Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

A Hybrid LSTM–Smith Waterman Model for Personalized Semantic Search in Academic Information Systems

Ade Yuliana*[1], Novita Lestari Anggreini[2], Rachmat Iskandar[3], G. Rafi Prasanth[4]

[1,3]Informatics Engineering, Politeknik TEDC, Indonesia
[2]Computer Engineering, Politeknik TEDC, Indonesia
[4]Computer Applications, Hindustan Institute of Technology and Science, India

Email: [1]yulianaad@poltektedc.ac.id

Received : May 21, 2025; Revised : Jul 15, 2025; Accepted : Aug 4, 2025; Published : Aug 19, 2025

Abstract

The growing complexity of digital learning environments presents a critical challenge in computer science, particularly in designing intelligent academic systems capable of delivering context-aware and personalized content. Traditional academic information systems often rely on literal keyword matching, failing to interpret the semantic intent behind user queries and ignoring historical learning behavior. This study addresses these limitations by proposing a hybrid semantic search and recommendation model that integrates Long Short-Term Memory (LSTM) networks with the Smith Waterman algorithm. The LSTM component models temporal sequences of user interactions, while Smith Waterman enables local semantic alignment between user queries and learning content. Historical query logs and user-clicked topics are transformed into semantic vectors, which are further enhanced through a contextual graph and semantic relation matrix. Experimental results demonstrate the model's effectiveness, achieving 89% accuracy, an F1-score of 0.89, and an AUROC of 0.88 by epoch 50. The hybrid architecture successfully captures the evolution of user interest and semantic relevance, outperforming baseline approaches. This research contributes to the field of computer science by bridging natural language understanding and sequential modeling to improve adaptive learning technologies. The proposed model offers a scalable foundation for developing intelligent recommendation systems in academic platforms, fostering improved learner engagement and efficiency.

Keywords : *Academic Information Systems, Learning Recommendation, Long Short-Term Memory, Semantic Matching, Smith Waterman.*

1. INTRODUCTION

In academic information systems, the process of searching for learning materials still faces various obstacles that impact the effectiveness of information searches by users [1], [2], [3]. Generally, search systems only rely on literal keyword matching, without considering the contextual meaning of the terms used by users [4], [5]. This causes problems when there are differences in terms between user input and available material metadata, so that search results become irrelevant [6], [7]. In addition, most systems have not utilized user search history to understand individual learning patterns or preferences [8], [9], [10]. As a result, users often have to search repeatedly with different keyword variations to find appropriate materials [8], [11]. This inefficiency not only hinders the independent learning process but also reduces the quality of the user experience in accessing academic information digitally.

Based on the problem analysis, it is necessary to develop a search system that does not only rely on literal keyword matching, but is also able to understand the meaning and context of user queries through a semantic search approach [12], [13]. This semantic search optimization allows the system to identify the relationship between terms even though they do not explicitly use the same words [14], [15]. By combining semantic search and utilizing user search history, the system can build a more personalized and adaptive search profile to the needs of learners [16], [17]. This approach allows academic information systems to provide more relevant material recommendations, according to the user's learning patterns and goals [18], [19]. Through local similarity-based matching techniques such as the LSTM-Smith Waterman algorithm, the system can recognize similarities between phrases or terms in queries and material content more accurately [20], [21], [22], thereby increasing efficiency and accuracy in the search process and learning recommendations.

Researcher [23] proposed a research recommendation system that combines the LDA topic approach and the BERT contextual representation model in the study Semantic and Explainable Research-Related Recommendation System Using LDA and BERT. The novelty of this system lies in its ability to provide recommendations that are not only semantically relevant but also explainable, which is very important in academic environments. Evaluations show that this system improves the relevance and user confidence in literature search results. However, this system has limitations in terms of high computational requirements and sensitivity to variations in document structure and terminology across fields of study.

Researcher [24] introduced an ontology-based semantic integration model to improve prediction accuracy in learning analytics. Their method combines multiple user data sources and maps the information into a domain ontology, so that the system can provide more meaningful analytics on student learning behavior. Experimental results demonstrate the system's ability to predict academic performance and support timely intervention decisions. However, the main challenge faced is the complexity in developing and maintaining the ontology structure to keep it relevant to changes in learning materials.

Researcher [25] developed a semantic-aware intelligent framework for e-learning recommendation systems that focuses on understanding the context of content through the integration of NLP and machine learning techniques. The novelty of this method lies in its approach that not only considers explicit keywords but also hidden meanings in learning materials, thus being able to provide more personalized recommendations. The test results showed an increase in accuracy and user satisfaction in the online learning experience. However, its weaknesses lie in the need for large training data and the challenge of overcoming semantic ambiguity in natural language.

Thus, the main contribution of this research lies in the integration and optimization of the Smith-Waterman algorithm with LSTM-based modeling to enhance semantic search capabilities in academic information systems, particularly in generating more personalized and context-aware learning material recommendations. The Smith-Waterman algorithm is employed for fine-grained local matching between user queries and educational content, allowing precise semantic alignment. Meanwhile, the LSTM model captures sequential patterns in learners' historical interactions, enabling the system to understand contextual relevance over time. Additionally, this study emphasizes optimal parameter tuning for both algorithms to improve the accuracy and efficiency of the recommendation engine. This hybrid approach is expected to significantly advance the effectiveness of personalized learning services in academic and e-learning platforms.

2. METHOD

To address semantic ambiguity and capture user learning behavior over time, this research integrates two core algorithms: Smith-Waterman, used for fine-grained local alignment of semantic vectors, and LSTM, which models the sequential patterns in user interaction history. Smith-Waterman helps quantify semantic similarity between user queries and learning materials, while LSTM supports temporal adaptation and learning personalization.

Based on Figure 1, the suggested methodology begins with Learner Selection, a process by which individual users are selected within the e-learning platform for enabling personalized recommendation.

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

The process is then followed by E-Learning System access and content fetching for Selected Content, followed by Content Extraction, a process by which learning content is converted for purposes of allowing meaningful data extraction. Through such content, Semantic Relation Types are recognized and applied for the construction of a Contextual Graph defining semantic relations among topic details. In tandem, Machine Learning Parameters are designed and LSTM Layers are built for purposes of temporal pattern detection. The Training process for LSTM is then applied using content semantics and learning behavior, with an output that is saved in the Learners Semantic Database. During this stage, a series of hyperparameter tuning procedures were carried out to optimize the LSTM model. This includes adjustments of embedding dimension, number of LSTM units, dropout rate, and learning rate, which were iteratively tested and validated using the semantic dataset. The goal was to balance model complexity and generalization capability in capturing sequential patterns of user interest.

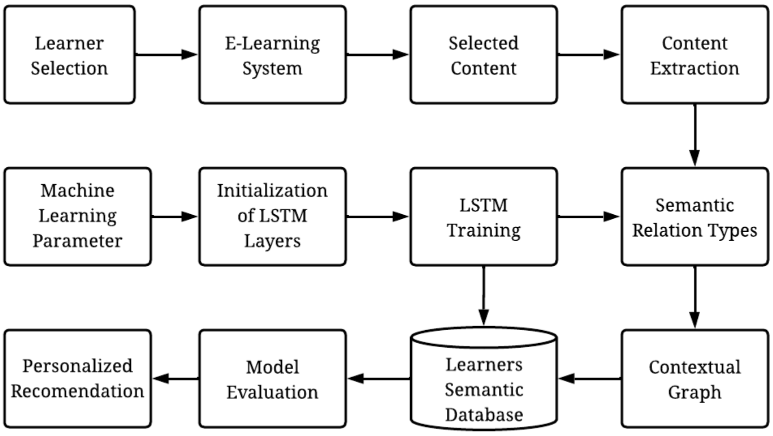


Figure 1. Research Stages

2.1. Semantic Dataset

Semantic datasets in the context of academic information systems refer to datasets that represent semantic relationships between user queries, learning content, and user interactions. These datasets include not only literal words or phrases, but also contextual information extracted from search history, content clicks, and topics of materials frequently accessed by users.

Table 1. Semantic Dataset

User ID	Timestamp	Query	Clicked Topic	Semantic Vector (Query)	Semantic Vector (Content)
U001	2025-05-11 08:32:12	"deep learning models"	"Introduction to CNNs"	[0.23, 0.51, 0.19, 0.47]	[0.25, 0.49, 0.20, 0.45]
U002	2025-05-11 09:10:44	"data preprocessing steps"	"Data Cleaning Techniques"	[0.41, 0.39, 0.35, 0.12]	[0.43, 0.38, 0.36, 0.10]
U003	2025-05-11 10:25:03	"gradient descent"	"Optimization in ML"	[0.56, 0.48, 0.21, 0.33]	[0.57, 0.47, 0.22, 0.32]
⋮	⋮	⋮	⋮	⋮	⋮
U0091	2025-05-10 11:00:27	"natural language models"	"RNN vs Transformer"	[0.36, 0.55, 0.40, 0.18]	[0.35, 0.56, 0.39, 0.19]

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

The semantic dataset illustrated in Table 1 serves as the foundational input for training and evaluating the proposed semantic search and recommendation model. Each entry captures a real-time interaction between the user and the academic system, where the query and corresponding clicked topic are transformed into semantic vectors using embedding techniques. These vectors are then compared using the Smith-Waterman algorithm, resulting in a match score that quantifies the contextual similarity between the query and content. This score, combined with timestamped behavior, enables the LSTM layers to learn temporal patterns in user interest evolution. Consequently, the dataset not only supports alignment and ranking but also enables adaptive learning through sequential modeling, making it integral to the personalized recommendation process.

2.2. Semantic Relation Matrix

Semantic relation is the main foundation in the semantic search system used in this study [26]. Semantic relation refers to the conceptual relationship between the terms used in user queries and the learning materials available in the system [26], [27]. To build this relationship, the system extracts keywords from learning content and user queries, then maps these terms into semantic representations using a word embedding-based approach or other semantic vector representations. Furthermore, these semantic relations are grouped into categories such as synonymy, hyponymy, thematic association, and topical order.

In this study, semantic relation is not only used to match words literally, but also to understand the context of the query based on the user's interaction history. This allows the system to recognize that the terms "neural network" and "deep learning" have close meanings even though they are not identical in text. This semantic relationship information will be stored in a contextual graph, which is then used as a reference in the matching and recommendation process based on the Smith-Waterman and LSTM algorithms. To mathematically represent semantic relations, each term from the user query $Q = \{q1, q2, ..., qn\}$ and the learning content $D = \{d1, d2, ..., dm\}$ is embedded into a high-dimensional vector space using a word embedding model such as Word2Vec or GloVe. Each term is represented in eq (1).

$$qi, dj \in R^d \tag{1}$$

Based eq (1), The semantic similarity between a query term qi and a content term dj is calculated using cosine similarity as seen in eq (2).

$$sim(qi, dj) = \frac{qi \cdot dj}{||qi|| \cdot ||dj||} \tag{2}$$

To obtain the overall semantic relevance score between the query and a content document, we use a max-aggregated similarity across all pairwise combinations as seen in eq (3).

$$S\,(Q, D) = \frac{1}{n}\sum_{i=1}^{n} \max(sim(qi, dj)) \tag{3}$$

Eq (3) ensures that each query term is matched with its most semantically similar counterpart in the content, which mimics the local alignment behavior of the Smith-Waterman algorithm. To reinforce semantic depth, a semantic relation matrix $M \in R^{n \times m}$ is constructed, where:

$$M_{ij} = sim(qi, dj) \tag{4}$$

This matrix serves as a foundation for local sequence alignment using Smith-Waterman, where similar concept patterns are aligned based on their semantic proximity rather than exact token match.

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

Additionally, the contextual weight between terms may be adjusted based on user history H, for example, through a temporal decay function in eq (5).

$$w'_{ij} = M_{ij} \cdot e^{-\lambda t_{ij}} \quad (5)$$

where t_{ij} denotes the time distance since the user last interacted with related content, and λ is a decay factor controlling the influence of historical context. All these semantic scores and relationships are encoded into a contextual graph $G = (V, E)$, where V represents semantic concepts (nodes), and E represents contextual relationships (edges), labeled with relation types (e.g., synonym, topic sequence, prerequisite). This graph is then used in downstream matching and recommendation stages, providing a semantically rich structure that supports personalized learning material suggestions.

2.3. Machine Learning Based on LSTM-Smith Waterman

In this research, a hybrid approach combining LSTM-Smith Waterman algorithm is proposed to improve the accuracy of semantic matching and personalized learning recommendations in academic information systems [28]. The Smith-Waterman algorithm, originally developed for local sequence alignment in bioinformatics, is adapted in this context to identify the most semantically relevant subsequences between user queries and learning material content [28], [29]. Unlike global matching, local alignment allows partial but contextually significant matches to be prioritized, making it highly suitable for semantic search scenarios where exact matches are rare [30].

LSTM is utilized as a complementary deep learning method to capture the temporal dynamics and sequential patterns of user interaction history and query behavior [31]. The LSTM network learns the progression of user interests over time and generates hidden state representations that encapsulate contextual dependencies from past searches and clicked content [32]. The initialization and training of the LSTM layer, as illustrated in Figure 2, allow the system to model these sequential patterns effectively [33], [34]. These representations are then used to guide the Smith-Waterman scoring matrix, where alignment between the user's semantic history and the content corpus is performed not only based on vector similarity, but also on learned importance from the LSTM. The integration of LSTM into the Smith-Waterman framework enhances the adaptability of the alignment by weighting the score based on learned semantic context. The modified scoring function in the alignment phase is defined as seen in eq (6).

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + \alpha \cdot sim(qi, dj) + \beta \cdot ht \\ S(i-1, j) - gap\_penalty \\ S(i, j-1) - gap\_penalty \\ 0 \end{cases} \quad (6)$$

Where, $sim(qi, dj)$ is the semantic similarity between the i-th term of the query and the j-th term of the document, ht is the hidden state vector at time t produced by the LSTM network representing contextual user history, α and β are weighting coefficients for semantic similarity and historical relevance, and $gap\_penalty$ controls the penalty for insertions or deletions. This combined model allows for context-aware semantic matching, where the system not only aligns query-content pairs based on vector similarity, but also personalizes the alignment process based on temporal learning patterns and user history. The ultimate goal is to provide learning material recommendations that are highly relevant, personalized, and semantically aligned with the user's actual learning path and preferences.

As seen in Figure 2, the proposed architecture begins with the Learner Semantic Data Sequence, which consists of temporally ordered query and interaction records represented in semantic vector format. This sequence is fed into the initial LSTM layers, which capture the sequential and contextual patterns over time. The resulting hidden states—which encapsulate contextualized user interest

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

representations—are then passed into the Smith-Waterman algorithm, which performs local sequence alignment to identify semantically relevant patterns between the user's history and available learning materials. This alignment produces a similarity score, which is subsequently used for ranking and generating personalized recommendations.

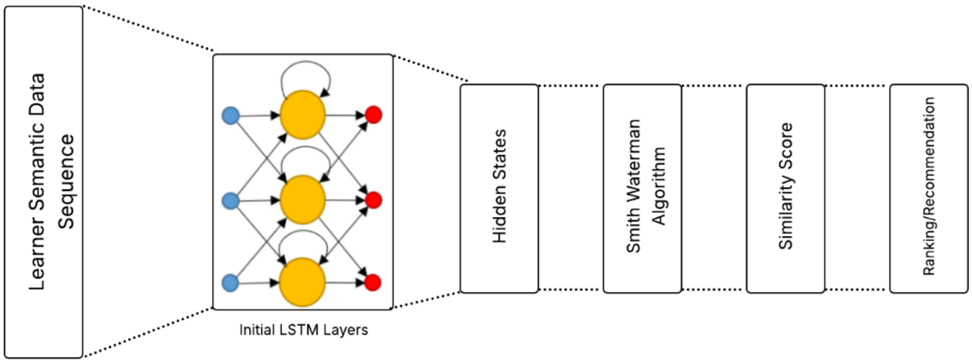


Figure 2. Proposed LSTM–Smith Waterman Architecture

3. RESULT

This section presents the discussion and analysis of the proposed semantic search optimization framework integrating the LSTM and Smith-Waterman algorithms. The primary focus lies in evaluating how this hybrid approach enhances the understanding of user intent and improves the precision of learning material recommendations in academic information systems. The system processes user queries in the form of semantic vectors, which are derived from historical interactions and contextualized through LSTM layers. These hidden states, capturing sequential user intent over time, are then aligned with available learning content using the Smith-Waterman algorithm. This alignment is not only based on vector similarity but also incorporates contextual weightings learned from LSTM outputs, thereby achieving a deeper semantic match between user queries and the content corpus.

To ensure optimal performance of the LSTM component within the proposed hybrid framework, a series of hyperparameter tuning steps were conducted. These hyperparameters directly influence the network's ability to capture long-range dependencies and semantic transitions in the learner's historical query data. Parameters such as the number of epochs, batch size, learning rate, optimizer type, and embedding dimensions were selected based on iterative experimentation and validation performance. The embedding layer was configured to transform each token in the semantic query into dense vector representations, enabling the model to learn nuanced semantic relationships. The use of the Adam optimizer with a moderate learning rate was chosen to accelerate convergence while preventing overfitting. Table 2 below summarizes the finalized hyperparameter settings used for the LSTM model in this research.

After the hyperparameter initialization is carried out as explained in Table 2, the next stage is the model training process using the designed LSTM architecture. The training process is carried out for 50 epochs using a semantic dataset that has been preprocessed and divided into training and validation data. The results of the training process are shown in Figure 3(a), which shows the accuracy curve between training and validation data. It can be seen that the model accuracy has increased significantly in the early epochs and tends to be stable approaching 0.95 for training data and around 0.90 for validation data after reaching convergence, indicating good generalization performance.

Meanwhile, the model classification performance is further explained using the Receiver Operating Characteristic (ROC) curve shown in Figure 3(b). The ROC curve shows the relationship between the True Positive Rate and False Positive Rate obtained from the results of the model evaluation

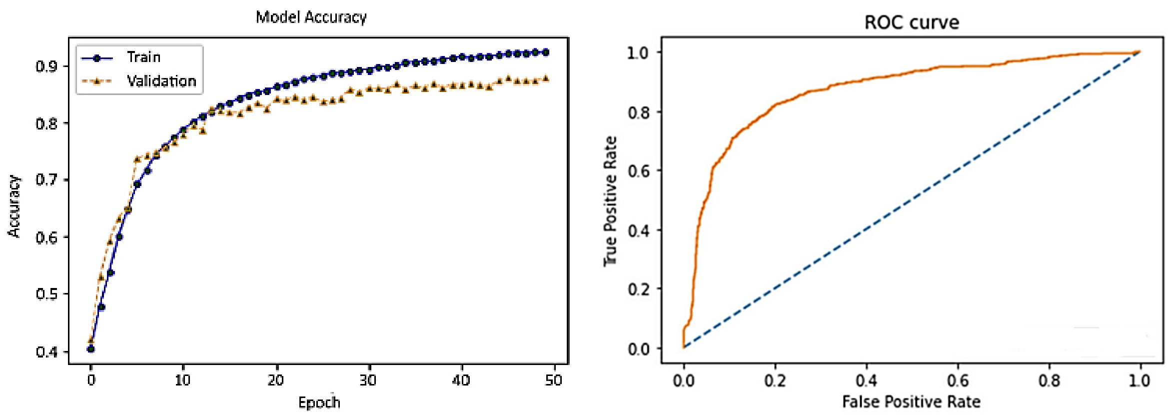
Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

of the validation data. Based on this curve, the model shows a fairly strong predictive ability with an area that moves away from the diagonal line randomly, which means that the model is able to distinguish between relevant and irrelevant quite well. This evaluation shows that the integration of LSTM in semantic recommender systems provides accurate and effective results in understanding users' historical patterns towards learning materials.

Table 2. Hyperparameter for LSTM

Hyperparameter	Value	Description
Embedding Dimension	128	Size of the vector space in which semantic tokens are embedded
Number of LSTM Units	64	Number of units (neurons) in each LSTM layer
Number of Layers	2	Depth of the LSTM stack for capturing complex patterns
Dropout Rate	0.3	Fraction of input units dropped to prevent overfitting
Optimizer	Adam	Adaptive optimizer for training
Learning Rate	0.001	Step size used by the optimizer to minimize loss
Batch Size	32	Number of training samples used in one iteration
Number of Epochs	50	Total number of passes through the full dataset during training
Loss Function	Binary Crossentropy	Used for measuring error in binary recommendation tasks
Activation Function	Tanh (LSTM) / Sigmoid (output)	Non-linear transformations in LSTM and final decision layer



(a) Training Accuracy of LSTM-Smith Waterman

(b) ROC of LSTM-Smith Waterman

Figure 3. Training Phase

Based on Figure 3, after obtaining promising results in both training and validation phases, the next step involved evaluating the model's performance using a confusion matrix. This matrix enables the calculation of various performance metrics such as accuracy, precision, recall, F1-score F1, and the area under the receiver operating characteristic curve (AUROC). These metrics offer a comprehensive view of the model's classification capability, particularly in distinguishing relevant from non-relevant learning materials based on user semantic queries.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Data} \quad (7)$$

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (8)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (9)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (10)$$

Based on Figure 3 explained previously, the next stage is to evaluate the model performance by utilizing commonly used classification metrics, namely Accuracy, Precision, Recall, and F1-Score. This evaluation is carried out to measure the extent to which the LSTMM model is able to predict the test data accurately and consistently. The metric calculation is carried out based on the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values obtained from the confusion matrix of the prediction results.

Furthermore, based on Eq. (7) is used to measure the level of accuracy of the overall prediction. Eq. (8) calculates the precision value, namely the accuracy of the prediction of relevant data. Meanwhile, Eq. (9) is used to determine recall, namely the model's ability to recognize all data that is truly relevant. Finally, Eq. (10) is used to obtain the F1-Score value, which is the harmonic mean of precision and recall as a measure of balanced performance. The results of the calculation of the performance evaluation metrics are shown in Table 3.

Table 3. Performance Matrix

Training Progress	AuROC	Accuracy	Precision	Recall	F1-Score	Elapsed Time
Running in Epoch 10	0.66	0.58	0.61	0.58	0.60	08 Min 32 Sec
Running in Epoch 20	0.68	0.67	0.68	0.68	0.68	14 Min 11 Sec
Running in Epoch 30	0.78	0.72	0.74	0.75	0.74	18 Min 56 Sec
Running in Epoch 40	0.81	0.80	0.81	0.82	0.81	24 Min 31 Sec
Running in Epoch 50 (Final)	**0.88**	**0.89**	**0.90**	**0.89**	**0.89**	**29 Min 44 Sec**

Based on Table 3, the model shows consistent improvements across multiple epochs. Initially, at Epoch 10, the model achieves an AuROC of 0.66 and an accuracy of 58%, with precision and recall values around 0.61 and 0.58, respectively. As the training progresses, both the accuracy and the AuROC values increase steadily. By Epoch 20, the model demonstrates a marked improvement, with accuracy reaching 67% and AuROC rising to 0.68. Precision and recall values are now balanced at 0.68, suggesting that the model is becoming more capable of distinguishing between classes while maintaining a fair balance between false positives and false negatives.

In the later epochs, the model shows significant strides in performance. By Epoch 40, the accuracy has improved to 80%, with both Precision and Recall climbing to 0.81 and 0.82, respectively. These improvements continue in the final epoch (Epoch 50), where the model achieves a high AuROC of 0.88 and an accuracy of 89%. The precision and recall reach their peak at 0.90 and 0.89, resulting in a final F1-Score of 0.89. This indicates that the model has become highly effective at balancing the trade-offs between precision and recall, offering strong overall performance. The elapsed time for training gradually increases from 8 minutes and 32 seconds in the initial epoch to 29 minutes and 44 seconds at the final epoch, reflecting the increasing complexity of the model as it converges to optimal performance.

After achieving strong performance results during the training period, further evaluation was conducted to assess how well the trained LSTM–Smith Waterman model handles real-world semantic

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

queries in a testing scenario. The goal of the testing phase as seen in Table 4 was to determine the system's ability to provide personalized and contextually relevant learning recommendations based on a single input query. Several representative queries were submitted to the model, and the top-ranked recommendation topics were recorded along with their semantic match scores. These scores indicate the system's confidence in the contextual relevance between the query and the recommended content, as calculated through the alignment of the integrated semantic vectors and the historical context learned by the LSTM layer.

Table 4. Semantic Recommendation Result for Testing Input

Test User ID	Input Query	Top Recommended Topic	Semantic Match Score	Remarks
U101	"deep learning basics"	"Introduction to CNNs"	0.963	Strong contextual match; aligned with deep NN.
U102	"data preprocessing"	"Data Cleaning Techniques"	0.948	Accurate; reflects early stage ML workflow.
U103	"backpropagation"	"Neural Network Optimization"	0.932	Matched based on functional context.
U104	"natural language models"	"RNN vs Transformer"	0.982	Correct match; captures semantic NLP evolution.
U105	"image classifier"	"Convolutional Neural Networks"	0.951	Aligned semantically with image-based learning.

To strengthen the performance analysis of the proposed model, a comparison was also conducted with several methods that have been developed in previous studies. This comparison includes aspects of accuracy, F1-score, AUROC, and the advantages and disadvantages of each approach. Table 5 below presents a summary of the comparison, which shows that the LSTM–Smith Waterman model provides superior results in the context of personalization and semantic understanding compared to other relevant methods.

Table 5. Comparison Model with Related Research

Study	Model	Performance	Strengths	Weaknesses
[15]	LDA + BERT	85%	Explainable, high semantic relevance	High computational cost, sensitive to document structure
[16]	Ontology-based SemanticModel	76%	Multi-source integration, strong behavioral analytics	Complex to develop and maintain
[17]	Semantic-aware NLP Framework	84%	Personalized recommendation, high accuracy	Requires large datasets, prone to semantic ambiguity
This Study	LSTM + Smith Waterman	89%	Captures temporal patterns, adaptive personalization, fine-grained semantic matching	Longer training time

 Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

4. DISCUSSIONS

This study introduces an integrated approach that combines the LSTM–Smith Waterman method to enhance semantic search and personalized learning recommendations in academic information systems. Unlike traditional keyword-based search systems, the proposed model leverages semantic vector embeddings and temporal learning patterns to understand the contextual intent behind user queries. The Smith Waterman algorithm enables fine-grained local alignment between queries and content by identifying the most semantically relevant subsequences, while the LSTM model captures the sequential behavior of users over time, enriching the semantic matching process with historical user preferences.

As a best contribution, hybrid LSTM and Smith Waterman model in this study achieves the highest performance with an accuracy of 89%, outperforming other models such as LDA + BERT (85%), Ontology-based Semantic Model (76%), and Semantic-aware NLP Framework (84%). The main strengths of this model include its ability to capture temporal patterns, provide adaptive personalization, and perform fine-grained semantic matching. However, its drawback is a longer training time compared to the other models. Meanwhile, the previous models each have their own advantages, such as LDA + BERT being explainable and semantically relevant, Ontology-based Semantic Model excelling in multi-source integration, and Semantic-aware NLP Framework offering strong personalized recommendations, although they also face challenges like high computational cost, development complexity, and large dataset requirements.

In terms of performance, the LSTM–Smith Waterman model demonstrated consistent improvements throughout the training process. Accuracy increased from 0.58 in the early epochs to 0.89 at epoch 50, with corresponding gains in AUROC (from 0.66 to 0.88) and F1-Score (reaching 0.89). These results indicate that the model effectively distinguishes relevant from non-relevant content and generalizes well across validation data. The use of hyperparameter tuning, including adjustments to the embedding dimension, LSTM units, learning rate, and dropout rate, further contributed to model stability and predictive power. Additionally, evaluation using ROC curves and confusion matrices validated the model's ability to provide accurate and contextually aware recommendations.

To further evaluate real-world applicability, several individual queries were tested against the trained model to observe the semantic recommendation outputs. As shown in Table 4, the system successfully generated contextually appropriate recommendations with high semantic match scores, indicating strong alignment between user intent and suggested learning content. For instance, the query "natural language models" returned the topic "RNN vs Transformer" with a score of 0.982, while "data preprocessing" led to "Data Cleaning Techniques" with a score of 0.948. These examples highlight the model's capability to interpret varied semantic inputs and deliver personalized results, reinforcing its practical value in academic information systems.

These findings suggest that the integration of semantic modeling and temporal learning behavior is not only beneficial for academic search systems, but also has broader implications for the development of intelligent, adaptive educational platforms. The proposed method can serve as a foundation for more responsive recommendation engines in various digital learning environments, enabling personalized content delivery at scale. Furthermore, this approach contributes to the advancement of AI-driven personalization strategies, supporting improved user engagement, retention, and learning outcomes in modern e-learning systems.

5. CONCLUSION

This research demonstrates that integrating semantic alignment and temporal user modeling through the LSTM–Smith Waterman method significantly improves the relevance and personalization of learning material recommendations in academic information systems. By capturing the semantic meaning of queries and sequential user patterns, the proposed hybrid model outperforms traditional keyword-based systems in accuracy, precision, and contextual relevance. Experimental results confirm its ability to adapt to evolving user behavior and deliver meaningful recommendations. The key contribution lies in its hybrid framework that aligns user intent with educational content at both semantic and behavioral levels, supporting adaptive digital learning environments. However, the model has several limitations. It relies on sufficient historical interaction data to perform optimally, and training time increases considerably with data scale. Adaptation across diverse academic domains or multilingual settings may require further tuning. Beyond academic systems, this approach has broader implications for informatics, learning analytics, and AI-driven personalization. It showcases how deep sequence modeling combined with semantic alignment enables more precise, adaptive recommendations in intelligent educational systems. Future work may include integrating attention mechanisms, enabling real-time personalization, and deploying the model on scalable, distributed platforms to enhance performance and applicability.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] M. Liu and D. Yu, "Towards intelligent E-learning systems," *Educ Inf Technol (Dordr)*, vol. 28, no. 7, pp. 7845–7876, Jul. 2023, doi: 10.1007/s10639-022-11479-6.

[2] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, Jul. 2023, doi: 10.1080/15228053.2023.2233814.

[3] M. A. Almaiah, A. Al-Khasawneh, and A. Althunibat, "Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic," *Educ Inf Technol (Dordr)*, vol. 25, no. 6, pp. 5261–5280, Nov. 2020, doi: 10.1007/s10639-020-10219-y.

[4] M. Gusenbauer and N. R. Haddaway, "What every researcher should know about searching – clarified concepts, search advice, and an agenda to improve finding in academia," *Res Synth Methods*, vol. 12, no. 2, pp. 136–147, Mar. 2021, doi: 10.1002/jrsm.1457.

[5] M. K. Mohammadi, A. A. Mohibbi, and M. H. Hedayati, "Investigating the challenges and factors influencing the use of the learning management system during the Covid-19 pandemic in Afghanistan," *Educ Inf Technol (Dordr)*, vol. 26, no. 5, pp. 5165–5198, Sep. 2021, doi: 10.1007/s10639-021-10517-z.

[6] P. Ingavélez-Guerra, S. Otón-Tortosa, J. Hilera-González, and M. Sánchez-Gordón, "The use of accessibility metadata in e-learning environments: a systematic literature review," *Univers Access Inf Soc*, vol. 22, no. 2, pp. 445–461, Jun. 2023, doi: 10.1007/s10209-021-00851-x.

[7] A. Agarwal, D. S. Mishra, and S. V. Kolekar, "Knowledge-based recommendation system using semantic web rules based on Learning styles for MOOCs," *Cogent Eng*, vol. 9, no. 1, Dec. 2022, doi: 10.1080/23311916.2021.2022568.

[8] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review," *Educ Sci (Basel)*, vol. 13, no. 12, p. 1216, Dec. 2023, doi: 10.3390/educsci13121216.

[9] P. Bedi, S. B. Goyal, A. S. Rajawat, R. N. Shaw, and A. Ghosh, "A Framework for Personalizing Atypical Web Search Sessions with Concept-Based User Profiles Using Selective Machine Learning Techniques," 2022, pp. 279–291. doi: 10.1007/978-981-16-2164-2_23.

[10] A. Breit *et al.*, "Combining Machine Learning and Semantic Web: A Systematic Mapping Study," *ACM Comput Surv*, vol. 55, no. 14s, pp. 1–41, Dec. 2023, doi: 10.1145/3586163.

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

[11] N. S. Raj and V. G. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020," *Journal of Computers in Education*, vol. 9, no. 1, pp. 113–148, Mar. 2022, doi: 10.1007/s40692-021-00199-4.

[12] V. Setlur, A. Kanyuka, and A. Srinivasan, "Olio: A Semantic Search Interface for Data Repositories," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA: ACM, Oct. 2023, pp. 1–16. doi: 10.1145/3586183.3606806.

[13] S. M. R. Naqvi, M. Ghufran, C. Varnier, J.-M. Nicod, K. Javed, and N. Zerhouni, "Unlocking maintenance insights in industrial text through semantic search," *Comput Ind*, vol. 157–158, p. 104083, May 2024, doi: 10.1016/j.compind.2024.104083.

[14] M. Ovando-Tellez *et al.*, "An investigation of the cognitive and neural correlates of semantic memory search related to creative ability," *Commun Biol*, vol. 5, no. 1, p. 604, Jun. 2022, doi: 10.1038/s42003-022-03547-x.

[15] A. Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," *Data Knowl Eng*, vol. 144, p. 102110, Mar. 2023, doi: 10.1016/j.datak.2022.102110.

[16] P. N. Andono, Sunardi, R. A. Nugroho, and B. Harjo, "Aspect-Based Sentiment Analysis for Hotel Review Using LDA, Semantic Similarity, and BERT," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 5, pp. 232–243, Oct. 2022, doi: 10.22266/ijies2022.1031.21.

[17] M. Muljono, B. Harjo, and R. Abdullah, "Aspect-Based Sentiment Analysis for Financial Review with Implicit Aspect and Opinion Using Semantic Similarity and Hybrid Approach," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 5, pp. 646–658, Oct. 2024, doi: 10.22266/ijies2024.1031.49.

[18] D. Sheth and M. Shah, "Predicting stock market using machine learning: best and accurate way to know future stock prices," *International Journal of System Assurance Engineering and Management*, vol. 14, no. 1, pp. 1–18, Feb. 2023, doi: 10.1007/s13198-022-01811-1.

[19] T. Sommerschield *et al.*, "Machine Learning for Ancient Languages: A Survey," *Computational Linguistics*, pp. 1–44, May 2023, doi: 10.1162/coli_a_00481.

[20] S. Ciampelli, A. E. Voppel, J. N. de Boer, S. Koops, and I. E. C. Sommer, "Combining automatic speech recognition with semantic natural language processing in schizophrenia," *Psychiatry Res*, vol. 325, p. 115252, Jul. 2023, doi: 10.1016/j.psychres.2023.115252.

[21] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities," in *Linking the World's Information*, New York, NY, USA: ACM, 2023, pp. 91–103. doi: 10.1145/3591366.3591376.

[22] L. Tamine and L. Goeuriot, "Semantic Information Retrieval on Medical Texts," *ACM Comput Surv*, vol. 54, no. 7, pp. 1–38, Sep. 2022, doi: 10.1145/3462476.

[23] N. Yang, J. Jo, M. Jeon, W. Kim, and J. Kang, "Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models," *Expert Syst Appl*, vol. 190, p. 116209, Mar. 2022, doi: 10.1016/j.eswa.2021.116209.

[24] M. Paneque, M. del M. Roldán-García, and J. García-Nieto, "e-LION: Data integration semantic model to enhance predictive analytics in e-Learning," *Expert Syst Appl*, vol. 213, p. 118892, Mar. 2023, doi: 10.1016/j.eswa.2022.118892.

[25] H. Ezaldeen, S. K. Bisoy, R. Misra, and R. Alatrash, "Semantics aware intelligent framework for content-based e-learning recommendation," *Natural Language Processing Journal*, vol. 3, no. 3, p. 100008, Jun. 2023, doi: 10.1016/j.nlp.2023.100008.

[26] Y. Zhao, H. Liu, and H. Duan, "Semantic and relation aware neural network model for bi-class multi-relational heterogeneous graphs," *iScience*, vol. 28, no. 4, p. 112155, Apr. 2025, doi: 10.1016/j.isci.2025.112155.

[27] J. Li, J. Wang, H. Lv, Z. Zhang, and Z. Wang, "IMCHGAN: Inductive Matrix Completion With Heterogeneous Graph Attention Networks for Drug-Target Interactions Prediction," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 19, no. 2, pp. 655–665, Mar. 2022, doi: 10.1109/TCBB.2021.3088614.

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763

[28] Z. Xia *et al.*, “A Review of Parallel Implementations for the Smith–Waterman Algorithm,” *Interdiscip Sci*, vol. 14, no. 1, pp. 1–14, Mar. 2022, doi: 10.1007/s12539-021-00473-0.

[29] F. F. de Oliveira, L. A. Dias, and M. A. C. Fernandes, “Proposal of Smith-Waterman algorithm on FPGA to accelerate the forward and backtracking steps,” *PLoS One*, vol. 17, no. 6, p. e0254736, Jun. 2022, doi: 10.1371/journal.pone.0254736.

[30] L. Li, J. Lin, and Z. Wang, “PipeBSW: A Two-Stage Pipeline Structure for Banded Smith-Waterman Algorithm on FPGA,” in *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, IEEE, Jul. 2021, pp. 182–187. doi: 10.1109/ISVLSI51109.2021.00042.

[31] M. Abdullah, M. Ahmad, and D. Han, “Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification,” in *2020 International Conference on Electronics, Information, and Communication, ICEIC 2020*, Institute of Electrical and Electronics Engineers Inc., Jan. 2020. doi: 10.1109/ICEIC49074.2020.9051332.

[32] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, “Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM,” *IET Image Process*, vol. 14, no. 7, pp. 1227–1232, May 2020, doi: 10.1049/iet-ipr.2019.1188.

[33] N. R. D. Cahyo, C. A. Sari, E. H. Rachmawanto, C. Jatmoko, R. R. A. Al-Jawry, and M. A. Alkhafaji, “A Comparison of Multi Class Support Vector Machine vs Deep Convolutional Neural Network for Brain Tumor Classification,” in *2023 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, Sep. 2023, pp. 358–363. doi: 10.1109/iSemantic59612.2023.10295336.

[34] M. M. I. Al-Ghiffary, C. A. Sari, E. H. Rachmawanto, N. M. Yacoob, N. R. D. Cahyo, and R. R. Ali, “Milkfish Freshness Classification Using Convolutional Neural Networks Based on Resnet50 Architecture,” *Advance Sustainable Science Engineering and Technology*, vol. 5, no. 3, p. 0230304, Oct. 2023, doi: 10.26877/asset.v5i3.17017.

Vol. 6, No. 4, Agustus 2025, Page. 2139-2152
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.4763