applied sciences

MDPI

Article

TF-IDF-Based Classification of Uzbek Educational Texts

Khabibulla Madatov [1,†], Sapura Sattarova [1,†] and Jernej Vičič [2,*]

[1] The Department of Computer Science, Urgench State University Named after Abu Rayhan Biruni, 14 Kh. Alimdjan Str., Urgench City 220100, Uzbekistan; habi1972@mail.ru (K.M.); sprsattarova@gmail.com (S.S.)
[2] Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska, 6000 Koper, Slovenia
* Correspondence: jernej.vicic@upr.si
† These authors contributed equally to this work.

Abstract

This paper presents a baseline study on automatic Uzbek text classification. Uzbek is a morphologically rich and low-resource language, which makes reliable preprocessing and evaluation challenging. The approach integrates Term Frequency–Inverse Document Frequency (TF–IDF) representation with three conventional methods: linear regression (LR), k-Nearest Neighbors (k-NN), and cosine similarity (CS, implemented as a 1-NN retrieval model). The objective is to categorize school learning materials by grade level (grades 5–11) to support improved alignment between curricular texts and students' intellectual development. A balanced dataset of Uzbek school textbooks across different subjects was constructed, preprocessed with standard NLP tools, and converted into TF–IDF vectors. Experimental results on the internal test set of 70 files show that LR achieved 92.9% accuracy (precision = 0.94, recall = 0.93, F1 = 0.93), while CS performed comparably with 91.4% accuracy (precision = 0.92, recall = 0.91, F1 = 0.92). In contrast, k-NN obtained only 28.6% accuracy, confirming its weakness in high-dimensional sparse feature spaces. External evaluation on seven Uzbek literary works further demonstrated that LR and CS yielded consistent and interpretable grade-level mappings, whereas k-NN results were unstable. Overall, the findings establish reliable baselines for Uzbek educational text classification and highlight the potential of extending beyond lexical overlap toward semantically richer models in future work.

Keywords: Uzbek language; text classification; low-resource languages; TF-IDF; cosine similarity; linear regression; k-Nearest Neighbors

Academic Editor: Pedro Couto

Received: 13 August 2025
Revised: 3 October 2025
Accepted: 6 October 2025
Published: 8 October 2025

Citation: Madatov, K.; Sattarova, S.; Vičič, J. TF-IDF-Based Classification of Uzbek Educational Texts. *Appl. Sci.* **2025**, *15*, 10808. https://doi.org/10.3390/app151910808

1. Introduction

Text classification, also referred to as text categorization, is the automated process of assigning predefined categories or topics to textual content based on linguistic and statistical features. With the exponential growth of digital information—ranging from news articles and scientific publications to social media posts and product reviews—manual categorization has become both impractical and inefficient. To address this challenge, a variety of automatic methods have been developed to classify and filter text, thereby improving retrieval, search, and indexing efficiency [1]. Beyond general information management, text classification also plays a central role in domain-specific applications such as educational content analysis, digital libraries, and intelligent tutoring systems.

In resource-rich languages such as English, text classification has been extensively studied using both traditional machine learning methods and modern deep learning

techniques. Early approaches typically relied on bag-of-words or Term Frequency–Inverse Document Frequency (TF–IDF) representations in combination with algorithms such as Naïve Bayes, Support Vector Machines (SVMs), and k-Nearest Neighbors (k-NN) [2,3]. These methods offered simplicity, interpretability, and relatively low computational cost, making them effective baselines across a variety of domains.

With the advent of deep learning, neural architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were introduced for document and sentence classification, achieving substantial improvements in accuracy [4]. The field was further transformed by the emergence of transformer-based models, most notably BERT (Bidirectional Encoder Representations from Transformers) [5]. BERT and its multilingual variants, including mBERT and XLM-R, have achieved state-of-the-art results on a wide range of natural language processing tasks by leveraging large-scale pre-training on massive text corpora [6,7].

While transformer-based models deliver superior performance, they demand substantial computational resources and large annotated datasets for fine-tuning. This limits their applicability in low-resource settings, where such datasets and infrastructure are often unavailable. Consequently, traditional models remain relevant as lightweight, interpretable, and reproducible baselines, particularly in educational and domain-specific applications.

Recent works on Kazakh have explored LSTM- and BERT-based models for named entity recognition [8], introduced a new sentiment analysis dataset [9], and developed corpora to support multilingual NLP in low-resource settings [10]. These examples demonstrate the growing attention to Turkic languages but also underline that systematic text classification research for Uzbek remains scarce.

Recent advances also highlight the potential of lightweight, parameter-free methods. For instance, Jiang et al. [11] propose a simple compressor-based k-NN classifier, which outperforms BERT on several out-of-distribution datasets, including four low-resource languages. Such approaches demonstrate that effective solutions for underrepresented languages do not necessarily require large pre-trained models but can emerge from resource-efficient alternatives.

Recent work on Turkish has explored novel augmentation strategies to address the scarcity of annotated data. Onan and Balbal [12] propose an ensemble data augmentation approach that combines task-specific and universal transformations, achieving significant improvements in sentiment classification accuracy.

Uzbek, as a morphologically rich and low-resource Turkic language, still lacks sufficient computational resources to support advanced NLP applications. While recent years have seen progress in basic tasks such as tokenization, morphological analysis, and lexicon construction, systematic studies of higher-level tasks remain limited. However, comprehensive evaluations of text classification—particularly in educational domains—are still missing, underscoring the importance of establishing reliable baselines.

Madatov and Bekchanov [13] have proposed a TF–IDF-based summarization model with adaptations to the structure and lexical density of Uzbek texts. They focus on extractive summarization, where the ranking of the most informative sentences is carried out using a normalized sentence-weighting scheme based on the TF–IDF weights of unique Uzbek words. While their objective was summarization, the underlying principle of representing texts through TF–IDF and assigning higher weights to more informative lexical units is closely related to text classification tasks. In both cases, the discriminative power of words is leveraged—either to select the most representative sentences or to distinguish documents by category. This demonstrates that TF–IDF remains a flexible and effective foundation for a range of NLP applications in low-resource languages such as Uzbek, including both summarization and classification.

Madatov et al. [14,15] developed an innovative dataset for text classification issues in Uzbek through the extraction and comparison of vocabulary from 35 primary school textbooks. Their developed corpus—the Uzbek Primary School Corpus (UPSC)—contains graded lists of vocabulary gathered based on a tailored lemma extraction approach. The study offers a beneficial linguistic resource for upcoming NLP issues like automatic classification of educational content in low-resource languages such as Uzbek.

A new resource presented an electronic dictionary of Uzbek word endings to assist in tasks like morphological analysis and machine translation. The resource, which was developed by a combinatorial approach, contains suffixes of different parts of speech. Although it does not deal with classification specifically, it is an important infrastructure for the development of Uzbek NLP [16].

While most low-resource languages lack linguistic infrastructure for sentiment analysis, a recent study introduced the first annotated corpora for Uzbek polarity classification [17]. The study combined a manually labeled dataset with an automatically translated corpus and experimented with both traditional machine learning and deep learning models.

To fill this void, in this research, we provide a systematic comparative analysis of three computationally lightweight and interpretable machine learning approaches to thematic classification of Uzbek school textbooks. All algorithms use Term Frequency–Inverse Document Frequency for text vectorization and continue with either cosine similarity, logistic regression, or k-Nearest Neighbors algorithms [18]. These algorithms are computationally lightweight and more appropriate for settings with limited computational resources. Unlike deep learning-based models—which require huge training datasets, Graphics Processing Units, and extensive tuning—these classical models enjoy high interpretability and lower resource demands, making them especially suitable for real-world application in schools and universities of Uzbekistan. By systematically comparing categorization algorithms with actual Uzbek educational text data, this paper presents a practical evaluation for future Uzbek NLP research. The techniques suggested herein can not only automate cataloging of curriculum content but also support intelligent tutoring systems, digital library indexing, and even analysis of student performance via automatic genre or topic identification.

Ultimately, this research demonstrates that combining Uzbek natural language processing with educational needs makes it possible to effectively identify learning materials that match the intellectual abilities of school students from grades 5 to 11. The TF–IDF-based text classification algorithms proposed by the authors prove to be a promising solution not only for educational content selection but also for classifying large-scale texts of any type in the Uzbek language.

2. Materials and Methods

2.1. Data Description and Preprocessing

In this study, we compiled and prepared a corpus of Uzbek texts consisting of both school textbooks and external literary sources.

2.1.1. Dataset

A total of 96 official Uzbek-language textbooks for grades 5 through 11 were collected from the "Maktab darsliklari" Android app https://play.google.com/store/apps/details?id=dev.mobile.books (accessed on 15 September 2025). These textbooks cover diverse subjects including literature, mathematics, physics, chemistry, biology, Uzbek language, history, and geography.

From this collection, we constructed two distinct datasets:

- **Internal dataset**: For each grade, balanced samples were created by selecting 10 excerpts from different textbooks. Each excerpt was limited to approximately

5000 words, resulting in a total of 70 plain-text files. This ensured grade-level balance across all classes $C \in \{5, 6, 7, 8, 9, 10, 11\}$.

- **External dataset**: To evaluate model robustness, we additionally included texts from various literary genres (novels, stories, and essays) outside the textbook corpus. These were treated as independent test cases for classification experiments.

2.1.2. Preprocessing

All documents were processed through the following pipeline:

1. **Encoding normalization**: All files were converted to UTF-8 encoding.
2. **Lowercasing**: Characters were transformed into lowercase to reduce vocabulary sparsity.
3. **Cleaning:** Non-textual elements such as headers, footers, and page numbers were removed.
4. **Punctuation removal**: Non-alphanumeric symbols were deleted while preserving sentence delimiters.
5. **Tokenization**: Text was split into tokens using whitespace and punctuation rules.
6. **Stopword removal**: A manually curated list of Uzbek stopwords was applied to eliminate function words.
7. **Vectorization**: TF–IDF representations of each document were constructed. For each token t in document d, the TF–IDF weight was computed as follows:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

where

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_k f_{k,d}}, \qquad \text{IDF}(t) = \log\left(\frac{N}{1 + n_t}\right).$$

Here, $f_{t,d}$ is the frequency of term t in document d; $\sum_k f_{k,d}$ is the total number of terms in d; N is the total number of documents; and n_t is the number of documents containing t.

The final vocabulary size after preprocessing was restricted to the 5000 most informative features, ensuring a compact and comparable representation across documents. Each textbook excerpt was mapped to a TF–IDF vector of fixed dimensionality, yielding labeled instances $V_1, V_2, \ldots, V_{70}$ for supervised learning.

2.1.3. Implementation Details

All experiments were implemented in Python 3.9. We employed `scikit-learn` for TF–IDF vectorization and classification, `NumPy` for numerical operations, and `matplotlib` for visualization of results. Cross-validation was conducted using 5-fold stratified splitting, ensuring balanced class distributions across folds. The accuracy and standard deviation of performance were reported to capture model reliability.

This unified dataset and preprocessing workflow ensures consistent input representation for all classification methods described in Sections 2.2–2.4.

2.2. TF–IDF + Linear Regression Algorithm

Although logistic regression is a standard classifier, we include linear regression as a simple multi-output baseline trained on one-hot labels.

Suppose 96 texts are given, each text is tokenized, a common vocabulary of the texts is constructed, the irrelevant words are removed from each text, and the TF–IDF values of the remaining words are calculated so that the corresponding vectors of the texts are brought to the same dimension:

$$\vec{V}_1, \vec{V}_2, \ldots, \vec{V}_{96} \in \mathbb{R}^d$$

Each vector belongs to one of the classes C_m, $m \in \{1, 2, 3, 4, 5, 6, 7\}$. Let us assume that the TF–IDF vector corresponding to the new 97th text is also adjusted to the same dimension as the vectors $\vec{V}_1, \vec{V}_2, \ldots, \vec{V}_{96} \in \mathbb{R}^d$. We hypothesize that there is a linear regression relationship between $\vec{V}_1, \vec{V}_2, \ldots, \vec{V}_{96}$ and their classes Y, and we construct the linear regression model as follows.

Step 1

$$Y = XW + \varepsilon \tag{1}$$

Here,

- $X \in \mathbb{R}^{96\times d}$ is the matrix of TF–IDF vectors of the 96 texts;
- $Y \in \mathbb{R}^{96\times 7}$ represents the classes corresponding to the 96 vectors;
- $W \in \mathbb{R}^{d\times 7}$ is the regression coefficient;
- ε is the error term.

Step 2

In linear regression, we convert the classes into numerical form. For example, we apply one-hot encoding.

Since there are 7 classes,

$$\begin{aligned} &\text{class } 1 \to (1,0,0,0,0,0,0),\\ &\text{class } 2 \to (0,1,0,0,0,0,0),\\ &\qquad\vdots\\ &\text{class } 7 \to (0,0,0,0,0,0,1). \end{aligned}$$

Thus, each $\vec{y}_i$ $(i = 1, \ldots, 96)$ will be represented in vector form.

Step 3

Now, by applying the **Least Squares Method**, we obtain

$$\min_W \|\varepsilon\|^2 \;=\; \min_W \|Y - XW\|^2$$

from which we derive

$$W = (X^T X)^{-1} X^T Y.$$

Step 4

Now, we classify the new document:

$$\hat{y}_{97} = \vec{x}_{97} W.$$

As a result, we obtain a 7-dimensional vector (each coordinate corresponds to a class). We select the class corresponding to the largest coordinate. That is,

$$\text{class}(\vec{x}_{97}) = \arg\max_{1\le j\le 7} \hat{y}_{97,j}.$$

In this expression, the class corresponding to the column with the largest value is selected.

2.3. TF–IDF + K-Nearest Neighbors Algorithm

We also evaluated the k-Nearest Neighbors (k-NN) classifier as a simple baseline. This classical non-parametric method assigns a test instance to the majority class among its k nearest training neighbors in the TF–IDF feature space [19].

Given a new text with TF–IDF vector $\vec{\mathbf{V}}_{97}$, the goal is to assign it to one of the seven grade levels $\{5, 6, 7, 8, 9, 10, 11\}$ by comparing it to the 96 textbook vectors $\vec{\mathbf{V}}_1, \vec{\mathbf{V}}_2, \ldots, \vec{\mathbf{V}}_{96}$.

1. **Distance computation**:
   For each $i = 1, 2, \ldots, 96$, the distance between $\vec{\mathbf{V}}_i$ and $\vec{\mathbf{V}}_{97}$ is computed. Following standard practice in text mining, we considered both Euclidean and cosine distance metrics:

$$d_{\text{euclidean}}(\vec{V}_i, \vec{V}_{97}) = \sqrt{\sum_{j=1}^{d} (V_i^{(j)} - V_{97}^{(j)})^2}, \qquad d_{\cos}(\vec{V}_i, \vec{V}_{97}) = 1 - \frac{\vec{V}_i \cdot \vec{V}_{97}}{\|\vec{V}_i\| \, \|\vec{V}_{97}\|}.$$

2. **Model selection**:
   Instead of fixing K heuristically (e.g., $\sqrt{n}$), we performed a grid search with $K \in \{1, 3, 5, 7, 9, 11\}$ using 5-fold stratified cross-validation on the 96 textbooks. Both Euclidean and cosine distances were tested, together with uniform and distance-based voting schemes. Dimensionality reduction via Truncated SVD (100–300 components) was also included in the search space.
3. **Sorting and neighbor selection**:
   For a given test vector, all distances are sorted in ascending order, and the top K nearest neighbors $\mathcal{N}_K$ are selected.
4. **Classification rule**:
   The predicted class $\hat{C}$ is obtained by majority voting among the labels of the K neighbors:

$$\hat{C} = \arg \max_{m \in \{5, \ldots, 11\}} \sum_{i \in \mathcal{N}_K} \mathbf{1}[C_i = m],$$

   where C_i is the grade label of $\vec{\mathbf{V}}_i$. In case of ties, the label of the closest neighbor is selected.
5. **Evaluation**:
   The final model was trained on all 96 textbook vectors and evaluated on two separate test sets: (i) an internal balanced set of 70 texts (10 per grade) and (ii) an external set of 7 literary texts. We report accuracy, precision, recall, F1-score, and confusion matrices.

2.4. TF–IDF + Cosine Similarity Algorithm

In this section, we present a method for classifying Uzbek texts using TF–IDF vectors and the cosine similarity measure.

Given a new document represented by its TF–IDF vector $\vec{\mathbf{V}}_{97} \in \mathbb{R}^d$, and a set of 96 existing documents represented by $\vec{\mathbf{V}}_1, \vec{\mathbf{V}}_2, \ldots, \vec{\mathbf{V}}_{96} \in \mathbb{R}^d$, the classification process proceeds as follows.

1. **Cosine similarity definition**:
   For each existing vector $\vec{\mathbf{V}}_i$, where $i = 1, 2, \ldots, 96$, we compute the cosine similarity between the new vector and $\vec{\mathbf{V}}_i$ as follows:

$$\text{sim}_i = \cos(\theta) = \frac{\vec{\mathbf{V}}_i^{\top} \vec{\mathbf{V}}_{97}}{\|\vec{\mathbf{V}}_i\|_2 \cdot \|\vec{\mathbf{V}}_{97}\|_2}$$

   where

- $\vec{\mathbf{V}}_i^\top \vec{\mathbf{V}}_{97}$ is the dot product between the two vectors;
- $\|\vec{\mathbf{V}}_i\|_2$ and $\|\vec{\mathbf{V}}_{97}\|_2$ are their Euclidean norms.

2. **Similarity interpretation**:
Cosine similarity provides a normalized measure of directional alignment between two TF–IDF vectors. A value closer to 1 indicates stronger textual similarity. In our case, this measure is used to compute the similarity between the new document vector $\vec{\mathbf{V}}_{97}$ and each of the existing vectors $\vec{\mathbf{V}}_1, \vec{\mathbf{V}}_2, \ldots, \vec{\mathbf{V}}_{96}$. The vectors with the highest similarity value, ideally approaching 1, are considered the best match and determine the classification outcome.
3. **Assign class**:
Each vector $\vec{\mathbf{V}}_i$ is labeled with a class C_m, where $m \in \{1, 2, \ldots, 7\}$, corresponding to school grade levels.
4. **Final classification step**:
Let

$$i^* = \arg \max_{1 \le i \le 96} sim_i$$

be the index of the most similar vector. The class name of the new document is assigned based on that of the most similar existing document:

$$\text{Class}(\vec{\mathbf{V}}_{97}) = \text{Class}(\vec{\mathbf{V}}_{i^*})$$

If multiple vectors have the same maximum similarity result, those classes are taken as the final result.

3. Results

This section presents the results of our grade-level classification experiments on Uzbek texts. The evaluation was carried out using TF–IDF-based feature representations combined with three classification algorithms: linear regression, k-Nearest Neighbors (KNN), and a 1-NN baseline with cosine similarity.

To address reviewer concerns, we report not only overall accuracy but also precision, recall, and F1-scores together with confusion matrices to provide deeper insight into class-wise performance. Random and majority-class baselines are included for context.

The experiments were conducted on two test sets: (i) an internal balanced corpus of 70 literary texts (10 per grade, drawn from 96 school textbooks) and (ii) an external corpus of 7 texts from diverse genres. Summary tables for each method are provided below.

Table 1 presents the statistics of the school textbooks used in this study. Since listing all 96 textbooks individually would be impractical, they are grouped by grade level, with the number of documents, total token count, and unique vocabulary size provided for each grade.

Table 1. Statistics of the school textbook corpora per grade.

№	File Name	Grade	Source Type	Number of Textbooks	Tokens	Unique Words
1	5_merged.txt	5	Internal	13	268,189	46,791
2	6_merged.txt	6	Internal	12	253,608	45,740
3	7_merged.txt	7	Internal	15	386,479	57,387
4	8_merged.txt	8	Internal	15	403,241	58,630
5	9_merged.txt	9	Internal	11	275,343	47,407
6	10_merged.txt	10	Internal	15	365,454	56,396
7	11_merged.txt	11	Internal	15	355,897	44,864
			Total	**96**	**2,303,150**	**221,036**

Table 2 presents a list of seven literary works in the Uzbek language obtained from an open electronic library resource of Uzbekistan—ZiyoUz [20]. These texts were selected to identify literary materials that align with the intellectual and linguistic capabilities of school students. The chosen works represent a variety of genres, including fantasy, crime fiction, historical chronicles, and classical poetry.

Table 2. List of external literary texts used in the classification experiments.

№	File Name	Source Description
1	`garri.txt`	Uzbek translation of "Harry Potter" novel by J.K. Rowling
2	`shaytanat.txt`	Crime novel "Shaytanat" by Tohir Malik
3	`kuhnadunyo.txt`	"Kuhna dunyo", a novel by Odil Yoqubov
4	`xorazm.txt`	"Xorazm tarixi" by Bayoniy, a historical chronicle
5	`attor.txt`	"Mantiq-ut-Tayr" by Fariduddin Attar
6	`mehrob.txt`	"Mehrobdan chayon", a novel by Abdulla Qodiriy
7	`avazxon.txt`	"Avazxon", an Uzbek folk epic

Table 3 presents the predictions of three TF–IDF-based methods on the external corpus of seven literary texts. Unlike the internal dataset of school textbooks, these external texts do not have predefined grade labels. Their role in this study is not to measure accuracy, but rather to explore which grade level each literary work may be most appropriate for.

The classification is performed by comparing each external literary text against the balanced TF–IDF representations of school textbooks (grades 5–11). The grade assigned is the one whose vocabulary and frequency distribution are most similar to the given text. It is important to emphasize that TF–IDF captures only lexical overlap and word frequency patterns, rather than deep semantic meaning. Thus, the mapping indicates how closely the word usage of an external text resembles that of a particular grade-level corpus, which we interpret as an indicator of suitability for students' intellectual and linguistic capabilities at that grade.

The results show that linear regression (LR) and cosine similarity (CS) provided consistent and interpretable predictions, often aligning with one another. By contrast, the KNN method frequently defaulted to grade 10, revealing its limitations in high-dimensional TF–IDF spaces. These findings support the conclusion that LR and CS are more reliable for recommending extracurricular literary works that match the linguistic and cognitive capacity of students in grades 5–11.

Table 3. Classification results based on external sources.

№	File Name	Source	TF–IDF + LR	TF–IDF + KNN	TF–IDF + CS
1	`garri.txt`	external	7	10	7
2	`shaytanat.txt`	external	7	10	7
3	`kuhnadunyo.txt`	external	7	7	8
4	`xorazm.txt`	external	11	8	10
5	`attor.txt`	external	11	10	11
6	`avazxon.txt`	external	11	10	7
7	`temuriy.txt`	external	7	7	7

Table 4 shows the structure of the internal balanced dataset. Each grade (5–11) contains 10 balanced excerpts of 5000 tokens each, drawn from different subjects (e.g., literature, sciences, and history). Here, grade 5 is listed in full and two representative files from grade 11 (#69 and #70) are displayed to illustrate dataset boundaries.

Table 4. Examples from the internal balanced dataset.

№	File Name	Subject (English)	Grade	Tokens
1	5.1_adabiyot.txt	Literature	5	5000
2	5.2_biologiya.txt	Biology	5	5000
3	5.3_geografiya.txt	Geography	5	5000
4	5.4_informatika.txt	Informatics	5	5000
5	5.5_jismoniy_tarbiya_6.txt	Physical Education	5	5000
6	5.6_matematika_1_2.txt	Mathematics	5	5000
7	5.7_musiqa.txt	Music	5	5000
8	5.8_onatili.txt	Mother Tongue	5	5000
9	5.9_tarixdan_hikoyalar.txt	History Stories	5	5000
10	5.10_tasviriy_sanat.txt	Fine Arts	5	5000
	. . . (balanced files for Grades 6–10 not shown for brevity) . . .			
69	11.9_jahon_tarixi.txt	World History	11	5000
70	11.10_adabiyot.txt	Literature	11	5000

As shown in Table 5, internal test results on 70 balanced files show strong performance (precision = 0.94, recall = 0.93, F1 = 0.93, and accuracy = 92.9%). At the same time, five-fold cross-validation on the same dataset yielded much lower performance (accuracy = 22.9% and macro F1 = 0.19). This contrast demonstrates that although the TF–IDF + linear regression model can successfully fit the balanced test set, its generalization ability across folds remains limited. Therefore, these results should be interpreted with caution regarding the model's broader applicability.

Table 5. Internal classification results using TF–IDF + linear regression.

Grade	5	6	7	8	9	10	11
5	9	0	0	0	0	1	0
6	0	10	0	0	0	0	0
7	0	0	10	0	0	0	0
8	0	0	0	10	0	0	0
9	0	0	0	0	9	1	0
10	1	0	1	1	0	7	0
11	0	0	0	0	0	0	10

Precision = 0.27, recall = 0.29, F1 = 0.25, and accuracy = 28.6% on 70 files (Table 6). The confusion matrix shows that KNN predictions are heavily biased toward grade 10, with many texts from grades 6–11 misclassified into this category. For example, grade 6 and grade 7 texts were mostly predicted as grade 10, while grade 11 was confused with both grade 7 and grade 10. Although some grade 5 texts were recognized correctly (8 out of 10), overall performance was weak and inconsistent across classes. This demonstrates that TF–IDF + KNN lacks generalization power in high-dimensional Uzbek text space, and it is not a reliable method for grade-level classification.

Cross-validation analysis. For completeness, we also evaluated KNN using five-fold cross-validation on the balanced internal dataset. The average accuracy was only 11.4%, with macro-averaged precision = 0.11, recall = 0.11, and F1 = 0.11. The confusion matrices across folds show that predictions collapsed into a few classes (mainly grade 10), with almost no correct matches for other grades. This confirms that the weak performance observed on the fixed internal test set is consistent under cross-validation as well, indicating that TF–IDF + KNN is unsuitable for Uzbek grade-level text classification.

Table 6. Internal classification results using TF–IDF + KNN.

Grade	5	6	7	8	9	10	11
5	8	0	1	0	0	1	0
6	0	1	2	0	0	7	0
7	0	1	2	1	0	6	0
8	0	0	1	4	0	5	0
9	0	0	1	1	1	7	0
10	0	0	1	1	1	7	0
11	0	0	2	0	0	7	1

Precision = 0.92, recall = 0.91, F1 = 0.92, and accuracy = 91.4% on 70 files. The TF–IDF + cosine similarity (1-NN) method shows high reliability on the balanced dataset. Most predictions are correct, with minor confusions between neighboring grades (e.g., 5 vs. 7, 7 vs. 10, 9 vs. 10, and 11 vs. 7), reflecting lexical overlap in adjacent curricula. Cross-validation is not applicable here since the model relies on a fixed set of 96 reference textbooks; removing subsets would reduce the index and distort retrieval. Thus, results are reported in the fixed-reference setting, matching the intended use case.

The internal classification results obtained using the TF–IDF representation combined with cosine similarity are summarized in Table 7; these results demonstrate that the majority of texts were accurately assigned to their respective grade levels, suggesting the effectiveness of TF–IDF features with cosine similarity for distinguishing among the seven grade categories.

Table 7. Internal classification results using TF–IDF + cosine similarity.

Grade	5	6	7	8	9	10	11
5	8	0	1	0	0	1	0
6	0	10	0	0	0	0	0
7	0	0	9	0	0	1	0
8	1	0	0	9	0	0	0
9	0	0	0	1	9	0	0
10	0	0	1	0	1	8	0
11	0	0	1	0	0	0	9

As summarized in Table 8, we compare three TF–IDF–based methods on the internal balanced dataset.

Table 8. Comparison of three TF–IDF-based methods on the internal balanced dataset (70 files).

Method	Accuracy	Precision	Recall	F1-Score
TF–IDF + Linear Regression	92.9%	0.94	0.93	0.93
TF–IDF + KNN	28.6%	0.27	0.29	0.25
TF–IDF + Cosine Similarity	91.4%	0.92	0.91	0.92

4. Discussion

In this study, three TF–IDF-based approaches were evaluated for grade-level classification of Uzbek educational texts: **TF–IDF + linear regression (LR)**, **TF–IDF + k-NN**, and **TF–IDF + cosine similarity (CS, 1-NN retrieval)**.

On the **internal balanced corpus (70 files)**, LR achieved 92.9% accuracy with a macro-averaged precision of 0.94, recall of 0.93, and F1-score of 0.93. CS performed comparably with 91.4% accuracy and a precision of 0.92, recall of 0.91, and F1-score of 0.92. Both methods demonstrated robustness across grades, with only minor confusions between

adjacent grade levels (e.g., 5 vs. 7, 7 vs. 10, 9 vs. 10, and 11 vs. 7), which reflect the natural lexical overlaps across curricula. In contrast, k-NN with optimized hyperparameters achieved only 28.6% accuracy, heavily misclassifying many texts into grade 10. This weakness can be explained by the curse of dimensionality in high-dimensional sparse TF–IDF space, where Euclidean distance becomes ineffective.

On the **external corpus (seven literary works)**, the goal was not accuracy but suitability assessment by mapping each text to the most similar grade-level corpus. Here, LR and CS provided consistent and interpretable predictions. For example, *Harry Potter* and *Shaytanat* mapped to grade 7, *Kuhna dunyo* to grades 7–8, and *Mantiq-ut-Tayr* to grade 11. k-NN, however, frequently defaulted to grade 10, confirming its poor generalization ability.

From a methodological standpoint, **cross-validation** was performed for LR and k-NN, validating the reliability of results. For CS, cross-validation is not meaningful because its model consists of a fixed set of 96 reference textbooks: removing folds would reduce the index set and artificially lower retrieval quality. Therefore, CS performance is reported only in the fixed-reference evaluation setting, which corresponds to its intended use case.

It is also important to acknowledge the **limitations** of this work. Uzbek is a low-resource, agglutinative language; hence, no lemmatization, stemming, or morphological normalization was applied. This may have led to feature sparsity due to multiple word forms. Furthermore, only baseline TF–IDF approaches were tested. Future work should incorporate more semantically powerful models, such as word embeddings or transformer-based architectures, to capture deeper linguistic patterns.

Overall, the findings show that LR and CS are reliable baselines for grade-level classification of Uzbek educational texts, while k-NN fails to handle the lexical distribution effectively. This study establishes baseline results for Uzbek educational corpora and provides a methodological foundation for future extensions toward semantically richer models for low-resource language applications.

5. Conclusions

This study established TF–IDF-based classification as a first baseline for Uzbek educational texts across grades 5–11. Among the evaluated methods, linear regression and cosine similarity consistently achieved reliable results, both exceeding 90% accuracy on the balanced internal dataset. In contrast, k-NN showed very weak performance due to the curse of dimensionality in sparse TF–IDF space, confirming its limited applicability.

The findings also demonstrate that simple lexical overlap is sufficient to approximate grade-level differences in Uzbek curricula, although minor confusions appear between adjacent grades. Nevertheless, the approach remains limited by the absence of linguistic preprocessing such as lemmatization and stemming, which are particularly important for an agglutinative, low-resource language like Uzbek.

Future work will expand beyond TF–IDF baselines toward semantically richer representations, including word embeddings, transformer-based architectures, and cross-lingual transfer learning. These methods will not only address current limitations but also enable more robust and interpretable models for educational applications in Uzbek and other low-resource languages.

Author Contributions: Conceptualization, K.M. and J.V.; methodology, K.M.; software, S.S.; validation, J.V.; investigation, K.M., J.V. and S.S.; resources, S.S.; data curation, S.S.; writing—original draft preparation, K.M. and S.S.; writing—review and editing, K.M. and J.V.; visualization, K.M.; supervision, J.V.; project administration, K.M.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data Availability Statement: The dataset supporting the findings of this study is openly available on Zenodo and is cited in the references as [21]. *TF-IDF Matrix for Grades 5–11 Uzbek Textbooks*.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
TF-IDF	Term Frequency–Inverse Document Frequency
KNN	K-Nearest Neighbors
LR	Linear Regression
CS	Cosine Similarity
UPSC	Uzbek Primary School Corpus
GPU	Graphics Processing Unit

References

1. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
2. McCallum, A.; Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, USA, 26–27 July 1998; Volume 752, No. 1, pp. 41–48.
3. Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of the European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142.
4. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882. [CrossRef]
5. Cesar, L.B.; Manso-Callejo, M.Á.; Cira, C.I. BERT (Bidirectional Encoder Representations from Transformers) for Missing Data Imputation in Solar Irradiance Time Series. *Eng. Proc.* **2023**, *39*, 26.
6. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual Is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502. [CrossRef]
7. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv* **2019**, arXiv:1911.02116.
8. Oralbekova, D.; Mamyrbayev, O.; Zhumagulova, S.; Zhumazhan, N. A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-Classification Approach. In *Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies*; Springer: Cham, Switzerland, 2024; pp. 116–128.
9. Yeshpanov, R.; Varol, H. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes. *arXiv* **2024**, arXiv:2403.19335.
10. Tleubayeva, A.; Aubakirov, S.; Tabuldin, A.; Shomanov, A. Development and Evaluation of a Small Kazakh Language Corpus to Improve the Efficiency of Multilingual NLP Systems in Low-Resource Environments. In Proceedings of the 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), Astana, Kazakhstan, 14–16 May 2025; IEEE: Piscataway Township, NJ, USA, 2025; pp. 1–6.
11. Jiang, Z.; Yang, M.; Tsirlin, M.; Tang, R.; Dai, Y.; Lin, J. "Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 6810–6828.
12. Onan, A.; Balbal, K.F. Improving Turkish Text Sentiment Classification through Task-Specific and Universal Transformations: An Ensemble Data Augmentation Approach. *IEEE Access* **2024**, *12*, 4413–4458. [CrossRef]
13. Madatov, K.A.; Bekchanov, S.K. The Algorithm of Uzbek Text Summarizer. In Proceedings of the 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russia, 28 June–2 July 2024; IEEE: Piscataway Township, NJ, USA, 2024; pp. 2430–2433.
14. Madatov, K.; Sattarova, S.; Vičič, J. Dataset of Vocabulary in Uzbek Primary Education: Extraction and Analysis in Case of the School Corpus. *Data Brief* **2025**, *59*, 111349. [CrossRef] [PubMed]

15. Madatov, K.A.; Sattarova, S. Creation of a Corpus for Determining the Intellectual Potential of Primary School Students. In Proceedings of the International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM) Altai, Russia, 28 June–2 July 2024; IEEE: Piscataway Township, NJ, USA, 2024; pp. 2420–2423. [CrossRef]
16. Rabbimov, I.M.; Kobilov, S.S. Multi-Class Text Classification of Uzbek News Articles Using Machine Learning. *J. Phys. Conf. Ser.* **2020**, *1546*, 012097. [CrossRef]
17. Kuriyozov, E.; Matlatipov, S.; Alonso, M.A.; Gómez-Rodríguez, C. Construction and Evaluation of Sentiment Datasets for Low-Resource Languages: The Case of Uzbek. In *Language and Technology Conference*; Springer: Cham, Switzerland, 2019; pp. 232–243.
18. Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Pearson Education: Delhi, India, 2016.
19. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Series: Springer Texts in Statistics; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-7137-0. [CrossRef]
20. ZiyoUz Library—Digital Collection of Literary Works. Available online: https://n.ziyouz.com/kutubxona/category/1-ziyouz (accessed on 15 September 2025).
21. Sattarova, S. *TF–IDF Matrix for Grades 5–11 Uzbek Textbooks [Data Set]*; Zenodo: Geneva, Switzerland, 2025. [CrossRef]