Topic Modeling for Constructing Learning Profiles using Latent Dirichlet Allocation and Coherence Evaluation

Andika Dwi Arko [1,*)], Muhamad Yusril Helmi Setyawan [2)], and Roni Andarsyah [3)]

[1, 2, 3)] Informatics Engineering, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

E-mail: adwiarko@gmail.com[1)], yusrilhelmi@ulbi.ac.id[2)], and roniandarsyah@ulbi.ac.id[3)]

ABSTRACT

Understanding individual learning patterns is essential for designing effective strategies in digital education environments. This study introduces a topic modeling framework using the Latent Dirichlet Allocation (LDA) algorithm to construct student learning profiles from the EdNet-KT1 dataset, comprising 153,824 interactions across 11,613 questions. Semantic tag-based pseudotexts were generated and modeled into 20 topics, selected as an optimal balance between semantic coherence and model interpretability. The semantic quality of topics was evaluated using the c_v coherence metric, which combines Pointwise Mutual Information (PMI) and cosine similarity to assess the semantic consistency of top keywords within each topic. The selected 20-topic configuration yielded a coherence score of 0.6688, indicating a high level of internal consistency and interpretability. Each question was assigned a dominant topic, enabling the computation of a student × topic accuracy matrix. Results reveal that 66% of students mastered more than five topics, reflecting broad conceptual exposure. Performance patterns were visualized using heatmaps, radar charts, and PCA-based clustering, with K-Means segmentation yielding four distinct student groups. Adaptive learning recommendations were generated for students with topic-level accuracy below 0.5 and more than 10 interactions. Topics topic_13, topic_10, and topic_12 were identified as the most frequently problematic. These findings highlight the potential of LDA and clustering techniques, reinforced by semantic validation through coherence metrics, to support personalized and context-aware learning strategies. Future research may incorporate temporal modeling and experimental validation to further enhance educational recommendations.

Keywords: Clustering, EdNet, Latent Dirichlet Allocation (LDA), learning profile, topic modeling

1. Introduction

The development of online learning platforms has generated large amounts of student interaction data, which holds significant potential for analyzing individual learning patterns. One effective method for extracting information from large-scale data is topic modeling, particularly with the Latent Dirichlet Allocation (LDA) algorithm [1]. LDA is a generative model that assumes that each document is a mixture of several hidden topics, where each topic is characterized by a specific word distribution. In the context of education, LDA has been used to group questions based on semantic similarity and evaluate the relationship between topics and student performance [2]. This process enables the formation of topic representations that are linked to learning performance, resulting in topic-based learning profiles [3].

Online learning systems today face several challenges, including the inability to deliver truly personalized content, limited contextual adaptation, and inadequate mechanisms to interpret learners' conceptual mastery. Despite the availability of large-scale interaction data, most current approaches fail to fully leverage semantic structures or student-specific conceptual gaps, limiting their effectiveness in supporting differentiated learning pathways. Despite its promising potential of Latent Dirichlet Allocation (LDA) in analyzing online learning behavior, the application of this method still faces several challenges. One of the main issues lies in the development of accurate and personalized learning recommendations tailored to the specific needs and competencies of each learner. Most existing recommendation systems still heavily rely on aggregate accuracy metrics, without considering topic context

* Corresponding author.
Received: April 21st, 2025. Revised: June 3rd, 2025. Accepted: June 16th, 2025.

Available online: July 8th, 2025.

DOI: https://doi.org/10.12962/j24068535.v23i1.a1301

or learners' conceptual strength [4]. Recent approaches highlight that these systems often overlook individual learning styles and semantic relationships between learning sources, resulting in suboptimal recommendation personalization [5]. Therefore, integrating topic modeling techniques such as LDA with adaptive recommendation strategies is crucial for aligning learning resources with learners' conceptual needs. Additionally, while LDA can identify topic structures, many previous studies have not included topic cohesion evaluation as a form of semantic validation, which can undermine the reliability of model interpretations [6].

A feasibility study conducted at MTs Swasta Asih Putera [7] emphasizes the importance of a more adaptive approach in LDA-based learning recommendation systems. The findings indicate that the effectiveness of the model is greatly influenced by the readiness of digital infrastructure and student characteristics, particularly in terms of engagement and personalization. The need for a more adaptive and context-aware education system is also evident in practical implementation. This feasibility study highlights the importance of developing systems that are responsive not only to student diversity but also to technological readiness at the institutional level [7]. These findings emphasize the relevance of solutions that can adapt to heterogeneous learning contexts while leveraging data-driven methods. In addition, many previous approaches did not include topic coherence evaluation as a form of semantic validation, which implies low interpretability and reliability of modeling results [6]. However, the integration of coherence metrics such as topic coherence scores has been proven to improve semantic consistency and ensure that the resulting topic structure truly represents meaningful learning content [1], [6].

Nevertheless, there are still many recommendation systems that rely on aggregate metrics such as overall accuracy, without considering conceptual mastery at the topic level [4]. This results in suboptimal personalization, as the model fails to capture differences in learning needs between individuals. Therefore, LDA should be integrated with adaptive recommendation approaches capable of dynamically and contextually mapping learning profiles, so that the system does not solely rely on the established topic structure but also considers the pedagogical relevance for each learner.

Although Artificial Intelligence (AI) and Machine Learning (ML) technologies have driven significant advances in adaptive learning systems, the integration of semantic dimensions in these processes has received little attention. Many current implementations do not link learning content structures to students' deep conceptual understanding, ultimately limiting the potential for personalized learning interventions [4], [8]. However, strong semantic representations are crucial for building learning profiles that reflect students' conceptual strengths and weaknesses. Previous studies have shown that semantic topic modeling, such as LDA, has high potential for improving learner representations, both on MOOC platforms and in K-12 environments [9]. These limitations underline the need for a more semantically aware and context-sensitive modeling approach that not only captures topic structures but also adapts to individual learners' conceptual profiles.

To address these limitations, this study proposes a systematic LDA-based framework design by mapping student learning patterns from large-scale interaction data (EdNet) and integrating topic cohesion evaluation to ensure semantic validity [6]. In addition, this approach will be complemented by a clustering algorithm to group students based on their performance on each topic. This approach is expected to support the development of more adaptive, personalized, and data-driven learning systems.

Initial analysis of the EdNet-KT1 dataset shows more than 131 million interactions involving 784,309 students, with more than 10,000 semantic tags representing the concepts being tested. On average, each student engaged in approximately 167 interactions, with substantial variation across individuals and topics. This distribution indicates a high degree of diversity in learning patterns, making topic modeling a relevant and useful approach.

Accordingly, this study aims to design and implement a semantically validated topic modeling framework based on Latent Dirichlet Allocation (LDA) to construct personalized and adaptive learning profiles. The framework seeks to address the limitations of existing systems by incorporating topic coherence evaluation and clustering to better capture conceptual learning patterns. To achieve this, the study utilizes the EdNet-KT1 dataset and applies

a combination of LDA, coherence scoring, and K-Means clustering to analyze large-scale learner interaction data and generate topic-based performance profiles.

The main contributions of this study include:

- Integration of topic coherence evaluation as a semantic validation mechanism for LDA modeling results [8].
- Mapping of student performance based on topics, obtained from the distribution of dominant topics in the questions that have been answered.
- Visualization of learning profiles using heatmaps and radar charts to provide a more granular understanding of strengths and weaknesses per topic.
- Application of the clustering algorithm (K-Means) to identify groups of students with similar performance patterns, to support the development of more accurate and contextual learning recommendations.

Overall, this study seeks to evolve this approach into a comprehensive and responsive topic modeling framework to address the challenges in current online learning systems.

To provide a clear understanding of the proposed framework and its evaluation, the remainder of this paper is organized as follows. Section 2 reviews the literature on the use of the EdNet dataset, Latent Dirichlet Allocation (LDA), and clustering techniques in adaptive learning systems. Section 3 describes the research methodology, including data preprocessing, pseudotext generation, topic modeling using LDA, topic cohesion evaluation, and student segmentation through clustering. Section 4 presents the experimental results, including topic structure, student accuracy mapping, and learning profile visualization. Section 5 discusses implications for adaptive learning, with a particular focus on topic difficulty and recommendation strategies. Finally, Section 6 concludes this paper and outlines future research directions, including temporal modeling and empirical validation.

2. Literature Review

This section discusses the literature used as references in this study, EdNet dataset, topic modeling using LDA, and clustering for profiling.

2.1. EdNet Dataset

The EdNet dataset is one of the largest publicly available datasets and has been widely used in research in the field of educational data mining. This dataset includes more than 131 million interactions from approximately 784,309 students in South Korea, collected over two years through Android, iOS, and web-based online learning platforms [10].

EdNet consists of data logs of student interactions with learning content, including practice questions, answers, completion times, and topic/tag metadata that can be used for context-based analysis. Due to its large scale and diversity of interactions, this dataset is suitable for the application of modern analysis techniques such as deep learning, performance prediction, and semantic approaches such as topic modeling [11], [12]. Several studies have utilized EdNet for various purposes, such as predicting correct-incorrect answers using a sequential approach and classifying learning behaviors [13].

However, the application of Latent Dirichlet Allocation (LDA) to EdNet to form semantic learning profiles of students has not been explored in depth. Therefore, this study attempts to fill this gap by developing a topic modeling-based framework that can capture the distribution of concepts in questions and link them directly to student performance.

2.2. Topic Modeling using LDA

Topic modeling has continued to play a central role in educational data mining over the past few years, particularly in analyzing learning behavior and uncovering conceptual structures in data generated by students. Latent Dirichlet Allocation (LDA) remains widely used due to its probabilistic basis and effectiveness in discovering hidden topics in large text corpora. Studies such as those conducted by Ding et al. [6] and Ji et al. [8] have applied LDA to classify learning content and explore student engagement patterns on online platforms. However, LDA is

Table 1: Contribution of Related Studies.

Author	Year	Contribution Topics
Ding et al. [6]	2024	Integration of LDA with semantic enhancement in the educational domain
Ji et al. [8]	2024	Personalized learning systems using educational data mining (EDM)
Shahbazi & Byun [14]	2021	Hybrid topic modeling for short texts
Nguyen et al. [15]	2024	Comparative evaluation of topic coherence metrics in educational modeling
Arko et al. (Ours)	2025	Integration of LDA-based topic modeling with topic coherence evaluation and clustering to construct adaptive learning profiles using EdNet-KT1 data

limited by several constraints, including its reliance on bag-of-words models, its inability to capture word order or semantics, and its sensitivity to hyperparameter settings such as the number of topics.

To address these limitations, recent models have integrated word embeddings and neural language models to enhance semantic representation. For example, BERTopic combines BERT-based embeddings, dimension reduction (UMAP), and density-based clustering (HDBSCAN) to generate cohesive and context-aware topic structures, as demonstrated by Shahbazi and Byun [14]. A summary of contributions from related studies is presented in Table 1, highlighting efforts to improve LDA with semantic embedding and coherence evaluation in constructing meaningful learning profiles in the context of education, including our proposed method. Nevertheless, LDA remains a favorite choice in many educational settings due to its interpretability and ease of integration with structured student performance data. Additionally, topic cohesion evaluation has become an essential validation step, with metrics such as c_v and UMass frequently used to ensure the semantic reliability of extracted topics [15].

Nevertheless, studies that connect LDA-derived topic structures with student performance indicators remain sparse, especially in large-scale datasets such as EdNet. This study seeks to bridge this gap by proposing an LDA-based framework enhanced with topic coherence evaluation and student clustering to construct semantically grounded and performance-informed learning profiles.

2.3. Clustering for Profiling

The clustering method is increasingly used in mapping student characteristics based on their learning performance. One example is research by George and Sumathy, which explored the formation of clusterable student profile attributes to reveal the relationship between profiles and academic grades [16]. However, these studies generally do not integrate topic-performance analysis that could enrich learning recommendation strategies.

The combination of LDA and *clustering* offers a more targeted approach to building adaptive learning systems. This combination enables the formulation of personalized recommendations based on topic distribution and student performance clusters [14]. Alternative models such as Non-negative Matrix Factorization (NMF) can be compared in further studies to measure the effectiveness of LDA versus NMF in this context [17]. In addition, alternative topic modeling techniques such as BERTopic offer a different perspective in educational content analysis.

In conclusion, this study aims to provide conceptual and methodological contributions to understanding student learning profiles through the exploration of large-scale interaction data. Using LDA as the main approach, this study integrates topic coherence evaluation, topic-based accuracy mapping, and student performance clustering to develop a framework that supports the development of more personalized, responsive, and data-driven online learning systems.

3. Methodology

This section discusses the research design used, including the experimental apparatus, variables observed, and procedural stages in conducting the study.

3.1. Research Design

This study uses an exploratory approach with the main objective of forming a student learning profile based on the distribution of topics from the questions they worked on. An exploratory approach was chosen because of its

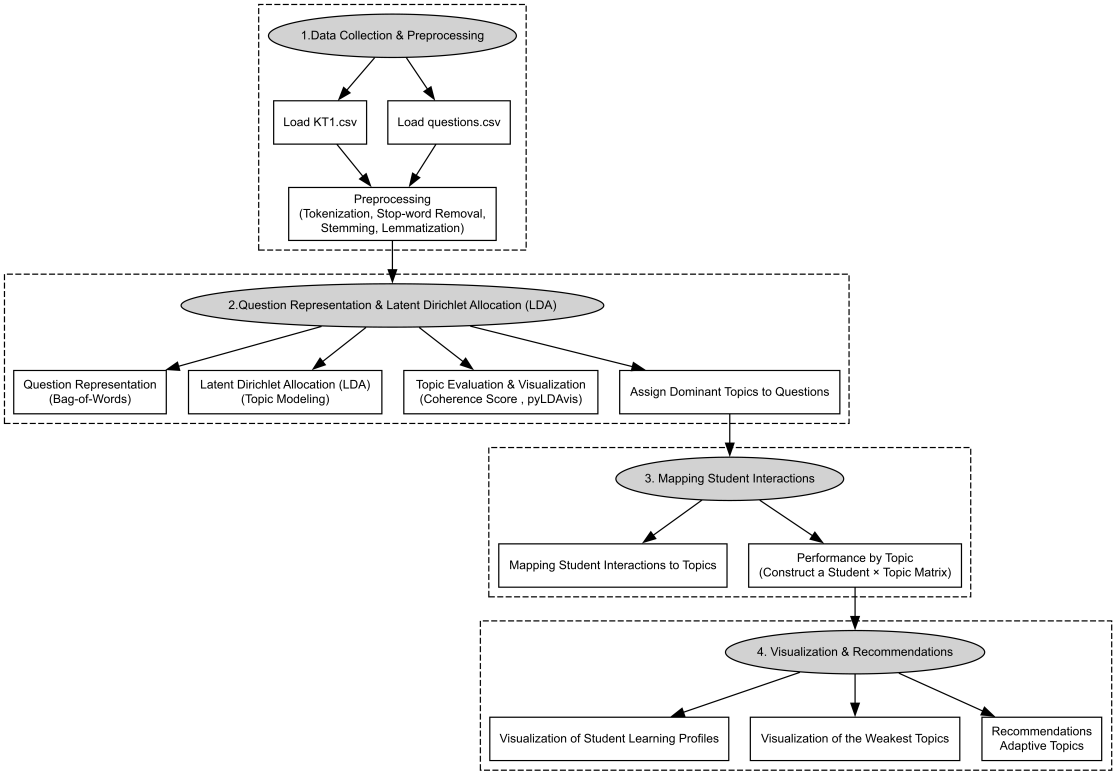

Fig. 1: Methodology Flowchart.

flexibility and relevance in the context of unstructured big data analysis, where researchers can identify patterns and latent relationships without relying on strict initial hypotheses. This approach has proven effective in the context of data-driven research, especially when the research objective is to discover hidden structures and relationships within complex data sets [18], [19].

In this study, topic modeling was performed using the Latent Dirichlet Allocation (LDA) algorithm, which is one of the main methods in topic extraction from text. LDA enables the representation of topic distributions in documents and provides a probabilistic basis for identifying dominant topics in the content of questions answered by students. This provides an opportunity to map student performance based on topics in a more contextual and granular manner.

The experiment was conducted on the Google Colab platform, with specifications of 20 GB RAM and an 8-core CPU, which was considered sufficient to handle large-scale data processing. All analyses were performed using Python 3.9, utilizing the Gensim library for LDA modeling, Scikit-Learn for feature extraction and clustering, and Pandas for data manipulation.

The main objective of this study is to develop a framework that can optimize student performance modeling through analysis of the distribution of topics in the questions they work on. By integrating topic modeling, semantic evaluation, and performance mapping, this study aims to support the development of more adaptive and data-driven learning systems.

The entire methodological process in this study is summarized visually in Fig. 1, which shows the main stages starting from data collection and preprocessing, question representation and LDA, mapping of student interactions, as well as visualization and recommendations. Overall, this flow depicts the integration of topic modeling, topic-based performance analysis, and personalized recommendation systems in an effort to form a data-driven adaptive learning framework.

Table 2: Dataset File Details.

File Name	Description	Size	Data Row
EdNet-KT1.zip	Dataset of student interactions with questions from EdNet (KT1)	1,11 GB	784.309
Question.csv	Question metadata, containing question_id and tags	541 KB	13.169
combined_question_interaction.csv	Combination of 1000 student interaction files with question content via question_id	11,2 MB	153.824
pseudotexts_question.csv	Pseudotext representation of questions based on “tag_XXX” in the clean_text column	348 KB	11.613
topic_question.csv	List of dominant topics for each question based on LDA distribution of pseudotext	97,3 KB	11.613
student_topic_representation.csv	Student accuracy for each topic based on the distribution of correct/incorrect answers	216 KB	9.697
student_grade_matrix_topic.csv	Topic-level accuracy matrix for students, used for calculating student performance per Topic	94,7 KB	1.000
user_topic_matrix.csv	Student × Topic matrix constructed from tagged interaction frequencies	94,8 KB	1.000
student_topic_cluster.csv	Results of student grouping based on topic performance using the K-Means method	126 KB	1.0000

3.2. Dataset Description

The main dataset used in this study is EdNet-KT1, which is part of the EdNet dataset that records student interactions in answering educational questions. This dataset is available in ZIP and CSV formats and can be accessed publicly through the EdNet GitHub repository. EdNet-KT1 includes question content metadata and semantic attributes tagged in the question.csv file. Overall, the dataset records more than 131 million interactions from 784,309 students, as well as 13,169 questions with semantic tags [10]. The scale and diversity of this data provide a great opportunity to explore learning behavior and map learning topics systematically.

However, for a more efficient analysis process, this study only used a subset of 1,000 interaction files from a total of approximately 700,000 log files available. These subsets were combined and processed to form a single file named interaksi_soal_gabungan.csv, which contains approximately 153,824 rows of student interactions with questions. Table 2 summarizes the file structure, size, and main contents of each dataset component used in this study.

The use of this data subset allows the analysis process to be carried out with controlled computation time, without sacrificing concept representativeness, because the subset selection maintains the variation of question topics/tags. The questions.csv file is the starting point in the process of creating pseudotexts for each question through the extraction and transformation of semantic tags into text format. This pseudotext is then used in the topic modeling stage with Latent Dirichlet Allocation (LDA).

3.3. Class Distribution in Datasets

Frequency distribution analysis of tags or question topics is conducted to understand student performance across various topics. An analysis was conducted on the frequency distribution of tags or question topics to explore how each student was distributed in terms of topic mastery, which is very important in contextualizing learning. It is expected that each student will master at least five topics as a basis for forming performance representations per topic [20].

In this study, distribution analysis was conducted through tag frequency visualization prior to topic formation, which provided an initial overview of the distribution of content and topic coverage in the dataset. This visualization shows how often each tag or topic appears and helps identify the correlation between the frequency of a topic's appearance and student performance on that topic.

Table 3: clean_text tags

No	question_id	tags	Pseudotext (clean_text)
1	q8098	5;2;182	tag_5 tag_2 tag_182
2	q8074	11;7;183	tag_11 tag_7 tag_183
3	q176	6;7;183	tag_6 tag_7 tag_183
4	q1279	24;26;182;184	tag_24 tag_26 tag_182 tag_184
5	q6363	130	tag_130

The distribution of these tags not only provides initial insights into the scope of the material but also serves as an important foundation in the process of forming topic models using LDA, which relies on patterns of tag co-occurrence as input for pseudotext representations of questions.

If the technical details regarding the visualization method and the frequency distribution of tags are not explained in detail here, further explanation can be found in Chapter 4, which discusses the analysis and processing of the experimental steps in more detail.

With this approach, class distribution in the dataset provides a deeper insight into student learning patterns, which can ultimately inform recommendations for the development of more adaptive and relevant learning systems.

3.4. Feature Extraction and Design

The main features used in this study include user_id, question_id, tags, user_answer, and correct_answer. The data extraction and transformation process begins with the construction of pseudotexts from the tags column, which involves concatenating all tag elements into a single textual string, each prefixed with "tag_XXX". This transformation is intended to simulate natural language documents from structured categorical metadata. Although the resulting text does not contain syntactic structure or grammar, it serves as an artificial textual representation that approximates the document format expected by topic modeling algorithms such as LDA

Pseudotext in this context refers to a deliberately constructed string of tokens that encode semantic concepts (tags) in a textual format. Unlike natural language sentences, pseudotext does not convey meaning through syntax, but rather through token co-occurrence patterns. This approach allows structured data (e.g., numerical tags or categorical labels) to be processed by natural language modeling pipelines.

The next step is text tokenization and basic text cleaning, including the removal of stopwords. This is performed using the Natural Language Toolkit (NLTK) library, applying stopword lists from both Indonesian and English to accommodate the potential multilingual nature of tag labels. Although the pseudotexts are artificial and lack conventional linguistic elements, stopword removal is still applied to ensure consistency and reduce token redundancy. An example of the transformation performed in the notebook is "algebra,linear_equation" → "tag_algebra tag_linear_equation". After stopwords are removed, only relevant tokens are kept.

Each entry in the dataset is uniquely identified by question_id (e.g., q8098), which serves as the primary reference for individual questions recorded in the question.csv file. This file also includes a tags column, where each question is annotated with one or more numeric semantic identifiers (e.g., 179;53) representing the conceptual topics discussed by the question. These numeric tags are then systematically converted into text tokens by adding the prefix “tag_” to each value, resulting in representations such as tag_179 and tag_53. The converted tokens are then combined into a single text sequence, called pseudotext, and stored in a new column labeled clean_text.

As shown in Table 3, question_id denotes a unique question instance in the dataset, while each token tag_XXX corresponds to a concept label generated from the original tag field. The generated pseudotext serves as a proxy document and forms the input corpus for topic modeling using the Latent Dirichlet Allocation (LDA) algorithm. This transformation method enables the conversion of structured educational metadata into a text format compatible with natural language processing pipelines. As a result, this facilitates the discovery of hidden semantic structures within large-scale learning datasets, thereby supporting more interpretable and concept-based topic modeling analysis.

Table 4: LDA Training Parameters.

Parameter	Value	Description
num_topics	20	Number of topics to be extracted
passes	30	Number of iterations (*passes*) over the entire corpus
alpha	auto	Automatic adjustment of document-topic distribution

3.5. Data Initialization and Preprocessing

The initialization process begins with the creation of pseudotext from the tags column, as explained in the previous section. Next, the transformed text is tokenized and cleaned, resulting in a list of tokens that are ready for use in further analysis.

To convert text into numerical representations, the Gensim library was used to form a dictionary and corpus, and Scikit-Learn (via CountVectorizer) was used to validate the sparsity level of the corpus. The dictionary contains unique words extracted from the entire corpus, while the corpus contains word pairs and their frequencies in each document, which become the main input for the Latent Dirichlet Allocation (LDA) model training process, as used by Aslantas et al. [21]

To improve the quality of the Bag-of-Words representation, frequency-based token filtering was applied using the following parameters:

- no_below=5: retains only tokens that appear in at least five of the included documents.
- no_above=0.8: removed tokens that appear in more than 80% of documents, as they are considered too general and uninformative.

This strategy aims to reduce sparsity (the degree of data emptiness) and avoid overfitting, which is a condition where the model adapts too much to the training data and loses its ability to generalize to new data. Validation was performed using CountVectorizer, which showed that the sparsity ratio was at a reasonable level and that the number of unique tokens available was sufficient to build a rich and representative model.

Next, the LDA model training process was carried out using the Gensim library with parameters as summarized in Table 4. Parameter selection was performed based on considerations of balance between model accuracy and computational efficiency. If necessary, more detailed information about alternative parameter testing can be found in the experiment section or appendix [22].

These initialization and preprocessing steps are fundamental because they determine the quality of the input that will be used in topic modeling, as well as supporting the formation of student learning profiles that are more accurate, meaningful, and interpretable.

3.6. LDA Model Architecture and Training Process

The main model used in this study is Latent Dirichlet Allocation (LDA), a probabilistic topic modeling method used to discover hidden structures (latent topics) in a collection of documents. In this context, LDA is used to map topics that emerge from students' practice questions, thereby forming learning profiles based on topic distribution.

Probabilistically, the Latent Dirichlet Allocation (LDA) model represents the probability of a word *w* occurring in a document *d* as a weighted mixture over latent topics z_k , as shown in (1).

$$P(w \mid d) = \sum_{k=1}^{K} P(w \mid z_k) \cdot P(z_k \mid d) \tag{1}$$

where:

- *K* is the total number of topics;
- $P(w|z_k)$ denotes the topic-word distribution;

- $P(z_k|d)$ is the document-topic distribution;
- $P(w|d)$ is the resulting word-document probability.

The evaluation of the number of topics was conducted by calculating the coherence score using the c_v metric from the Gensim library, with the number of topics varying between 5 and 30. Based on the evaluation results, 20 topics were selected because they showed the first significant increase in the coherence value, from around 0.58 to 0.6688. Although the coherence score continued to increase for the number of topics ranging from 25 to 30, the increase was marginal and risked producing topics that were too narrow, redundant, or semantically meaningless.

This phenomenon is known as semantic overfitting, which is a condition where a model breaks down a large topic into several very similar subtopics, thereby reducing the interpretability and practical usefulness of the model, especially in the context of learning applications. Therefore, selecting 20 topics is considered an optimal compromise between semantic quality, model readability, and analysis efficiency.

The LDA training parameters were configured as follows:

- num_topics = 20 – the number of latent topics to be extracted.
- passes = 30 – the total number of full iterations performed on the corpus during the training process.
- alpha = auto – allow the algorithm to automatically learn the prior density of relevant document topics (Dirichlet hyperparameter).

The model was trained using the Variational Bayes inference approach, which is the default method implemented in the Gensim library. This approach was chosen because it has proven to be efficient in handling large corpora and capable of producing stable and reproducible estimates of topic distributions [23].

Through a combination of selecting the optimal number of topics and setting appropriate training parameters, the LDA model is expected to effectively capture the latent semantic structure of the questions. This semantic representation forms the basis for constructing context-aware and adaptive student learning profiles, thereby supporting a more in-depth analysis of the relationship between topics and individual cognitive characteristics in digital learning environments.

3.7. Representation of Student Learning Profiles

To evaluate student performance at the topic level, accuracy is defined as the ratio between the number of questions answered correctly by student u on topic t divided by the total number of questions attempted by student u on topic t, as shown in (2).

$$Accuracy_{u,t} = \frac{Correct_{u,t}}{Attempted_{u,t}} \tag{2}$$

where:

- $Correct_{u,t}$ is the number of questions answered correctly by student *u* on topic *t*;
- $Attempted_{u,t}$ is the total number of questions attemted by student *u* on topic *t*;
- $Accuracy_{u,t}$ represents the proportion of correct answers for student *u* on topic *t* [15].

This calculation produces a student × topic accuracy matrix that quantitatively describes each student's performance across various topics. This matrix provides a structured basis for analyzing individual knowledge levels and identifying specific learning gaps.

3.8. Topic Quality Evaluation

Internal evaluation of the LDA model was performed by calculating the topic coherence score (c_v), which combines the average Pointwise Mutual Information (PMI) and cosine similarity to assess the semantic consistency

of topics. The topic coherence score *T* is defined based on the paired Pointwise Mutual Information (PMI) among the top *N* words, as shown in (3).

$$Coherence(T) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} PMI(w_i, w_j) \tag{3}$$

where :

- $T = \{w1, w2, \ldots, w_n\}$ denotes the set of the top *N* words representing a topic;
- *PMI* (w_i,w_j) is the pointwise mutual information between words w_i and w_j, reflecting how often they co-occur in the corpus relative to chance;
- *N* is the number of top words used for coherence evaluation;
- The coefficient $\frac{2}{N(N-1)}$ normalizes the score by the number of unique word pairs int the set *T*.

This metric measures the extent to which the top words in a topic appear together across the corpus, compared to the frequency of their expected random co-occurrence [24]. This assessment reflects the strength of semantic relationships among words within a topic. Higher coherence scores indicate topics that are more semantically consistent and interpretable.

Interactive visualization was performed using PyLDAvis to illustrate the distribution of topics and the relationships between topics visually, thereby facilitating interpretation of the model results. PyLDAvis is an interactive tool specifically designed to support the interpretation of topic models, particularly those generated by Latent Dirichlet Allocation (LDA). This tool visualizes topics as circles in a two-dimensional space, where the size of each circle reflects its relative prevalence, and the distance between circles indicates semantic differences. In this study, PyLDAvis was used to validate the readability of the LDA model by showing that the extracted topics were distinct and semantically cohesive.

3.9. Evaluation Scenario and Visualization

Performance visualization was performed using a heatmap to display accuracy per topic. Student performance clustering was performed using K-Means, with cluster distribution visualization using Principal Component Analysis (PCA) [25]. Individual student profiles could be visualized through the use of radar charts. Topic recommendations were given to students with low accuracy as an adaptive learning strategy [26]. Internal validation of the number of clusters was performed using common metrics such as the Silhouette Score and the Davies-Bouldin Index, which indicated that selecting 4 clusters provided adequate separation between student groups based on topic distribution.

With this approach, the study is expected to reveal deeper patterns of student learning and support the development of an adaptive learning system based on topic distribution and individual student performance on each topic.

4. Experiment and Result

4.1. Experiment in Question Topic Modeling

This experiment aims to explore the topics contained in the EdNet-KT1 dataset. This dataset includes more than 1,000 student interaction files that have been merged based on the question_id column, resulting in a combined dataset called interaksi_soal_gabungan.csv. This dataset stores information about interactions such as student responses, completion times, and tags for each question. To prepare for topic modeling, preprocessing was performed on the tags column, where each tag was converted into a pseudotext representation in the format “tag_XXX”. The results of this preprocessing are stored in the file soal_pseudo_text.csv and used as input for the topic modeling stage using the Latent Dirichlet Allocation (LDA) approach.

After preprocessing, a Bag-of-Words (BoW) corpus was successfully formed from 11,613 unique questions, with a total of 189 unique tokens and a sparsity level of 98.82%. The high sparsity value indicates that the text

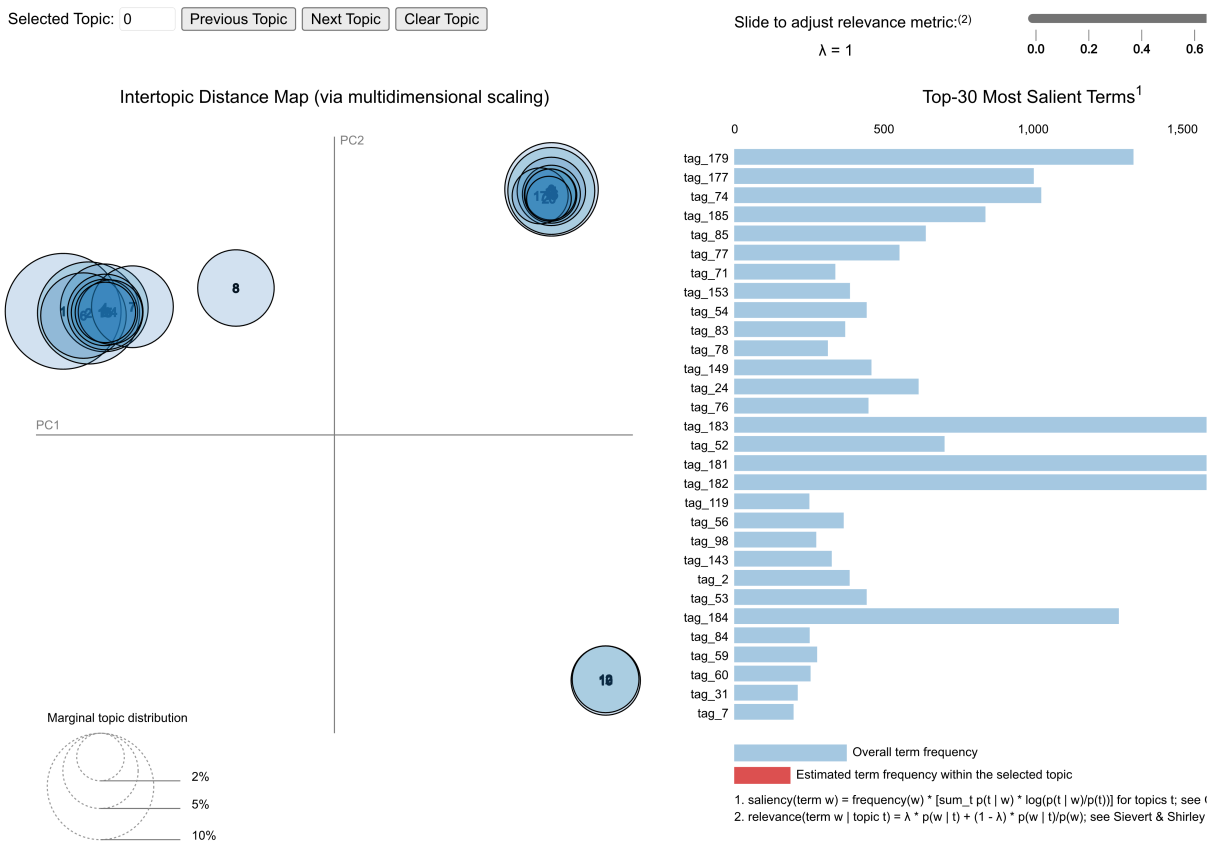

Fig. 2: PyLDAviz Visualization (LDA Topics)

characteristics are suitable for the LDA approach. The LDA model was trained with 20 topics and produced a Coherence Score of 0.6688. This value indicates a good level of semantic consistency between topics. The distribution of topics generated is visualized using PyLDAvis, which shows that the topics are non-overlapping and spatially separated, providing confidence in the quality of the model generated [27]. The visualization is presented in Fig. 2.

4.2. Representation and Mapping of Topics to Questions and Students

Each question is labeled based on the dominant topic inferred from the topic distribution generated by the LDA model. These labels are stored in the topic_question.csv file. Student interactions are subsequently mapped to these topics by associating each answered question with its corresponding dominant topic. The accuracy for each topic is calculated by dividing the number of correctly answered questions by the total number of attempts made on that topic.

The results of this mapping are stored in student_topic_representation.csv and are further summarized into a user × topic matrix (user_topic_matrix.csv). This matrix serves as the foundation for visualizing student performance and conducting segmentation analysis. For illustration purposes, Table 5 presents five representative questions, along with their semantic tags, transformed pseudotexts, and the dominant topics assigned through LDA.

Each entry in Table 5 corresponds to a unique question (question_id) from the EdNet dataset. Each question is accompanied by one or more numeric semantic tags recorded in the tags column. These numeric values represent latent conceptual identifications defined by the dataset, although their explicit meanings are not provided. To facilitate topic modeling, these numeric tags are systematically converted into prefixed string tokens (e.g., tag_5, tag_182) and concatenated into a single text sequence known as *pseudotext*. This artificial text is stored in the clean_text column and serves as document input for the Latent Dirichlet Allocation (LDA) model.

Although pseudotexts do not follow natural language grammar, they encode patterns of co-occurrence of conceptual tags, which can be effectively modeled using topic modeling algorithms. This transformation bridges the

Table 5: Sample Questions with Semantic Tags and Assigned Dominant Topics.

No	question_id	tags	pseudotext	dominant_topic
1	q8098	5;2;182	tag_5 tag_2 tag_182	topic_12
2	q8074	11;7;183	tag_11 tag_7 tag_183	topic_14
3	q176	6;7;183	tag_6 tag_7 tag_183	topic_14
4	q1279	24;26;182;184	tag_24 tag_26 tag_182 tag_184	topic_17
5	q6363	130	tag_130	topic_2

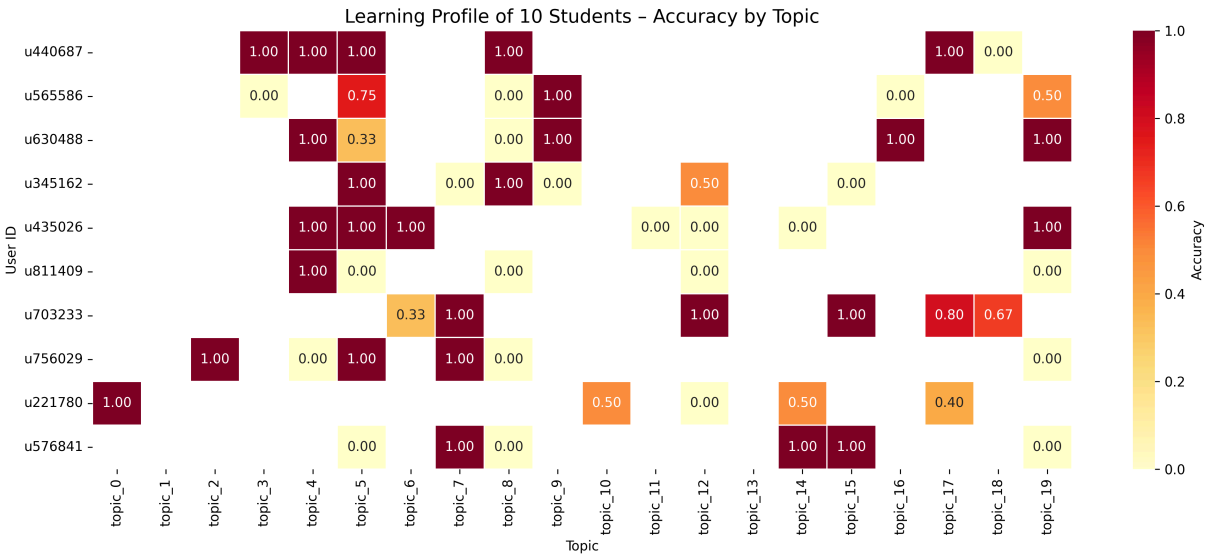

Fig. 3: Heatmap Visualization (Student Accuracy by Topic)

gap between structured metadata and text-based semantic analysis, enabling the LDA model to associate dominant topics with each question. These associated topics are then used to align student interaction data with thematically grouped content, facilitating the formation of interpretable and context-sensitive learning profiles.

The analysis shows that 66% of students were involved in more than five different topics, highlighting diverse conceptual exposure. These results reinforce the usefulness of topic-based models in capturing multidimensional learning behavior and supporting contextual student profiling.

4.3. Visualization of Student Learning Profiles

To better understand patterns in student performance, we visualize the results of the topic-wise accuracy analysis using multiple methods. First, a heatmap is used to display student accuracy across different topics, highlighting both strengths and weaknesses in a visually intuitive format. Second, a radar chart is employed to compare the topic-wise performance of 10 randomly selected students who engaged with at least five topics. The visualizations are presented in Fig.3 and Fig.4. These visualizations enable the identification of topics that are strengths or weaknesses for students in a visual and intuitive manner.

As shown in Fig. 3, the heatmap provides an overview of how individual students perform across all LDA-derived topics. Lighter shades represent higher accuracy, while darker tones indicate lower mastery levels. This visualization enables quick identification of commonly difficult topics as well as students who may require targeted pedagogical interventions.

As illustrated in Fig. 4, radar charts display the topic-wise accuracy of individual students who engaged with at least five different topics. Each axis represents an LDA-generated topic (e.g., topic_4, topic_17), and the radial value (ranging from 0 to 1.0) indicates the student's level of mastery for that topic. The varying shapes of the charts reveal distinct learning profiles. For example, student u703233 exhibits high and consistent accuracy across multiple topics—indicating broad conceptual mastery—while students such as u565586 or u576841 show strong

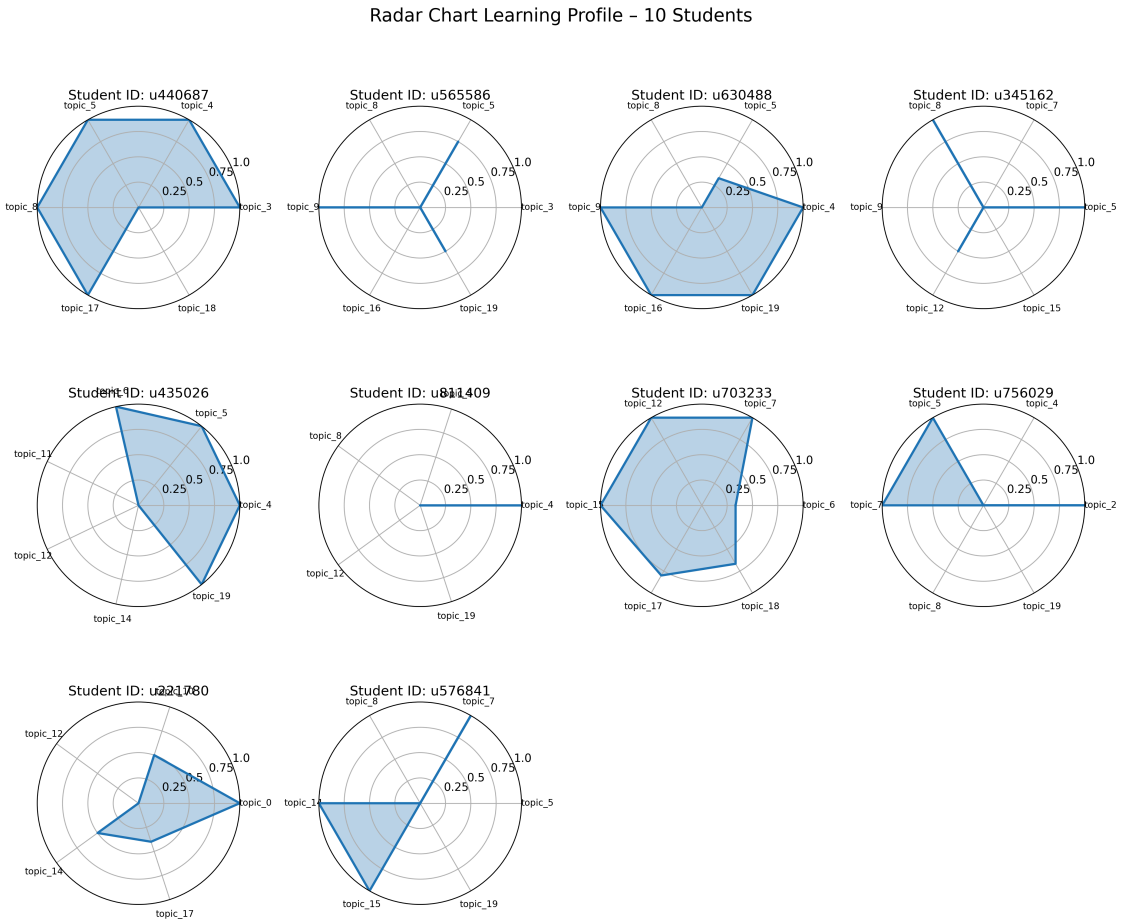

Fig. 4: Radar Chart Visualization (Student Learning Profile)

performance in only a few areas. These differences underscore the importance of implementing adaptive learning strategies that address specific weaknesses while reinforcing individual strengths.

4.4. Clustering and Segmentation of Student Profiles

After obtaining student profile representations, further segmentation was carried out to group students based on similarities in learning performance characteristics. The user_id × topic matrix that had been formed was used as the basis for applying the K-Means clustering algorithm, with a predetermined number of clusters of 4. The results of this clustering process were visualized using Principal Component Analysis (PCA). The visualization shows that students can be grouped quite clearly into 4 clusters:

- Cluster 0 – Students with average performance across most topics.
- Cluster 1 – Students with low and uneven performance.
- Cluster 2 - Students with high and consistent performance.
- Cluster 3 - Outliers with irregular or highly fluctuating performance.

As shown in Fig. 5, the clustering results are visualized using Principal Component Analysis (PCA), which reduces the high-dimensional topic space into a 2D representation. This visualization highlights the separability of the clusters, confirming the effectiveness of topic-based profiling in capturing variations in student mastery levels and learning characteristics.

The diversity of cluster distribution reinforces the potential for implementing differentiated instructional strategies. For instance, students in Cluster 1 may benefit from targeted remedial support, while those in Cluster 2

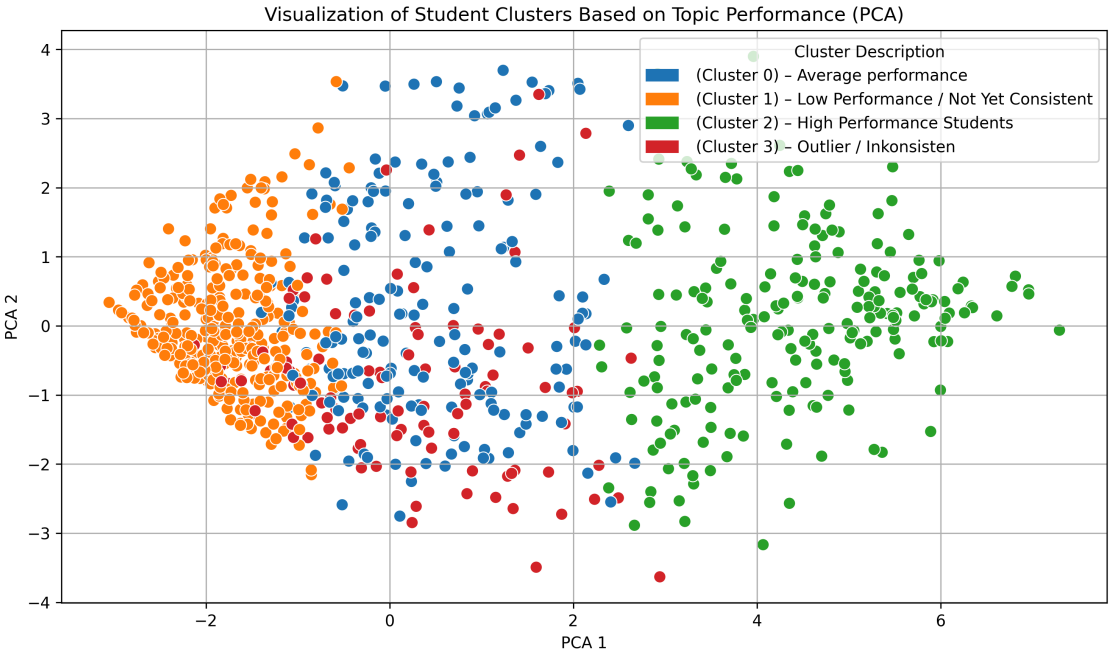

Fig. 5: Visualization of Student Cluster Based on Topic Performance (PCA)

could be offered enrichment activities or advanced material [28]. This adaptive approach enables more personalized learning pathways based on actual topic-level performance rather than aggregate scores alone.

The PCA projection shows a clear separation between the 4 clusters, each representing specific performance patterns. This confirms that topic-based accuracy features are effective in clustering students and can serve as a solid foundation for adaptive learning system design.

4.5. Adaptive Learning Recommendations

Based on the student × topic accuracy matrix analysis, a filtering threshold was applied to identify learning gaps. Specifically, topics were flagged for recommendation if student accuracy on that topic was below 0.5 and students had attempted more than 10 questions related to that topic. This criterion ensures that recommendations are based on sufficient interaction data while targeting areas of significant weakness.

The recommendations generated are compiled and stored in the file rekomendasi_pembelajaran_adaptif.csv. To identify common patterns, the frequency of each recommended topic is collected from all students and displayed in a bar chart. As shown in Fig. 6, topics such as topic_13, topic_10, and topic_12 appear as the most frequently recommended topics, indicating that these topics pose the greatest challenges for most students.

This visualization highlights specific learning difficulties in certain topics on a large scale and supports the formulation of targeted, data-driven intervention strategies for adaptive learning systems.

4.6. Model Coherence Evaluation

To evaluate the semantic quality of the Latent Dirichlet Allocation (LDA) model, an internal coherence analysis was conducted using the c_v coherence score, a commonly used metric for measuring the semantic relationship between words in each topic. This evaluation was performed on various topic configurations, ranging from 5 to 30 topics, to identify the optimal model structure that balances cohesion, readability, and analytical utility.

The results show a significant increase in cohesion scores—from around 0.58 on 15 topics to a peak of 0.6688 on 20 topics. After this point, the increase in cohesion became marginal and was accompanied by a tendency toward semantic overfitting, in which broader conceptual topics were fragmented into overly similar and narrowly defined subtopics. This fragmentation reduced the readability of topics and made it difficult to map topics to question content or student responses.

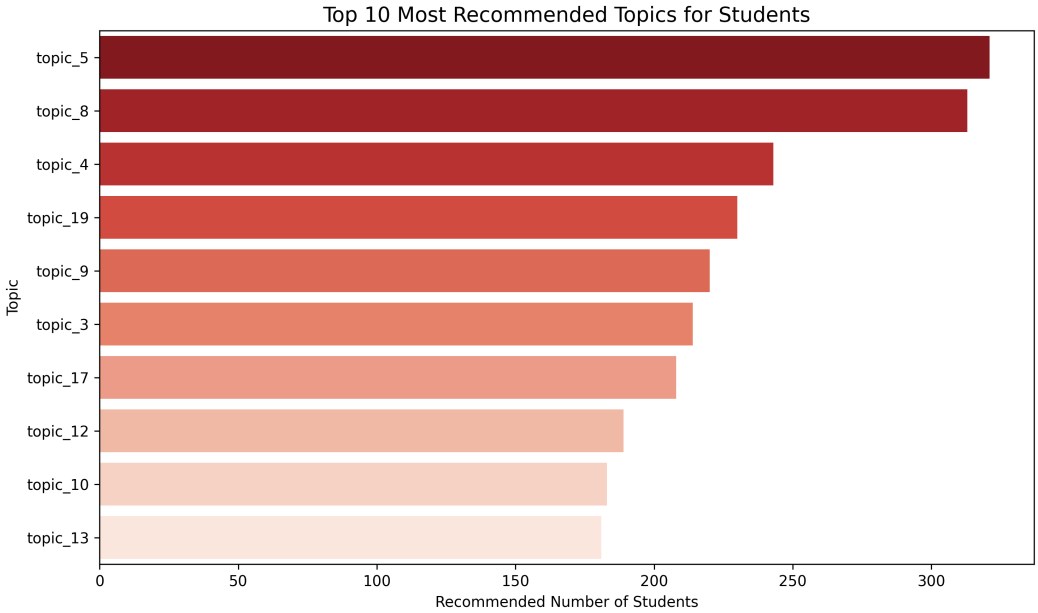

Fig. 6: Bar Chart Visualization (Top 10 Most Recommended Topics for Student)

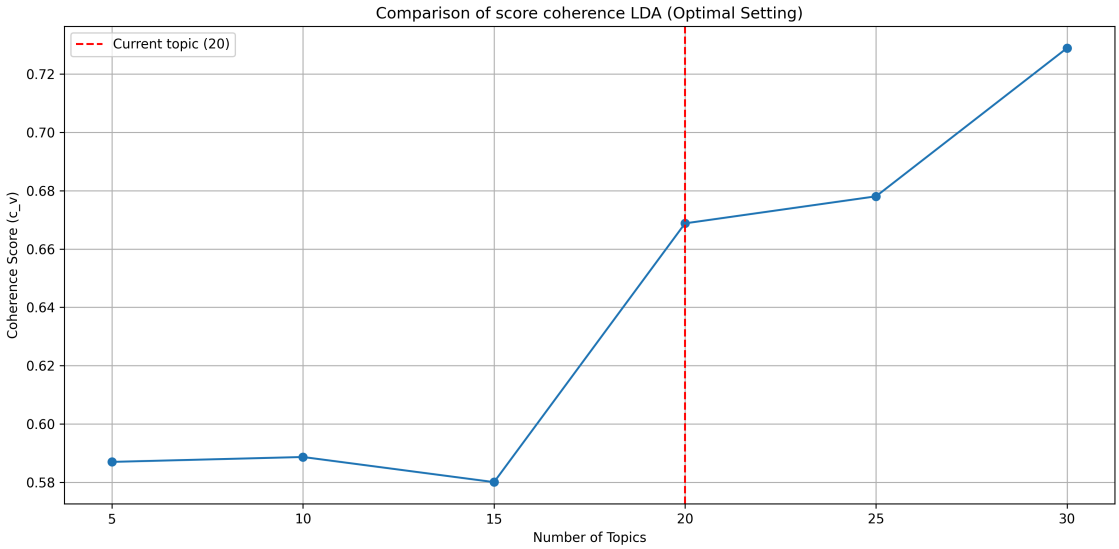

Fig. 7: Comparison Graph of Score Coherence LDA.

Considering semantic coherence, topic uniqueness, and analysis efficiency, a configuration of 20 topics was selected as the optimal model. Fig. 7 illustrates the coherence trend across different numbers of topics, with the red vertical line marking 20 topics as the equilibrium point between model quality and readability.

In addition to topic cohesion evaluation, cluster quality analysis was also conducted to determine the optimal number of clusters in the segmentation of student learning profiles based on topic accuracy. The K-Means clustering algorithm was applied with values of K=2 to K=6, and clustering performance was evaluated using three common internal metrics: Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The results are summarized in Table 6.

Numerically, the configuration with K=2 produces the best clustering quality, as indicated by the highest Silhouette Score (0.296), the lowest *DBI* (1.41), and the highest *CHI* (367.4). However, for pedagogical interpretation and adaptive intervention purposes, this study opts to use K=4 in subsequent analysis and visualization. Although it has lower clustering quality (Silhouette Score = 0.198, *DBI* = 2.78, *CHI* = 168.1), K=4 allows for more detailed and educationally meaningful segmentation of student performance, resulting in 4 distinct groups:

Table 6: Clustering Evaluation Results.

Number of Clusters (*K*)	Silhouette Score	Davies-Bouldin Index (DBI)	Calinski-Harabasz Index (CHI)
2	0.296	1.41	367.4
3	0.204	2.62	228.2
4	0.198	2.78	168.1
5	0.136	2.70	141.8
6	0.129	2.63	123.0

- High-performing students
- Low-performing student
- Average performers
- Inconsistent or fluctuating performers

This structure supports more targeted adaptive strategies, such as providing remedial support to low-performing groups and enrichment material to high-performing groups. Therefore, selecting K=4 represents a considered compromise—trading statistical optimality for practical relevance and readability in the context of personalized learning. This aligns with the educational objectives of this study, which prioritize actionable insights over quantitative performance alone.

Through this series of evaluation experiments, this study shows that LDA-based topic modeling is capable of extracting meaningful latent patterns from student interaction data, thereby supporting a framework for personalized and adaptive learning in the context of digital education. In addition, this analysis reinforces the importance of cohesion-based validation to ensure the semantic reliability of the generated topics.

In addition to quantitative metrics, qualitative validation was also performed using PyLDAvis, an interactive visualization tool that allows examination of topic separation and keyword cohesion. The visualization confirmed that most topics were well separated with minimal overlap, indicating strong thematic differences. Representative keywords within each topic demonstrate high internal consistency, confirming that the LDA model successfully captures semantically cohesive clusters. These findings are consistent with the relatively high c_v score (0.6688), providing additional support for the statistical and conceptual integrity of the model.

Additionally, the dominant topics assigned to questions were found to align meaningfully with the original semantic tags in the EdNet dataset, enhancing confidence in the topic tagging process and recommendations. Overall, these results validate the robustness of the learned profiles and highlight the practical utility of topic-based modeling in an educational context.

5. Conclusion

This study introduced an approach to construct semantic-based student learning profiles using Latent Dirichlet Allocation (LDA) on large-scale EdNet-KT1 data, comprising over 131 million interactions from approximately 784,000 students. Tags were transformed into pseudotexts to enable thematic representation, and an LDA model trained with 20 topics achieved a coherence value of 0.6688. PyLDAvis visualization revealed strong topic separation, affirming the interpretability of the model. This configuration was selected as optimal due to its balance between semantic clarity and model simplicity. Results indicate that excessive topic granularity may diminish support for adaptive learning strategies.

Each question was labeled with a dominant topic, allowing student accuracy to be calculated per topic and aggregated into a user-topic matrix. The analysis showed that 66% of students mastered more than five topics, reflecting broad conceptual coverage. Heatmaps and radar charts enabled fine-grained insights into student strengths and weaknesses. Recommendations were generated for students with topic-level accuracy below 0.5 and at least 10 attempted questions, with topic_13, topic_10, and topic_12 emerging as the most challenging. Clustering via

K-Means yielded four distinct learning profiles, visualized using PCA, highlighting the diversity of student characteristics.

While this framework demonstrates strength in semantic representation and performance segmentation, it does not yet incorporate temporal or affective learning dimensions. Future work should consider sequential modeling (e.g., LSTM) and empirical validation through pre-post testing or A/B experiments. Nevertheless, this study confirms the utility of LDA-based topic modeling, semantic coherence evaluation, and cluster-based profiling as a robust analytical framework for understanding student learning behavior. Leveraging large-scale educational data, this model advances the development of adaptive, personalized, and evidence-based online learning environments.

The findings underscore that topic modeling offers more conceptually meaningful performance insights than traditional aggregate scores. By mapping each question to its dominant topic and measuring topic-specific accuracy, the system enables detailed profiling of student understanding. These individualized and context-aware profiles support the design of targeted interventions. Moreover, semantic validation through coherence metrics and interactive visualization ensures the pedagogical relevance of topics.

This approach has significant potential to improve online learning systems. When integrated into educational platforms, it enables customized content delivery, informed student segmentation, and data-driven decision making. This framework can be integrated into existing e-learning platforms to facilitate real-time profiling, targeted interventions, and automatic topic recommendations based on individual learning patterns and mastery gaps. Transparent visualization of learning profiles also supports instructional monitoring and progress tracking. Overall, this study contributes to the development of more responsive and student-centered digital education systems.

CRediT Authorship Contribution Statement

Andika Dwi Arko: Writing – review and editing, Writing – original draft, Validation, Software, Methodology, Conceptualization, Investigation. **Muhamad Yusril Helmi Setyawan**: Writing – review and editing, Formal analysis, Data Curation, Conceptualization. **Roni Andarsyah**: Writing – review and editing, Methodology.

Declaration of Competing Interest

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

Data Availability

The dataset used in this study was openly provided via [https://github.com/riiid/ednet], and additional experimental materials and scripts are available at [https://github.com/dikadwi/Topic-Modelling-EdNet].

Declaration of Generative AI and AI-assisted Technologies in The Writing Process

The authors used generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and take full responsibility for the final publication.

References

[1] X. Pan and Y. Xu, "Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject—A Case Survey of Latent Dirichlet Allocation Method," *IEEE Access*, vol. 11, pp. 132627–132640, 2023, doi: 10.1109/access.2023.3334619.

[2] H. Chen, C. Yin, R. Li, W. Rong, Z. Xiong, and B. David, "Enhanced learning resource recommendation based on online learning style model," *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 348–356, Jun. 2020, doi: 10.26599/tst.2019.9010014.

[3] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, R. M. G. García-Castelán, and D. Escobar-Castillejos, "Learning Analytics to Determine Profile Dimensions of Students Associated with Their Academic Performance," *Applied Sciences*, vol. 12, no. 20, p. 10560, Oct. 2022, doi: 10.3390/app122010560.

[4] X. Li, "Building a Machine Learning Algorithm-Based Model to Suggest Tourist Attractions in Response to Travelers' "Slow Life" Requirements," *Journal of Cases on Information Technology*, vol. 27, no. 1, pp. 1–16, Mar. 2025, doi: 10.4018/jcit.371409.

[5] L. Yan, C. Yin, H. Chen, W. Rong, Z. Xiong, and B. David, "Learning Resource Recommendation in E-Learning Systems Based on Online Learning Style," in *Knowledge Science, Engineering and Management*, Springer International Publishing, 2021, pp. 373–385. doi: 10.1007/978-3-030-82153-1_31.

[6] R. Ding, P. Huang, S. Chen, J. Zhang, J. Huang, and Y. Zheng, “An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic,” in *2024 26th International Conference on Advanced Communications Technology (ICACT)*, IEEE, Feb. 2024, pp. 1–6. doi: 10.23919/icact60172.2024.10471952.

[7] M. Y. H. Setyawan and M. I. C. R., “Feasibility Study on the Implementation of a Personalized Learning System for Optimizing Individual Engagement and Potential at MTs SWASTA ASIH PUTERA, Cimahi.” Bandung, 2024.

[8] X. Ji, L. Sun, X. Xu, and X. Lei, “Construction and Innovative Exploration of Personalized Learning Systems in the Context of Educational Data Mining,” *International Journal of Information and Communication Technology Education*, vol. 20, no. 1, pp. 1–14, Jul. 2024, doi: 10.4018/ijicte.346992.

[9] H. Yan, F. Lin, and Kinshuk, “Including Learning Analytics in the Loop of Self-Paced Online Course Learning Design,” *International Journal of Artificial Intelligence in Education*, vol. 31, no. 4, pp. 878–895, Dec. 2020, doi: 10.1007/s40593-020-00225-z.

[10] Y. Choi *et al.*, “EdNet: A Large-Scale Hierarchical Dataset in Education,” in *Artificial Intelligence in Education*, Springer International Publishing, 2020, pp. 69–73. doi: 10.1007/978-3-030-52240-7_13.

[11] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi, “SAINT+: Integrating Temporal Features for EdNet Correctness Prediction,” in *LAK21: 11th International Learning Analytics and Knowledge Conference*, in LAK21. ACM, Apr. 2021, pp. 490–496. doi: 10.1145/3448139.3448188.

[12] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Jan. 2020, doi: 10.1002/widm.1355.

[13] D. J. Lemay, C. Baek, and T. Doleck, “Comparison of learning analytics and educational data mining: A topic modeling approach,” *Computers and Education: Artificial Intelligence*, vol. 2, p. 100016, 2021, doi: 10.1016/j.caeai.2021.100016.

[14] Z. Shahbazi and Y.-C. Byun, “Topic prediction and knowledge discovery based on integrated topic modeling and deep neural networks approaches,” *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 1, pp. 2441–2457, Aug. 2021, doi: 10.3233/jifs-202545.

[15] L. T. Nguyen *et al.*, “Evaluating the Performance of Topic Modeling Techniques for Bibliometric Analysis Research: An LDA-based Approach,” *HighTech and Innovation Journal*, vol. 5, no. 2, pp. 312–330, Jun. 2024, doi: 10.28991/hij-2024-05-02-07.

[16] L. George and P. Sumathy, “An Integrated Clustering and BERT Framework for Improved Topic Modeling,” Aug. 2022, doi: 10.21203/rs.3.rs-1986180/v1.

[17] A. Nanyonga, H. Wasswa, and G. Wild, “Topic Modeling Analysis of Aviation Accident Reports: A Comparative Study between LDA and NMF Models,” in *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, IEEE, Dec. 2023, pp. 1–2. doi: 10.1109/smartgencon60755.2023.10442471.

[18] S. Heras, J. Palanca, P. Rodriguez, N. Duque-Méndez, and V. Julian, “Recommending Learning Objects with Arguments and Explanations,” *Applied Sciences*, vol. 10, no. 10, p. 3341, May 2020, doi: 10.3390/app10103341.

[19] K. Limiansi and S. Hadi, “Students Self-Efficacy Profile in Online Learning: Basic Information to Improve the Quality of Learning,” in *Proceedings of the 5th International Conference on Current Issues in Education (ICCIE 2021)*, in iccie-21. Atlantis Press, 2022. doi: 10.2991/assehr.k.220129.016.

[20] A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, “Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques,” *Computers*, vol. 14, no. 3, p. 83, Feb. 2025, doi: 10.3390/computers14030083.

[21] G. ASLANTAŞ, M. GENÇGÜL, M. RUMELLİ, M. ÖZSARAÇ, and G. BAKIRLI, “Customer Segmentation Using K-Means Clustering Algorithm and RFM Model,” *Deu Muhendislik Fakultesi Fen ve Muhendislik*, vol. 25, no. 74, pp. 491–503, May 2023, doi: 10.21205/deufmd.2023257418.

[22] V. Pasupuleti, B. Thuraka, C. S. Kodete, and S. Malisetty, “Enhancing Supply Chain Agility and Sustainability through Machine Learning: Optimization Techniques for Logistics and Inventory Management,” *Logistics*, vol. 8, no. 3, p. 73, Jul. 2024, doi: 10.3390/logistics8030073.

[23] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational Inference: A Review for Statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Apr. 2017, doi: 10.1080/01621459.2017.1285773.

[24] D. O’Callaghan, D. Greene, J. Carthy, and P. Cunningham, “An analysis of the coherence of descriptors in topic modeling,” *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, Aug. 2015, doi: 10.1016/j.eswa.2015.02.055.

[25] I.-C. Chang, T.-K. Yu, Y.-J. Chang, and T.-Y. Yu, “Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals,” *Sustainability*, vol. 13, no. 19, p. 10856, Sep. 2021, doi: 10.3390/su131910856.

[26] Y. Chen, X. Li, J. Liu, and Z. Ying, “Recommendation System for Adaptive Learning,” *Applied Psychological Measurement*, vol. 42, no. 1, pp. 24–41, Mar. 2017, doi: 10.1177/0146621617697959.

[27] X. Deng, H. Hou, M. Jin, and L. Zhai, “Forecasting Students’ Employment Rate Under the OBE Model,” in *Machine Learning & Applications*, in CMLA. Academy & Industry Research Collaboration, Jun. 2023, pp. 37–47. doi: 10.5121/csit.2023.131004.

[28] G. S. Maguate, J. S. Odango, J. N. D. Dela Cruz, M. A. Cornel, A. M. Abule, and F. T. Uy, “Analyzing Student Performance: A Clustering Approach for Academic Intervention,” 2024, doi: 10.5281/zenodo.12200364.