

Article

Not peer-reviewed version

---

# TF-IDF Based Classification of Uzbek Educational Texts

---

[Khabibulla Madatov](#)<sup>†,‡</sup>, [Sapura Sattarova](#)<sup>‡</sup>, [Jernej Vičič](#)<sup>\*</sup>

Posted Date: 15 August 2025

doi: 10.20944/preprints202508.1137.v1

Keywords: Uzbek language; text classification; low-resource languages; TF-IDF; Cosine Similarity; Logistic Regression; k-Nearest Neighbors





Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

## Article

# TF-IDF Based Classification of Uzbek Educational Texts

Khabibulla Madatov <sup>1,†,‡</sup> , Sapura Sattarova <sup>2,‡</sup> and Jernej Vičič <sup>2,\*</sup> 

<sup>1</sup> Urgench State University named after Abu Rayhan Biruni, 14 Kh. Alimdjani Str., Urgench City, 220100, Uzbekistan

<sup>2</sup> Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska, 6000 Koper

\* Correspondence: jernej.vicic@upr.si

† All authors are currently affiliated with the institutions listed above.

‡ These authors contributed equally to this work.

**Abstract:** This paper presents an approach to automatic Uzbek text classification. Uzbek language is a morphologically rich and low-resource language. The approach integrates Term Frequency–Inverse Document Frequency (TF-IDF) representation with conventional machine learning and similarity-based approaches. The aim is to categorize learning materials at the school grade level to support improved alignment of materials and student learning outcomes. In order to carry out the research, a dataset of 5th–11th grade school textbooks in different subjects was collected. The texts were preprocessed using standard natural language processing (NLP) tools and were transformed into TF-IDF vectors. These were used to train three common classification models: Logistic Regression (LR), k-Nearest Neighbors (k-NN), and Cosine Similarity (CS). Each new input text is compared with the grade-level textbook corpus, and the grade with the highest similarity is selected. It provides an estimate of the appropriate intellectual level for the material. The experimental findings indicate that Logistic Regression achieved 82% accuracy, and Cosine Similarity performed slightly better at 85.7%. Conversely, the k-NN method achieved only 22% accuracy, indicating its low applicability for Uzbek text classification. Overall, the proposed approach demonstrates practical value for pedagogical purposes and potential applicability to wider document analysis issues.

**Keywords:** Uzbek language; text classification; low-resource languages; TF-IDF; Cosine Similarity; Logistic Regression; k-Nearest Neighbors

## 1. Introduction

Text classification, or text categorization, is the computer-assisted assignment of pre-established categories or topics to textual content based on its features. As text data on the internet has grown exponentially in recent years—whether in the form of news articles, social media updates, or product reviews—manual processing has become increasingly impractical. As a result, various automatic approaches have been constructed for classifying and filtering such information in terms of topic or category, making it more suitable for retrieval, search, and indexing [1].

In particular, text classification is a fundamental NLP task that involves the classification of text data into pre-defined categories [2]. Its usage in numerous industries stretches from digital libraries, e-learning websites, legal document indexing, to automated content moderation. Text classification is required in educational settings for the automatic categorization of teaching materials by subject so that learning resources can be managed. It is also applied to the analysis of student essays—vocabulary richness, coherence, and relevance to subject matter—which also enables personalized guidance and advice [3].

Text classification is a highly investigated NLP task and has witnessed immense growth in the past twenty years. The early approaches were mainly statistical and rule-based, which later changed to machine learning algorithms trained on human-engineered features like term frequency and document structure. More recent advancements have brought in deep learning architectures,

such as convolutional and recurrent neural networks, and transformer models with state-of-the-art performance on most tasks. Simple approaches like TF-IDF combined with traditional classifiers have nonetheless remained very much relevant, particularly in situations where computational resources or large-scale labeled corpora are scant.

In recent years, text classification research has primarily been aimed at high-resource languages like English, Chinese, and Spanish [4–6]. Such languages enjoy enormous annotated corpora and pre-trained models. By contrast, the Uzbek language as a low-resource language is underrepresented in the NLP literature. Although some preliminary work using rule-based or keyword-matching approaches has been conducted, comparative large-scale, methodologically intensive work is lacking.

Madatov and Bekchanov [7] 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. This approach enables the successful determination of important content in texts and contribution to automatic summarization and text analysis tools for low-resource languages such as Uzbek.

Madatov et al. [8,9] 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 education content in low-resource languages like 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 [10].

Although most languages possess enormous lexical resources such as dictionaries and thesauri, creating such resources for Uzbek has only just started. Another more recent effort was to create an Uzbek WordNet from a structural adaptation of the Turkish WordNet. This study helps to make more semantic resources available for Uzbek to enable a broad variety of NLP tasks [11].

While most low-resource languages lack linguistic infrastructure for sentiment analysis, recent studies have introduced the first annotated corpora for Uzbek polarity classification [12]. The study combined a manually labeled dataset with an automatically translated corpus and experimented with both traditional machine learning and deep learning models [13].

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 and k-Nearest Neighbors algorithms [16]. 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. Although transformer-based models such as BERT or mBERT obtained state-of-the-art results on most NLP tasks [14], they are less accessible for low-resource languages with no large-scale corpora or annotated datasets.

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 school textbooks for grades 5–11 as follows.

#### 2.1.1. Dataset

Official Uzbek-language textbooks for grades 5 through 11 were downloaded from the “Maktab darsliklari” Android app<sup>1</sup>. A total of 96 textbooks covering literature, mathematics, physics, chemistry, biology, Uzbek language, history, and geography were collected and converted to plain-text format for further processing.

#### 2.1.2. Preprocessing

All corpora were processed through the following pipeline.

1. **Encoding normalization:** Converted every file to UTF-8.
2. **Lowercasing:** Transformed all characters to lowercase to reduce sparsity.
3. **Cleaning:** Stripped non-textual elements (headers, footers, page numbers).
4. **Punctuation removal:** Removed non-alphanumeric symbols while preserving sentence delimiters.
5. **Tokenization:** Split text into tokens using whitespace and punctuation rules.
6. **Vectorization:** Constructed TF-IDF representations of each grade corpus. For each token  $t$  in document  $d$ , the TF-IDF weight was computed as:

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

where the term frequency (TF) is defined as:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

and the inverse document frequency (IDF) is given by:

$$\text{IDF}(t) = \log\left(\frac{N}{1 + n_t}\right)$$

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  $t$ .

Let each of the 96 textbooks belong to one of the grades in  $\{5, 6, 7, 8, 9, 10, 11\}$ . We first construct a vocabulary from the tokens extracted from these textbooks. For each of the 96 documents, we generate a TF-IDF vector whose dimensionality corresponds to the size of the vocabulary. In our case, the number of unique tokens is 221036. Each vector is labeled with its respective grade. Let these vectors be denoted as  $V_1, V_2, \dots, V_{96}$ , with class labels drawn from 7 categories, i.e.,  $C \in \{5, 6, 7, 8, 9, 10, 11\}$ .

If the incoming text does not contain any terms present in the school textbook corpus, its TF-IDF vector will consist entirely of zeros. Consequently, TF-IDF-based models will fail to classify the input or compute meaningful similarity, as no informative features are represented in the resulting vector.

<sup>1</sup> <https://play.google.com/store/apps/details?id=dev.mobile.books>

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

## 2.2. TF-IDF + Linear Regression Algorithm

In this section, we present a method for classifying Uzbek texts using TF-IDF vectorization combined with a linear regression approach.

Let the new document be transformed into a TF-IDF vector  $\vec{V}_{97} \in \mathbb{R}^d$ , and suppose we have a set of 96 existing documents represented as vectors  $\vec{V}_1, \vec{V}_2, \dots, \vec{V}_{96} \in \mathbb{R}^d$ .

### 1. Regression model formulation:

For each existing vector  $\vec{V}_i$ , where  $i = 1, 2, \dots, 96$ , we assume a linear relationship of the form:

$$\vec{V}_{97} \approx \beta_i \cdot \vec{V}_i$$

where  $\beta_i \in \mathbb{R}$  is a scalar coefficient representing how closely  $\vec{V}_i$  aligns with the new vector  $\vec{V}_{97}$ .

### 2. Compute optimal scalar $\beta_i$ :

The optimal scalar  $\beta_i$  is computed as:

$$\beta_i = \frac{\vec{V}_i^\top \vec{V}_{97}}{\vec{V}_i^\top \vec{V}_i}$$

where:

- $\vec{V}_i^\top \vec{V}_{97}$  is the dot product between the existing and new vectors,
- $\vec{V}_i^\top \vec{V}_i$  is the squared norm of the existing vector.

### 3. Residual error :

The squared residual error for each  $i$  is defined as:

$$r_i = \left\| \vec{V}_{97} - \beta_i \vec{V}_i \right\|^2$$

Expanding and substituting the optimal  $\beta_i$ , we obtain:

$$\begin{aligned} r_i &= \left\| \vec{V}_{97} - \beta_i^* \vec{V}_i \right\|_2^2 \\ &= \vec{V}_{97}^\top \vec{V}_{97} - 2\beta_i^* \vec{V}_i^\top \vec{V}_{97} + (\beta_i^*)^2 \vec{V}_i^\top \vec{V}_i \\ &= \vec{V}_{97}^\top \vec{V}_{97} - \frac{(\vec{V}_i^\top \vec{V}_{97})^2}{\vec{V}_i^\top \vec{V}_i} \end{aligned} \quad (1)$$

### 4. Assign class:

Each existing vector  $\vec{V}_i$  has an associated class label  $C_m$ , where  $m \in \{1, 2, \dots, 7\}$ , representing the school grade level of the document.

### 5. Select best match:

We identify the index  $i^*$  corresponding to the minimum residual error:

$$i^* = \arg \min_{1 \leq i \leq 96} r_i$$

### 6. Final classification step:

We determine the class of the vector corresponding to  $i^*$ , and this class will be the most similar to the given text:

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



### 2.3. TF-IDF + K-Nearest Neighbors Algorithm

Suppose a new text is given. Let its TF-IDF vector be denoted as  $\vec{V}_{97}$ . We present the K-Nearest Neighbors algorithm to determine which of the existing 96 vectors  $\vec{V}_1, \vec{V}_2, \dots, \vec{V}_{96}$  is most similar to the new vector  $\vec{V}_{97}$ :

1. **Determine K:**

Since we have a corpus of 96 school textbooks, we compute:

$$K = \sqrt{96} \approx 9$$

2. **Compute similarity (Euclidean distance):**

For each  $i = 1, 2, \dots, 96$ , compute the similarity between  $\vec{V}_i$  and  $\vec{V}_{97}$  using the following formula:

$$\text{sim}(\vec{V}_i, \vec{V}_{97}) = \sqrt{\sum_{j=1}^n (\vec{V}_i^{(j)} - \vec{V}_{97}^{(j)})^2}$$

where  $\vec{V}_i^{(j)}$  and  $\vec{V}_{97}^{(j)}$  are the  $j$ -th components of the respective vectors.

3. **Assign class:**

Each vector  $\vec{V}_i$  is associated with a class label  $C_m$ , where  $m \in \{1, 2, \dots, 7\}$ , representing the school grades.

4. **Sorting step:**

Sort the computed similarities in ascending order:

$$\text{sort}(\text{sim}(\vec{V}_i, \vec{V}_{97})), \quad 1 \leq i \leq 96$$

Then select the top  $K = 9$  vectors.

5. **Select top K neighbors:**

Selected 9 vectors and their corresponding class  $C_m$  will be defined, where  $m = 1, \dots, 7$ .

6. **Final classification step:**

Identify the most frequently occurring class  $C_m$  according to the 9 selected vectors. If there are many such classes, the one is selected whose corresponding vector has the greatest value from the vector of the given text. We consider that the given text belongs to that class.

### 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  $\vec{V}_{97} \in \mathbb{R}^d$ , and a set of 96 existing documents represented by  $\vec{V}_1, \vec{V}_2, \dots, \vec{V}_{96} \in \mathbb{R}^d$ , the classification process proceeds as follows:

1. **Cosine similarity definition:**

For each existing vector  $\vec{V}_i$ , where  $i = 1, 2, \dots, 96$ , we compute the cosine similarity between the new vector and  $\vec{V}_i$  as:

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

where:

- $\vec{V}_i^\top \vec{V}_{97}$  is the dot product between the two vectors,
- $\|\vec{V}_i\|_2$  and  $\|\vec{V}_{97}\|_2$  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  $\vec{V}_{97}$  and each of the existing

- vectors  $\vec{V}_1, \vec{V}_2, \dots, \vec{V}_{96}$ . The vector with the highest similarity value, ideally approaching 1, are considered the best match and determine the classification outcome.
3. **Assign class:**  
Each vector  $\vec{V}_i$  is labeled with a class  $C_m$ , where  $m \in \{1, 2, \dots, 7\}$ , corresponding to school grade levels.
4. **Final classification step:**  
Let

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

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:

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

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 conducted on Uzbek texts. The evaluation is carried out using TF-IDF-based feature representations combined with multiple classification algorithms, including Linear Regression , k-Nearest Neighbors, and Cosine Similarity . The analysis covers both internal (school textbooks) and external (literary and informational texts) corpora. Summary tables are provided below.

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

Note: 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.

Based on the list of texts provided in Table 2, the outcomes of three classification tasks can be observed. According to the results, the Uzbek-language version of Harry Potter corresponds to the intellectual level of a 5th-grade student, while literary works titled Shaytanat and Temuriy match the intellectual capacity of 7th-grade students. Additionally, Kuhna Dunyo, a novel by Uzbek author Odil Yoqubov, is suitable for 9th-grade students, whereas the historical novel Khorezm by Bayoniy aligns with the intellectual level of 10th-grade students. The famous prose work The Conference of the Birds (Mantiq ut-Tayr) by Fariduddin Attar and the Uzbek folk epic Avazxon are found to be appropriate for 11th-grade students. It is notable that the results obtained using TF-IDF-based LR and CS methods are consistent with each other. In contrast, the analysis performed using the KNN method yielded inaccurate results, demonstrating that it is not suitable for classifying Uzbek texts.

**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

*Note:* Table 2 presents a list of seven literary works in the Uzbek language obtained from an open electronic library resource of Uzbekistan—ZiyoUz [15]. 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 3.** Classification results based on external sources.

№	File name	Source	TF-IDF + LR	TF-IDF + KNN	TF-IDF + CS
1	garri.txt	external	5	7	5
2	shaytanat.txt	external	7	10	7
3	kuhnadunyo.txt	external	9	7	9
4	xorazm.txt	external	10	8	10
5	attor.txt	external	11	10	11
6	avazxon.txt	external	11	10	11
7	temuriy.txt	external	7	7	7

*Note:* Table 3 presents the results of identifying literary works appropriate to the intellectual abilities of students in grades 5–11.

School textbooks are considered official and standardized sources that align with the national educational curriculum. For this reason, the authors of this study selected school textbooks from grades 5 to 11 as the primary basis for identifying suitable text classification algorithms for the Uzbek language. A total of 96 processed .txt files were compiled across these seven grade levels. From each grade, five representative literary texts were chosen, resulting in a curated set of 35 documents. Each text was evaluated using three TF-IDF-based methods—Linear Regression, K-Nearest Neighbors, and Cosine Similarity—to determine which grade level each document is most appropriate for, based on linguistic and cognitive complexity.

Table 5 presents the final results of the classification of literary texts based on school textbooks. For this, 5 literary texts were taken from each of the 5-11 grade textbooks. The classification problem on a total of 35 literary texts was carried out in the Python programming language. Each algorithm shows how suitable it is for Uzbek texts. Of the three methods based on TF-IDF, Logistic Regression and Cosine Similarity showed their suitability for classifying Uzbek texts, achieving classification accuracies of 82% and 85.7%, respectively. On the contrary, the TF-IDF + KNN method showed poor results with an accuracy of only 22%, which proves that it is not suitable for text classification in the Uzbek language.



**Table 4.** List of internal literary texts used in the classification experiments.

#	File Name	Grade	Source Description
1	dunyo.txt	5	“Dunyoning ishlari”, novel, by O’tkir Hoshimov
2	ezop.txt	5	Ezop fables
3	guliston.txt	5	“Guliston”, by Sa’diy Sheroziy
4	hadislar.txt	5	Imom Buxoriy hadisth
5	hellados.txt	5	“Hellados”, story by Nodar Dumbadze
6	mahbub.txt	6	“Mahbub ul-qulub”,by Alisher Navoiy
7	muzqaymoq.txt	6	“Muzqaymoq” story,by Odil Yoqubov
8	nasihatlar.txt	6	Nasihatlar, by Abay
9	shumbola.txt	6	“Shum bola” by G’afur Gulom
10	yulduz.txt	6	“Yulduzli tunlar”, by P. Qodirov
11	mehrob.txt	7	“Mehrobdan chayon”, by Abdulla Qodiriy
12	memor.txt	7	“Me’mor” novel by Mirmuhsin
13	oq_kema.txt	7	“Oq kema” asari, by Chingiz Aytmatov
14	qiyomat.txt	7	“Qiyomat qarz” novel,by O’lmas Umarbekov
15	ravshan.txt	7	“Ravshan” epic, folklore
16	chinor.txt	8	“Chinor” novel, Asqad Muxtor
17	kuntugmish.txt	8	“Kuntug’mish” epic,folklore
18	lutfiy.txt	8	“G’azallar”, by Lutfiy
19	nilvarim.txt	8	“Nil va Rim”, prose,by Usmon Nosir
20	qochoq.txt	8	“Qochoq”, novel,by Said Ahmad
21	asr.txt	9	“Asrga tatigulik kun”, Chingiz Aytmatov
22	farhodshirin.txt	9	“Farhod va Shirin” epic, Alisher Navoiy
23	navoiy.txt	9	“Navoiy” excerpt, Oybek
24	ulugbek.txt	9	“Ulug’bek xazinasi” novel by Odil Yoqubov
25	xoja.txt	9	Xoja story’s
26	atoyi.txt	10	Atoiy ghazals
27	bobur.txt	10	“Boburnoma”,by Zahiriddin Muhammad Bobur
28	hayot.txt	10	“Hayotga muxabbat”, Djek London
29	ikkieshik.txt	10	“Ikki eshik orasi” novel,by O’tkir Hoshimov
30	rustamxon.txt	10	“Rustamxon” epic, folklore
31	kechakunduz.txt	11	“Kecha va kunduz” by Abdulhamid Cho’lpon
32	mashrab.txt	11	Ghazals, Boborahim Mashrab
33	qutadgu.txt	11	“Qutadg’u bilig” epic, Yusuf Xos Hojib
34	rabguziy.txt	11	“Qisasi Rabg’uziy”, Nosiriddin Rabg’uziy
35	choliqushi.txt	11	“Choliqushi” novel by Rashod Nuri Guntekin

*Note:* Table 4 provides a list of literary texts and their authors used to evaluate TF-IDF-based classification methods for Uzbek language materials. The dataset includes five texts per grade level, selected from a variety of genres such as epics, stories, novels, and ghazals. The texts represent authors of diverse national and cultural backgrounds, enabling a comprehensive assessment of the classification methods across literary styles and complexity.

**Table 5.** Evaluation results of three methods based on TF-IDF.

№	File name	Grade	TF-IDF + LR	TF-IDF + KNN	TF-IDF + CS
1	dunyo.txt	5	5	7	5
2	ezop.txt	5	5	8	5
3	guliston.txt	5	10	10	10
4	hadislar.txt	5	5	10	5
5	hellados.txt	5	5	10	5
6	Sariqdev.txt	6	6	7	6
7	muzqaymoq.txt	6	6	10	6
8	nasihatlar.txt	6	5	8	5
9	shumbola.txt	6	5	7	6
10	yulduz.txt	6	6	10	6
11	mehrob.txt	7	7	7	7
12	memor.txt	7	7	8	7
13	oq_kema.txt	7	7	7	7
14	qiyomat.txt	7	7	7	7
15	ravshan.txt	7	7	8	7
16	chinor.txt	8	8	7	8
17	kuntugmish.txt	8	8	7	8
18	lutfiy.txt	8	8	7	8
19	nilvarim.txt	8	8	6	8
20	qochoq.txt	8	8	7	8
21	asr.txt	9	9	7	9
22	farhodshirin.txt	9	9	11	9
23	navoiy.txt	9	9	10	9
24	ulugbek.txt	9	9	7	9
25	xoja.txt	9	11	10	11
26	turdifarogiy	10	10	7	10
27	bobur.txt	10	10	7	10
28	hayot.txt	10	10	7	10
29	ikkieshik.txt	10	10	10	10
30	rustamxon.txt	10	10	10	10
31	kechakunduz.txt	11	11	10	11
32	mashrab.txt	11	10	10	10
33	qutadgu.txt	11	11	8	11
34	rabguziy.txt	11	11	10	11
35	choliqushi.txt	11	5	7	5
<b>Accuracy</b>			<b>82%</b>	<b>22%</b>	<b>85.7%</b>

#### 4. Discussion

This article presents a study on the classification of educational resources in the Uzbek language, the authors consider the issue of classifying educational texts and determining which grade the given educational material corresponds to in terms of the student's intellectual abilities. To put it more clearly, if a schoolchild is not provided with educational materials that match their intellectual abilities — that is, if they read books that do not match their vocabulary level — the student will not be able to absorb that information. The student's vocabulary is formed and expanded based on school textbooks. Therefore, the authors of the article used various textbooks for grades 5–11 as the object of the research.

Based on the results of this research work, the authors compared three classification algorithms in the field of Natural Language Processing. The study demonstrates not only the effectiveness, but also the limitations of the methods used, as well as their relevance in a broader educational context. The results show that TF-IDF-based Logistic Regression and Cosine Similarity methods gave high performance in classifying Uzbek texts, while the K-Nearest Neighbors method, with only 22% accuracy, was found to be unsuitable as an alternative solution for Uzbek language classification.

5. Conclusions

In this study, three classification methods based on TF-IDF—Logistic Regression, K-Nearest Neighbors, and Cosine Similarity—were tested for their applicability in classifying Uzbek-language texts. A selected set of 35 literary samples from school textbooks for grades 5 to 11 was used to evaluate how well each approach could accurately identify the corresponding intellectual level of students.

Experimental analysis shows that Logistic Regression and Cosine Similarity significantly outperform KNN in terms of classification accuracy and robustness. While Logistic Regression achieved 82% accuracy, Cosine Similarity achieved the highest score of 85.7%, indicating their suitability for the morphological and syntactic structure of the Uzbek language. On the other hand, KNN recorded only 22% accuracy, which proved that it was not suitable for classifying Uzbek-language texts.

The results confirm that combining TF-IDF with Logistic Regression or Cosine Similarity is a reliable and effective method for classifying Uzbek texts. These models are not limited to educational materials; they also apply to literary and thematic texts in Uzbek. The findings show how NLP-based classification strategies can support educational decisions, in particular in selecting texts that are appropriate for students’ cognitive levels.

Future research may focus on the thematic classification of Uzbek texts using more advanced and modern models. Additionally, developing comprehensive deep classification models through neural-network-based vectorization techniques will significantly contribute to the automatic analysis of Uzbek texts, a low-resource language. This, in turn, will open up broader opportunities for identifying, categorizing, and integrating educational materials that align with the intellectual abilities of school students. Integrating education with natural language processing will have a powerful impact on ensuring quality education for the youth, who represent the future of the nation.

**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:** The dataset supporting the findings of this study is openly available at Zenodo: Sattarova, S. (2025). *TF-IDF Matrix for Grades 5–11 Uzbek Textbooks* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.15705517> (accessed on 20 June 2025).

**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	Logistic Regression
CS	Cosine Similarity
UPSC	Uzbek Primary School Corpus
GPU	Graphics Processing Unit

References

1. Deng, X.; Li, Y.; Weng, J.; Zhang, J. Feature Selection for Text Classification: A Review. *Multimedia Tools and Applications* **2019**, *78*(3), 3797–3816. <https://doi.org/10.1007/s11042-018-6293-7>

2. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. <https://doi.org/10.3390/info10040150>

3. Page, E.B. Project Essay Grade: PEG. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*; Shermis, M.D., Burstein, J., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2003; pp. 43–54.
4. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine* **2018**, 13(3), 55–75. <https://doi.org/10.1109/MCI.2018.2840738>
5. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001.
6. Schwenk, H.; Li, X. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, 7–12 May 2018; pp. 3548–3551.
7. Madatov, K.A.; Bekchanov, S.K. The Algorithm of Uzbek Text Summarizer. In *Proceedings of the International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM)*, Erlagol, Russia, 30 June–4 July 2024; pp. 2430–2433. <https://doi.org/10.1109/EDM61683.2024.10615191>
8. Madatov, K.; Sattarova, S.; Vičič, J. Dataset of Vocabulary in Uzbek Primary Education: Extraction and Analysis in Case of the School Corpus. *Data in Brief* **2025**, 59, 111349. <https://doi.org/10.1016/j.dib.2025.111349>
9. 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)*, Erlagol, Russia, 1–5 July 2024; pp. 2420–2423. <https://doi.org/10.1109/EDM61683.2024.10615103>
10. Matlatipov, S.; Tukeyev, U.; Aripov, M. Towards the Uzbek Language Endings as a Language Resource. In *Proceedings of the International Conference on Computational Collective Intelligence*; Springer: Cham, Switzerland, 2020; pp. 729–740.
11. Madatov, K.A.; Khujamov, D.J.; Boltayev, B.R. Creating of the Uzbek WordNet Based on Turkish WordNet. *AIP Conference Proceedings* **2022**. <https://doi.org/10.1063/5.0089532>
12. Rabbimov, I.M.; Kobilov, S.S. Multi-class Text Classification of Uzbek News Articles Using Machine Learning. *Journal of Physics: Conference Series* **2020**, 1546(1), 012097. <https://doi.org/10.1088/1742-6596/1546/1/012097>
13. Kuriyozov, E.; Salaev, U.; Matlatipov, S.; Gómez-Rodríguez, C. Construction and Evaluation of Sentiment Datasets for Low-Resource Languages: The Case of Uzbek. In *Proceedings of the Language and Technology Conference*; Springer International Publishing: Cham, Switzerland, 2019; pp. 232–243.
14. 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*, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
15. Ziyoz Library – Digital Collection of Literary Works. Available online: <https://n.ziyouz.com/kutubxona/category/1-ziyouz-com-kutubxonasi>
16. Tan, Pang-Ning and Steinbach, Michael and Kumar, Vipin (2016). *Introduction to data mining*. Pearson Education India.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.