris, and R. Girometti, "Beyond kappa: An informational index for diagnostic agreement in dichotomous and multivalue ordered-categorical ratings," vol. 58, no. 12, pp. 3089–3099, December,2020, https://doi.org/10.1007/s11517-020-02261-2.

[41] G. Rau and Y.-S. Shih, "Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data," vol. 53, p. 101026, September,2021, https://doi.org/10.1016/j.jeap.2021.101026.

[This page intentionally left blank.]