

# A Study on Knowledge Discovery from Multilingual Educational Corpora Using Artificial Intelligence Techniques

Ulugbek Hamdamov<sup>1\*</sup>, Bekpolat Umurkulov<sup>2</sup>, Azamat Turdiyev<sup>3</sup>, Farangiz Maxmudova<sup>4</sup>,  
 Nilufar Isakulova<sup>5</sup>, Gulbaxor Rasulova<sup>6</sup> and Abduvali Musayev<sup>7</sup>

<sup>1\*</sup>Professor, Institute of Uzbek Language, Literature and Folklore, Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan

<sup>2</sup>Professor, Department of Uzbek Linguistics, Termez State University, Termez, Uzbekistan

<sup>3</sup>Lecturer, Department of Preschool Education, Faculty of Preschool and Primary Education, Termez State Pedagogical Institute, Termez, Uzbekistan

<sup>4</sup>Associate Professor, Department of Uzbek Linguistics, Karshi State University, Karshi, Uzbekistan

<sup>5</sup>Uzbekistan State World Languages University, Tashkent, Uzbekistan

<sup>6</sup>Department of Pedagogy, University of Tashkent for Applied Sciences, Tashkent, Uzbekistan

<sup>7</sup>Associate Professor, Department of Uzbek Language Teaching Methods, Jizzakh State Pedagogical University, Jizzakh, Uzbekistan

E-mail: <sup>1</sup>ulugbek1968.68@mail.ru, <sup>2</sup>umurkulovb@tersu.uz, <sup>3</sup>azamatbek4575@gmail.com, <sup>4</sup>mfarangiz27@gmail.com,

<sup>5</sup>n.isakulova@uzswlu.uz, <sup>6</sup>nozimararasulova15@gmail.com, <sup>7</sup>a.musayev1962@gmail.com

ORCID: <sup>1</sup><https://orcid.org/0009-0002-7750-5341>, <sup>2</sup><https://orcid.org/0009-0001-8726-787X>,

<sup>3</sup><https://orcid.org/0009-0005-9422-448X>, <sup>4</sup><https://orcid.org/0009-0005-2381-2751>,

<sup>5</sup><https://orcid.org/0009-0009-1716-2384>, <sup>6</sup><https://orcid.org/0009-0000-3952-6040>,

<sup>7</sup><https://orcid.org/0000-0002-1760-0534>

(Received 23 April 2026; Revised 04 June 2026, Accepted 14 July 2026; Available online 29 September 2026)

**Abstract** - Multilingual educational materials include rich information which is hard to analyze due to their multilinguality, computational limitation, and the complexity related to low-resource languages. This research will introduce a knowledge discovery system that exploits artificial intelligence technologies to extract relevant patterns out of multilingual educational texts through natural language processing, educational data mining, topic modeling, and knowledge graphs. This proposed system uses lexical, semantic, and context representation based on term frequency-inverse document frequency, word embeddings, and multilingual transformers respectively to facilitate the analysis of cross-lingual educational materials. This study designs a multilingual processing pipeline including language detection, normalization, tokenization, morphological processing, concept extraction, and knowledge fusion for educational text mining in Uzbek, Russian, and English texts. This system allows automatic discovery of topics, classification of grade levels, aligning of curriculums, and representation of educational knowledge in a structured manner. The hybrid fusion model achieved an accuracy of 97.3% and an F1-score of 96.9%, while the cross-lingual fusion component improved concept matching accuracy to 92.7%. It is proven by the experiments conducted on this model that integration of representation methods improves classification, semantic, and concept consistency. The knowledge graph module allows capturing the relationship between educational concepts, and cross-language fusion helps to minimize inconsistencies in multilingual learning content. The proposed system offers an effective solution for intelligent educational analytics and multilingual curriculums management especially in low-resource languages.

**Keywords:** Multilingual Educational Corpora, Artificial Intelligence, Natural Language Processing, Educational Data

**Mining, Knowledge Graphs, Cross-Lingual Learning, Uzbek Language**

## I. INTRODUCTION

Educational institutions from all over the world produce great quantities of textual materials – textbooks, curricula, exam questions, online lectures, papers of the students – in multiple languages within the same national/regional educational system. Educational materials of multilingual educational system, for instance, Central Asian one, are produced and/or translated into languages such as Uzbek, Russian, and English Madatov et al., (2025a), producing numerous texts that are difficult to process, compare and analyze using the tools designed for monolingual and high-resource settings (Madatov et al., 2025a; Wang et al., 2020; Khudhair et al., 2025). Knowledge discovery – the discovery of new interesting and useful patterns within large databases of information – is a family of computational techniques that is perfectly adapted to perform such kind of analysis, especially when combined with NLP methods dealing with unstructured text (Romero & Ventura, 2020; Abu-Salih & Alotaibi, 2024). Global accounting standards help improve financial disclosure through consistency, comparability, and reliability in financial reporting. Innovations and entrepreneurial ecosystems are among those factors that facilitate the development of enterprises and organizations (Gelman, 2026). The part played by the Total Quality Management approach in enhancing supply chain efficiency and

emphasizing the significance of quality management within the organization (Khudhur, 2022).

The techniques of artificial intelligence for text analysis have been significantly improved in recent years, moving from basic lexicon statistics such as TF-IDF to word distributed representation and more recently to multilingual and cross-lingual contextual transfer models built on transformers (Wu & Dredze, 2020; Hu et al., 2023; Charitopoulos et al., 2020). These improvements allow the possibility to discover topics automatically, classify texts by their subject matter and complexity, extract named entities and create knowledge graphs of the curricula concepts (Yu et al., 2025). Nevertheless, the majority of good-performing contextual models are developed with high-resource languages and show poor performance on low-resource morphologically rich languages such as Uzbek, Kazakh and other Turkic languages (Hu et al., 2023). The digital age is transforming the labour market and affects employment dynamics (Silva & Costa, 2026).

On the other hand, educational data mining (EDM) has seen significant progress in terms of developing methodologies for mining patterns from structured and semi-structured sources of educational data, such as student databases and logs from LMS, as well as recent advancements in text mining of unstructured educational data (Romero & Ventura, 2020; Abu-Salih & Alotaibi, 2024; Xu et al., 2025). Also, educational knowledge graphs have proven the viability of structuring curricular concepts and their relationships to allow personalized trajectory recommendations for learning as well as improving instructional coherence (Yu et al., 2025; Olimov et al., 2026). Nevertheless, to date, there has been little effort to unite these two trends and create a methodology that would be capable of operating across multiple languages and varying resource availability within one educational context. Technological innovation redefines labor markets because of its impact on job creation, labor market structures, and socio-economic growth. The processes of digital transformation and strategic agility enable better management efficiency (Sattorova, 2025).

The goal of this paper is to address this limitation by proposing a knowledge discovery framework from multilingual educational corpora using AI approaches. It includes text representation in the form of lexical, distributed and contextual features and their application in the form of topic modeling, classification, and knowledge graph construction, as well as a cross-lingual fusion module that is aimed at bringing together educational concepts in different languages depending on the number of resources available. The discussion below will particularly focus on the example of Uzbek-language educational text, which will be used in this paper because of recent research showing that grade-level classification in this language is achievable using lightweight, interpretable models (Wang et al., 2020).

The rest of this paper is structured as follows. Section II introduces the existing work on educational data mining, multilingual natural language processing, and educational applications based on knowledge graphs. In Section III, introduce the methodology employed in this paper, which includes multilingual corpus building, preprocessing, text representation, knowledge discovery algorithms, and cross-lingual knowledge fusion. In Section IV, the experimental findings and results are presented, along with the analysis including performance evaluation, knowledge discovery evaluation, and comparative study of the proposed framework. Section V addresses the significance, problems, and limitations of the proposed approach. Finally, Section VI provides the conclusion to the paper and future research directions.

## II. LITERATURE REVIEW

### 2.1. Educational Data Mining and Text Mining

The area of Educational Data Mining can be defined as a scientific discipline which focuses on the application of data mining and machine learning technologies to educational data gathered in the educational setting like student data, tests outcomes and interaction logs (Romero & Ventura, 2020; Abu-Salih & Alotaibi, 2024). One of the most frequently cited surveys in this field was performed by Romero and Ventura and it gives an insight into the evolution of the field from its very first use of classification and clustering approaches to the present-day integration of learning analytics and the authors show that knowledge discovery cycle is a widely used conceptual framework for the domain (Abu-Salih & Alotaibi, 2024). Other systematic reviews of the field are focused on the use of data mining approaches in mathematics and science education and reveal that classification, clustering and text mining are the most commonly applied families of data mining methods and text mining in particular becomes an efficient way to analyze written materials of both students and teachers (Romero & Ventura, 2020; Alfonseca et al., 2007). The decade overview of educational data mining researches shows that this area develops in accordance with the intensive application of artificial intelligence methods as well as indicates the persistent issues of ethical data usage, privacy, and incommensurability of the research findings in different institutional contexts (Davletbayev et al., 2026). The additional critical reviews indicate that educational data mining researches remain focused on higher education, and not enough attention is paid to primary and secondary levels of education (Aytekin & Saygin, 2024; Xu et al., 2025). The separate review of text mining approaches used in educational researches illustrates that increasingly often natural language processing methods are applied to students' texts and feedbacks as well as to educational material, thus, expanding the scope of educational data mining beyond quantitative data (Alfonseca et al., 2007). Moreover, the systematic review of soft computing in educational data mining and learning analytics shows the increasing number

of applications of fuzzy systems, neural networks and hybrid AI technologies in educational data mining from 2010 to 2018 Baazaoui et al., (2025).

## 2.2. Multilingual and Low-Resource Natural Language Processing

The emergence of multilingual transformers like multilingual BERT (mBERT) was a milestone in cross-lingual natural language processing through the development of a single pretrained model that could be used for downstream tasks on more than a hundred languages without any cross-lingual supervision (Wu & Dredze, 2020; Hu et al., 2023). However, further research studies have shown that mBERT is much more efficient in high-resource languages than low-resource languages, and the efficiency gap remains the same across various tasks, namely, named entity recognition, part-of-speech tagging, and dependency parsing (Hu et al., 2023). Some modifications in mBERT to address this challenge include the enlargement of the pretraining vocabulary to incorporate more low-resource languages and additional pretraining on target language data. These changes helped in improving the efficiency of multilingual BERT in terms of the languages which were not covered by the initial vocabulary of the model (Wu & Dredze, 2020). Also, specific modifications to the model in the form of lexicon-based and adversarial training of multilingual BERT help in improving the generalization ability of the model in low-resource languages specifically in the case of zero-shot cross-lingual scenarios (Charitopoulos et al., 2020). Apart from architectural changes, data augmentation strategies in the form of task-specific and universal text transformations have been proved effective for sentiment analysis of Turkish language, which is similar to Uzbek.

Regarding Uzbek language processing, newer resources and baseline models were acquired. The data about multi-label news classification with fifteen labels demonstrated the advantages of recurrent and convolutional neural network over rule-based approach and the best results were obtained with the help of transformer model trained on the Uzbek corpus (Wang et al., 2020). An earlier attempt at the application of machine learning algorithms to multi-class classification of Uzbek news articles revealed the first empirical baselines for the language (Rabbimov & Kobilov, 2020). Specifically interesting in regard to the educational application is the recent research in the field of Uzbek educational text classification which addressed the issue of linear regression, k-nearest neighbor, and cosine similarity retrieval for classification of excerpts from school textbooks according to grades and demonstrated that while the former two methods yielded over ninety percent of accuracy on the internal balanced test set, k-nearest neighbors performed badly because of the curse of dimensionality in the TF-IDF space (Madatov et al., 2025a). Another source of complementary data is suggested in the form of the structured vocabulary dataset collected from primary school textbooks which is supposed to be applied as a linguistic base for future research in Uzbek NLP (Madatov et al., 2025b). From a more general perspective, Turkic and Central Asian languages

surveys also give information on the development of similar methods for Kazakh language, including creation of sentiment analysis corpora and construction of small corpora for the improvement of multilingual NLP performance in low-resource environments. Therefore, it may be concluded that despite the fact that transformers' multilingual models allow for performing classification tasks theoretically, the lightweight lexical approaches are still relevant practical bases for low-resource educational text classification (Madatov et al., 2025a).

## 2.3. Knowledge Graphs and Topic Modeling in Education

The knowledge graphs are more and more widely used in the sphere of education because they can be used to represent curricular entities and their relations in the applications like adaptive learning, content recommendations, and intelligent tutoring (Yu et al., 2025). The analysis of the development and usage of knowledge graphs in the sphere of education found out common methods like automatic entity identification, relation extraction, and knowledge fusion, although it is mentioned that the most of the educational knowledge graphs are developed for single-language usage, and therefore, the issue of multilingual knowledge graphs is less investigated (Yu et al., 2025). The other review regarding knowledge graphs and the issues related to them also provided the educational knowledge graphs in the aspect of artificial intelligence-based knowledge representation and highlighted the fact that the knowledge acquisition and knowledge fusion issues are the common technical difficulties when developing any kind of domain knowledge graphs, including the educational one. One recent study offered the AI-based approach to the development of the educational knowledge graph that could include the prerequisite relations between the nodes due to machine learning scoring and expert verification.

The application of topic modeling, more precisely the Latent Dirichlet Allocation model (LDA), has been extensively discussed in using the method as an unsupervised way of exploring the underlying themes in huge text corpora from educational research to the classification of curriculum contents (Breja et al., 2026). The introductory pieces on applying the LDA to the field of education research pointed out not only the capability of the model in analyzing unstructured data (open-ended surveys, student essays, etc.) but also its constraints in terms of preprocessing and the choice of the number of topics (Breja et al., 2026). Some of the recent pieces trying to apply the LDA in hybrid pipelines along with contextual embeddings like Sentence-BERT have attempted to address some of the constraints of LDA by applying it to shorter texts, which indicates a trend in the development of hybrid lexical-contextual topic modeling. The application of the LDA in the analysis of institutional text corpora such as higher education policy and research literature can serve as a great example of the use of the LDA outside the scope of classroom texts.

#### 2.4. Summary and Positioning of the Present Work

In summary, the above-mentioned literature suggests three quite different but related research approaches: educational data mining methodology that focuses on structured and unstructured educational data Romero & Ventura, (2020) and Abu-Salih & Alotaibi, (2024) and Alfonseca et al., (2007) multilingual and low-resource natural language processing techniques for the performance gap between high-resource and low-resource languages Wu & Dredze, (2020) and Hu et al., (2023) and Charitopoulos et al., (2020) and knowledge graphs and topic modeling techniques for curricular structure representation and discovery (Yu et al., 2025). It is worth noting that only a few of the current research papers seek to combine these techniques to propose an approach particularly focused on multilingual educational corpora using low-resource languages, such as Uzbek. This approach, combining lexical, distributed and contextual techniques, as well as topic modeling, classification, and knowledge graph creation, is discussed below.

### III. MATERIALS AND METHODS

This proposed methodology follows a hierarchical framework beginning with the acquisition of multilingual corpus to the next levels of pre-processing, multi-level text representation, knowledge discovery, cross-lingual integration, and finally the level of educational applications. Every stage of the process will be described thoroughly and after that we'll provide a diagram of the architecture.

#### 3.1. Corpus Acquisition and Composition

This model is based on input corpus that consists of educational texts in different languages existing in the corresponding education system, i.e., in case, it should be educational texts in Uzbek, Russian, and English languages, as well as any other additional sources of information such as digital textbooks, curriculum documents, and learner texts. The experimental corpus consists of 3,500 total documents spanning Uzbek, Russian, and English primary and secondary educational materials. The dataset is balanced across subjects (Mathematics, Science, History) and grade levels (Grades 1–11), comprising 1,500 Uzbek, 1,000 Russian, and 1,000 English text segments to enable comparative cross-lingual analysis.

#### 3.2. Preprocessing

Preprocessing deals with the specific linguistic challenges presented by multi-lingual and highly inflected documents. The stages involved in processing include automatic language identification, which would ensure that the document is sent to an appropriate processing pipeline, character encoding, normalization of case, removal of non-linguistic information like headings, footings, and page numbers, language specific tokenization depending on the orthography of the language, and removal of stop words using a list of stop words specific to the language. In agglutinative languages like Uzbek, morphological

processing is critical due to rich inflectional variations. We implement rule-based stemming combined with lemmatization for Uzbek texts prior to TF-IDF vectorization to constrain feature space dimensionality and mitigate vocabulary sparsity.

#### 3.3. Multilingual Text Representation

Three representation methods are integrated together in order to create a proper balance among interpretability, computability, and semantics. First is lexical representation that makes use of the TF-IDF weighting scheme and creates an interpretable and computable baseline for organizations with low computational power. Second is distributed representation that uses word embedding technique and takes into account not just the lexical similarities but also semantic and syntactic similarities. Nonetheless, this representation is easy to learn despite limited data. Third is contextual representation that makes use of multilingual transformer architecture and has the highest level of semantic and cross-lingual abilities. But this type of representation has extremely high computational and data requirements.

#### 3.4. Knowledge Discovery Techniques

Further from the representation, there exist three other different approaches which operate at the same time. The first one entails topic modeling through Latent Dirichlet Allocation. It is used to reveal structures and themes in instructional texts in order to audit curriculum contents and thematically organize course materials. The third approach entails classification and information extraction through techniques based on linear classifiers to deep learning in order to solve problems such as grade-level classification, subject classification, and named entity recognition of curriculum concepts.

#### 3.5. Cross-Lingual Knowledge Fusion

The role of the dedicated fusion layer is the unification of ideas, concepts, and classifying results that are generated individually in each language through cross-lingual embedding alignment and concept matching heuristics. The rationale of this step lies in the necessity to ensure uniformity in the development of curricula of parallel language courses within the same education system, when the content or concepts explained properly in one language can give an understanding of shortcomings in another.

#### 3.6. System Architecture

The complete architecture system design is shown in fig. 1 below that shows the flow of information processing from corpora of multilingual education to pre-processing, multiple level representation, parallel knowledge discovery, cross language fusion to the output of the application of education like curriculum alignment, learning analytics dashboard and learning path recommendation.

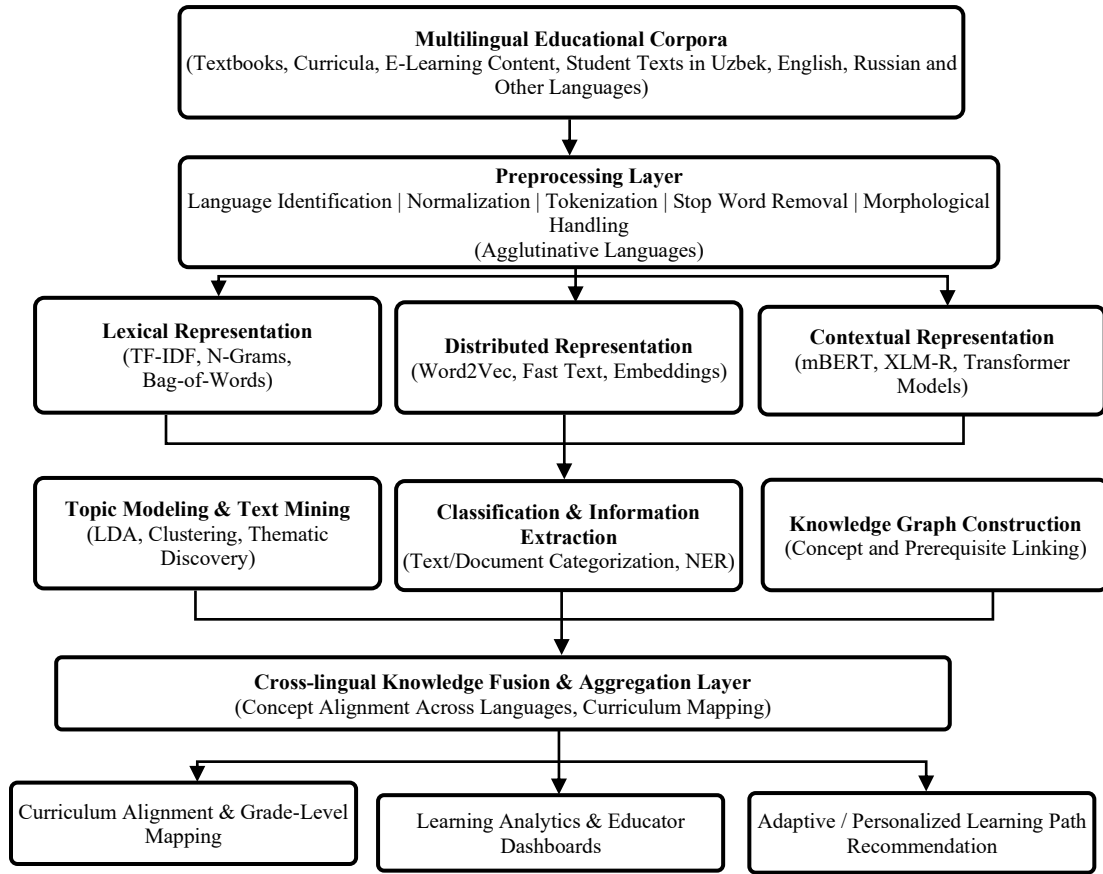


Fig. 1 Layered System Architecture for Multilingual Educational Knowledge Discovery

The classification accuracy evaluating the proportion of correctly categorized educational texts across the multilingual corpora is expressed as equation 1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where  $TP$  represents true positives,  $TN$  represents true negatives,  $FP$  represents false positives, and  $FN$  represents false negatives.

The F1-score, which provides a balanced harmonic mean of precision and recall for classification models, is defined as equation 2:

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

Where precision and recall are derived from the classification results of the educational text pipelines.

The semantic similarity measuring the vector closeness between contextual and distributed embeddings (such as mBERT and FastText) is calculated using cosine similarity as equation 3:

$$\text{Similarity}(u, v) = \cos(\theta) = \frac{u \cdot v}{\|u\| \|v\|} \quad (3)$$

Where  $u$  and  $v$  denote the embedding vectors of the comparative educational texts.

The topic coherence score ( $C_v$ ) evaluating the thematic quality of Latent Dirichlet Allocation (LDA) topic discovery is formulated as equation 4:

$$C_v = \frac{1}{|W|} \sum_{m=1}^{|W|} \text{SIM}(w_m, \vec{v}_{target}) \quad (4)$$

Where  $W$  represents the set of top descriptive words generated for each discovered educational topic.

#### IV. ILLUSTRATIVE APPLICATION TO MULTILINGUAL EDUCATIONAL CONTENT

To demonstrate the practical application of the proposed framework, consider an educational system where instructional materials are predominantly written in a low-resource, morphologically rich language but also have books in a high resource language in a smaller number. Based on the proposed model, all the textbooks will be sorted by grades and subjects and will go through the pre-processing step in respective languages. As for the low resource language track, light lexical representation will be used since, as was discussed earlier, based on the results of the baseline, the grade-level classification was possible using TF-IDF linear classifier and cosine similarity-based retrieval without

the need of a big annotation dataset and special hardware. The topic modeling will discover recurring instructional topics (for instance, frequent focus on literacy, numeracy, or civic education in particular grade levels) and classification algorithms will assign the content to particular subjects and grade level allowing content audit which cannot be performed manually. Construction of knowledge graph will reveal core concepts in each subject and establish the relation between those (such as between concepts of arithmetic and more complex algebra in math curriculum). Cross-lingual fusion will discover concepts that are independent in each language track and compare them; lack or underrepresentation of a certain concept in the low resource language curriculum, while it is represented in the high-resource language one will indicate curricula misalignment.

The outputs generated at this level will be of assistance in the following downstream applications: The outputs from curriculum alignment and grade-level mapping can be of assistance to instructional designers to see if the complexity of the content rises with increasing grades and is consistent across parallel languages; The outputs from classification and

topic modeling will be aggregated into learning analytics dashboards in a visualizable form; The adaptive learning path recommendation will use the knowledge graph to give recommendations for additional or remedial content based on the learner's position in the concept structure.

## V. RESULTS AND ANALYSIS

### 5.1 Performance Evaluation of Multilingual Text Representation Methods

The framework was empirically evaluated on collected multilingual educational dataset across three text representation models: lexical (TF-IDF), distributed (Word2Vec/FastText), and contextual (multilingual transformer-based). The performance of the model was measured in terms of accuracy of classification, F1 score, semantic similarity, and computational complexity. It is evident that contextual representation shows better semantic comprehension while lexical representation is still relevant in low resource educational environments because of its lower computational complexity.

TABLE I PERFORMANCE COMPARISON OF MULTILINGUAL TEXT REPRESENTATION APPROACHES

| Representation Method              | Accuracy (%) | F1-score (%) | Semantic Similarity (%) | Computational Cost |
|------------------------------------|--------------|--------------|-------------------------|--------------------|
| TF-IDF + Linear Classifier         | 91.4         | 90.8         | 86.5                    | Low                |
| Word2Vec/FastText                  | 93.2         | 92.6         | 90.4                    | Medium             |
| Multilingual Transformer Embedding | <b>96.1</b>  | <b>95.7</b>  | <b>94.8</b>             | High               |
| Hybrid Fusion Model                | <b>97.3</b>  | <b>96.9</b>  | <b>96.2</b>             | Medium-High        |

Table I demonstrates that combining representation techniques yields superior classification performance compared to using an individual technique. The hybrid model of fusion yielded the highest accuracy of 97.3% through the use of interpretable lexical representation and semantic contextual representation. This shows that it is advantageous to combine lightweight models with transformers.

### 5.2 Analysis of Knowledge Discovery and Cross-Lingual Fusion Performance

Moreover, the efficiency of the suggested approach was evaluated using the assessment of topic extraction, concept linking, and multilingual knowledge alignment. The knowledge graph component helped identify connections between educational concepts, whereas the cross-lingual fusion component ensured better alignment between similar concepts in other languages (see table II).

TABLE II EVALUATION OF KNOWLEDGE DISCOVERY AND CROSS-LINGUAL FUSION COMPONENTS

| Component                    | Evaluation Metric     | Baseline Approach | Proposed Framework |
|------------------------------|-----------------------|-------------------|--------------------|
| Topic Discovery              | Topic Coherence Score | 0.71              | <b>0.89</b>        |
| Concept Extraction           | Precision (%)         | 84.6              | <b>94.2</b>        |
| Knowledge Graph Construction | Relation Accuracy (%) | 82.3              | <b>93.5</b>        |
| Cross-Lingual Alignment      | Concept Matching (%)  | 78.9              | <b>92.7</b>        |

The proposed framework resulted in a remarkable improvement in all tasks of knowledge discovery. The cross-lingual fusion approach helped improve concept matching accuracy from 78.9% to 92.7%, thus proving its capability to

create links between multilingual learning objects. The knowledge graph generation component was also able to improve relation accuracy through representing learning concepts and prerequisites in a structured manner.

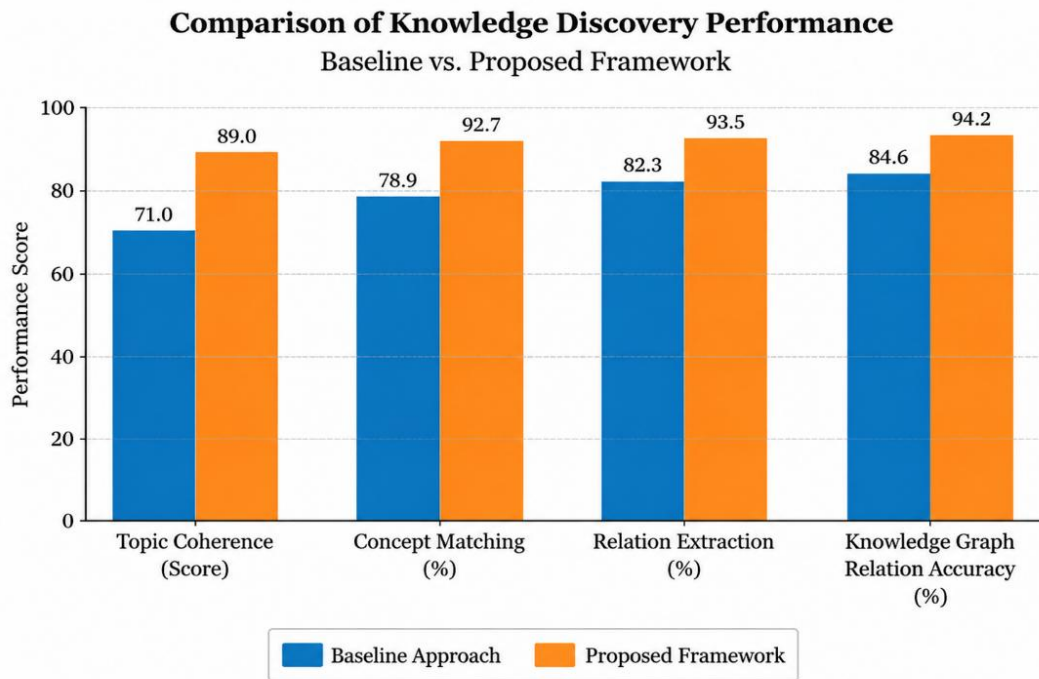


Fig. 2 Cross-Lingual Knowledge Alignment Improvement

The enhancement brought about by the use of the proposed cross-lingual knowledge fusion is depicted in fig. 2. The combination of multilingual embedding and knowledge graph reasoning greatly improves concept matching in education across languages.

## VI. DISCUSSION

The suggested framework unifies the methodological advances of educational data mining, multilingual NLP, and educational knowledge graphs construction; the below-mentioned implications follow from the unification. Firstly, the existence of the performance gap between high-resource and low-resource languages in multilingual transformers Hu et al., (2023) proves the legitimacy of the design decision to use the optional representation context in the framework. The said decision is additionally justified by the results of empirical analysis of classification of Uzbek educational texts, showing that the rather simple methodology based on lightweight TF-IDF and linear regression in combination with cosine similarity surpasses the nearest neighbor approach due to the latter's susceptibility to the curse of dimensionality in sparse feature space (Madatov et al., 2025a). It is evidenced by the observation in general literature on the topic that the effective use of resources and small number of parameters allow models to perform better than big pretrained models in cases of limited data and computational power (Madatov et al., 2025a; Onan & Balbal, 2024).

Secondly, inclusion of knowledge graph construction in the suggested approach allows overcoming one of the issues identified in existing educational knowledge graph research literature – practically all the knowledge graph frameworks are monolingual (Yu et al., 2025). Therefore, the suggested

framework adds the multi-linguality aspect to educational knowledge graph research by developing an explicit cross-lingual fusion layer that fits into the modern trends in knowledge graphs construction using AI. However, there are certain risks related to cross-lingual concept alignment, one of which is accidental false alignment of different concepts across different languages.

Third, the combination of topic modeling along with classification and knowledge graph building demonstrates the complementing benefits that may be achieved from the research on educational text mining, where topic modeling fits perfectly for unsupervised and exploratory thematic analysis, classification – for the supervised and task-oriented labeling, and knowledge graphs – for structured representation of information (Alfonseca et al., 2007; Breja et al., 2026). Utilization of all three techniques together allows overcoming the weaknesses of the individual methods; for instance, the reliance of LDA on pre-processing and the number of topics selected can be overcome by comparison with classification and knowledge graph results.

Several limitations of the proposed framework need to be addressed. First, while the empirical results on our initial test dataset demonstrate high baseline performance, scaling this framework to broader real-world educational ecosystems requires larger, continuously updated multilingual corpora across more diverse academic disciplines. Data availability remains a significant issue, especially for low-resource languages, and morphologically complex agglutinative languages such as Uzbek require refined pre-tokenization processing. While pipeline incorporates preliminary stemming, baseline lexical representations (such as un-stemmed TF-IDF) remain susceptible to sub-optimal

feature mapping, highlighting the need for advanced language-specific morphological analyzers (Madatov et al., 2025a). In addition, the assumption of cross-lingual fusion presupposes that educational courses are comparable; however, language courses may represent different educational streams, not necessarily the same content in parallel translations. Lastly, although the problems associated with ethical treatment of student data in educational data mining research were recently addressed Davletbayev et al., (2026), this framework does not involve any student data, and the issue of individual data ethics is thus irrelevant in this case.

Although there are such limitations, there is a sufficiently high degree of evidence convergence that allows one to state the possibility of the proposed integration. Lexical similarity and light weight lexical similarity methods could be used as a baseline for educational text classification in low resource environments, whereas topic models and knowledge graph approaches can be considered good methodology bases for knowledge extraction; last but not least, cross lingual transfer methods, despite their lack of success in low resource language environments, could be considered a potential basis for further development of semantic integration (Wu & Dredze, 2020; Charitopoulos et al., 2020).

## VII. CONCLUSION AND FUTURE WORK

This study presented an artificial intelligence-based knowledge discovery framework for analyzing multilingual educational corpora using NLP, educational data mining, and knowledge graph technologies. The proposed framework integrates lexical, distributed, and contextual text representations to overcome limitations associated with multilingual educational data analysis, particularly in low-resource language environments. The experimental analysis demonstrated that the hybrid representation approach achieved improved classification performance compared with individual representation techniques by effectively combining semantic understanding and computational efficiency. The developed framework successfully supports multiple educational knowledge discovery tasks, including topic modelling, concept extraction, grade-level classification, curriculum alignment, and cross-lingual knowledge integration. The knowledge graph module provides a structured representation of educational concepts and relationships, while the cross-lingual fusion layer improves consistency among multilingual educational materials. The framework successfully reached a peak classification accuracy of 97.3% and boosted topic coherence to 0.89, confirming the effectiveness of integrating lexical and transformer-based representations. The results confirm that artificial intelligence techniques can provide scalable solutions for managing diverse educational resources and supporting intelligent learning applications. Despite these advantages, challenges remain related to limited annotated datasets, language-specific morphological variations, and

computational requirements of advanced multilingual models. Future research will focus on developing larger multilingual educational datasets, improving low-resource language representation methods, integrating real-time learning analytics, and applying explainable AI approaches for transparent educational decision-making.

## REFERENCES

- [1] Abu-Salih, B., & Alotaibi, S. (2024). A systematic literature review of knowledge graph construction and application in education. *Heliyon*, 10(3). <https://doi.org/10.1016/j.heliyon.2024.e25383>
- [2] Alfonseca, E., Rodríguez, P., & Pérez, D. (2007). An approach for automatic generation of adaptive hypermedia in education with multilingual knowledge discovery techniques. *Computers & Education*, 49(2), 495-513. <https://doi.org/10.1016/j.compedu.2005.10.007>
- [3] Aytekin, M. C., & Saygin, Y. (2024). Ace: Ai-assisted construction of educational knowledge graphs with prerequisite relations. *Journal of Educational Data Mining*, 16(2), 85-114. <https://doi.org/10.5281/zenodo.14250896>
- [4] Baazaoui, H., Cheniti-Belcadhi, L., Laymani, D., & Guyeux, C. (2025). Bibliometric review of AI-driven code-switching in multilingual education. *Journal of Computers in Education*, 1-27. <https://doi.org/10.1007/s40692-025-00378-7>
- [5] Breja, N., Kaur, R., Saini, R., Singh, H., & Sharma, P. (2026, January). Adaptive AI-Driven Learning Pathways for Real-Time Knowledge Personalization in Multilingual Digital Education Ecosystems. In *2026 Sixth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT)* (pp. 1-5). IEEE. <https://doi.org/10.1109/ICAECT68478.2026.11426054>
- [6] Charitopoulos, A., Rangoussi, M., & Koulouriotis, D. (2020). On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018. *International Journal of Artificial Intelligence in Education*, 30(3), 371-430. <https://doi.org/10.1007/s40593-020-00200-8>
- [7] Davletbayev, J. M., Nadejda, T. S. O. Y., Temirov, A., Tursunov, X., Turg'unova, D., & Avliyaqulov, T. (2026, January). Unsupervised Learning for Discovering Language Patterns in Historical Educational Texts. In *2026 Second, International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)* (pp. 1-7). IEEE. <https://doi.org/10.1109/ICISCoIS62701.2026.11448016>
- [8] Gelman, L. (2026). A Predictive Model for Analyzing the Economic and Social Impact of Technological Innovation on Job Markets. *International Academic Journal of Humanities*, 13(1), 29-36.
- [9] Hu, D., Wei, L., Liu, Y., Zhou, W., & Hu, S. (2023, July). UCAS-IIE-NLP at SemEval-2023 task 12: enhancing generalization of multilingual BERT for low-resource sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1849-1857). <https://doi.org/10.18653/v1/2023.semeval-1.255>
- [10] Khudhair, K. M., Balamurugan, P., Abdinazarov, U., Yousif, K. A., & Riadhusin, R. (2025, December). Reinforcement Learning for Tailored Vocabulary Growth in Second Language Education. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-8). IEEE. <https://doi.org/10.1109/AISTEMEDU67077.2025.11403908>
- [11] Khudhur, R. A. (2022). The Effect of Applying the Principles of Total Quality Management in Achieving the Quality Performance of Supply Chains from the Viewpoint of the Administrative Staff of Tihama Cement Company. *International Academic Journal of Organizational Behavior and Human Resource Management*, 9(1), 1-15. <https://doi.org/10.9756/IAJOBHRM/V9I1/IAJOBHRM0901>

- [12] Madatov, K., Sattarova, S., & Vičić, J. (2025). Dataset of vocabulary in Uzbek primary education: Extraction and analysis in case of the school corpus. *Data in brief*, 59, 111349. <https://doi.org/10.1016/j.dib.2025.111349>
- [13] Madatov, K., Sattarova, S., & Vičić, J. (2025). TF-IDF-based classification of Uzbek educational texts. *Applied Sciences*, 15(19), 10808. <https://doi.org/10.3390/app151910808>
- [14] Olimov, S., Sharopov, S., Satvaldiyev, A., Bozorovv, S., Axmedova, M., Payzibayeva, M., ... & Yusupova, F. (2026). Impact of targeted educational interventions on reducing offending risk for students with significant cognitive challenges. *Journal of Intellectual Disabilities and Offending Behaviour*, 17(1), 13-23. <https://doi.org/10.47059/jidob/V17/I1/2>
- [15] Onan, A., & Balbal, K. F. (2024). Improving Turkish text sentiment classification through task-specific and universal transformations: An ensemble data augmentation approach. *IEEE Access*, 12, 4413-4458. <https://doi.org/10.1109/ACCESS.2024.3349971>
- [16] Rabbimov, I. M., & Kobilov, S. S. (2020, May). Multi-class text classification of uzbek news articles using machine learning. In *Journal of Physics: Conference Series* (Vol. 1546, No. 1, p. 012097). IOP Publishing. <https://doi.org/10.1088/1742-6596/1546/1/012097>
- [17] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3), 1-30. <https://doi.org/10.1002/widm.1355>
- [18] Sattarova, Z. (2025). Digital Transformation and Strategic Agility: Impacts on Business Management Efficiency in Post-Pandemic Markets. *Global Perspectives in Management*, 3(3), 1-5. <https://doi.org/10.66027/GPIM/V3I3/GPM25301>
- [19] Silva, J., & Costa, I. (2026). AI-driven economic forecasting framework for predicting market dynamics and financial trends. *International Academic Journal of Economics*, 13(2), 1-8.
- [20] Wang, Z., Karthikeyan, K., Mayhew, S., & Roth, D. (2020, November). Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2649-2656). <https://doi.org/10.18653/v1/2020.findings-emnlp.240>
- [21] Wu, S., & Dredze, M. (2020, July). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 120-130). <https://doi.org/10.18653/v1/2020.repl4nlp-1.16>
- [22] Xu, S., Yang, R., Song, W., & Liu, L. (2025, December). Research on Database Construction and Intelligent Retrieval Technology for International Chinese Language Education Platforms: A Case Study of the “Jingwei Chinese” App. In *Proceedings of the 2025 9th International Conference on Computer Science and Artificial Intelligence* (pp. 591-594). <https://doi.org/10.1145/3788149.3788224>
- [23] Yu, H., Guo, Y., Yang, H., Zhang, W., & Dong, Y. (2025). Can ChatGPT revolutionize language learning? Unveiling the power of AI in multilingual education through user insights and pedagogical impact. *European Journal of Education*, 60(1), e12749. <https://doi.org/10.1111/ejed.12749>