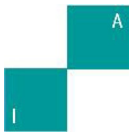
Inteligencia Artificial 28(75), 46-62
doi: 10.4114/intartif.vol28iss75pp46-62



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Semantic alignment in disciplinary tutoring system: leveraging sentence transformer technology

Rosana Abdoune[1,2, *], Lydia Lazib[2], Farida Dahmani-Bouarab[2], Nada Mimouni[3]
[1] LARI Laboratory, Mouloud Mammeri University of Tizi-Ouzou, Tizi-Ouzou, Algeria.
[2] Computer Science Department, Mouloud Mammeri University of Tizi Ouzou, Tizi Ouzou, Algeria.
[*] rosana.abdoune@ummto.dz
[3] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, 75003 Paris, France

Abstract. In this work, we present a disciplinary e-tutoring system that integrates ONTO-TDM, an ontology designed for teaching domain modeling, with advanced transformer technology. Our primary objective is to enhance semantic similarity tasks within the system by fine-tuning a Sentence Transformer model. By carefully adjusting training parameters with a curated dataset of question-answer pairs focused on algorithms and data structures, we achieved a notable improvement in system performance. The Sentence Transformer model, combined with domain ontology, achieved an accuracy of 91%, a precision of 93%, a recall of 89%, and an F1-score of 90%, significantly surpassing the results of existing works. This methodology highlights the potential to deliver personalized support and guidance in tutoring scenarios. It effectively addresses the evolving needs of modern education by offering tailored answers and reducing the necessity for constant learner-tutor interaction, thereby improving the efficiency of educational support systems.

Keywords: Disciplinary e-tutoring system, domain ontology, sentence transformer, question-answering system, education.

1 Introduction

Since the inception of COVID-19, academic institutions have significantly shifted from traditional classroom teaching to online education or e-learning. This transition was driven by the need to continue providing quality education, using electronic resources to ensure access to a broad range of educational materials and services [1], and to facilitate online communication [2]. However, despite the widespread adoption of e-learning platforms, many still lack dynamic question-answering (QA) system, leaving students without adequate support to clarify doubts in specific disciplines. For instance, a computer science student grappling with complex concepts like sorting algorithms and linked lists within algorithms and data structures field may find existing course materials insufficient. Inconsistent responses from online forums further compound the issue, leading students to hesitate when reaching out to instructors for fear of appearing burdensome or asking basic questions.

Addressing these obstacles, the implementation of a robust QA system emerges as a promising solution to enhance the learning experience for students. Such a system would provide continuous support by promptly responding to learners' inquiries about specific disciplines and topics, while also alleviating the workload of instructors by efficiently addressing individual student queries. Integrating such a system into educational institutions can ensure a more dynamic, efficient, and learner-centric approach to online education.

QA systems are becoming increasingly common, serving as robust platforms for responding to human queries in natural language. Leveraging document repositories or pre-structured ontologies, they provide accurate and concise answers. With significant contributions from academic research, the QA field is rapidly expanding globally to meet growing demand [3]. Their popularity stems from their ability to provide accurate, contextually relevant answers tailored to specific questions. This capability is growing interest in multiple contexts. The

ISSN: 1137-3601 (print), 1988-3064(on-line)

convenience of retrieving targeted information efficiently contributes to their adoption. As a result, QA systems are recognized as valuable tools across multiple domains.

In the recent surge of advancements within artificial intelligence, language representation models have become a pivotal force, pushing the boundaries of state-of-the-art performance across a variety of NLP tasks. At the heart of this revolution lies the Transformer [4] network architecture, which employs self-attention mechanisms rather than traditional methods. This architecture is foundational to several language models, including BERT [5], and its variants such as RoBERTa [6]. These models are distinguished by their ability to discern context bidirectionally within text sequences. Furthermore, sentence transformers such as SBERT [7] advance the field by encapsulating higher linguistic structures, ranging from sentences to entire documents, through sophisticated embedding techniques. Pre-trained on extensive corpora of unlabeled text, these models are fine-tuned to excel in downstream tasks like text classification and sentiment analysis. The integration of these advanced techniques into QA systems reflects a broader trend of adopting cutting-edge methods for effective responses to questions expressed in natural language across various domains.

This study aims to build upon a previous system tailored to answer learners' queries regarding specific subjects such as algorithms and data structures. The initial system employed a domain ontology called ONTO-TDM (ontology for teaching domain modeling), alongside a combination of NLP techniques, including TF-IDF [8] and word embedding [9]. Our goal is to elevate the performance of semantic similarity analysis by integrating transformer architecture technologies. This enhancement targets the optimization of the process for determining the most suitable response to a given query and seeks to overcome the limitations of the prior method by refining word representation and embeddings. The paper is structured as follows: Section 2 provides a literature review and discusses related works. Section 3 outlines the proposed method, focusing on architecture design and component implementation. Subsequently, Section 4 presents the experiments and analysis conducted on the proposed approach. Section 5 delves into a discussion of the results. Finally, Section 6 concludes by summarizing our findings and discussing future works.

2 Literature review

Ontologies play a central role in information retrieval and knowledge representation, acting as structured frameworks that capture domain-specific concepts, relationships, and semantics [10]. These frameworks provide a formalized way to organize information, which, when integrated into QA systems, enhances the accuracy of responses by ensuring consistency and efficient reasoning. For instance, [11] presented an automated question-answering system that operates within a domain ontology focused on natural language processing courses. This system uses the Jena framework to translate user query intents into core elements of the ontology, forming tailored SPARQL queries that enable the efficient retrieval of accurate answers. Similarly, [12] introduced an e-learning bot that leverages an ontology-based knowledge base to provide material content in response to student queries. This knowledge base comprises “users” and “learning objects”, including content items, practice items, and assessment items, achieving a 71% accuracy rate in offering relevant suggestions. While these studies have shown promise in incorporating ontologies into QA systems, they are limited by their reliance on basic NLP techniques that do not fully capture the context and semantic depth of user queries. As a result, they run the risk of generating responses that, while technically correct, may not fully capture the nuanced meanings or contextual subtleties intended by users.

Artificial intelligence (AI) techniques, including machine learning and deep learning algorithms, have revolutionized information retrieval by enabling QA systems to effectively understand and generate human language [13]. By encoding words and phrases in a way that captures their contextual meaning, these models can recognize subtle linguistic cues and infer implicit context. For example, [14] proposed a hybrid chatbot that merges educational course material with everyday conversation, using GloVe and QANet models trained on diverse datasets such as Chinese Wikipedia and Ministry of Education textbooks. While GloVe converts user queries into word vectors, QANet processes these queries to extract relevant context for generating customized responses. Transformers models, on the other hand, have significantly enhanced traditional word embeddings such as GloVe, offering dynamic, context-sensitive representations that capture the subtleties of language, extending their influence beyond natural language processing to domains such as computer vision [15], audio analysis [16], speech processing [17], and the Internet of Things [18]. In a notable study [19], researchers developed a QA system specifically for Bangla reading comprehension, addressing the educational needs of Bangladeshi students. The study focused on fine-tuning transformer-based models, with particular emphasis on BERT and ELECTRA. These advanced models outperformed traditional architectures, with BERT achieving an impressive 87.78% accuracy in testing and 99% in training, demonstrating its effectiveness in the educational context.

The integration of ontologies with AI, particularly transformer models, represents a significant leap in educational NLP systems. These models bring a nuanced understanding of language context, which, when combined with the structured knowledge of ontologies, allow for a more sophisticated interpretation of educational content. For instance, [20] proposes a model for learning embeddings in educational knowledge graphs that combines TransE, a translation-based technique for structural embeddings of entities and relations, and a pre-trained BERT for encoding literal information associated with entities and relations. These embeddings are then combined, using three trained Gated Recurrent Units (GRU), a type of recurrent neural network architecture. Experimental results demonstrate the effectiveness of the model in processing pedagogical knowledge graphs, outperforming other baselines. A sophisticated article recommendation system is proposed in [21] using a knowledge graph enriched with key article data such as titles, publication years, etc. The system uses BERT to provide contextual semantic analysis within the knowledge graph. RippleNet, an end-to-end model, is then applied to navigate the graph, capturing user preferences and forming a targeted set of pre-recommendation nodes. A prediction layer then ranks these nodes and provides a personalized top-N paper recommendation list. Experimental evaluations show a significant increase in accuracy over baseline methods.

Merging transformer models with ontologies represents a novel approach in QA systems, showing promise in fields like medicine [22]. Yet, its application in education is still in its infancy, with few studies [23] [24] providing preliminary insights through limited testing. Our work aims to bridge this gap by employing a sentence transformer model and a domain ontology to refine question analysis and answer generation in educational QA systems. This fusion of cutting-edge transformer technology with structured ontologies promises to enhance the accuracy and responsiveness of tutoring frameworks, marking a significant advancement in educational technology.

3 The proposed approach

In this section, we describe the proposed QA system, called DETuto (Disciplinary E-Tutoring System), which uses an extended domain ontology ONTO-TDM2 to improve knowledge representation. It includes a fine-tuned sentence transformer model that converts queries and ontology entities into vectors to facilitate semantic similarity computations.

We begin by constructing a knowledge graph, denoted as $G = \{(h, r, t)\} \subseteq E \times R \times E$ which is instantiated from the ONTO-TDM2 ontology and expressed as RDF triplets $t = (h, r, t)$. Each triplet consists of a head h and a tail t from set E, along with a relation r from set R. Collectively, E and R represent all entities and relationships within G. Subsequently, we fine-tune a pre-trained sentence transformer model using a dataset of question-answer pairs, $D = \{(q_i, a_i) \mid q_i \in Q, a_i \in A\}$, where Q represent questions and A their corresponding answers, focusing on the algorithms and data structures discipline. The resulting model, M', is then employed to generate sentence embeddings. These embeddings are dense vectors that encapsulate the semantic information of both the learner's query and the entities within the knowledge graph. We denote the embedding of the learner's query as $Q_{vec} = M'(Q)$ and the embeddings of the knowledge graph entities as E_i. To assess contextual similarity, we evaluate the similarity between the generated embeddings, represented as $S(Q_{vec}, E_i)$. This comparison helps determine if the query aligns with the context encapsulated by the knowledge graph. Upon detecting a context match, the system proceeds to search for answers within pertinent tutoring question-answer pairs. This search involves comparing the query's embedding Q_{vec} with different question embeddings within the knowledge graph, denoted as QA_{vecj}, where j indexes these questions. The answer from the most similar question is then returned. Finally, if the learner is not satisfied with the response, the query is directed to a domain expert for a more detailed answer, who then updates the ontology accordingly.

Figure 1 provides a visual representation of the DETuto architecture illustrating the system's components and their interactions. The system preprocesses a domain-specific QA dataset to fine-tune a sentence transformer model, which generates embeddings for learner queries and ontology entities. It performs a two-step similarity analysis, first matching the query with ontology vectors and then comparing it to stored QA pairs to retrieve the most relevant answer. If unsatisfied, the learner's query is escalated to a domain expert, who updates the ontology, enhancing the system over time.

The detailed description of each module and its functionality is provided in the following sections after an overview of the ONTO-TDM2 domain ontology.

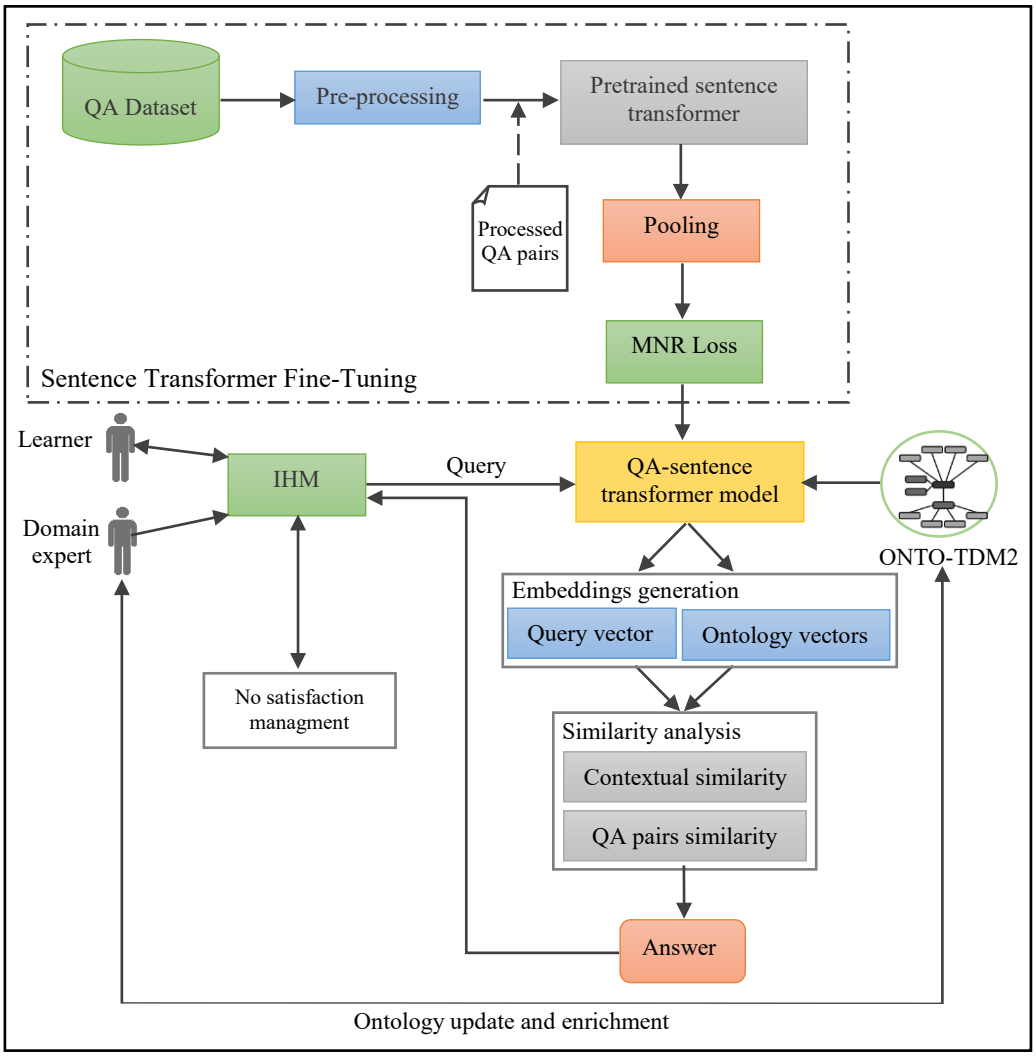


Figure 1. The proposed system process.

3.1 The domain ontology ONTO-TDM2

Ontologies serve as organized frameworks for structuring and presenting knowledge within specific domains. They function as conceptual models, delineating relationships among entities and concepts in a specific area. Fundamental components within ontologies encompass classes (representing concepts), properties (defining relationships between classes), and instances (individual entities falling under classes). In this study, we leverage the ONTO-TDM domain ontology [25] [26] that proposes metadata for different characterizations of a discipline (notion, knowledge item, exercise, error, semantic links, etc.) and can be used by various learning systems. Figure 2 presents the hierarchical structure of the ONTO-TDM ontology, visualized using the OntoGraf feature of the Protégé[1] tool.

[1] https://protege.stanford.edu/

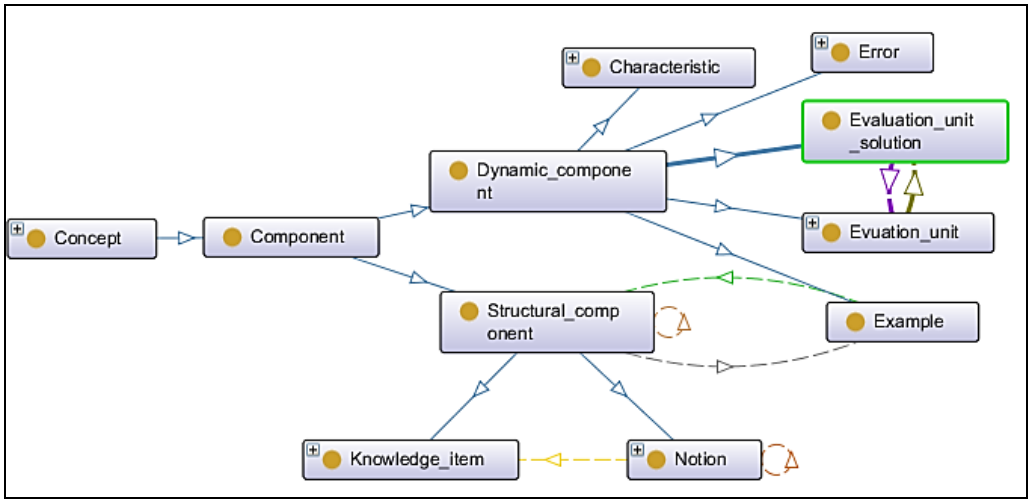


Figure 2. ONTO-TDM Ontology Hierarchy Visualization using Protégé's OntoGraph tool.

A significant aspect of our research involves extending this ontology to ONTO-TDM2 by introducing new concepts and associated relations to align with the requirements of our proposed disciplinary e-tutoring system, known as DETuto. The additional concepts and their links are mentioned below:

- *Tutoring question*: encapsulates targeted inquiries designed for educational purposes, addressing topics within a specific discipline.
- *Tutoring answer*: is a conceptual entity that serves as the response to a specific "Tutoring Question". It encapsulates the information or explanation provided to address the inquiry posed in the question.
- *Has-answer*: serves as a link between a "Tutoring Question" and its corresponding "Tutoring Answer". This relationship denotes that a specific question is associated with a particular answer within the ontology.
- *Has-question*: the inverse of "has-answer", it establishes a connection between a "Tutoring Answer" and its associated "Tutoring Question".
- *Connected-to*: links the concept "Notion" to its respective "tutoring questions" enabling the representation of distinct topics within a discipline. For instance, the notion "Linked List" may include tutoring questions like "What is a linked list?" As specificity increases, the ontology introduces "Knowledge Items", related to a "Notion" such as "Linked List Implementation," each featuring targeted questions like "How to implement a linked list?"

Furthermore, in addition to structuring and representing knowledge within specific domains, the ontology plays a critical role in organizing question-answer pairs within a structured framework. For example, questions about the definition, benefits, or limitations of a concept such as "array" would be directly linked to the corresponding notion "array" within the ontology. This direct linkage ensures that questions about specific aspects of a concept are seamlessly linked to the relevant concept within the ontology, thereby enhancing the clarity and accessibility of domain-specific knowledge within the disciplinary e-tutoring system. Moreover, the ontology further improves the organization of question-answer pairs by associating detailed questions with specific knowledge items related to the concept. For example, questions about detailed operations or functionalities, such as "How do I delete an element within an array?" or "How do I sort an array?" are linked to corresponding knowledge items such as "deletion" and "sorting" within the notion "array". Similarly, comparative questions, such as comparing two different types of data structures, are linked to broader concepts of data structure within the ontology, ensuring a structured representation of domain-specific knowledge across different levels of granularity. In addition, when comparing two linear data structures, the comparison question would be linked to the notion of linear data structure, figure 3 illustrates the relationships between questions, answers, and various concepts within the ontology. This systematic linking of questions to relevant concepts and knowledge items within the ontology enables a structured representation of question-answer pairs, thereby improving the organization and accessibility of domain-specific knowledge within the disciplinary e-tutoring system.

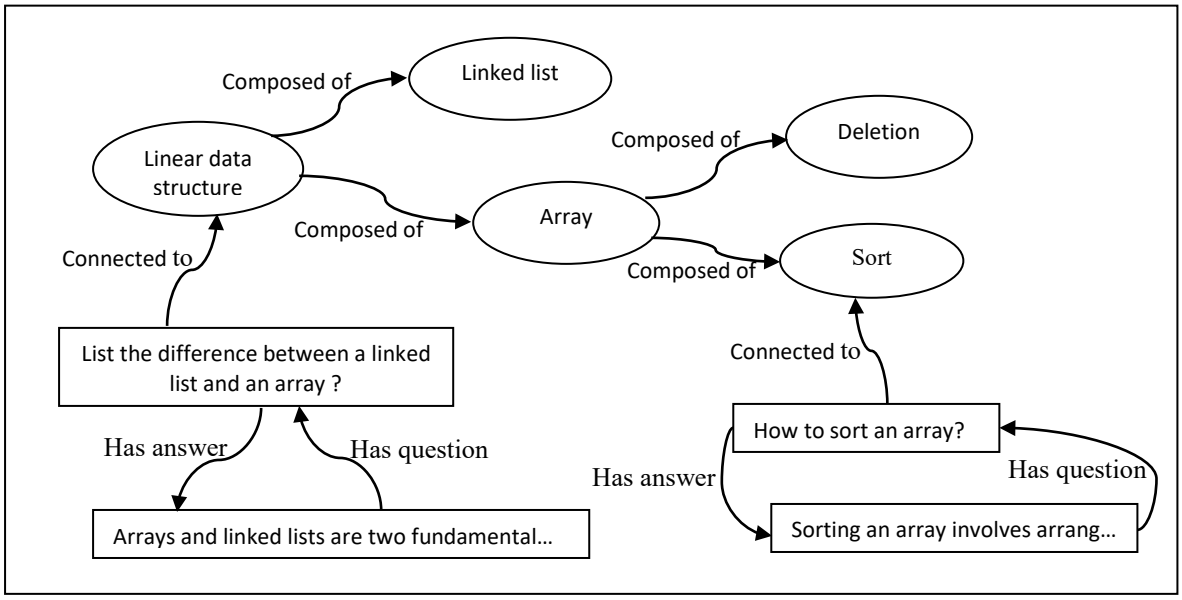


Figure 3. Extract of the relationships between questions, answers and key concepts within the knowledge graph of the 'Algorithms and Data Structures' domain.

3.2 Fine-tuning process of the sentence transformer

Fine-tuning is crucial for optimizing a pre-trained model, such as sentence transformers, for a specific task. When applied to a question-answering system focused on specific disciplines, it is essential to align the model's understanding to their complexity. This involves refining the model's grasp of technical vocabulary, enabling it to navigate the specific context of queries and respond adeptly to nuanced questions.

We selected the "all-MiniLM-L12-v2"[2] sentence transformer for our experiment, a BERT-like pre-trained model. With a maximum sequence length of 256 tokens, this lightweight model, compact at 120 MB, excels in both efficiency and delivering high-quality semantic similarity. It has demonstrated promising results across a range of natural language processing tasks, including semantic similarity [27] and clustering [28].

The fine-tuning process uses the Multiple Negative Ranking (MNR) loss function. This loss function plays a crucial role in improving the learning process and has been proven effective in various fields such as information retrieval, natural language processing, and recommender systems. It is designed to focus specifically on positive pairs, each of which consists of a question q_i and the corresponding answer a_i in the training dataset. In each batch, it randomly samples N-1 negative answers a_j (where $i \neq j$). The main goal of training is to maximize the cosine similarity between a question and its positive answers $S\ (q_i, a_i)$ while minimizing the similarity between a question and the negative answer $S\ (q_i, a_j)$.

Where N is the number of negative samples per positive pair. Upon fine-tuning with the MNR loss function, the model is prepared for deployment. It can efficiently generate dense vector representations for textual data, encapsulating semantic representations for downstream tasks such as semantic search. This paves the way for more accurate and efficient natural language processing applications.

3.3 Embedding generation

In this section, we introduce the process of embedding generation for both the ontology and the query. Ontology Vectorization involves transforming each entity and relation into numerical embeddings to enhance semantic understanding. Additionally, Query Vectorization outlines the methodology for acquiring a dense vector representation of the query, capturing its context and semantic essence through token embeddings.

3.3.1 Ontology vectorization

To streamline computational operations and enhance semantic understanding, we employed our fine-tuned model for knowledge graph embeddings. This process involves converting each triple *(h, r, t)* into numerical

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

embeddings, where $h, r, t \in E_i$ are the embeddings of words and sentences representing entities in G. These embeddings are generated by inputting the textual representations and labels of entities and relationships into the corresponding fine-tuned model $E_i = M'(h, r, t)$. By embedding the ontology entities, we establish a high-dimensional representation space where semantic similarities and relationships among the entities can be effectively analyzed.

3.3.2 Query vectorization

To obtain a dense vector representation of the query, we first pre-process the text by lowercasing and tokenizing it. The tokens are denoted $T_1, T_2 \ldots T_n$, where n represents the number of tokens in the query, and together they form the tokenized query Qt. The fine-tuned model then processes each token independently and computes embeddings $Qte = M'\{T_{e1}, T_{e2} \ldots T_{en}\}$. Qte represents the embeddings for each token in the query, denoted as $\{T_{e1}, T_{e2} \ldots T_{en}\}$, where each T_{ei} corresponds to the embedding of the i^{th} token. Through this process, the model captures the semantic and contextual nuances inherent to each token in the query. Finally, the fine-tuned model aggregates the embeddings for the tokenized query, represented by Qte, into one vector, Q_{vec}, which represents the entire sentence query. This aggregation is achieved through mean pooling, where semantic information from each token is integrated to form a comprehensive representation of the entire query. The pooling is defined in equation 1:

$$Q_{vec} = \frac{1}{N} \sum_{i=1}^{N} Qte_i \tag{1}$$

3.4 Similarity analysis

The last module of the proposed system architecture is the identification of responses based on learners' queries and the knowledge graph through contextual similarity and QA pairs similarity. This module will use the generated embeddings and cosine similarity to generate the most appropriate response to the learner's query.

3.4.1 Contextual similarity

Contextual similarity aims to understand the context of the query and its relevance within the domain covered by the ontology. By analyzing contextual information embedded in the query, the system evaluates its alignment with the ontology's knowledge domain, effectively distinguishing between queries that require further processing and those that are unrelated to the domain. In this process, the vectorized form of the query represented as Q_{vec}, is compared with the embedded representation E_i of each entity in the ontology using cosine similarity. When the contextual similarity $S(Q_{vec}, E_i)$ exceeds a predefined threshold (set to 0,8), it indicates a substantial alignment with the domain of the ontology.

3.4.2 QA pairs similarity

If the query is in the context of the knowledge graph, the system explores the similarity between the query Q_{vec} and the questions of the QA pairs QA_{vecj} within the knowledge graph. Equation 2 defines this process:

$$Q_{best} = \begin{cases} \arg\max_{Q_j \in G} S(Q_{vec}, QA_{vecj}) & If \quad S(Q_{vec}, E_i) \rangle \; Threshold \\ Out\text{-}of\text{-}context\ message & Else \end{cases} \tag{2}$$

Where Q_{best} is the question in the knowledge graph G with the highest similarity score to the query. However, if the query is unrelated to the context, the system returns an out-of-context message.

Furthermore, if the generated answers do not meet the user's expectations or requirements, the system takes a flexible approach to address the query effectively. This may include referring the query to a domain expert for a more nuanced response tailored to the user's needs. By leveraging the expertise of domain specialists, the system strives to provide comprehensive and satisfying answers to user queries, thereby improving the overall user experience.

4 Experimental analysis

4.1 Dataset characteristics

In this study, we developed a dataset focused on algorithms and data structures, that we have named AlgDataQA, by integrating online scraping techniques—using tools such as Beautiful Soup and Scrapy—with manual data collection methods. We chose this domain because algorithms and data structures are widely studied globally, yet lack a comprehensive and valid dataset. This area presents significant challenges for learners, making it essential to create a dedicated dataset to improve QA systems and facilitate education in this discipline. The dataset includes 10,053 question-answer pairs gathered from various sources, including forums, FAQs[3], and specialized websites[4], where qualified mentors and domain experts actively contribute valuable knowledge. Additionally, experts within our university faculty have carefully examined and validated our corpus, ensuring its accuracy and reliability.

The AlgDataQA dataset offers a varied set of theoretical question-answer pairs, carefully chosen to delve into the essential elements of algorithmic problem solving and data structures. The questions cover various aspects of algorithms and data structures. Some focus on fundamental principles, like the efficiency of "sorting algorithms" or the systematic process of navigating through elements in "linked lists". Others touch on implementation, guiding through the creation and management of data structures. Comparative questions may explore the strengths and weaknesses of using a "binary search tree" versus a "hash table" for a specific application. The dataset also includes historical and evolution perspectives offering a comprehensive exploration of algorithmic domains.

Table 1 provides a summary of essential characteristics extracted from our dataset, shedding light on the linguistic nuances within questions and answers. The metrics encompass the average length of questions and answers, the vocabulary size, and the maximum length of responses, reaching up to 254 words. This comprehensive overview encapsulates the varied textual features embedded in our dataset.

Table 1: Characteristics of the AlgDataQA dataset

Characteristics of AlgDataQA	Value
Number of QA pairs	10,053
Vocabulary size	11,809
Average length answer	61.35
Average length question	12.05
Max length answer	254

The AlgDataQA dataset underwent a series of initial pre-processing steps to ensure consistency and quality in the text data. These steps included:

- Lowercasing: All text was converted to lowercase to maintain uniformity and avoid discrepancies due to case variations.
- Removal of Uninterpretable Characters: Unwanted characters, such as newline symbols and excessive whitespace, were removed. This step helps in cleaning the text and preventing potential issues during subsequent processing stages.
- Tokenization: For tokenization, we employed the pre-trained tokenizer associated with the sentence transformer model. Tokenization is the process of splitting text into smaller units, such as words or subwords, which are then used for further analysis or model training. Using the pre-trained tokenizer ensures that the text is processed in a manner consistent with the model's requirements and optimizes the representation of the data for model fine-tuning and evaluation.

These pre-processing steps were crucial in preparing the dataset for effective use in model training and evaluation, ensuring that the text data is clean, standardized, and compatible with the model's input requirements.

[3] https://www.quora.com/
[4] https://www.geeksforgeeks.org/

After this pre-processing phase, we divided our dataset into two segments: the training set, which comprises 80% of the dataset that is used for fine-tuning the model, and the evaluation set, which comprises 20% of the dataset, is reserved for evaluating the performance of the model.

4.2 Fine-tuning hyper-parameters

We fine-tuned the pre-trained sentence transformer model "all-MiniLM-L12-v2", which is designed to generate dense vector representations of sentences. This compact variant of the MiniLM series is optimized for both efficiency and performance making it well-suited for tasks such as semantic search, clustering, and sentence similarity. Our fine-tuning process utilized a training dataset of 8043 question-answer pairs. During training, we employed a batch size of eight (8), ran the model for nine (9) epochs, and used the Multiple Negatives Ranking (MNR) loss function. Additionally, a learning rate warm-up over 1000 steps was applied to gradually adjust the learning rate, with a maximum sequence length of 256 tokens. The experiments were conducted on a single NVIDIA Tesla T4 GPU using the Google Collaboratory server[5].

4.3 Performance evaluation and comparison

4.3.1 Fine-tuning evaluation

In our study, we first compared the performance of our fine-tuned model over the original to predict the semantic similarity between the questions and their answers embeddings within the QA pairs of our test dataset. This comparison aims to demonstrate the effectiveness of the fine-tuned model to capture complex nuances within a specific context.

To evaluate the semantic similarity, we used several metrics:

- The average cosine similarity reflects the typical alignment between pairs of vectors.
- The median cosine similarity indicates the central tendency of similarity within the dataset.
- The standard deviation of the cosine similarity measures the variability of the similarity scores around the mean.

Table 2 illustrates the performance enhancements achieved with the fine-tuned model compared to the original model, highlighting a significant improvement in the performance of the fine-tuned model. Specifically, the average cosine similarity jumped from 0.77 in the original model to 0.82 in the fine-tuned model, and the median cosine similarity improved from 0.80 to 0.84. In addition, the standard deviation of the cosine similarity decreased from 0.12 to 0.09, indicating a reduction in variability and greater consistency in the predictions of the fine-tuned model.

Table 2: Performance comparison between the Original Model and the Fine-Tuned Model across Various Cosine Similarity Metrics

Metrics	Original model	Fine-tuned model
Average cosine similarity	0.77	0.82
Median cosine similarity	0.80	0.84
Standard Deviation of Cosine Similarity	0.12	0.09

We also perform an independent student's t-test to statistically validate the significance of these improvements. The t-test is a widely accepted statistical method for comparing the means of two independent groups and determining the statistical significance of the observed differences. It assumes that the data follow a normal distribution and that both groups have similar variances. The test calculates a t-value using the means of the samples. An α -level, often set at 0.05, serves as the threshold for deciding whether the observed differences are not due to random chance. If the p-value, which represents the probability of observing the calculated t-value, is less than α, it indicates a significant difference between the groups means [29].

[5] https://colab.research.google.com/

In our analysis, we set the α–level at 0.01, which indicates a 1% risk of falsely concluding a significant difference between the models when none exists. The results of our t-test ($p = 2.5e\text{-}28$, $p < 0.01$), indicate a significant difference between the two models, providing robust evidence of the superior efficacy of our fine-tuned model over the original.

4.3.2 Comparison with our previous works

To evaluate the effectiveness of the distributed sentence representation using the fine-tuned sentence transformer model in improving the proposed system, we have implemented an evaluation phase using our domain ontology. We created five (05) reformulations for each 150 QA pairs contained in our ontology, for a total of 750 questions, to serve as an evaluation set, ensuring that our fine-tuned model was evaluated on unseen data and showed its ability to process and understand variations in natural language. This approach mirrors real-world usage where users may phrase their questions in diverse ways. The fine-tuned model's responses to these reformulated queries were then compared to the expected answers from the ontology.

For instance, consider the question "What is a binary tree?" The following reformulations exemplify the diverse ways in which the same concept can be articulated:

- Could you define a binary tree?
- What does binary tree mean when we are talking about data structures?
- In computer science terms, what is the definition of a binary tree?
- Could you explain what a binary tree is in the realm of data structures?
- Can you provide an overview of the concept of binary trees and their significance in data structure?

The evaluation set covers a wide range of topics in algorithms and data structures, including questions about fundamental concepts like sorting algorithms, searching algorithms, and various data structures such as arrays, binary trees, etc. Additionally, it encompasses different types of questions within this domain, including definitions, advantages, real-world applications, and operations such as implementation, deletion, and addition. Despite the size of the evaluation data, the breadth and depth of the questions ensure a comprehensive evaluation of the system's capabilities across a diverse range of algorithms and data structure concepts and scenarios.

We performed a comparative analysis between the fine-tuned sentence transformer model and methods used in previous research [8] [9], namely TF-IDF and Word2Vec to build the initial tutoring QA system, using Accuracy, Precision, Recall, and F1-score measures.

- **Accuracy**: assesses the overall proportion of correctly answered questions out of the total number of questions. The Total Number of Questions represents the entire set of QA pairs used in the evaluation. Accuracy is computed using the following formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \tag{3}$$

- **Precision**: measures the proportion of correct answers among all the answers produced by the system.

$$\text{Precision} = \frac{\text{Number of Correct Answers}}{\text{Number of Correct Answers} + \text{Number of Incorrect Answers}} \tag{4}$$

The Number of Correct Answers refers to those answers provided by the system that match the correct answers from the evaluation set, while the Number of Incorrect Answers denotes the answers that do not align with the expected correct answers.

- **Recall:** evaluates the proportion of correct answers identified by the system out of all possible correct answers.

$$\text{Recall} = \frac{\text{Number of Correct Answers}}{\text{Number of Correct Answers} + \text{Number of Missed Answers}} \tag{5}$$

Where the Number of Missed Answers denotes the correct answers from the evaluation set that the system failed to provide.

- **F1-score:** provides a balanced measure of Precision and Recall by computing their harmonic mean. It is computed as:

$$F1-score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Figure 4 illustrates the comparative performance of the proposed fine-tuned model against TF-IDF and Word2Vec methods. The proposed system, leveraging the fine-tuned model, demonstrates promising results based on preliminary assessments. Additionally, it is worth noting that the system efficiently handles questions related to algorithms and data structures, further emphasizing its relevance and utility in this domain.

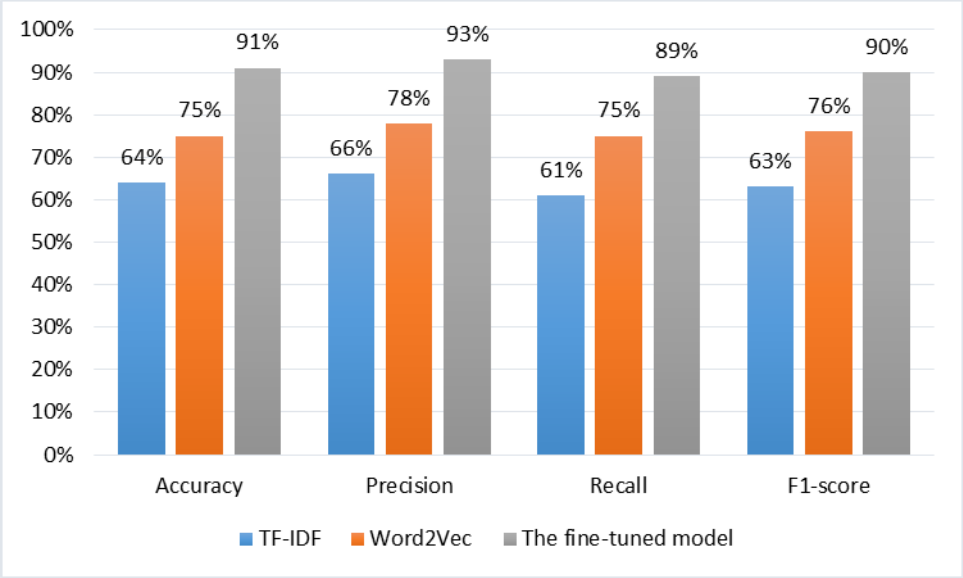


Figure 4. Comparison of the TF-IDF, Word2Vec, and the Fine-tuned model on our proposed system.

4.3.3. Comparison with existing approaches

We conducted a comparative analysis of our proposed method, which integrates sentence transformer technology with a domain ontology, against established question-answering systems used in educational and tutoring contexts. Our evaluation utilized a test dataset derived from the ontology, specifically designed with QA pairs, as outlined in section 4.3.2.

In our comparative analysis, we evaluated a range of methodologies, including word embedding techniques such as Word2Vec and GloVe, as detailed in [14]. Specifically, we trained a Word2Vec model using the same dataset that was employed for fine-tuning the sentence transformer. The dataset underwent comprehensive pre-processing, which included word tokenization and stop words removal. The Word2Vec model was configured with the following parameters: a vector size of 100, a context window of 5, and a minimum word frequency of 1. Following the training of the Word2Vec model, we utilized it to generate word vector representations for the QA pairs within the test set. These representations were then used to match queries to the most relevant answers based on cosine similarity scores.

Additionally, we examined the use of transformer-based models as detailed in [19]. For this purpose, we employed a pre-trained Bert model specifically designed for question answering tasks. In this framework, the answers from the QA pairs in the dataset are treated as context for their corresponding questions. Following tokenization, BERT utilizes its transformer architecture to encode both the question and the context into contextual word embeddings. It then generates a response by determining the best match between the query and the available contexts. Consequently, the model can identify the most relevant answer based on the correspondence probabilities.

Performance was evaluated using a range of metrics, including accuracy, precision, recall, and F1-score, as detailed in Section 4.3.2. This comprehensive assessment ensured a consistent and thorough comparison across all methodologies.

Table 3 and figure 5 present a comparative analysis of existing methods versus our proposed approach. It summarizes key performance metrics, including accuracy, precision, recall, and F1-score, for each method.

Table 3: Comparative analysis: existing methods vs. our proposed approach

Methods	Accuracy	Precision	Recall	F1-score
Word2vec model	47%	49%	45%	47%
Transformer model	56%	57%	53%	55%
Sentence transformer + Domain ontology	91%	93%	89%	90%

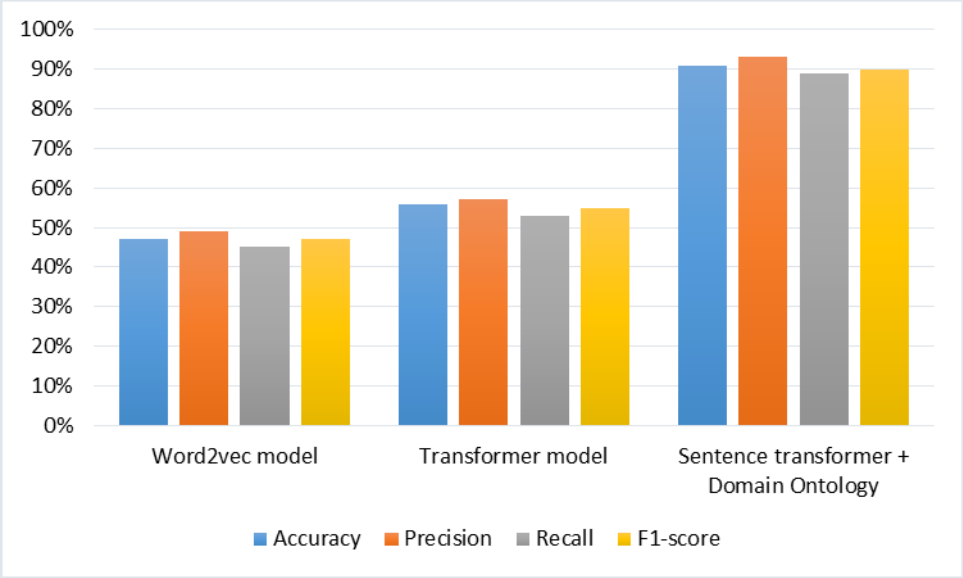


Figure 5: Visual Comparison: Existing Methods vs. Our Proposed Approach

Table 3 and Figure 5 provide a comparative analysis of various methods based on key metrics: Accuracy, Precision, Recall, and F1-score. The data show that the Sentence Transformer combined with Domain Ontology delivered the highest performance, achieving an accuracy of 91%, Precision of 93%, Recall of 89%, and an F1-score of 90%. This demonstrates the exceptional effectiveness of our proposed approach relative to the existing methods.

4.3.4. Impact of ontology integration with sentence transformer in QA system

To evaluate the effect of integrating ontology with sentence transformers, we compared the performance of the sentence transformer model alone against the model enhanced with the ontology. Using a consistent evaluation dataset, we measured accuracy, precision, recall, and F1-score to assess how the addition of the ontology influences the model's ability to produce accurate and contextually relevant responses. Table 4 shows the difference between using the sentence transformer alone compared to using the model combined with the domain ontology.

Table 4: comparison between using the sentence transformer alone and using the model combined with the domain ontology

Method	Accuracy	Precision	Recall	F1-score
Sentence transformer alone	86%	88%	85%	86%
Sentence transformer + domain ontology	91%	93%	89%	90%

The results demonstrate that integrating the Sentence Transformer with domain ontology significantly enhances performance across all metrics compared to using the model alone. This improvement is attributed to the direct association of each question with specific concepts and knowledge items within the ontology, which has substantially increased the accuracy of our QA system. By linking questions to their corresponding concepts and detailed knowledge, such as delete operations within targeted data structures, the system benefits from a comprehensive framework for understanding and responding to queries. This approach not only allows for more precise contextualization of queries but also ensures that generated answers are highly relevant to the intended topic or task. Furthermore, mapping questions to their corresponding answers within the ontology enhances the system's capability to provide accurate and contextually appropriate responses, thereby maximizing its interpretive effectiveness. Overall, this meticulous mapping process underscores the robustness of our approach, demonstrating a marked improvement in QA performance.

5 Results discussion

Our research has shown that the fine-tuned model significantly outperforms the original in terms of algorithms and data structures. The improved similarity scores between the embeddings of questions and their respective answers, indicating a deeper conceptual understanding, evidence this. For example, QA pairs related to specific notions or concepts such as “Matrix”, “Array”, and “Queue” have shown more accurate embeddings. Table 5 and Figure 6 present a comparison of similarity scores between question-answer pairs for the original and fine-tuned models. The data reveal significant improvements in semantic alignment for key QA pairs, with similarity scores increasing notably: for “matrix” from 0.53 to 0.71, for “tree” from 0.69 to 0.82, and for “graph” from 0.68 to 0.80. These enhancements highlight the refined model's enhanced capability to capture the complex nuances and subtleties of the domain, demonstrating its superior performance in matching and understanding domain-specific QA pairs.

Table 5: Comparison of Similarity Scores of QA pairs between Original and Fine-Tuned Models

QA pair	Original model similarity	Fine-tuned model similarity
How to multiply two matrices?	0.53	0.71
What is a tree data structure?	0.69	0.82
What are the different traversal types of a graph?	0.68	0.80

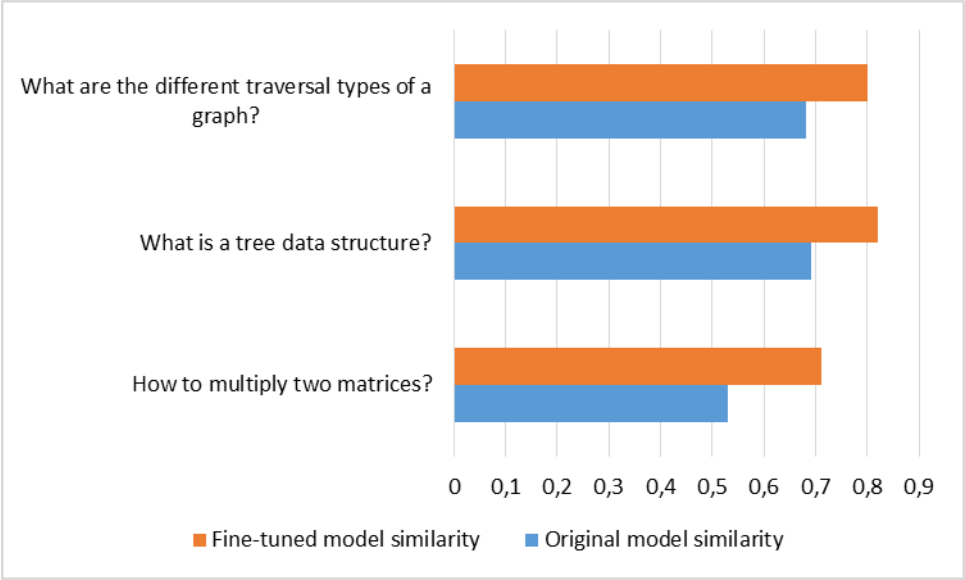


Figure 6. Visual Comparison of QA Pair Similarity Scores: Original Model vs. Fine-Tuned Model

Our proposed approach represents a significant advancement over existing methods, including word embeddings and transformer models, for question-answering tasks, particularly within the field of education.

Word embeddings like Word2Vec have several limitations. Their static nature results in the same vector representation for a word regardless of context, which is critical in QA tasks. The model cannot also capture sentence-level semantics, reducing its effectiveness in understanding complex queries and answers. Furthermore, Word2Vec is less effective in specialized domains, such as algorithms and data structures, as it fails to inherently capture domain-specific nuances and specialized knowledge, leading to decreased accuracy in these contexts.

Similarly, while BERT produces dynamic contextual embeddings, it is less effective at generating high-quality sentence-level representations, which can hinder its ability to fully capture the meaning of longer texts or complex queries. Additionally, BERT's substantial computational requirements pose challenges for scaling with large datasets or real-time applications. In specialized domains, BERT often requires significant fine-tuning to achieve optimal performance, as its pre-training on general data may not adequately capture the specific nuances of domain-specific language compared to more specialized models.

Furthermore, despite the previous works employing domain ontology with methods such as TF-IDF, which is constrained by its inability to capture semantic relationships and context beyond term frequency, and Word2Vec with aggregation, which often struggles to accurately represent complex relationships and domain-specific nuances, these methods have inherent limitations that impact their performance.

In contrast, the fine-tuned sentence transformer excels in capturing language nuances, as demonstrated by its markedly higher precision, recall, and F1-score in the algorithmic and data structure domain. This model's superior performance is attributed to its ability to generate highly accurate and contextually relevant responses. When combined with a domain-specific ontology, the sentence transformer's capabilities are further enhanced. The ontology provides a structured framework that enriches the model's understanding of domain-specific concepts, leading to improved accuracy and contextual relevance in question-answering tasks. This integrated approach not only surpasses existing methods but also represents a significant advancement over our previous works, offering a more precise and effective tool for handling complex queries and generating accurate answers.

While our results are promising, it is worth noting that for certain specific concepts within this domain, the similarity scores were not as high. This can be attributed to the limited amount of data available in the fine-tuned dataset, particularly for notions such as "Weak AVL tree". It should be noted that the experiment only involved one specific model and did not explore various sets of hyper-parameters. This suggests that further optimizations and refinements of hyper-parameters could lead to enhanced model capabilities and improved performance outcomes. Additionally, expanding the evaluation to encompass a larger dataset would enhance the robustness and reliability of the system's performance evaluation. Evaluating on a broader range of queries and scenarios can provide valuable insights into the system's effectiveness across various contexts and usage patterns.

6 Conclusion

In this paper, we have enhanced our disciplinary e-tutoring system by integrating a fine-tuned sentence transformer with domain-specific ontology for question-answering tasks, utilizing the AlgDataQA dataset of 10,000 question-answer pairs focused on algorithms and data structures. Our results show that this approach effectively combines the contextual power of transformer technology with specialized domain knowledge, resulting in significantly improved precision and relevance compared to existing methods.

However, due to limited computational resources, we were unable to explore additional models or conduct in-depth research on the effects of different hyperparameters on performance. Future research should focus on exploring additional models and optimizing hyperparameters to enhance performance further. Expanding computational resources would enable more comprehensive experimentation, including the training of complex models and larger datasets. Investigating how the approach scales with real-time applications and diverse domains could provide valuable insights into its broader applicability. Additionally, incorporating user feedback could offer practical insights into improving the model's effectiveness based on real-world interactions.

Furthermore, disciplinary tutoring systems would be invaluable in schools and universities as student population increase, alleviating the challenge of limited tutor availability. By facilitating tailored responses and reducing the need for constant learner-tutor interaction, it enables students to receive personalized support and guidance, ultimately improving learning outcomes in the face of increasing educational demands.

Acknowledgements

We would like to express our gratitude to the experts whose valuable insights and feedback significantly contributed to the validation and improvement of our results.

References

[1] R. D. Alsaffar. Critical Performance Parameters of Distance Education through e-Learning: A Case Study in Kuwait. *International Journal of Information and Education Technology*, *8* (11), 804-808, 2018. doi: 10.18178/ijiet.2018.8.11.1143

[2] G. Sakkir, S. Dollah, & J. Ahmad. E-learning in covid-19 situation: Students' perception. *EduLine: Journal of Education and Learning Innovation*, *1*(1), 9-15, 2021. doi: 10.35877/454RI.eduline378

[3] M. A. C. Soares, & F. S. Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, *32*(6), 635-646, 2020. doi: 10.1016/j.jksuci.2018.08.005

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,...& I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, *30,* 2017.

[5] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805,* 2018.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D.Chen, ... & V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] N. Reimers, & I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084,* 2019.

[8] R. Abdoune, L. Lazib, & F. Dahmani-Bouarab. Disciplinary e-tutoring based on the domain ontology ONTO-TDM. In *2022 4th International Conference on Computer Science and Technologies in Education (CSTE)* (pp. 143-147). IEEE, 2022, May.

[9] R. Abdoune, L. Lazib, & F. Dahmani-Bouarab. Word Embeddings for a Disciplinary Tutoring System. In *2022 5th International Symposium on Informatics and its Applications (ISIA)* (pp. 1-6). IEEE, 2022, November.

[10] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood & H. M. Abbasi. A survey of ontology learning techniques and applications. *Database, 2018*, bay101, 2018. doi: 10.1093/database/bay101

[11] X. Xie, W. Song, L. Liu, C. Du, & H. Wang. Research and implementation of automatic question answering system based on ontology. In *The 27th Chinese Control and Decision Conference (2015 CCDC)* (pp. 1366-1370). IEEE, 2015, May. doi: 10.1109/CCDC.2015.7162131

[12] F. Clarizia, F. Colace, M.Lombardi, F. Pascale, & D. Santaniello. Chatbot: An education support system for student. In *Cyberspace Safety and Security: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings 10* (pp. 291-302). Springer International Publishing, 2018. doi:10.1007/978-3-030-01689-0_23

[13] K. A. Hambarde, & H. Proenca. Information retrieval: recent advances and beyond. *IEEE Access*, 2023. doi: https://doi.org/10.48550/arXiv.2301.08801

[14] E. H. K. Wu, C. H. Lin, Y. Y. Ou, C. Z. Liu, W. K. Wang, & C. Y. Chao. Advantages and constraints of a hybrid model K-12 E-Learning assistant chatbot. *Ieee Access*, *8*, 77788-77801, 2020. doi:10.1109/ACCESS.2020.2988252

[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, & M. Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, *54*(10s), 1-41, 2022. doi: https://doi.org/10.1145/3505244

[16] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, & G. Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021. doi: https://doi.org/10.21437/Interspeech.2022-227

[17] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, & J. Qadir. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*, 2023. doi: https://doi.org/10.48550/arXiv.2303.11607

[18] M. Wang, N. Yang, & N. Weng. Securing a smart home with a transformer-based iot intrusion detection system. *Electronics, 12*(9), 2100, 2023. doi: https://doi.org/10.3390/electronics12092100

[19] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, & A. S. Ali. Reading comprehension based question answering system in Bangla language with transformer-based learning. *Heliyon*, *8*(10), 2022. doi:10.1016/j.heliyon.2022.e11052

[20] S. Yao, R. Wang, S. Sun, D. Bu, & J. Liu. Joint embedding learning of educational knowledge graphs. *Artificial Intelligence Supported Educational Technologies*, 209-224, 2020. doi: https://doi.org/10.1007/978-3-030-41099-5_12

[21] L. Gao, Y. Lan, Z. Yu, & J. M. Zhu. A personalized paper recommendation method based on knowledge graph and transformer encoder with a self-attention mechanism. *Applied Intelligence*, *53*(24), 29991-30008, 2023. doi:10.1007/s10489-023-05108-z

[22] Q. Guo, S. Cao, & Z. Yi. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, *37*(11), 8548-8564, 2022. doi:10.1002/int.22955

[23] L. S. Nair, & M. K. Shivani. Knowledge graph based question answering system for remote school education. In *2022 International Conference on Connected Systems & Intelligence (CSI)* (pp. 1-5). IEEE, 2022, August. doi: 10.1109/CSI54720.2022.9924128

[24] G. Agrawal, K. Pal, Y. Deng, H. Liu, & Y. C. Chen. CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 21, pp. 23164-23172), 2024, March. doi:10.1609/aaai.v38i21.30362

[25] F. Bouarab-Dahmani, M. Si-Mohammed, C. Comparot, & P. J. Charrel. Learners automated evaluation with the ODALA approach. In *Proceedings of the 2009 ACM symposium on Applied Computing* (pp. 98-103), 2009, March. doi:10.1145/1529282.1529303

[26] F. Bouarab-Dahmani, & N. Hammid. Metadata modelling disciplines for e-learning by doing systems. *International Journal of Metadata, Semantics and Ontologies*, *10*(4), 261-280, 2015. doi:10.1504/IJMSO.2015.074750

[27] C. Galli, N. Donos, & E. Calciolari. Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis. *Information*, *15*(2), 68, 2024. doi: https://doi.org/10.3390/info15020068

[28] L. Wang, J. Chou, D. Rouck, A. Tien, & D. Baumgartner. Adapting Sentence Transformers for the Aviation Domain. In *AIAA SCITECH 2024 Forum* (p. 2702), 2024. Doi: https://doi.org/10.48550/arXiv.2305.09556

[29] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, & G. Pandey. Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, *22*(4), 407-411, 2019. doi: 10.4103/aca.ACA_94_19