Using Topic Modeling to Improve the Quality of Age-Based Text Classification

Anna Glazkova[a]

[a] *University of Tyumen, 6, Volodarskogo street, Tyumen, 625003, Russia*

Abstract
The prediction of the age audience of the text plays a crucial role in selecting information suitable for children, book publishing, and editing. In this paper, we evaluate the impact of document topic distribution vectors on the quality of age-based text classification. We formulated this problem as a binary classification task and developed a topic-informed machine learning classifier for resolving this problem. We compared three common topic modeling techniques to obtain document topic distribution vectors, including Latent Dirichlet Allocation, Gibbs Sampled Dirichlet Multinomial Mixture, and BERTopic. In most cases, our topic-informed classifier achieved improvements on a dataset of Russian fiction abstracta over baseline approaches.

Keywords 1
Topic model, text classification, LDA, GSDMM, BERTopic.

1. Introduction

Text difficulty assessment is one of the main tasks in computational linguistics and natural language processing. The difficulty of a text is determined by the combination of all text aspects that affects the reader's understanding, reading speed, and level of interest in the text [1]. There is evidence that the tools for text difficulty assessment play a crucial role in regulating children's access to suitable information, selecting relevant literature, or automating some aspects of editorial and publishing activities.

There is a large volume of published studies describing the role of various linguistic features in determining the reading levels of text. The first serious discussions and analyses of text difficulty emerged in the middle of the last century with the creating of readability indices [2-3]. In recent years, researchers explored the impact of lexical [4-6], morphological [6-7], semantic [7], syntactic [6, 8-9], and psycholinguistic [10-11] features on the quality of text difficulty assessment. The study [12] presented a comparison of Russian book abstracts assigned to different age ratings using unsupervised topic modeling. Another recent study [13] explores the problem of assessing the complexity of Russian educational texts.

In this work, we evaluate the effectiveness of topic modeling features for age-based text classification of Russian books. The age-based classification is a specific task of determining the text difficulty. Its goal is to predict the estimated age audience of the text. We use the corpus of abstracts of fiction books [14]. Each abstract has a reader age label: adult or children's. We use these labels as indicators of text difficulty. Further, we obtain document topic distribution vectors for abstracts using three common topic modeling approaches, such as a) Latent Dirichlet Allocation (LDA); b) Gibbs Sampled Dirichlet Multinomial Mixture (GSDMM); c) BERTopic, an algorithm for generating topics using state-of-the-art embeddings. We evaluate the impact of topic modeling features on several machine learning methods, including Logistic Regression (LR), Linear Support Vector Classifier

VI International Conference Information Technologies and High-Performance Computing (ITHPC-2021),
September 14–16, 2021, Khabarovsk, Russia
EMAIL: a.v.glazkova@utmn.ru
ORCID: 0000-0001-8409-6457

(LSVC), and Multilayer Perceptron neural network (MLP). In most cases, our topic-informed classifiers outperform the baselines.

The rest of the paper is structured as follows. In Section 2, we describe our methodology. Section 3 provides evaluation results. Section 4 concludes this paper. Section 5 contains acknowledgments.

2. Methods

We apply three machine learning classifiers based on Logistic Regression, Linear Support Vector Classifier, and Multilayer Perceptron with lexical features. Lexical features were obtained only from the 5000 top words ordered by term frequency across the corpus. We produced a sparse representation of the word counts (the bag-of-words model) and used it as an input for each classifier. The text preprocessing for the bag-of-words model consisted of the four steps, which are: a) removing special characters and digits; b) converting to lowercase; c) lemmatization using pymorphy2 [15]; d) removing stopwords and short words containing fewer than 3 characters. The methods were implemented with classes from the Scikit-learn library [16] using the following parameters:

1. LR: “l2” penalization, tolerance for stopping criteria is 1e-5.
2. LSVC: “l2” penalization, tolerance for stopping criteria is 1e-5.
3. MLP: 2 hidden ReLU layers of 2000 and 1000 neurons respectively, the solver is an L-BFGS method [17]. We trained the model for 10 epochs.

The classifiers described above were used as baselines. Further, we obtained topic distribution vectors for each document in the corpus. The document topic distribution vector represents the topic distribution in the text by the word frequency. We concatenated the topic distribution vector with a corresponding lexical vector (Figure 1) and evaluated the benefits of topic-informed models. Document topics distribution vectors were obtained using three common types of probabilistic topic models:

1. Latent Dirichlet Allocation [18]. LDA is a two-level Bayesian generative model, which assumes that topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions [19]. In this work, the LDA topic model was implemented using Gensim [20].
2. Gibbs Sampled Dirichlet Multinomial Mixture [21]. GSDMM is a short text clustering model. This technique is essentially a modified LDA assuming that a document encompasses only one topic. This differs from LDA which assumes that a document can have multiple topics.
3. BERTopic [22], which is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters. This approach performs three main steps: a) extracting document embeddings using state-of-the-art language models; b) clustering document embeddings to create groups of similar documents with UMAP [23] and HDBSCAN [24] algorithms; c) extracting topics by getting the most important words per cluster with class-based TF-IDF (c-TF-IDF).

To preprocess texts for LDA and GSDMM, we first performed the four preprocessing steps mentioned above and then built bigrams for collocated words with a total collected count of more than 5 and a threshold equal to 100. When applying the BERTopic technique, we used a multilingual version of BERT (Bidirectional Encoder Representations from Transformers)[2] [25] to produce document embeddings.

[2] https://huggingface.co/bert-base-multilingual-cased

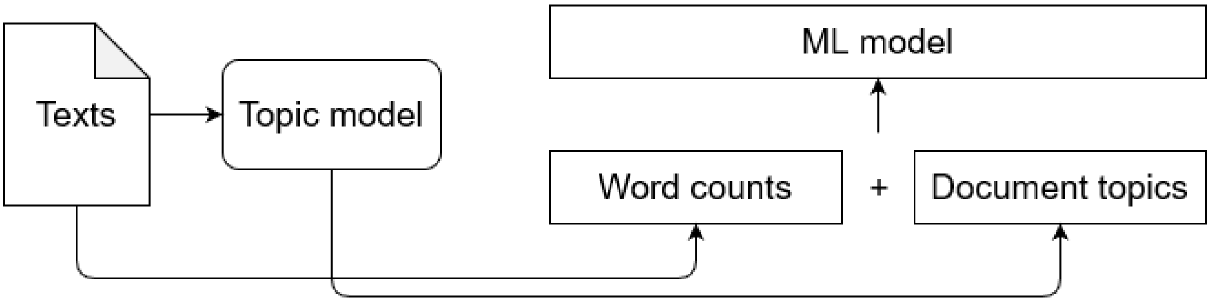

Figure 1: Scheme of topic-informed model

3. Experiments

In this section, we describe our experiments with baseline classifiers and topic-informed models.

3.1. Evaluation dataset

We conducted experiments on the corpus of abstracts of fiction books[3] which is a part of the Russian corpus for age-based text classification [14]. The corpus consists of annotated fiction abstracts from online libraries. Table 1 presents the summary statistics for our data. The number of tokens and sentences is evaluated using the NLTK tokenizer [26].

Table 1
Characteristics of data

Sample	Number of texts	Avg length of texts (tokens)	Avg number of sentences
Train	4646 Adult: 2688 Children's: 1958	106,38	5,52
Test	800 Adult: 189 Children's: 611	110,14	5,66

3.2. Results

We performed model training on the training sample and tested our models on the test sample. We computed recall (R), precision (P), and F1-scores (F), weighted by the number of true instances for each label (weighted recall, precision, and F1-score). The results are shown in Table 2. In brackets, we clarified the increase in F1-scores for topic-informed models relative to the relevant baselines. For each classifier, we evaluated LDA and GSDMM topic models with a number of requested latent topics equal to 25, 50, 75, and 100. We also estimated document topic vectors obtained by BERTopic varying the minimum topic size from 2 to 10 in increments of 2.

As can be seen from the table below, the classification results mainly indicate the advantage of topic-informed machine learning classifiers. The best result was obtained by the MLP classifier using BERTopic vectors with minimum topic sizes equal to 8 and 10. Moreover, the topic-informed Logistic Regression and MLP classifiers both achieved their best results using BERTopic document topics. In most cases, the classifiers also benefit from GSDMM topics. The LSVC classifier showed its best result using 100-dimensional GSDMM topic vectors. For our data, we did not identify a clear benefit of LDA topics for the LR and LSVC classifiers.

[3] https://www.kaggle.com/oldaandozerskaya/fiction-corpus-for-agebased-text-classification

Table 2
Results for our topic-informed models and baselines, %

Method	Topic model	F	P	R
LR	-	77,44	86,07	75,63
LR	LDA, 25 topics	77,33 (-0,11)	86,04	75,55
LR	LDA, 50 topics	76,75 (-0,69)	85,69	74,88
LR	LDA, 75 topics	77,33 (-0,11)	86,04	75,54
LR	LDA, 100 topics	77,68 (+0,24)	86,48	75,88
LR	GSDMM, 25 topics	77,67 (+0,23)	86,31	75,88
LR	GSDMM, 50 topics	78,57 (+1,13)	86,29	76,88
LR	GSDMM, 75 topics	77,43 (-0,01)	85,74	75,63
LR	GSDMM, 100 topics	78,12 (+0,68)	86,3	76,38
LR	BERTopic, n=2	78,24 (+0,8)	86,33	76,5
LR	BERTopic, n=4	78,01 (+0,57)	86,26	76,25
LR	BERTopic, n=6	78,58 (+1,14)	86,61	76,88
LR	BERTopic, n=8	79,37 (+1,93)	86,72	77,75
LR	BERTopic, n=10	**79,59 (+2,15)**	**86,64**	**78**
LSVC	-	78,14	86,79	76,38
LSVC	LDA, 25 topics	78,82 (+0,68)	87,01	77,13
LSVC	LDA, 50 topics	78,02 (-0,12)	86,76	76,27
LSVC	LDA, 75 topics	78,93 (+0,79)	87,05	77,25
LSVC	LDA, 100 topics	77,91 (-0,23)	86,72	76,13
LSVC	GSDMM, 25 topics	79,49 (+1,35)	86,76	77,88
LSVC	GSDMM, 50 topics	79,26 (+1,12)	86,84	77,63
LSVC	GSDMM, 75 topics	78,92 (+0,78)	86,56	77,25
LSVC	GSDMM, 100 topics	**79,61 (+1,47)**	**87,11**	**78**
LSVC	BERTopic, n=2	78,36 (+0,22)	86,86	76,63
LSVC	BERTopic, n=4	78,25 (+0,11)	86,66	76,56
LSVC	BERTopic, n=6	78,82 (+0,68)	86,85	77,13
LSVC	BERTopic, n=8	78,82 (+0,68)	86,85	77,13
LSVC	BERTopic, n=10	78,82 (+0,68)	86,85	77,13
MLP	-	79,05	87,08	77,38
MLP	LDA, 25 topics	79,61 (+0,56)	87,27	78
MLP	LDA, 50 topics	80,06 (+1,01)	87,11	78,5
MLP	LDA, 75 topics	80,17 (+1,12)	87,31	78,63
MLP	LDA, 100 topics	79,39 (+0,34)	87,2	77,75
MLP	GSDMM, 25 topics	80,5 (+1,45)	87,12	79
MLP	GSDMM, 50 topics	79,26 (+0,21)	86,84	77,63
MLP	GSDMM, 75 topics	79,6 (+0,55)	86,8	78
MLP	GSDMM, 100 topics	79,72 (+0,67)	87,15	78,13
MLP	BERTopic, n=2	79,71 (+0,66)	86,84	78,13
MLP	BERTopic, n=4	80,17 (+1,12)	87,31	78,63
MLP	BERTopic, n=6	79,95 (+0,9)	87,39	78,38
MLP	BERTopic, n=8	**80,84 (+1,79)**	**87,25**	**79,38**
MLP	BERTopic, n=10	**80,84 (+1,79)**	**87,25**	**79,38**

4. Conclusion

In this paper, we have focused on the age-based classification task. We have explored Logistic Regression, Linear Support Vector Classifier, and Multilayer Perceptron classifiers with a set of document topic features obtained using LDA, GSDMM, and BERTopic topic modeling techniques. We tested our baselines and topic-informed classifiers on the corpus of fiction abstract to predict the age of readers.

We demonstrated the superiority of topic-informed models as compared to baselines. The most improvement for age-based classification gave BERTopic and GSDMM document topics. We also showed that the usage of LDA topics does not significantly increase the results for the LR and LSVC classifiers for our dataset. The possible explanation is that LDA topic models are aimed at working with longer texts. Therefore, in further work, we plan to evaluate the impact of topic modeling features on the corpus of fiction texts that are much longer and multi-thematical than book abstracts.

5. Acknowledgements

This study is supported by the grant of the President of the Russian Federation no. MK-637.2020.9.

6. References

[1] Chen, Xiaobin, and Detmar Meurers. "Characterizing text difficulty with word frequencies." Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (2016): 84-94.

[2] R. Flesch A new readability yardstick. Journal of applied psychology 32(3) (1948) 221.

[3] E. Dale, J. S. Chall, The concept of readability. Elementary English 26(1) (1949) 19-26.

[4] Heilman, Michael, et al. "Combining lexical and grammatical features to improve readability measures for first and second language texts." Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (2007): 460-467.

[5] Mukherjee, Partha, Gondy Leroy, and David Kauchak. "Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study." IEEE journal of biomedical and health informatics 23.5 (2018): 2164-2173.

[6] Hancke, Julia, Sowmya Vajjala, and Detmar Meurers. "Readability classification for German using lexical, syntactic, and morphological features." Proceedings of COLING 2012 (2012): 1063-1080.

[7] Salesky, Elizabeth, and Wade Shen. "Exploiting morphological, grammatical, and semantic correlates for improved text difficulty assessment." Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (2014): 155-162.

[8] Sheehan, Kathleen M., Irene Kostin, and Yoko Futagi. "When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty." Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington DC (2008): 1978-1983.

[9] Poulsen, Mads, and Amalie KD Gravgaard. "Who did what to whom? The relationship between syntactic aspects of sentence comprehension and text comprehension." Scientific Studies of Reading 20.4 (2016): 325-338.

[10] Crossley, Scott A., Hae Sung Yang, and Danielle S. McNamara. "What's so Simple about Simplified Texts? A Computational and Psycholinguistic Investigation of Text Comprehension and Text Processing." Reading in a Foreign Language 26.1 (2014): 92-113.

[11] Howcroft, David M., and Vera Demberg. "Psycholinguistic models of sentence processing improve sentence readability ranking." Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (2017): 958-968.

[12] Glazkova, Anna. "Exploring Book Themes in the Russian Age Rating System: a Topic Modeling Approach." Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2020) (2020): 304-314.

[13] Sakhovskiy, Andrey, Valery Solovyev, and Marina Solnyshkina. "Topic Modeling for Assessment of Text Complexity in Russian Textbooks." 2020 Ivannikov Ispras Open Conference (ISPRAS). IEEE (2020): 102-108.
[14] Glazkova, Anna, Yury Egorov, and Maksim Glazkov. "A Comparative Study of Feature Types for Age-Based Text Classification." Analysis of Images, Social Networks and Texts. Lecture Notes in Computer Science (2020): 120-134.
[15] Korobov, Mikhail. "Morphological analyzer and generator for Russian and Ukrainian languages." International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham (2015): 320-332.
[16] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.
[17] Zhu, Ciyou, et al. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization." ACM Transactions on Mathematical Software (TOMS) 23.4 (1997): 550-560.
[18] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
[19] Vorontsov, Konstantin, and Anna Potapenko. "Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization." International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham (2014): 29-46.
[20] Řehůřek, Radim, and Petr Sojka. "Gensim—statistical semantics in python." Retrieved from genism. org (2011).
[21] Yin, Jianhua, and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014): 233-242.
[22] M. Grootendorst, BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics, 2020. URL: https://doi.org/10.5281/zenodo.4381785.
[23] McInnes, Leland, et al. "UMAP: Uniform Manifold Approximation and Projection." Journal of Open Source Software 3.29 (2018): 861.
[24] McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." Journal of Open Source Software 2.11 (2017): 205.
[25] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[26] Loper, Edward, and Steven Bird. "NLTK: the Natural Language Toolkit." Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1 (2002): 63-70.