Text Complexity And Linguistic Features: Is The Relationship Multilingual?

Anonymous ACL submission

Abstract

Text complexity assessment is a challenging task requiring various linguistic aspects to be taken into consideration. A large number of studies have been introduced in this field. Nevertheless, as the methods and corpora are quite diverse, it may be hard to draw general conclusions as to the effectiveness of linguistic information for evaluating text complexity due to the diversity of methods and corpora. Moreover, a cross-lingual impact of different features on various datasets has not been investigated. We experimentally assessed seven commonly used feature types on six corpora for text complexity employing four common machine learning models. We showed which feature types can significantly improve the performance and analyzed their impact according to the dataset characteristics, language, and origin.

1 Introduction

Text complexity is critical for a comprehension process. Too difficult texts may be hard to understand. In contrast, unnecessarily simple texts conflict with the reader's level of communicative skills. Hence, text complexity assessment is an essential task that represents a major challenge for developing natural language processing (NLP) tools.

Text complexity can be expressed in different ways, ranging from quantitative characteristics to semantic complexity represented by text topics. Numerous studies have been published on evaluating various features for text complexity assessment. The reported results were obtained from text corpora of widely differing sizes and domains. Moreover, the authors used different machine learning (ML) models and text representation techniques. This makes it complicated to achieve an objective evaluation of the impact of different types of features.

In this work, we perform an extensive evaluation of seven feature types for text complexity assessment that were frequently used in research on the subject. We utilized six text complexity corpora in both English and Russian and four ML models in order to answer the following research questions (RQ). **RQ1**: How do different types of features affect the performance of baselines? **RQ2**: Are these feature types the same for different languages? **RQ3**: Do feature-enriched models outperform fine-tuned state-of-the-art language models?

2 Related Work

A number of various text characteristics have been used as complexity markers. First approaches were intuitive and limited by computational resources. The most of these readability indices represented linear combinations of both average word length and sentence length (Cantos and Almela, 2019). Despite their simplicity, these algorithms are still in use in some spheres, including government requirements for insurance[1].

Rapid development of NLP tools, including neural networks, has made it possible to significantly expand the set of features. Many authors have studied the impact of features of different nature. Thus, Feng et al. (2010) considered the number and density of named entities, semantic chains, referential relations, language modeling, syntactic dependencies, and morphology. Ivanov et al. (2018) studied average lengths, frequencies, morphological, and syntactic features in Russian corpora. Another challenging question is the robustness of different features across various corpora with texts of different languages, styles, and genres. This issue was partly solved by Isaeva and Sorokin (2021), who studied three groups of features, namely, average lengths plus frequencies, morphological, and syntactic ones. The experiments on three corpora of educational texts showed that there is a core of features which are crucial for all texts.

[1] https://www.cga.ct.gov/current/pub/chap_699a.htm#sec_38a-297

As in other many areas of NLP, state-of-the-art results can be achieved by fine-tuning BERT[2] (Devlin et al., 2019) and similar models. Martinc et al. (2021) studied unsupervised and supervised approaches, comparing BERT, HAN[3], and BiLSTM[4]. The experiments were conducted on a few English and Slovenian corpora. The results suggested that BERT can be used as a high-level baseline for our research.

3 Linguistic Features

According to the related works, we identified seven types of features: 1) *readability indices*, e.g., the Flesch–Kincaid readability test and the SMOG grade; 2) *traditional*, e.g., the average word length and type/token ratio; 3) *morphological*, e.g., the proportion of nouns and verbs; 4) *punctuation*. e.g., the number of semicolons; 5) *syntactic*, e.g., the average syntactic tree depth and number of clauses; 6) *frequencies*, e.g., the percentage of tokens included in the list of the most frequent words; and 7) *topic modeling*. In total, we collected 128 features for English and 126 for Russian without topic modeling. Additionally, we evaluated 100 topics using Latent Dirichlet allocation (Blei et al., 2003). To our knowledge, such a wide range of features was considered for the first time in relation to Russian text complexity models. A full list of features can be found in Appendix A.

4 Datasets and Models

4.1 Datasets

For the Russian language, there are few corpora with complexity labels. Therefore, we decided to experiment with one of such corpora, *Fiction Previews* **(Fic)** presented by Glazkova et al. (2021), and collect two additional ones. The texts and labels were extracted for one of them, named *Recommended Literature* **(RL)**, from the list of recommended literature for schoolchildren created by the Russian Ministry of Education. For the second one, *Books Read By Students* **(BR)**, we conducted a survey of schoolchildren of different ages and collected the most frequently mentioned texts. All collected texts were divided into fragments 70 sentences each. This allowed us to considerably increase the size of datasets without significant loss of labeling quality (Isaeva and Sorokin, 2021).

Characteristics	**RL**	**Fic**	**BR**
Total texts	9230	58184	5795
Total categories	3	2	5
Total words	4888290	26252666	2897003
Total unique words	103875	304731	55577
Avg words/text	1053.28	918.64	984.75
Avg words/sentence	14.95	13.12	13.92
Avg sentences/text	70	70	70
	CC	**CL**	**OSE**
Total texts	219	2834	567
Total categories	6	1	3
Total words	84014	491944	381137
Total unique words	10007	24449	13611
Avg words/text	450.46	199.65	757.82
Avg words/sentence	22.24	24.94	22.04
Avg sentences/text	23.26	9.46	34.98

Table 1: Some statistics of the datasets.

For English, there are a couple of corpora with complexity labels; we used three of them. *Common Core State Standards* **(CC)**[5] is a corpus designed to represent text complexity levels for each grade in the USA. *OneStopEnglish* **(OSE)** corpus was specially created for training readability models (Vajjala and Lucic, 2018). It consists of 189 text samples, each in three complexity versions. *CommonLit* **(CL)** corpus was presented at a Kaggle competition[6]. Continuous labeling is used in this corpus instead of classifying in the rest ones.

An overview of the datasets is shown in Table 1.

4.2 Models

1. **Linear Support Vector Classifier (LSVC).** LSVC was built with the l2 penalty and the squared hinge loss. We fitted LSVC on bag-of-words (BoW) text representations with a maximum length of 10000. Scikit-learn (Pedregosa et al., 2011) was used for implementation.

2. **Random Forest (RF).** We used 100 estimators and the Gini impurity to measure the quality of a split. The implementation details are the same as those for LSVC.

3. **Feedforward Neural Network (FNN).** The hyperparameters used are identified in Table 2. We employed the Adam optimizer (Kingma and Ba, 2015). The model was implemented using Keras (Chollet et al., 2015). Each model was trained with early stopping for a maximum of 100 epochs and patience

[2]Bidirectional Encoder Representations from Transformers

[3]Hierarchical attention networks

[4]Bidirectional Long short-term memory networks

[5]http://www.corestandards.org/

[6]https://www.kaggle.com/c/commonlitreadabilityprize

Hyperparameters	FNN	CNN
Number of convolutional layers	-	2
Number of pooling layers	-	2
Number of convolutional filters	-	256
Filter size	-	256
Number of fully connected layers	3	1
Size of fully connected layers	1024	32
Activation (hidden layers)	tanh	relu
Dropout	0.5	

Table 2: Hyperparameters for neural baselines.

of 20. We utilized Sentence Transformers text representations obtained using the all-mpnet-base-v2 model (Reimers and Gurevych, 2019) for the English corpora and the distiluse-base-multilingual-cased model (Reimers and Gurevych, 2020) for the Russian ones.

4. **Convolutional Neural Network (CNN).** The training details are the same as for FNN. The model was implemented using FastText embeddings for English (Mikolov et al., 2018) and Russian (Kutuzov and Kuzmenko, 2016).

We randomly shuffled all the Russian corpora and the CL dataset and split them into train and test sets in the ratio of 3:1. The splitting was conducted in such a way that all fragments of one book belonged either to the train set or to the test one. Due to the small number of documents in OSE and CC corpora, we used five-fold cross-validation on these datasets to obtain more reliable results. For all of the models above, we systematically evaluated each type of linguistic features applying the Min-Max technique for normalization.

To compare the scores obtained with the results of a few state-of-the-art models, we used BERT-base and RuBERT (Kuratov and Arkhipov, 2019) for English and Russian corpora respectively. Each model was fine-tuned for 3 epochs using Transformers (Wolf et al., 2020).

Model	RL	Fic	BR	CC	CL	OSE
BERT	62.74	80.96	45.23	42.18	**0.453**	70.99
LSVC	63.16	76.66	32.31	28.22	0.673	70.41
RF	48.21	78.87	30.94	30.03	0.627	68.21
FNN	63.26	66.34	34.22	33.73	0.533	54
CNN	58.12	80.12	39.82	33.6	0.593	70.64
+ readability						
LSVC	63.16	76.67	32.12	32.43	0.663	70.49
RF	49.89	78.45	29.19	27.77	0.599	70.11
FNN	63.62	68.23	40.89	37.56	0.502	56.07
CNN	61.35	80.52	45.9	35.89	0.59	68.59
+ traditional						
LSVC	62.67	77.14	33.15	29.3	0.666	69.89
RF	46.53	78.26	30.03	28.57	0.609	**73.01**
FNN	**69.76**	70.51	32.12	34.7	0.482	58.76
CNN	65.19	80.68	44.32	**45.98**	0.604	64.82
+ morphological						
LSVC	63.22	77.03	32.55	31.99	0.662	71.75
RF	46.63	76.2	30.36	29.56	0.611	70.67
FNN	69.12	72.04	35.63	37.42	0.504	62
CNN	68.63	80.75	42.29	37.12	0.573	69.02
+ punctuation						
LSVC	62.87	76.73	32.26	30.44	0.664	70.41
RF	47.25	78.2	30.3	28.39	0.629	68.92
FNN	66.54	68.7	35.21	32.51	0.505	55.79
CNN	67.95	80.86	40.74	43.68	0.58	64.33
+ syntactic						
LSVC	61.91	76.88	32.66	29.27	0.674	70.54
RF	46.7	77.41	28.84	33.97	0.617	72.59
FNN	69.41	68.31	32.1	36.48	0.476	56.68
CNN	65.35	81.01	45.49	36.19	0.592	58.71
+ frequency						
LSVC	63.07	76.84	32.52	33.08	0.662	71.34
RF	45.87	77.76	30.01	26.02	0.64	67.63
FNN	67.46	67.58	31.45	35.33	0.729	63.01
CNN	65.08	**81.11**	**46.97**	38.65	0.597	56.38
+ topic modeling						
LSVC	62.14	76.92	35.36	29.97	0.669	67
RF	49.44	77.65	34.09	27.15	0.623	66.1
FNN	62.01	77.3	38.85	34.08	0.516	59.46
CNN	65.78	80.91	43.93	41.28	0.588	64.95

Table 3: F1 (%) and MAE for each type of features.

5 Results and Discussion

We report the results in terms of the mean absolute error (MAE, for the CL corpus) and weighted F1-score (for the other corpora) in Table 3. The gray highlight presents the values that outperform the baseline. The best results are shown in bold. Appendix B contains the overall results expressed by several common metrics.

Based on the results, we can identify four performance categories, see Table 4, that describe the impact of various linguistic features (**RQ1**). In most cases, all the considered features improved the model performance on all datasets. Meanwhile, it was only morphological features that gave a positive impact in most classifiers for all corpora. Readability features exceeded the baseline on most models for most datasets except the BR corpus. Punctuation, traditional, and syntactic features showed a performance growth at least for two models on each corpus. Frequency and topic modeling features produced mixed results. On the one hand, topic modeling features improved the performance of all classifiers on two corpora. Nevertheless, the score on the OSE corpus increased for only one model. This could be because the corpus contains parallel versions of the same papers. Although frequency features improved the performance in some cases, they demonstrated higher MAE in most classifiers on the CL dataset. Probably, it reflects the

Improvement	RL	Fic	BR
All classifiers	-	-	7
3 classifiers	1,3	1,2,3,4,5,6,7	3
2 classifiers	2,4,5,6,7	-	1,2,4,5,6
1 classifier	-	-	-
	CC	**CL**	**OSE**
All classifiers	5	1,3,7	-
3 classifiers	1,2,3,6,7	2,4,5	1,3,5
2 classifiers	4	-	2,4,6
1 classifier	-	6	7

Table 4: Performance categories on all corpora. Correspondence of linguistic feature types is in Sec.3.

Features	RL	Fic	BR	Avg Rus
Readability	2.4	0.71	7.13	**3.41**
Traditional	4.54	1.71	1.21	**2.49**
Morphological	6.04	1.62	2.3	**3.32**
Punctuation	4.91	0.93	0.75	**2.2**
Syntactic	4.26	0.63	0.58	**1.83**
Frequency	3.4	0.48	1.88	**1.92**
Topic modeling	3.04	4.07	10.87	**5.99**
	CC	**CL**	**OSE**	**Avg Eng**
Readability	6.39	3.05	0.96	**3.47**
Traditional	9.67	3.04	-1.33	**4.74**
Morphological	8.3	3.19	4.51	**5.33**
Punctuation	7.2	2.06	-1.14	**2.71**
Syntactic	8.18	3.04	-1.33	**3.3**
Frequency	5.91	-9.54	-0.76	**-1.46**
Topic modeling	5.13	1.3	-1.47	**1.65**

Table 5: Averaged performance growth, %.

Model	RL	Fic	BR	CC	CL	OSE
LSVC	-	78.09	34.5	33.12	0.633	71.44
RF	49.38	-	-	-	0.568	76.44*
FNN	62.99	78.7	40.88	39.71	0.466	74.24
CNN	65.29	81.06	43.85	43.58	0.541	-

Table 6: F1 (%) and MAE for feature combinations.

fact that short texts normally lack word frequency and context information because of word sparsity (Yan et al., 2013; Xun et al., 2016).

Table 5 illustrates the performance growth as a percentage averaged over all classifiers for Russian and English corpora (**RQ2**). The averaged results demonstrate that the models trained on Russian texts benefit more from topic modeling and frequency features in comparison with the models trained on English corpora. On the other hand, the results on the CC corpus indicate that this superiority is rather due to text length than language properties. Readability and punctuation features present similar results for both languages. Although morphological, traditional, and syntactic features show better performance on English texts, the results on specific corpora are strongly determined by the source of texts and the type of markup. Thus, any influence of syntactic features for the OSE corpus could not be identified during our experiments. However, there was a significant increase for the CC corpus containing fiction texts that characterized English as an analytic language. Overall, these results indicate that the impact of all feature types is mainly attributable to specific circumstances of a corpus. This enables one to use transfer-learning algorithms for cross-lingual analysis of text corpora having similar characteristics.

The performance of the models trained on feature combinations per dataset is presented in Table 6. The results are given only for those models whose performance was increased by two and more types of features. We enriched the baseline models with the concatenation of all features that showed a positive impact for the relevant models and datasets. The combination of features increased the F1 of RF on the OSE corpus outperforming all the results obtained for this dataset. This result is marked with an asterisk (*). Moreover, FNN trained on feature combinations showed the best result among all the feature-enriched models on the CL corpus. Taken together, the results presented in Table 3 and Table 6 demonstrate that feature-enriched models outperformed BERT on five out of the six corpora (**RQ3**). In some cases, significant increases were obtained, including 7.02% for the RL corpus and 3.8% for the CC corpus. By contrast, the performance of feature-enriched models depends on the features used and data specifics.

6 Conclusion

We have presented the first comparative analysis of various linguistic features on six corpora in terms of text complexity assessment. Each feature type was evaluated in four representative ML models. Our research demonstrated the superiority of some features over others. We also identified performance categories based on the scores obtained and estimated the impact of feature combinations. According to out study, the results depend more on the dataset specificity rather than on language. This provides an opportunity for exploring cross-lingual transfer learning and multi-lingual models for text complexity assessment. Finally, experimental results on most corpora showed that feature-enriched models can achieve significant improvements in comparison with the state-of-the-art ones. Here, future research may focus on evaluating more complex semantic and narrative features and on explaining text complexity in terms of each feature type.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yurii Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

Pascual Cantos and Ángela Almela. 2019. Readability indices for the assessment of textbooks: a feasibility study in the context of efl. *Vigo International Journal of Applied Linguistics*, pages 31–52.

Francois Chollet et al. 2015. Keras.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noemie Elhadad. 2010. A comparison of features for automatic readability assessment. pages 276–284.

Anna Glazkova, Yury Egorov, and Maksim Glazkov. 2021. A comparative study of feature types for age-based text classification. In *Analysis of Images, Social Networks and Texts*, pages 120–134, Cham. Springer International Publishing.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ulyana Isaeva and Alexey Sorokin. 2021. *Investigating the Robustness of Reading Difficulty Models for Russian Educational Texts*, pages 65–77.

V. V. Ivanov, M. I. Solnyshkina, and V. D. Solovyev. 2018. Efficiency of text readability features in russian academic texts. page 267–283.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 320–332. Springer.

Y. Kuratov and M. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 333–339.

Andrey Kutuzov and Elizaveta Kuzmenko. 2016. Webvectors: a toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 155–161. Springer.

Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. Word frequencies in written and spoken english: based on the british national corpus.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Matej Martinc, Senja Pollak, and Marko Robnik-Sikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47:1–39.

G. Harry McLaughlin. 1969. Smog grading - a new readability formula. *Journal of reading*, 12(8):639–646.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Radim Rehurek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from gensim.org*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

R.J. Senter and E. A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. Readability formula for russian texts: A modified version. In *Mexican International Conference on Artificial Intelligence*, pages 132–145. Springer.

Mildred C. Templin. 1957. Certain language skills in children; their development and interrelationships.

Sowmya Vajjala and Ivana Lucic. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304. Association for Computational Linguistics.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Guangxu Xun, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao, and Aidong Zhang. 2016. Topic discovery for short texts using word embeddings. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1299–1304. IEEE.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

A Evaluated features

For evaluation we used the following libraries: readability[7], pymorphy2 (Korobov, 2015), nltk (Loper and Bird, 2002), gensim (Rehurek et al., 2011), spacy (Honnibal and Montani, 2017), deeppavlov (Burtsev et al., 2018), and API of readability.io. The source code for our methods is available at: `https://github.com/pretty-nickname/readability`.

[7] https://pypi.org/project/readability/

A.1 Readability indices

1. Flesch–Kincaid readability test (Kincaid et al., 1975).
2. Coleman–Liau index (Coleman and Liau, 1975).
3. Automated readability (ARI) index (Senter and Smith, 1967).
4. SMOG grade (McLaughlin, 1969).
5. Dale-Chall index (Dale and Chall, 1948).

A.2 Traditional features

1. Average and mean sentence length.
2. Average and mean word length.
3. Long words (>4 syllables) proportion.
4. Type/token ratio (TTR) (Templin, 1957).
5. NAV: TTR for Nouns only plus TTR for Adjectives only divide by TTR for Verbs only (Solnyshkina et al., 2018).

A.3 Morphological features

1. Percentages of lexical categories.
2. Percentage of grammatical cases.
3. Proportion of animated nouns.
4. Proportion of grammatical aspects for verbs.
5. Proportion of grammatical tences for verbs.
6. Proportion of transitive verbs.

A.4 Punctuation

1. Punctuation/token ratio.
2. Semicolons/token ratio.

A.5 Syntactic features

Three features were extracted from each of the following characteristics: average, mean, and maximum.

1. Syntactic tree depth.
2. Distance between a node and its descendant.
3. Number of clauses.
4. Number of adverbial clause modifiers.

5. Number of adnominal clauses.
6. Number of clausal complements.
7. Number of open clausal complements.
8. Number of nominal modifiers.
9. Length of nominal modifiers sequence.

A.6 Frequencies

For evaluating frequencies of Russian and English words we used dictionaries based on Russian National Corpus[8] and British National Corpus (Leech et al., 2001) respectively.

1. Average and mean frequency.
2. Proportion of words, which are in the list of the most 100/200/.../1000 popular words, and similar features for nouns, verbs, adverbs, and adjectives separately.

B Overall Results

B.1 Russian Corpora

Model	F	P	R			
BERT	62.74	65.71	61.86			
LSVC	63.16	63.54	64.98			
RF	48.21	61.8	59.92			
FNN	63.26	79.19	53.76			
CNN	58.12	58.23	58.99			
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	63.16	63.22	64.89	62.67	62.83	64.56
RF	49.89	63.88	60.68	46.53	55.2	58.82
FNN	63.62	81.66	52.91	**69.76**	**93.52**	56.03
CNN	61.35	66.33	59.49	65.19	66.22	64.64
	+ morphological			+ punctuation		
LSVC	63.22	63.11	64.98	62.87	63.07	64.73
RF	46.63	58.54	59.07	47.25	62.9	59.58
FNN	69.12	92.34	55.78	66.54	87.1	54.43
CNN	68.63	72.84	**66.58**	67.95	71.19	66.33
	+ syntactic			+ frequency		
LSVC	61.91	61.88	63.88	63.07	62.93	64.64
RF	46.7	57.58	58.9	45.87	57.89	58.65
FNN	69.41	93.01	55.78	67.46	89.35	54.6
CNN	65.35	69.58	63.21	65.08	66.22	64.64
	+ topic modeling			combined		
LSVC	62.14	62.71	64.22	-	-	-
RF	49.44	65.98	60.68	49.38	62.93	60.34
FNN	62.01	65.98	59.66	62.99	68.99	58.99
CNN	65.78	67.24	64.89	65.29	68.18	63.54

Table 7: Results for the Recommended Literature corpus: F - F1-score weighted, P - precision weighted, R - recall weighted. %.

Model	F	P	R			
BERT	80.96	81.83	80.82			
LSVC	76.66	77.89	76.87			
RF	78.87	79.67	78.99			
FNN	66.34	72.31	65.01			
CNN	80.12	80.87	80.04			
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	76.67	77.84	76.87	77.14	78.29	77.34
RF	78.45	78.85	78.51	78.26	78.86	78.36
FNN	68.23	72.54	67.36	70.51	70.61	70.49
CNN	80.52	81.9	80.37	80.68	81.74	80.56
	+ morphological			+ punctuation		
LSVC	77.03	78.27	77.24	76.73	77.94	76.94
RF	76.2	77.16	76.38	78.2	78.93	78.32
FNN	72.04	72.09	72.04	68.7	68.75	68.69
CNN	80.75	81.73	80.65	80.86	81.84	80.75
	+ syntactic			+ frequency		
LSVC	76.88	78.08	77.09	76.84	78	77.04
RF	77.41	78.21	77.54	77.76	78.4	77.86
FNN	68.31	68.41	68.29	67.58	67.59	67.57
CNN	81.01	81.97	80.9	**81.11**	82.08	**81.01**
	+ topic modeling			combined		
LSVC	76.92	78.18	77.12	78.09	79.3	78.27
RF	77.65	78	77.71	-	-	-
FNN	77.3	78.28	77.17	78.7	79.06	78.66
CNN	80.91	82.07	80.78	81.06	**82.17**	80.93

Table 8: Results for the Fiction Previews corpus.

Model	F	P	R			
BERT	45.23	54.06	**41.32**			
LSVC	32.31	35.74	34.28			
RF	30.94	32.73	37.18			
FNN	34.22	39.06	31.75			
CNN	39.82	57.34	33.66			
	F	P	R	F	P	R
	+ readability			+ traditional		
LSVC	32.12	35.5	34.2	33.15	36.79	35.2
RF	29.19	26.87	36.04	30.03	32.49	36.34
FNN	40.89	61.23	31.83	32.12	44.37	27.08
CNN	45.9	66.18	37.8	44.32	64.88	36.27
	+ morphological			+ punctuation		
LSVC	32.55	37.52	36.5	32.26	35.79	34.35
RF	30.36	37.52	36.5	30.3	37.94	36.57
FNN	35.63	42.75	31.68	35.21	39.54	33.05
CNN	42.29	55.72	37.26	40.74	57.25	33.44
	+ syntactic			+ frequency		
LSVC	32.66	36.02	34.66	32.52	35.77	34.28
RF	28.84	31.26	34.74	30.01	32.14	35.88
FNN	32.1	40.95	28.46	31.45	37.54	28.39
CNN	45.49	67.47	36.57	**46.97**	**69.57**	38.41
	+ topic modeling			combined		
LSVC	35.36	38.63	36.88	34.5	37.36	35.88
RF	34.09	37.74	38.18	-	-	-
FNN	38.85	45.77	35.96	40.88	55.03	35.96
CNN	43.93	62.93	36.65	43.85	62.93	37.18

Table 9: Results for the Books Read By Students corpus.

B.2 English Corpora

[8] http://dict.ruslang.ru/freq.php

Model	F		P		R	
BERT	42.18		64.57		33.77	
LSVC	28.22		30.13		30.61	
RF	30.03		30.38		34.65	
FNN	33.73		37.93		32.9	
CNN	33.6		58.04		26.92	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			+ traditional		
LSVC	32.43	33.55	35.59	29.3	31.37	31.5
RF	27.77	26.95	31.95	28.57	28.68	32.88
FNN	37.56	42.34	36.53	34.7	38.48	34.28
CNN	35.89	56.83	29.25	**45.98**	**78.2**	36.12
	+ morphological			+ punctuation		
LSVC	31.99	35.29	33.33	30.44	32.07	33.32
RF	29.56	29.53	34.26	28.39	27.23	34.65
FNN	37.42	46.15	34.7	32.51	37.2	32
CNN	37.12	57.32	30.62	43.68	60.51	37.95
	+ syntactic			+ frequency		
LSVC	29.27	29.45	31.95	33.08	35.74	34.67
RF	33.97	34.75	38.33	26.02	23.17	31.55
FNN	36.48	41.42	35.64	35.33	40.79	34.27
CNN	36.19	62.18	28.3	38.65	54.04	32.45
	+ topic modeling			combined		
LSVC	29.97	31.38	32.42	33.12	35.21	34.67
RF	27.15	29.19	30.15	-	-	-
FNN	34.08	38.34	32.91	39.71	47.55	37.94
CNN	41.28	65.93	33.85	43.58	44.09	**39.44**

Table 10: Results for the Common Core State Standards corpus.

Model	MAE		MSE	
BERT	**0.4532**		**0.3159**	
LSVC	0.6728		0.695	
RF	0.6266		0.6199	
FNN	0.533		0.4421	
CNN	0.5926		0.555	
	MAE	**MSE**	**MAE**	**MSE**
	+ readability		+ traditional	
LSVC	0.6627	0.6742	0.6664	0.6819
RF	0.5986	0.5743	0.609	0.5831
FNN	0.5024	0.4045	0.4823	0.3832
CNN	0.5896	0.5496	0.6041	0.5813
	+morphological		+ punctuation	
LSVC	0.6621	0.6775	0.664	0.6785
RF	0.6113	0.5917	0.6288	0.6204
FNN	0.5042	0.4002	0.5053	0.4102
CNN	0.5728	0.5269	0.5803	0.5307
	+ syntactic		+ frequency	
LSVC	0.6741	0.6924	0.6619	0.6703
RF	0.6167	0.5853	0.6401	0.643
FNN	0.4759	0.3705	0.7293	0.7627
CNN	0.5923	0.5566	0.5973	0.5602
	+ topic modeling		combined	
LSVC	0.6686	0.6861	0.6334	0.6166
RF	0.623	0.5986	0.568	0.5174
FNN	0.5156	0.4149	0.4658	0.3542
CNN	0.5882	0.5403	0.5408	0.4726

Table 11: Results for the CommonLit corpus: MAE - mean absolute error, MSE - mean squared error.

Model	F		P		R	
BERT	70.99		78.15		69.34	
LSVC	70.41		72.15		72.03	
RF	68.21		70.44		69.85	
FNN	54		56.34		52.83	
CNN	70.64		**84.44**		65.23	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			+ traditional		
LSVC	70.49	72.17	72.02	69.89	71.76	71.69
RF	70.11	71.63	71.83	73.01	74.89	74.45
FNN	56.07	59.02	54.59	58.76	62.86	57.18
CNN	68.59	76.29	67.37	64.82	77.32	60.71
	+ morphological			+ punctuation		
LSVC	71.75	73.65	73.39	70.41	72.15	72.03
RF	70.67	72.22	72.25	68.92	70.24	70.4
FNN	62	65.37	60.19	55.79	57.56	54.8
CNN	69.02	78.87	66.33	64.33	75.55	60.33
	+ syntactic			+ frequency		
LSVC	70.54	72.61	72.37	71.34	73.1	73.04
RF	72.59	73.67	73.82	67.63	68.8	69.89
FNN	56.68	77.87	49.85	63.01	65.63	61.68
CNN	58.71	73.85	54.88	56.38	68.41	53.15
	+ topic modeling			combined		
LSVC	67	68.9	69.14	71.44	72.96	73.07
RF	66.1	68.1	66.45	**76.44**	77.18	**77.37**
FNN	59.46	61.84	58.38	74.24	75.71	74.17
CNN	64.95	76.98	62.17	-	-	-

Table 12: Results for the OneStopEnglish corpus.