Advances in Social Science, Education and Humanities Research, volume 437
International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)

Text Age Rating Methods for Digital Libraries

Glazkova A.V.

University of Tyumen, Tyumen, Russia
[*]Corresponding author. Email: a.v.glazkova@utmn.ru

ABSTRACT
The addressee plays a major role in communication. Text creating involves taking into account the features of the target audience, to which he refers in writing. In this article, the text addressee detection is considered from the point of view of natural language processing. The task of age classification deserves special attention. Its relevance is associated with the development of e-learning systems and digital libraries. Moreover, nowadays all information products in Russia must be marked by age rating. This article describes the first attempt to solve the automatic age rating prediction task by the example of Russian texts. In this work, we analyze the main factors affecting the text age rating and propose the first approximation classifier for determining the age of the textual target audience. Our approach is based on a range of features designed to capture readability, lexical and topic modeling characteristics. We use these features to train a Linear Support Vector Classifier. We trained and tested our classifier on a dataset of 1200 previews of fiction books in Russian annotated for age rating by books' publishers. Our performance evaluation suggests that proposed features are a good indicator for text age rating. However, in future work, we plan to add and evaluate other types of models and linguistic features.
Keywords: *content rating, age restrictions, Russian Age Rating System, text classification, text addressee, textual target audience, machine learning*

1. INTRODUCTION

The amount of textual information is constantly increasing. Some of this information is presented in electronic form, such as the content of web pages, electronic books, other natural language texts. Modern challenges contribute to the development and implementation of natural language processing tools in large repositories of electronic documents [1].
The active development of natural language processing technologies opens up many opportunities for e-learning and other activities in the field of education [2]. In connection with the constant growth of information flows, one of the key tasks is the development of methods and software for organizing textual information.
The participants of the modern educational process are faced with the need to quickly view and classify large volumes of text documents. This problem arises everywhere: for example, in the processes of information searching in the Internet, obtaining information in electronic libraries, working with text databases and other elements of the educational process. There is no doubt the need to improve search tools in electronic document space, which usually comes down to comparing specified text fragments to other texts in natural language. Recent developments in search engines are primarily aimed at expanding the capabilities of text processing tools, which leads to an increase in the relevance of queries. Thus, text classification and clustering systems help the user reduce the number of documents viewed.
An important task for developers of modern text classification systems is to bridge the gap between calculated text features and deeper characteristics: for example, is this text suitable for people with a certain level of education or is this text suitable for children. This article is devoted to the task of age rating text classification. The significance of this task is associated with the development of digital libraries, as well as the introduction of age restrictions on information resources in many countries [3-4]. This project provided an important opportunity to advance the understanding of the factors influencing the age rating of the text and possible classification features for the automatic age rating prediction task.
This paper begins by the brief review of related work. It will then go on to the analysis of the modern Russian Age Rating System. The third section is concerned with the methodology used for this study. The fourth section presents the results of the first approximation classifier for determining the age rating of the text.

2. RELATED WORK

2.1. Text Addressee Detection

The addressee factor is one of the key aspects of communication. This factor assumes that the author takes into account the text of the target audience [5]. The addressee can be a single person, a group of individuals or society as a whole, and the author consciously or instinctively creates a text that reflects the traits of the audience. There are various bases for classifying the audience of a text, including: size, gender, professional attributes, and age.

Although the age rating prediction task is an increasingly important area in applied linguistics and natural language processing, there is still very little scientific understanding of the reasons determining the age rating of the text. A search of the literature revealed few studies which consider the influence of the addressee factor on the text from the point of view of linguistics [6-11].The authors of these studies emphasize that any text contains traits that determine the image of its likely reader. Thus, the text focuses on the target audience, while the reader is interested in texts relevant to his needs and level of personal development. In addition, the content of the considered papers allows us to conclude that the age of the text is affected by two main factors:

1) textual simplicity (i. e. the ease of reading and perception);
2) semantic content (including topics, vocabulary, figurativeness).

2.2. Russian Age Rating System

An age rating system is a system of rules accepted for information classification based on their suitability for audiences due to their treatment of issues such as sex, violence, or substance abuse; their use of profanity; or other matters typically deemed unsuitable for children and adolescents [12-13]. Most countries have some form of rating system that issues age restrictions. Age rating recommendations may be mandatory or advisory. In some countries content sellers may have a legal obligation to enforce restrictive ratings.

In 2012, Russia issued a set of rules governing access to information harmful to the health and development of children. In accordance with international standards for protecting children from information that causes them mental, physical and moral harm, the corresponding federal law has been adopted in the Russian Federation [14]. According to the current law, the classification of information products is carried out by its producers independently (including with the participation of experts). Information assessment takes into account the following factors:

1) topic, genre, content and decoration;
2) features of perception of the information contained in it by children of a certain age category;
3) harm to the health and development of children.

The system includes the following categories of information:

1) for children under the age of six;
2) for children over the age of six;
3) for children who have reached the age of twelve years;
4) for children over the age of sixteen;
5) prohibited for children.

In today's Russia, all books and other information products must be provided with age marking.

3. METHODOLOGY

In this section, we consider methods of computer linguistics and natural language processing that could be used for age rating prediction.

3.1. Readability Indices

Readability indices are measures of determining the complexity of the text. The readability index can be calculated based on several parameters. As a rule, these are easily calculated quantities, such as: lengths of sentences, number of words, proportion of the most frequency (or rare) words, etc. The relations between the parameters are regulated by specially calculated coefficients.

Most readability indexes are designed for English. Russian is different from English in a number of ways. Russian words, for example, are usually longer than English, and Russian sentences, on the contrary, are shorter. Scientists have made attempts to revise the readability formulas for the Russian language. Thus, the study [15] proposed the coefficients for the Flesch–Kincaid formula for Russian texts. The project [16] offers the adaptation of several readability formulas.

3.2. Lexical Features

Lexical features include linguistic models simplifying text representation in natural language processing and special dictionaries using for age categories differentiation.

Classic examples of a text representation model are the bag-of words model and the TF-IDF model [17]. The bag-of-words model represented text as the bag (or multiset) of its words, disregarding grammar and even word order but keeping multiplicity. This model is usually presented in the form of a matrix in which the rows correspond to a single text, and the columns are the words included in it. The cell at the intersection is the number of occurrences for the particular word in the corresponding document. The TF-IDF model is similar to the previous one, but the intersection of lines and columns contains the TF-IDF measure for a given word in a specific document. In addition to models describing the quantitative characteristics and importance of words, lexical features of the text can be described by word embeddings based on distributive semantics or other language models [18-20].

Additionally, the age rating prediction task can use the lexicon-based features, such as curse words, hate-speech terms and abusive language features. For instance, a significant analysis and discussion on the abusive lexicon for violence rating prediction from movie scripts was presented in [21]. Unfortunately, currently we do not have a ready-made dictionary of similar vocabulary for Russian. A detailed study of the relation between the domain and the lexicon-based features is a part of our future work.

3.3. Topic Modeling Features

Topic modeling is a way to build a model for a text collection that determines which topics each document relates to. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a set of texts.

The approaches to topic modeling based on Bayesian networks are most used in modern applications. Probabilistic topic models are a relatively young area of research. One of the first methods proposed for topic modeling was probabilistic latent-semantic analysis (PLSA), based on the maximum likelihood principle, as an alternative to classical clustering methods based on the calculation of distance functions. Further studies also used the Dirichlet latent placement method and its many generalizations [22-23].

4. EXPERIMENTS AND DISCUSSION

4.1. Dataset

In our work, we used a dataset of previews for fiction books in Russian. We classify book's previews because the full texts of books are not available in public access in sufficient quantities. Since assigning an age rating is based on the full text of books, we understand that using previews can adversely affect the quality of classification.

We excluded the 0+ class from consideration due to the small number of examples and the short length of texts. The final dataset contained 1200 previews (300 previews per class) collected in electronic libraries. Age ratings were assigned to books by experts. The dataset statistics is given in Table 1. The dataset was divided into training and test samples in a ratio of 80 to 20.

4.2. Results

We trained a Linear Support Vector Classifier (LinearSVC) using the set of features to classify texts into one of four categories of age rating. We chose LinearSVC because it shows high results for one-dimensional vectors of features, as was shown in [21, 24].

The results of LinearSVC were compared with the values obtained using the feed-forward network (FNN) with two hidden layers.

We evaluated three types of features, such as the TF-IDF model, the values of readability indices, and texts' topic distributions (see Table 2). In this work, we used five readability values, including the Flesch–Kincaid test, the Coleman–Liau index, the automated readability index (ARI), the SMOG grade, the Dale-Chall formula. We implemented the adaptation of readability formulas for Russian available at [16]. To obtain topic distributions, we built an LDA topic model for 50 topics.

The experiments were carried out using the Python 3.6 programming language and the freely distributed libraries, such as Scikit-learn [25], Keras [26], Gensim [27].

Table 1 Dataset statistics

Characteristic	6+	12+	16+	18+
Average number of characters per document	28653	77960	82354	52036
Average number of words per document	3574	9188	9916	6458

Table 2 Age rating prediction feature set

Type	Number of features	Description
Readability indices	5	The values obtained with (1) the Flesch–Kincaid readability test, (2) the Coleman–Liau index, (3) the automated readability index, (4) the SMOG grade, (5) the Dale-Chall formula.
Lexical features	2000	The values of TF-IDF for 2000 most frequent words.
Topic modeling features	50	The values of topic distribution per document using the LDA topic model.

In Table 3, we present the evaluation metrics in terms of the macro-averaging F1-score. The symbol (-) indicates that the respective feature type is excluded.
Both methods showed similar results. The most influence on the quality of classification was provided by lexical features (TF-IDF). The addition of other features had a positive, but not so impressive impact on the quality of the classification.

Since current age rating systems are not based on the simplicity of the book's language, but on the safety of the content, the readability features have not had much impact. Despite this, the addition of these features contributed to the improvement of the classification quality. We also suggest that topic modeling features might be more meaningful when using full-text books, rather than previews.

Table 3 Results

Features	**F1-score (%, macro-averaging)**	
	LinearSVC	FNN
All features	70,76	69,71
(-) topic distributions	68,52 (-2,24)	68,49 (-1,22)
(-) readability	67,94 (-2,82)	69,12 (-0,59)
(-) TF-IDF	35,71 (-35,05)	32,6 (-37,11)
(-) topic distributions and readability	66,85 (-3,91)	67,25 (-2,46)

5. CONCLUSION

Approaches to automatic age rating prediction can be based on readability, lexical or topic features. Although the problem is relevant, no previous studies have compared different types of features. This study carries out an analysis of the dataset of books' previews in Russian to explore how different types of features affect automatic age rating prediction. We apply the conclusions from our analysis to an approach to age rating prediction and build a first approximation classifier for determining the age rating of the text.
In future work, we hope to consider other types of features and also compare different machine learning methods. The results of the study can be used in digital libraries for book search, as well as in search engines for web content filtering.

ACKNOWLEDGMENT

The work was performed under the grant of the President of the Russian Federation for the state support of young Russian scientists and candidates of sciences (project no. MK-637.2020.9). The subject of the study is "Methods for automatic age text classification".

REFERENCES

[1] V. B. Barakhnin, O. Y. Kozhemyakina, A. N. Duisenbayeva, Y. N. Yergaliev, R. I. Muhamedyev, The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area, Journal of Physics: Conference Series, IOP Publishing, 2018, vol. 1117, №. 1, p. 012001. DOI: 10.1088/1742-6596/1117/1/012001.

[2] I. G. Zakharova, Machine Learning Methods of Providing Informational Management Support for Students' Professional Development, The Education and science journal, 2018, №20 (9), pp. 91-114 (In Russian). DOI: 10.17853/1994-5639-2018-9-91-114.

[3] A. V. Glazkova, An approach to text classification based on age groups of addressees, Trudy SPIIRAN, 2017, vol. 52, pp. 51-69 (In Russian). DOI: 10.15622/sp.52.3.

[4] A. V. Glazkova, The evaluation of the proximity of text categories for solving electronic documents classification tasks, Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika, 2015, № 2 (31), pp. 18-25 (In Russian). DOI: 10.17223/19988605/31/2.

[5] I. A. Sternin, Adressee's Factor in Speech Influence, Bulletin of the Voronezh State University. Series: Philology. Journalism, 2004, vol. 1, pp. 171-178 (In Russian).

[6] E. V. Beloglazova, The addressee age as a factor determining the discursive structure of a work of fiction, Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Seriia 2, IAzykoznanie, 2014, vol. 13. – №. 5. DOI: 10.15688/jvolsu2.2014.5.15.

[7] K. Oleksiy, Specificity of expressing the communicative role of the addressee in artistic communication, Linguistics Studies, 2019, pp. 135-143 (In Ukranian). DOI: 10.29038/2413-0923-2019-10-135-143.

[8] L. A. Borisova, The Factor of Target Audience of a Legislative Text and Its Influence on the Language of Law, Vestnik NSU. Series: History and Philology, 2018, vol. 17, №. 2, pp. 129-137 (In Russian). DOI: 10.25205/1818-7919-2018-17-2-129-137.

[9] L. G. Kim, E. S. Belyaeva, Recipient's Pre-Text Expectations as a Factor of Various Interpretations of Political Discourse, Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya, 2019, vol. 57, pp. 48-62. DOI: 10.17223/19986645/57/3.

[10] E. I. Budaragina, Means of creating an addressing image in artistic text, abstract of dissertation for the degree of candidate of philological sciences / Moscow State Pedagogical University, Moscow, 2006.

[11] C. Nord, What do we know about the target-text receiver? Investigating Translation, 2000, vol. 32 , pp. 195-212. DOI: 10.1075/btl.32.24nor.

[12] G. Adomavicius, Y. O. Kwon New recommendation techniques for multicriteria rating systems, IEEE Intelligent Systems, 2007, vol. 22, №. 3, pp. 48-55. DOI: 10.1109/mis.2007.58.

[13] J. Gosselt, J. Van Hoof, M. De Jong Media rating systems: Do they work? Shop floor compliance with age restrictions in The Netherlands, Mass communication and society, 2012, vol. 15, №. 3, pp. 335-359.

[14] Federal Law of December 29, 2010 N 436-FZ (as amended on May 1, 2019) "On the Protection of Children from Information Harmful to Their Health and Development" (as amended and additional, entered into force on October 29, 2019), URL: http://www.consultant.ru/document/cons_doc_LAW_108808/ (accessed 13.02.2020) (In Russian).

[15] I. A. Oborneva, Automated assessment of the complexity of educational texts based on statistical parameters, thesis for the degree of candidate of pedagogical sciences, Moscow, 2006.

[16] Text readability rating, URL: http://readability.io/ (accessed 13.02.2020).

[17] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008, 504 p.

[18] T. Mikolov et al. Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 2013, pp. 3111-3119.

[19] M. E. Peters et al. Deep contextualized word representations, arXiv preprint arXiv:1802.05365, 2018.

[20] J. Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805, 2018.

[21] V. R. Martinez et al. Violence rating prediction from movie scripts, Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, vol. 671-678. DOI: 10.1609/aaai.v33i01.3301671.

[22] K. Vorontsov, A. 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, pp. 29-46. DOI: 10.1007/978-3-319-12580-0_3.

[23] A. V. Sukhareva, K. V. Vorontsov, Building a complete set of topics of probabilistic topic models, Intelligent systems. Theory and applications, 2019, vol. 23, №. 4, pp. 7-23.

[24] C. Nobata et al. Abusive language detection in online user content, Proceedings of the 25th international conference on world wide web, 2016, pp. 145-153. DOI: 10.1145/2872427.2883062.

[25] Scikit-learn. Machine learning in Python. URL: https://scikit-learn.org/stable/index.html (accessed 23.03.2020).

[26] Keras: The Python Deep Learning library. URL: https://keras.io/ (accessed 23.03.2020).

[27] Gensim: Topic modelling for humans. URL: https://radimrehurek.com/gensim/ (accessed 23.03.2020).