UDC 004.8:811.1/.2 **DOI: 10.18413/2313-8912-2023-9-1-0-4**

Andrei Paraschiv[1]
Mihai Dascalu[2]
Marina I. Solnyshkina[3]

Classification of Russian textbooks by grade level and topic using ReaderBench

[1] Polytechnic University of Bucharest
313 Splaiul Independentei, Sector 6, Bucharest 060042, Romania
E-mail: andrei.paraschiv74@upb.ro

[2] Polytechnic University of Bucharest
313 Splaiul Independentei, Sector 6, Bucharest 060042, Romania
E-mail: mihai.dascalu@upb.ro

[3] Kazan (Volga region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
E-mail: mesoln@yandex.ru

Received 10 February 2023; accepted 16 March 2023; published 30 March 2023

Acknowledgments. This work was supported by a grant of the Ministry of Research, Innovation and Digitalization, Project CloudPrecis, Contract Number 344/390020/06.09.2021, MySMIS code: 124812, within POC. We thank the Research Lab Text Analytics at Kazan Federal University for assisting in compiling the corpus of academic texts and cooperating to conduct the research.

Abstract. Textbooks are essential resources for classroom and offline reading, while the quality of learning materials guides the entire learning process. One of the most important factors to be considered is their readability and comprehensibility. Therefore, the correct pairing of textbook complexity and student grade level is paramount. This article analyzes automated classification methods for Russian-language textbooks on two dimensions, namely the topic of the text and its complexity reflected by its corresponding school grade level. The studied corpus is a collection of 154 textbooks from the Russian Federation from the second to the eleventh grade levels. Our analysis considers machine learning techniques on the textual complexity indices provided by the open-source multi-language framework ReaderBench and BERT-based models for the classification tasks. Additionally, we explore using the most predictive ReaderBench features in conjunction with contextualized embeddings from BERT. Our results argue that incorporating textual complexity indices improves the classification performance of BERT-based models on our dataset split using 2 strategies. More specifically, the F1 score for topic classification improved to 92.63%, while the F1 score for school grade-level classification improved to 54.06% for the Greedy approach in which multiple adjacent paragraphs are considered a single text unit up until reaching the maximum length of 512 tokens imposed by the language model.

Keywords: Text readability; Russian language; Textbook analysis; Topic classification; ReaderBench framework; Textual complexity indices; Transformer-based Language Model

How to cite: Paraschiv, A., Dascalu, M. and Solnyshkina, M. I. (2023). Classification of Russian textbooks by grade level and topic using ReaderBench, *Research Result. Theoretical and Applied Linguistics*, 9 (1), 50-63. DOI: 10.18413/2313-8912-2023-9-1-0-4

УДК 004.8:811.1/.2 **DOI: 10.18413/2313-8912-2023-9-1-0-4**

Параскив А.[1]
Даскалу М.[2]
Солнышкина М. И.[3]

Типология учебников русского языка на основе ReaderBench: уровневый и тематический подходы

[1] Политехнический университет Бухареста
313 Сплаиул Индепендетей, Сектор 6, Бухарест, 060042, Румыния
E-mail: andrei.paraschiv74@upb.ro

[2] Политехнический университет Бухареста
313 Сплаиул Индепендетей, Сектор 6, Бухарест, 060042, Румыния
E-mail: mihai.dascalu@upb.ro

[3] Казанский (приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: mesoln@yandex.ru

*Статья поступила 10 февраля 2023 г.; принята 16 марта 2023 г.;
опубликована 30 марта 2023 г.*

Информация об источниках финансирования или грантах, благодарности: Работа выполнена при поддержке гранта Министерства исследований, инноваций и цифровизации, проект CloudPrecis, номер договора 344/390020/06.09.2021, код MySMIS: 124812, в рамках POC. Выражаем благодарность НИЛ "Текстовая аналитика" Казанского (Приволжского) федерального университета за помощь в составлении корпуса и сотрудничество при проведении исследования.

Аннотация. Учебник является важным образовательным ресурсом для чтения в классе и самостоятельной работы, а качество учебных материалов определяет весь учебный процесс. Одним из наиболее важных факторов, которые следует учитывать, является их удобочитаемость и понятность. Поэтому правильное сочетание сложности учебника и уровня компетентности учащихся имеет первостепенное значение. В данной статье анализируются автоматизированные методы классификации русскоязычных учебников по двум измерениям, а именно по теме текста и его сложности, отражаемой соответствующим школьным уровнем (классом). Корпус исследования – 154 учебника, используемых для обучения в 2 – 11 классах Российской Федерации. Исследование осуществлено на основе методов машинного обучения с использованием индексов сложности текста, рассчитываемых при помощи многоязычной платформы с открытым исходным кодом ReaderBench и классификационными моделями на основе BERT. Кроме того, мы изучаем наиболее предиктивные функции ReaderBench в сочетании с контекстуальными вложениями от BERT. Наши результаты доказывают, что включение индексов сложности текста улучшает эффективность классификации моделей на основе BERT в нашем наборе данных, разделенном с использованием двух стратегий. В частности, показатель F1 для классификации по темам улучшился до 92,63 %, а показатель F1 для классификации по уровням обучения (классам) улучшился до 54,06 % для жадного алгоритма, при котором несколько смежных абзацев считаются единым текстовым блоком до тех пор, пока не будет достигнута максимальная длина абзаца, 512 токенов, для изучаемой языковой модели.

Ключевые слова: Читабельность текста; Русский язык; Анализ учебника; Тематическая классификация; Фреймворк ReaderBench; Индексы сложности текста; Языковая модель на основе преобразователя

Информация для цитирования: Параскив А., Даскалу М., Солнышкина М. И. Типология учебников русского языка на основе ReaderBench: уровневый и тематический подходы // Научный результат. Вопросы теоретической и прикладной лингвистики. 2023. Т. 9. № 1. С. 50-63. DOI: 10.18413/2313-8912-2023-9-1-0-4

Introduction

Despite the increased usage of electronic multimedia in education, such as video courses, audiobooks, or interactive online courses, textbooks are still among the most valuable and frequently employed educational materials, especially for early grade levels. Wakefield (2006) shows that 87-88% of US students report reading their textbooks in class at least once weekly. Textbooks are an essential resource for the classroom, where the teacher acts as a facilitator, and are also used in the home environment; as such, the entire learning process is guided by quality learning materials (Swanepoel, 2010). Since the content of the textbooks has a major impact on the education system's effectiveness (Khine, 2013), their content and complexity level need careful attention from the research community. One of the most important factors to be analyzed when assessing textbooks is their readability and comprehensibility (Bansiong, 2019). Therefore, the correct pairing of textbook complexity and student grade level has to be achieved. Using objective, computable parameters to rate the readability of textbooks in a highly competitive and saturated market is of invaluable help to any actor or stakeholder in the educational space.

The problem of readability and text complexity has long been outstanding in the research community. Early on, the approach was focused on using easy-to-obtain numerical quantification of the analyzed text and using them as input for an algorithmic approach. For example, Kincaid et al. (1975) introduced the Flesch–Kincaid readability index (FKI) as a measure of readability for English texts based on their structural features. More modern approaches, such as CohMetrix (Crossley et al., 2008), use text cohesion metrics to evaluate the readability of English texts for English as a second language (ESL) students. CohMetrix uses computational linguistic tools that provide a deeper insight into the structure of a text. Modern statistical approaches rely on many textual features, from simple shallow metrics and counts to more complex lexical, morphological, and syntactic features.

Nowadays, related works on text difficulty focus on various languages and topics. Chatzipanagiotidis et al. (2021) classified Greek as a second language corpus using a Sequential Minimal Optimization (SMO). For the Italian and English languages, DeepEva (Bosco et al., 2021) uses two long-short-term memory (LSTM) neural layers to classify English and Italian sentences according to their complexity. A simpler approach to compare Slovak and Canadian textbooks was used by Beníčková et al. (2021). The authors used simple formulas and counted to create syntactic and semantic text difficulty coefficients that were then aggregated to obtain a numerical score for text difficulty. Even so, the authors identified a larger proportion of technical terms in Slovak textbooks that unnecessarily increased the difficulty of those fragments.

For the Russian language, Batinic et al. (2017) used a Naive Bayes classifier to differentiate between three Russian levels as foreign language textbooks. They achieved 74% accuracy using several numerical features, such as FKI for the Russian language, counts of different parts of speech (POS), and the number of abstracta (i.e., abstract words per sentence) in texts. Solovyev et al. (2020a) considered a linguistic ontology, RuThes-Lite, to compute the complex features of the text on the Thesaurus graph for discrimination of the school grade level. Using the Russian Dictionary of Abstractness/Concreteness (Solovyev, Ivanov, and Akhtiamov, 2019), Solovyev et al. (2020b) the authors proposed an online tool, RusAC, to assess the abstractness score of a Russian text. The authors argued that these indices are a good indicator of the topic of the text, whereas scientific texts tend to be more concrete. The analysis of the complexity of academic texts using textual indices was also addressed by Solovyev et al. (2018, 2019); Churunina et al. (2020). Sakhovskiy et al. (2020) showed that text complexity, measured by the grade the textbook addresses, correlates with relevant topic features such as the coherence of topics, topics with semantically closer words, and the frequency of topic words. At the sentence level, Ivanov (2022) constructed a graph from the dependency tree and used BERT and Graph Neural Networks to predict the complexity at this level.

Numeric linguistic features have been proven reliable predictors of text complexity and text topic (Norris and Ortega, 2009; Santucci et al., 2020; Zipitria et al., 2012). One example of a framework providing textual complexity indices available in Russian is Readerbench (Dascalu et al., 2017; Corlatescu et al., 2022), presented in detail in the following section.

ReaderBench Textual Complexity Indices

Readerbench provides over 500 indices for the Russian language[1] divided into three levels of granularity: sentence level (Sent), paragraph level (Par), and document level (Doc). Additionally, there are three methods of aggregation: mean (M), maximum (Max), and standard deviation (SD). These indices are further referenced by abbreviations such as “M (UnqWd / Par),” representing the mean value of unique words per paragraph. The framework provides several classes of indices, including surface, word, morphology, syntax, and cohesion indices, as presented in Table 1.

Surface indices are simple normalized counts that do not consider the content of the text. These range from the number of words per document, paragraph, or sentence to the entropy of words (Shannon, 1948; Brown et al., 1992). Word indices provide insights into individual words, their length, how different their lemma is from the inflected form, or if the word stands for a specific named entity such as locations (LOC), persons (PER), or organizations (ORG).

More complex indices are based on part-of-speech (POS) tagging, such as the morphology category or the syntactic indices based on the results of syntactic dependency parsers. A key constituent when assessing text difficulty is text cohesion computed using the Cohesion Network Analysis (CNA) graph (Dascalu et al., 2018). This cohesion category helps distinguish between textbooks that pose a higher challenge to the reader than others by highlighting cohesion gaps or low cohesion segments.

Text pre-processing from ReaderBench, including POS tagging, dependency parsing, and named entity recognition, relies on spaCy, while the CNA graph is built using BERT-based models. For this analysis, we considered Russian spaCy models[2] (i.e., “ru_core_news_lg”) and RuBERT (Kuratov and Arkhipov, 2019), part of the DeepPavlov library[3].

[1] https://github.com/readerbench/ReaderBench/wiki/Textual-Complexity-Indices
[2] https://spacy.io/models/ru
[3] https://github.com/deeppavlov/DeepPavlov

Table 1. Available indices in the Readerbench framework for the Russian language
Таблица 1. Доступные для русского языка индексы Readerbench

Abbreviation	Description	Granularity
	Surface indices	
Wd	number of Words per granularity unit	Doc, Par, Sent
UnqWd	number of unique Words per granularity unit	Doc, Par, Sent
Comma	number of commas	Doc, Par, Sent
Punct	number of punctuation marks, including commas	Doc, Par, Sent
Sent	number of Sentences per granularity unit	Doc, Par
WdEnt	word entropy	Doc, Par, Sent
	Word indices	
Chars	number of characters per Word	Word
NgramEntr 2	Bigram entropy in words	Word
LemmaDiff	distance from a word and its lemmatized version	Word
Repetitions	number of occurrences of the same lemma	Doc, Par, Sent
NmdEnt	number of syllables in a word	Doc, Par, Sent
Syllab	number of Words	Doc, Par, Sent
	Morphology indices	
*Pos*Main	number of words with a specific POS	Doc, Par, Sent
UnqPosMain	number of unique words with a specific POS	Doc, Par, Sent
Pron	number of specific pronouns (first, second, third, intensive, indefinite)	Doc, Par, Sent
	Syntax indices	
Dep	Dependencies of specific type	Doc, Par, Sent
ParseTreeDpth	Depth of parse tree	Doc, Par, Sent
	Cohesion indices	
AdjSentCoh	Cohesion between two adjacent sentences	Doc, Par
AdjParCoh	Cohesion between two adjacent paragraphs	Doc
IntraParCoh	Cohesion between sentences contained within a given paragraph	Doc, Par
InterParCoh	Cohesion between paragraphs	Doc
StartEndCoh	Cohesion between first and last text element	Doc, Par
StartMiddleCoh	Cohesion between the start and all middle text elements	Doc, Par
MiddleEndCoh	Cohesion between all middle and last elements	Doc, Par
TransCoh	Cohesion between the last sentence of the current paragraph and the first sentence from the upcoming paragraph	Doc

Research Objective

Our research objective is to analyze the predictive power of state-of-the-art models on both the horizontal (i.e., topic) and the vertical dimensions (i.e., complexity derived from school grade level) of textbooks written in the Russian language. As such, we analyze 154 Russian language textbooks covering 10 school grades from the 2nd to 11th grades across 13 topics ranging from STEM subjects, such as Maths and Physics, to humanities and social sciences. We explore several textual complexity indices and Cohesion Network Analysis features from ReaderBench and their contribution to classifying the aforementioned textbooks. Additionally, we consider large Russian language models, such as RuBERT (Kuratov and Arkhipov, 2019), in conjunction with the selected features and compare their performance with the stand-alone model based only on the textual complexity indices.

We open-source our code at *https://github.com/readerbench/rus-textbooks* and our best models for both grade level and topic classifications at *https://huggingface.co/readerbench/ru-textbooks*.

Method

Textbooks Corpus

This study considers a stable version from February 2023 of the corpus elaborated by the linguistic experts from the Text Analytics Laboratory, Institute of Philology and Intercultural Communication, Kazan. The dataset consists of 154 Russian textbooks distributed across 10 school grade levels and 13 different subjects – this later task is also called topic classification. Table 2 shows that the topics are not evenly distributed across the grades. Subjects such as Arts, Music, Maths, Science, and Technology are not present above the 4th grade in the corpus. In contrast, Biology, Geography, History, Physics, and Social Studies are only present after the 4th grade. For Biology, a special case refers to the 10th and 11th grades, where all three textbooks were evenly spread across both grades; since a clear separation could not be established, we associated all these three books with the 10th grade.

Table 2. Textbook distribution over grades and topics
Таблица 2. Распределение учебников по классам и темам

	Arts	Biology	Ecology	Geography	History	IT	Maths	Music	Physics	Russian	Science	Social studies	Technology
2nd Grade	1		1			3	6	1		6	2		4
3rd Grade	2		1			3	3	2		6	2		3
4th Grade	2					4	7	1		5	3		5
5th Grade		6			3					4		2	
6th Grade		1		2	5					1		2	
7th Grade		5		1	1	2			2	2		2	
8th Grade		5	1	1	1				1			4	
9th Grade		3			1				3	1		2	
10th Grade				1	4							2	
10-11th Grade		3											
11th Grade				1	4							2	

The input sequence length is limited for RuBERT to 512 tokens, corresponding to an average of 300 to 350 words. Since the average number of tokens per textbook far exceeds this limit (see Table 3), the entire document cannot be used as the classification unit. For this, we experimented with two document-splitting strategies.

Table 3. Average number of paragraphs, words, and tokens per textbook and subject
Таблица 3. Среднее количество абзацев, слов и токенов на учебник и предмет

Subject	# paragraphs	Average # words	# tokens
Art	461	8,190	12,174
Biology	1,287	40,732	60,738
Ecology	749	16,790	24,334
Geography	1,418	43,998	60,030
History	1,117	55,880	75,978
IT	939	14,703	21,347
Maths	2,101	27,555	40,020
Music	303	5,579	8,444
Physics	615	30,235	41,000
Russian	2,552	33,492	55,053
Science	1,415	29,931	42,584
Social studies	990	36,928	50,663
Technology	419	7,653	11,746

The first strategy selects individual paragraphs in their occurrence order within each textbook, while the second approach appends subsequent paragraphs in a Greedy manner just before they exceed 512 tokens. Inherently, the extracted text units are coherent since they contain full paragraphs with fully expressed ideas. Given the distribution of paragraph lengths (M = 36.12 tokens per paragraph and SD = 38.55), both strategies produce classification units under 512 tokens. There were only 64 paragraphs in all textbooks that exceeded 512 tokens. These were split into smaller chunks that followed sentence splits and would fit into the model input. After splitting the data into classification units, we obtain the class distributions in Table 4 and in Table 5. We notice that the topic classes are highly imbalanced regardless of the splitting strategy, with limited data for Arts, Ecology, and Music. The text distribution across school grade levels is much more balanced but still has fewer examples for the 2nd and 6th grades.

Table 4. Class distribution for subjects after splitting the documents using the two approaches

Таблица 4. Распределение классов по темам после разделения документов с использованием двух подходов

Subject	Paragraph split	Greedy split
Art	2,306	143
Biology	29,612	3,247
Ecology	2,248	173
Geography	8,510	828
History	21,226	3,379
IT	11,277	608
Maths	33,618	1,493
Music	1,212	79
Physics	3,690	627
Rus	63,819	3,200
Science	9,906	685
Social studies	15,845	1,885
Technology	5,032	332

Table 5. Class distribution across grade levels after splitting the documents using the two approaches

Таблица 5. Распределение классов по уровням обучения после разделения документов с использованием двух подходов

	Paragraph split	Greedy split
2nd Grade	26,570	1,125
3rd Grade	31,535	1,538
4th Grade	42,854	2,225
5th Grade	21,107	1,593
6th Grade	10,187	1,041
7th Grade	23,539	2,132
8th Grade	16,356	1,826
9th Grade	11,541	1,455
10th Grade	14,027	2,190
11th Grade	10,585	1,554

We created 3 stratified folds with an 80-20% train-test split, used as independent evaluations that considered having different textbooks in the test set to avoid data contamination. As such, we ensured that each fold per subject or grade level had different books in the test set, with at least one for each subject, thus limiting the effect of artificially improving performance since similar paragraphs would have been encountered during training.

ReaderBench Feature Selection

We computed the textual indices for each classification unit using the previously mentioned text splits. Since none of the records represent an entire document, we discarded all document-based indices and kept only paragraph-, sentence-, and word-based values. Additionally, we removed any indices with a zero-standard deviation since these denote no variance and do not provide any valuable insights for classification.

Since the Shapiro-Wilk normality test (Shapiro and Wilk, 1965) did not confirm a normal distribution for most indices, we employed the Kruskal-Wallis (Kruskal and Wallis, 1952) to identify statistically significant indices for the target classes. Since our analysis is both horizontal based on the textbook topic and vertical, using the school grade for which the book was written, we employed two lists of indices, one for topic classification and another for grade classification. Also, we compile a subset for each list since the indices differ between the 2 splits (i.e., Paragraph and Greedy). As we can observe in Table 6 and in Table 7, splitting texts by paragraph leads to more indices being statistically non-significant.

Table 6. Statistically significant indices for Grade classification for each of the text-splitting strategies
Таблица 6. Статистически значимые индексы для уровневой (по классам) классификации для каждой из стратегий разделения текста

Topic Classification					
Paragraph Split			Greedy Split		
Textual Complexity Index	χ^2	*p*	Textual Index	χ^2	*P*
M(Dep amod / Par)	63733.98	<0.001	M(Dep amod / Par)	10676.51	<0.001
Max(Dep amod / Par)	63731.46	<0.001	M(UnqPOS adj / Par)	10522.63	<0.001
M(NmdEnt loc / Par)	62444.71	<0.001	M(UnqPOS noun / Par)	10481.48	<0.001
Max(NmdEnt loc / Par)	62443.97	<0.001	M(Dep nmod / Par)	10476.44	<0.001
Max(NmdEnt loc / Sent)	61867.77	<0.001	M(POS adj / Par)	10415.67	<0.001
M(NmdEnt loc / Sent)	61348.56	<0.001	M(POS noun / Par)	10172.37	<0.001
M(UnqPOS adj / Par)	60850.61	<0.001	M(UnqWd / Par)	10165.55	<0.001
Max(UnqPOS adj / Par)	60848.82	<0.001	M(Wd / Par)	9768.10	<0.001
M(POS adj / Par)	60014.28	<0.001	M(WdEntr / Par)	9629.86	<0.001
Max(POS adj / Par)	60012.54	<0.001	M(Dep amod / Sent)	9604.56	<0.001

Table 7. Statistically significant indices for Topic classification for each text splitting strategy
Таблица 7. Статистически значимые индексы для тематической классификации для каждой стратегии разделения текста

Grade Classification					
Paragraph Split			Greedy Split		
Textual Index	χ^2	*p*	Textual Index	χ^2	*p*
M(Dep amod / Par)	56631.90	<0.001	M(Dep amod / Sent)	10212.22	<0.001
Max(Dep amod / Par)	56628.59	<0.001	M(UnqPOS adj / Sent)	10055.40	<0.001
Max(Dep amod / Sent)	54922.33	<0.001	M(POS adj / Sent)	10007.35	<0.001
M(UnqPOS adj / Par)	54451.12	<0.001	M(Dep nmod / Par)	9953.53	<0.001
Max(UnqPOS adj / Par)	54447.95	<0.001	M(Dep amod / Par)	9718.38	<0.001
M(POS adj / Par)	53889.49	<0.001	M(UnqPOS adj / Par)	9630.25	<0.001
Max(POS adj / Par)	53886.36	<0.001	M(UnqWd / Sent)	9611.29	<0.001
Max(Chars / Word)	53661.26	<0.001	M(POS adj / Par)	9593.33	<0.001
Max(Syllab / Word)	52983.98	<0.001	M(Dep nmod / Sent)	9587.78	<0.001
Max(NgramEntr 2 / Word)	52260.14	<0.001	SD(Syllab / Word)	9429.66	<0.001

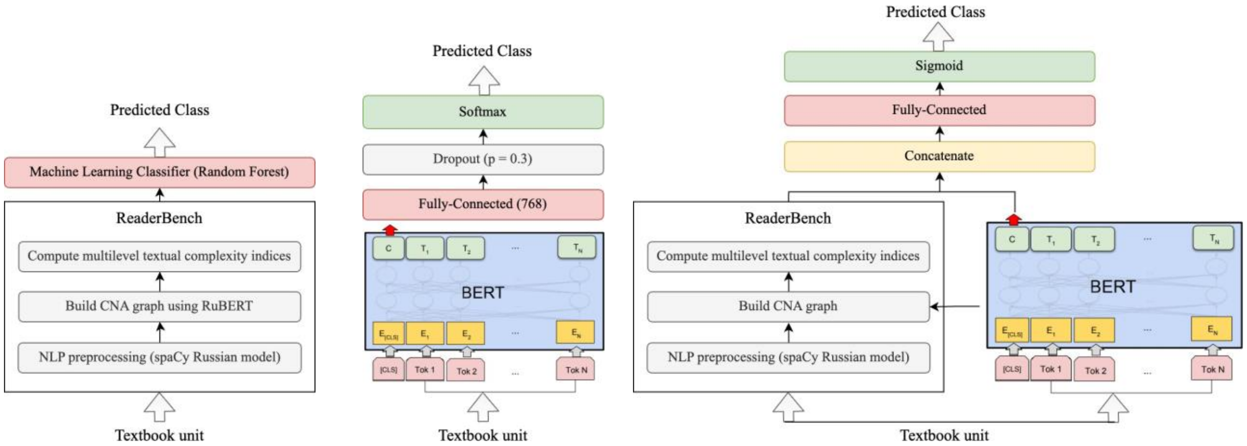
We observe that most nonpredictive indices are based on standard deviations, which is to be expected since this split leads to an increased number of shorter texts per document; thus, less variance is explained between the classification units. In contrast, the Greedy split leads to almost all indices being significantly different between the classes. According to the Kruskal-Wallis test, the most significant index was M (Dep amod / Par) for Topic and Grade classifications using both split text strategies. Afterward, Max (Dep amod / Par) for the Paragraph split and M (UnqPOS adj / Par) for the Greedy text split are the most predictive specific indices. Overall, differences are determined by the degree of descriptive elements from the text (i.e., NmdEnd_loc) and a more diverse vocabulary (i.e., WdEntr).

Classification Models

This study focuses on two types of multiclass classification for which we employ various methods (see Figure 1): based on the textbook topic (i.e., a 13-class classification) and the textbook school grade (i.e., a 10-class classification).

First, we consider a Random Forest classifier as a baseline to identify topics and grades based only on linguistic indices (see Figure 1.a). In order to identify a good set of hyperparameters, we perform a grid search over the number of estimators, minimum numbers of samples required to be a leaf, minimum number of samples required to split a node, number of features considered for the best split, and the maximum tree depth.

Figure 1. The three considered architectures
Рисунок 1. Три архитектуры исследования



(a) Random Forest Baseline

(b) RuBERT

(c) RuBERT+Features

Second, we use two BERT-based models. The first neural model uses only the RuBERT model and a linear, fully connected classifier with 768 dimensions over the pooled CLS token (see Figure 1.b) and a dropout layer with 0.3 probability. The second neural model concatenates the textual features for the classified samples to the pooled CLS token before being fed into the linear classifier (see Figure 1.c).

Experimental Setup

Both BERT-based models were trained using an AdamW optimizer with a learning rate of 1e-5. Since the classes were imbalanced, we used a weighted cross-entropy loss. Due to the difference in average text length between the Paragraph and the Greedy split, we used different maximum sequence lengths for each one, respectively 64 for the Paragraph and 512 for the Greedy split. The models were trained with early stopping for validation loss and patience of 2. Each model was trained on all three folds and we report the average values for accuracy, class-weighted average precision, recall, and F1 scores of all three runs.

Results

Table 8 and Table 9 present the classification results for each employed model. As expected, BERT-based models perform considerably better than classical machine learning models, such as Random Forest. Also, the way we pre-process the documents plays a significant role. The Greedy split performs notably better than a simple paragraph split for both topic and grade classifications. Since Transformer models create contextualized embeddings, providing larger windows for the classification of textbooks proves especially efficient. The performance gain for topic classification is more than 15% F1 score for the plain RuBERT model and over 12% for RuBERT enhanced with complexity indices. Similarly, the improvement is over 15% for both BERT-based models in the grade classification task.

Table 8. Classification results for topic classification. Metrics are computed as the average over three different folds
Таблица 8. Результаты классификации для тематической классификации. Метрики рассчитываются как среднее по трем архитектурам.

Model	Split strategy	Accuracy	Precision	Recall	F1
Random	Paragraph	50.27	51.28	50.27	50.39
Forest	Greedy	72.02	71.41	72.02	70.03
RuBERT	Paragraph	78.43	82.43	78.43	76.61
	Greedy	92.84	92.81	92.84	92.52
RuBERT+Features	Paragraph	78.93	82.20	78.93	79.91
	Greedy	**92.98**	**92.91**	**92.98**	**92.63**

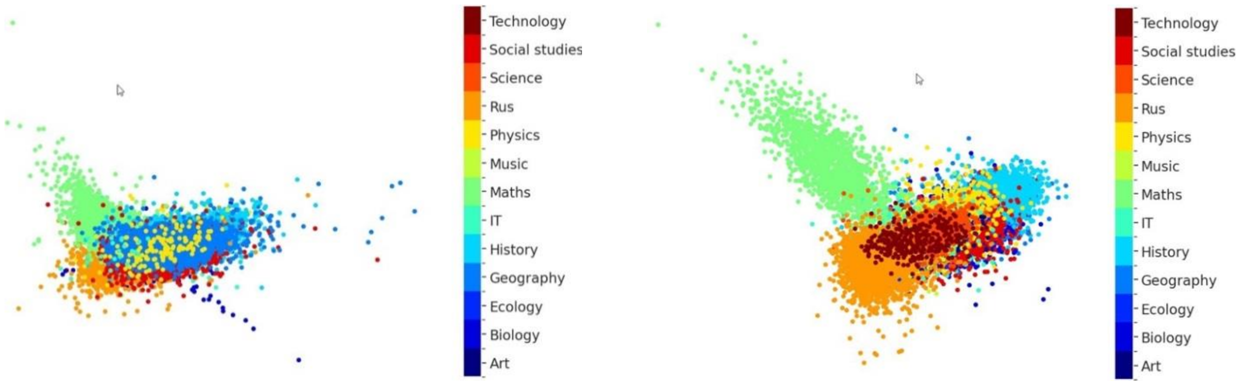
Table 9. Classification results for grade classification. Metrics are computed as the average over three different folds
Таблица 9. Результаты классификации по классам. Метрики рассчитываются как среднее по трем архитектурам.

Model	Split strategy	Accuracy	Precision	Recall	F1
Random	Paragraph	11.59	18.25	11.59	13.13
Forest	Greedy	13.42	48.48	13.42	20.54
RuBERT	Paragraph	35.96	41.24	35.96	36.79
	Greedy	52.07	55.93	52.07	51.99
RuBERT+Features	Paragraph	36.60	41.20	36.60	37.43
	Greedy	53.21	57.88	53.21	54.06

The difference between these two splitting strategies can also be observed in the Linear Discriminant Analysis (LDA) projection (Xanthopoulos et al., 2013) of the significant indices for topic classification. The classes are much better delimited for the Greedy text split (see Figure 2.b) than for the Paragraph split (see Figure 2.a).

Additionally, adding textual complexity indices greatly improves the grade classification performance, with over 2% weighted F1 score. For Topic classification, the indices improve the model only marginally since both BERT-based models perform exceptionally well, achieving an F1 score of over 92%.

Figure 2. LDA projection for the significant textual indices clustered for topic classification
Рисунок 2. Проекция LDA для значимых текстовых индексов, сгруппированных для тематической классификации.

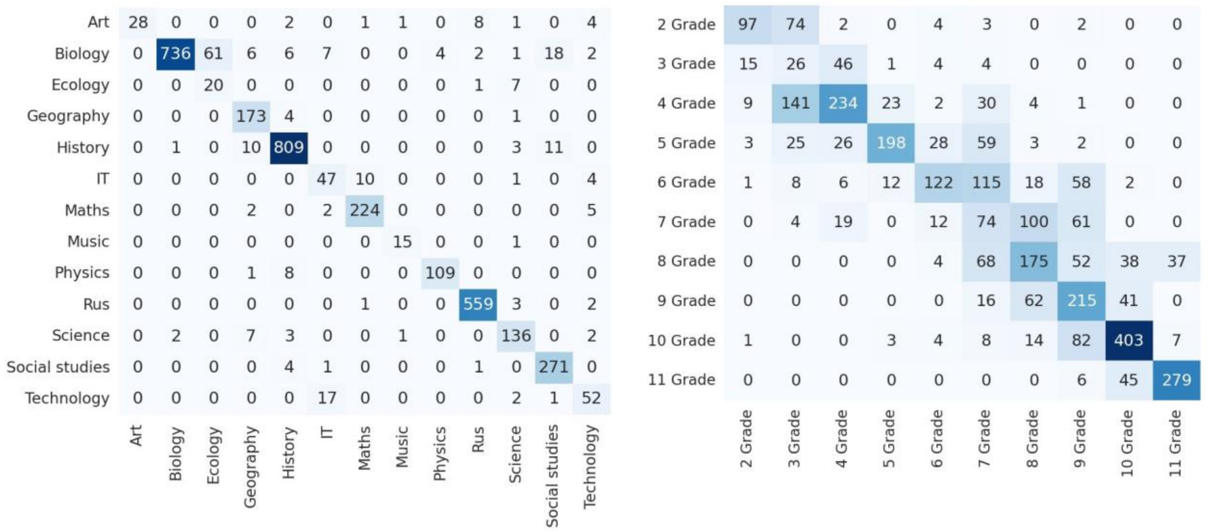


(a) Paragraph splitting
(a) Разбиение на абзацы

(b) Greedy text splitting
(b) разбиение текста при помощи жадного алгоритма

The confusion matrix for topic classification (see Figure 3.a) highlights that most errors are for Ecology, Geography, IT, and Social Studies. Since Ecology has only 3 textbooks, being the most imbalanced of the classes, a higher error rate was expected. Also, the model identified the Ecology fragments as Biology, which is plausible without a larger context. We also noticed the same type of error for IT, where the model placed 17 fragments as Technology. An interesting pattern of errors emerges in the grade classification confusion matrix in Figure 3.b. We can observe that the erroneous predictions tend to bleed into the neighboring grades, most going around the matrix diagonal. This is easily arguable since textbook complexity is on a continuum throughout school grades; as such, there should be no sudden jumps in complexity between consecutive grades. We can also observe that the 10th and 11th grade levels are best predicted since these levels have textbooks from only 4 of the 13 topics, and there is less noise due to the changing of domains between the fragments.

Figure 3. Confusion matrices for the best model considering RuBERT+Features
Рисунок 3. Матрицы смешивания для лучшей модели с учетом архитектуры RuBERT+Features



	Art	Biology	Ecology	Geography	History	IT	Maths	Music	Physics	Rus	Science	Social studies	Technology
Art	28	0	0	0	2	0	1	1	0	8	1	0	4
Biology	0	736	61	6	6	7	0	0	4	2	1	18	2
Ecology	0	0	20	0	0	0	0	0	0	1	7	0	0
Geography	0	0	0	173	4	0	0	0	0	0	1	0	0
History	0	1	0	10	809	0	0	0	0	0	3	11	0
IT	0	0	0	0	0	47	10	0	0	0	1	0	4
Maths	0	0	0	2	0	2	224	0	0	0	0	0	5
Music	0	0	0	0	0	0	0	15	0	0	1	0	0
Physics	0	0	0	1	8	0	0	0	109	0	0	0	0
Rus	0	0	0	0	0	0	1	0	0	559	3	0	2
Science	0	2	0	7	3	0	0	1	0	0	136	0	2
Social studies	0	0	0	0	4	1	0	0	0	1	0	271	0
Technology	0	0	0	0	0	17	0	0	0	0	2	1	52

(a) Topic classification
(а) Тематическая классификация

	2 Grade	3 Grade	4 Grade	5 Grade	6 Grade	7 Grade	8 Grade	9 Grade	10 Grade	11 Grade
2 Grade	97	74	2	0	4	3	0	2	0	0
3 Grade	15	26	46	1	4	4	0	0	0	0
4 Grade	9	141	234	23	2	30	4	1	0	0
5 Grade	3	25	26	198	28	59	3	2	0	0
6 Grade	1	8	6	12	122	115	18	58	2	0
7 Grade	0	4	19	0	12	74	100	61	0	0
8 Grade	0	0	0	0	4	68	175	52	38	37
9 Grade	0	0	0	0	0	16	62	215	41	0
10 Grade	1	0	0	3	4	8	14	82	403	7
11 Grade	0	0	0	0	0	0	0	6	45	279

(b) Grade classification
(b) Уровневая классификация (по классам)

Discussion

This study provides insights into the school textbook corpus of the Russian language that, in its current versions, spans 13 topics and 10 school grades. Our results show that Transformer-based models, such as RuBERT, can be used to identify the textbook topic with very high accuracy. We argue that textual complexity indices add to the robustness of the model and even increase its performance. Since the simple RuBERT model already achieves high accuracy (about 92.84%), any additional improvement was expected to be quite low.

The grade level classification has proven to be more difficult, with accuracy for the simple RuBERT model up to 52.07% over all the 10 classes. Here, we notice a much greater impact of the textual indices. Additionally, we show that our best-trained model struggles to differentiate between adjacent grade levels, with an adjacent accuracy for the best model reaching 85.61% when compared to its precise accuracy of 56.30%. This can be due to incremental increases in complexity between grades or to the practice of recapitulating some of the topics discussed in the previous grade at the beginning of the textbooks. This last assumption is supported by the fact that most erroneous classified fragments are from the first 50 paragraphs from all textbooks, with a median of 57, in contrast to the median of 69 paragraphs for the entire corpus. A Kruskal-Wallis test on paragraph identifiers with erroneous classification rejects the null hypothesis with $p < .05$ and $\chi2=24.84$, showing that paragraph identifiers with a higher error rate are lower on average is significant.

Conclusions and Future Work

In this study, we address the classification of Russian textbooks based on their topic and corresponding grade level. We show that using Transformer-based large language models supports the identification with very high accuracy of the school subject. Further, we present a classification method to predict the grade level of a text fragment with reasonably high accuracy. We show that both classification tasks achieved improved performance using the textual complexity indices from the open-source ReaderBench framework. Our best-performing BERT-based model enhanced with textual indices achieved a 92.63% F1 score on the 13 class topic classification and a 54.06% F1 score on the 10 class grade level classification.

In future work, we aim to improve the classification capabilities for grade-level detection by exploring further textbook datasets with more balanced coverage of the topics across all grade levels. Additionally, we will experiment with Graph Neural Networks like VGCN-BERT (Lu et al., 2020) that better capture the global information about the vocabulary, as well as large encoder-decoder language models like Flan-T5 (Chung et al., 2022) that were fine tuned on several tasks and achieved state-of-the-art performance. Lastly, we plan to expand the ReaderBench framework with indices proposed by (Solovyev et al., 2020 a, b) to cover Slavic languages better and enhance its multilingual capabilities.

References

Bansiong, A. J. (2019). Readability, content, and mechanical feature analysis of selected commercial science textbooks intended for third grade Filipino learners, *Cogent Education*, 6, 1706395. DOI: 10.1080/2331186X.2019.1706395 *(In English)*

Batinic, D., Birzer, S., Zinsmeister, H. (2017). Automatic classification of Russian texts for didactic purposes, *Trudy meždunarodnoj konferencii "Korpusnaja lingvistika"*, Sankt-Peterburg, Russia, 9-15. *(In English)*

Beníčková, Z., Vojíř, K. and Held, L., (2021). A comparative analysis of text difficulty in Slovak and Canadian science textbooks, *Chemistry-Didactics-Ecology-Metrology*, 26 (1-2), 89–97. DOI: 10.2478/cdem-2021-0007 (In English)

Bosco, G. L., Pilato, G. and Schicchi, D. (2021). Deepeva: A deep neural network architecture for assessing sentence complexity in Italian and English languages, *Array*, 12, 100097. *(In English)*

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C. and Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English, *Computational Linguistics*, 18, 31–40. *(In English)*

Chatzipanagiotidis, S., Giagkou, M. and Meurers, D. (2021). Broad linguistic complexity analysis for Greek readability classification, *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 48–58. *(In English)*

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S. et al. (2022). Scaling instruction- finetuned language models, arXiv preprint arXiv:2210.11416. https://doi.org/10.48550/arXiv.2210.11416 *(In English)*

Churunina, A., Solnyshkina, M., Gafiyatova, E. and Zaikin, A. (2020). Lexical features of text complexity: the case of Russian academic texts, *SHS Web of Conferences*, 88, 01009. https://doi.org/10.1051/shsconf/20208801009 *(In English)*

Corlatescu, D., Ruseti, S. and Dascalu, M. (2022). Readerbench learns Russian: Multilevel analysis of Russian text characteristics, *Russian Journal of Linguistics*, 26 (2), 342–370.

https://doi.org/10.22363/2687-0088-30145 *(In English)*

Crossley, S. A., Greenfield, J. and McNamara, D. S. (2008). Assessing text readability using cognitively based indices, *Tesol Quarterly*, 42, 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x *(In English)*

Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., Crossley, S. A. and Trausan-Matu, S. (2017). ReaderBench: a multi-lingual framework for analyzing text complexity, *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017*, Tallinn, Estonia, 495–499. https://doi.org/10.1007/978-3-319-66610-5_48 *(In English)*

Dascalu, M., McNamara, D. S., Trausan-Matu, S. and Allen, L. (2018). Cohesion network analysis of CSCL participation, *Behavior Research Methods*, 50, 604–619. https://doi.org/10.3758/s13428-017-0888-4 *(In English)*

Ivanov, V. V. (2022). Sentence-level complexity in Russian: An evaluation of BERT and graph neural networks, *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.1008411 *(In English)*

Khine, M. S. (2013). Analysis of science textbooks for instructional effectiveness, in Khine, M. S. (ed.), *Critical Analysis of Science Textbooks: Evaluating instructional effectiveness*, Springer, Dordrecht, Netherlands, 303-310. http://doi.org/10.1007/978-94-007-4168-3_15 *(In English)*

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog count and Flesch Reading Ease formula) for navy enlisted personnel, *Institute for Simulation and Training*, 56. *(In English)*

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American statistical Association*, 47, 583–621. https://doi.org/10.2307/2280779 *(In English)*

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language, arXiv preprint arXiv:1905.07213. https://doi.org/10.48550/arXiv.1905.07213 *(In English)*

Lu, Z., Du, P. and Nie, J. Y. (2020). VGCN-BERT: augmenting BERT with graph embedding for text classification, *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020*, Lisbon, Portugal, 12035, 369–382. https://doi.org/10.1007/978-3-030-45439-5_25 *(In English)*

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity, *Applied linguistics*, 30 (4), 555–578. https://doi.org/10.1093/applin/amp044 *(In English)*

Sakhovskiy, A., Solovyev, V. and Solnyshkina, M. (2020). Topic modeling for assessment of text complexity in Russian textbooks, *2020 Ivannikov Ispras Open Conference (ISPRAS)*, Moscow, Russia, 102–108. https://doi.org/10.1109/ISPRAS51486.2020.00022 *(In English)*

Santucci, V., Santarelli, F., Forti, L. and Spina, S., (2020). Automatic classification of text complexity, *Applied Sciences*, 10, 7285. https://doi.org/10.3390/app10207285 *(In English)*

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell System Technical Journal*, 27 (3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x *(In English*

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, 52 (3/4), 591–611. https://doi.org/10.2307/2333709 *(In English)*

Solovyev, V., Ivanov, V. and Solnyshkina, M. (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics, *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 34 (5), 3049–3058. https://doi.org/10.3233/JIFS-169489 *(In English)*

Solovyev, V. D., Ivanov, V V. and Akhtiamov, R. B. (2019). Dictionary of abstract and concrete words of the Russian language: a methodology for creation and application, *Research in Applied Linguistics*, 10, 218–230. https://doi.org/10.22055/RALS.2019.14684 *(In English)*

Solovyev, V., Solnyshkina, M., Gafiyatova, E., McNamara, D. and Ivanov, V. (2019). Sentiment in academic texts, *Proceedings of the 24th Conference of Open Innovations Association FRUCT*, IEEE Computer Society, Moscow, Russia, 408–414.

https://doi.org/10.23919/FRUCT.2019.8711900 *(In English)*

Solovyev, V., Ivanov, V. and Solnyshkina, M. (2020a). Thesaurus-based methods for assessment of text complexity in Russian, *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020*, Proceedings, Part II, Mexico City, Mexico, 152–166. https://doi.org/10.1007/978-3-030-60887-3_14 *(In English)*

Solovyev, V. D., Solnyshkina, M., Andreeva, M., Danilov, A. and Zamaletdinov, R. (2020b). Text complexity and abstractness: Tools for the Russian language, *Proceedings of the International Conference "Internet and Modern Society" (IMS-2020)*, St. Petersburg, Russia, 75–87. *(In English)*

Swanepoel, S. (2010). The assessment of the quality of science education textbooks: Conceptual framework and instruments for analysis, Ph.D. Thesis, University of South Africa, Pretoria, South Africa. *(In English)*

Wakefield, J. F. (2006). Textbook usage in the United States: The case of US history, *International Seminar on Textbooks*, Santiago, Chile, Online Submission. *(In English)*

Xanthopoulos, P., Pardalos, P. M. and Trafalis, T. B. (2013). Linear discriminant analysis, *Robust Data Mining*, Springer, 27–33. *(In English)*

Zipitria, I., Sierra, B., Arruarte, A. and Elorriaga, J. A. (2012). Cohesion grading decisions in a summary evaluation environment: A machine learning approach, *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34, 2615–2620. *(In English)*

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Andrei Paraschiv, Researcher, Computer Science and Engineering Department, Polytechnic University of Bucharest, Bucharest, Romania.

Андрей Параскив, исследователь, факультет информатики и инженерии, Политехнический университет Бухареста, Бухарест, Румыния.

Mihai Dascalu, Ph.D. (CS), Ph.D. (Edu), Professor, Dr., Department of Computers, Polytechnic University of Bucharest, Bucharest, Romania.

Михай Даскалу, доктор наук (Информационные технологии, Образование), профессор, профессор кафедры вычислительной техники, Бухарестский политехнический университет, Бухарест, Румыния.

Marina I. Solnyshkina, Doctor of Philology, Head and Chief Researcher, Text Analytics Research Laboratory, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia.

Марина Ивановна Солнышкина, доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, руководитель и главный научный сотрудник, НИЛ «Текстовая аналитика», Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Казань, Россия.