Exploring Book Themes in the Russian Age Rating System: a Topic Modeling Approach*

Anna Glazkova[0000−0001−8409−6457]

University of Tyumen, Tyumen 625003, Russia
a.v.glazkova@utmn.ru

Abstract. Age rating systems are created to indicate target ages of potential content users based on information security and text semantics. Age ratings are usually given as numbers, which tell us the youngest age the content is suitable for. A book or film with a 12+ rating has content which is suitable only for people aged 12 years and over, and a book or film with an 18+ rating is suitable for adults only. Currently, content assessment in terms of information security is carried out by experts. In this paper, we empirically compare book abstracts assigned to different age ratings using unsupervised topic modeling. We use an LDA model to discover topics from a collection of book abstracts. We then use statistical methods to study relations between the age rating categories assigned to books by experts and the topics obtained. We believe that our comparisons show interesting and useful findings for age rating automation.

Keywords: Topic modeling · Age restrictions· Age rating · Text classification · Statistical methods.

1 Introduction

Age-based ratings serve as a warning that the content may be unsuitable to children. Moreover, age ratings are used to ensure that entertainment content, such as books, but also films, games or mobile apps, is clearly labelled with a minimum age recommendation.

While some books are suitable for readers of all ages, others are only suitable for older children and young teenagers. A specific portion of books contain information that is only appropriate for an adult audience.

Age-based rating systems in different countries differ. Whereas the classification systems in Russia, Europe and Germany are based purely on age, the rating systems in the USA and Australia might be interpreted with consideration of factors other than age. For example, in Australia there are two different 18+ ratings applied to either adult content or pornographic materials [6].

The Russian Age Rating System (RARS) includes 5 categories of content:

* Supported by the grant of the President of the Russian Federation no. MK-637.2020.9.

– for children under the age of six (0+);
– for children over the age of six (6+);
– for children over the age of twelve (12+);
– for children over the age of sixteen (16+);
– prohibited for children (18+).

The RARS was introduced in 2012 when the Federal law of Russian Federation no. 436-FZ of 2010-12-23 «On Protection of Children from Information Harmful to Their Health and Development» was passed [3]. The law prohibits the distribution of «harmful» material that depicts violence, unlawful activities, substance abuse, or self-harm.

The aim of this article is to compare the topics of the texts assigned to different age rating categories (according to the RARS). Our findings can potential benefit many text classification applications, such as recommender systems and text filtering systems. First, through topic analysis, we can gain a deeper understanding of the structure of age rating systems. Since the age rating of a book is currently being assessed empirically, the topic analysis will be an important step towards formalizing this task. Based on the results of the analysis, it will become more clear which books on which topics most often contain information that is unsuitable for children or addressed to a particular age audience. In addition, the results of topic modeling will help to highlight specific topics for different age groups. This is the reason why topic distributions can be used as additional features for automatic age rating classifiers in our further research.

The paper is divided into six sections. The first section is introduction. The second section is concerned with the data preprocessing used for this study and the description of our topic model. The third part is related to empirical analysis of topics. The last section is conclusion.

2 Methodology

2.1 Data Preparation

We use a collection of abstracts for books in Russian. These abstracts was collected on the basis of public online libraries.

Text preprocessing included the following actions:

– standard steps, such as conversion into lower case letters, removing punctuations and digits, lemmatization, removing extra white spaces;
– excluding all the stop words using Natural Language Toolkit (NLTK) Python library [14] and words with fewer than 3 symbols;
– removing words with TF-IDF weights less than 0.15. TF-IDF (term frequency–inverse document frequency) is a statistical measure that shows how important a word is to a document in a text collection. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the text collection that contain the word [9]. So, this action allowed us to exclude words typical

of book abstracts (these are usually the words «book», «author», «reader», etc.). The threshold 0.15 was chosen empirically. During the study, we tested values in the range $[0.1, 0.3]$ with an increment of 0.05 and compared the coherence of the models;
- excluding personal names to delete the mentions of authors and characters. This allowed our topic model to form themes according to the semantic proximity of abstracts, and not according to the belonging of books to one author or the coincidence of the characters names. To recognize named entities, we used the Natasha Python library [13];
- we have combined common phrases (with a frequency of mutual occurrence of more than 5) into bigrams using the Gensim library [12].

Some statistics of the data are summarized in Table 1.

Table 1. Some characteristics of the data.

Category	Number of texts	Avg number of words per text
0+	53	45.33
6+	3107	72.03
12+	3110	72.03
16+	3989	91.76
18+	3986	76.36

2.2 LDA

To discover topics from the collection of abstracts, we choose to apply standard Latent Dirichlet Allocation (LDA) [15]. Topic modeling is a type of statistical modeling for recognizing main topics in a collection of documents. As a rule, topic modeling is based on LDA, a hierarchical network that relates words and documents through latent topics [1]. Topics are characterized by diverse frequency of words. The document is presented as a bag-of-words approach, and the topic looks like a set of words ranked in decreasing order of their probabilities. LDA topic model were applied to analyze various subject areas, such as social media analysis [17,19,18], analysis of emails [5], news [7,16], fiction texts [10], and others.

We designed a topic model for 100 topics, which reflect the main content of the collection of abstracts. Our approach to choosing the optimal number of topics was to build topic models with different values of number of topics and pick the number that gives the highest coherence value (Fig. 1).

2.3 Topic Distribution Estimation

We calculated the topic distribution for each document in the collection. It is a vector of length equal to the number of topics in the LDA model. The topic

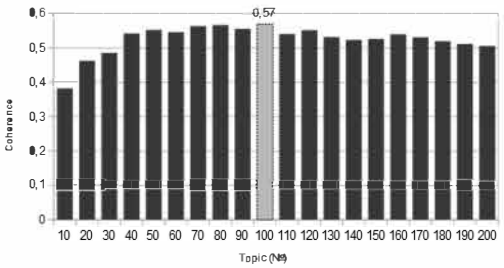

Fig. 1. Coherence values.

distribution vector shows how a document extracted from the collection of abstracts corresponds to each topic. Each element of the topic distribution vector is a number from 0 to 1, where 0 is a complete non-match and 1 is an absolute match. Then, we got the averaged topic distribution vector for each age rating category to obtain the average values of topic distribution for a group of documents:

$$a_j = \{w_1, w_2, \ldots, w_k\}, j \in [1, n], \tag{1}$$

where n is a number of age categories, k is a number of topics, w_x is a severity value of a particular topic for the age category, $x \in [1, k]$.

For the obtained averaged topic distribution vectors $a_1, a_2, \ldots, a_n$, we determined:

- the most typical topics for each age category. We calculated the standard deviations for each vector $a_1, a_2, \ldots, a_n$. Next, we marked values that exceeded the three standard deviations as corresponding to the most typical topics for the category (the three sigma rule).
- age-specific topics that are typical mainly for one age rating category. For each topic, we have a vector

$$t_i = \{v_1, v_2, \ldots, v_m\}, i \in [1, k], \tag{2}$$

where v_z ($z \in [1, m]$) from the vector t_i corresponds to the value w_i from the vector a_z.

In the case when the number of observations (in our case, the number of categories) is rather small, we cannot use the three-sigma rule to search for outliers. In this case, it is necessary to use other statistical techniques for small-sized samples [4]. We then applied the Dixon's Q-test [2] to determine age-specific topics. The Dixon's Q-test is the simpler test that allows us to examine if one observation from a small set of replicate observations (typically the number of observations is more than 3 and no more than 10) is an outlier or not [11].

First, we arrange the values $v_1, v_2, \ldots, v_m$ for each vector t_i in ascending order (from the lowest to the highest value):

$$v'_1 \leq v'_2 \leq \ldots \leq v'_m. \tag{3}$$

Therefore, we estimate the experimental Q-value:

$$Q_{exp} = \frac{v'_m - v'_{m-1}}{v'_m - v'_1}. \quad (4)$$

In the next step, we compare the calculated Q_{exp} value to the tabulated critical Q_{crit} value for a chosen confidence level.

3 Empirical Analysis of Topics

As mentioned earlier, the purpose of this study is to compare the topics of texts assigned to different age rating categories (according to the RARS). Understanding the differences between categories will help us highlight topics specific to the categories. Thus, in the future, we will be able to use the empirical results of topic modeling as a set of additional features for automatic text classification based on the age of the addressee.

3.1 Distribution of Topics

According to the Russian law, books of any genre in printed and electronic versions are subject to labeling in accordance with age restrictions. The law distinguishes five age categories: 0+, 6+, 12+, 16+ and 18+ (prohibited for children). The difference between the books of these age categories is determined by the presence of scenes of violence, cruelty, descriptions of antisocial actions, diseases, the mention of narcotic substances, alcoholic beverages, tobacco and swear words. Table 2 shows the most common topics for age categories based on our topic model.

Table 2: Top-5 topics per category.

№	Keywords
	0+
1	«Сказка», «волшебный», «сказочный», «серия», «японский», «украинский», «индийский», «арабский» («tale», «magic», «fairy tale», «series», «Japanese», «Ukrainian», «Indian», «Arabic»)
2	«Рассказ», «сказка», «тетрадь», «маленькие», «занятие», «пособие», «интеллект», «серия» («story», «fairy tale», «notebook», «small», «lesson», «manual», «intelligence», «series»)
3	«Любовь», «рассказ», «стих», «город», «сборник», «сказка», «пересказ», «перевод» («love», «story», «verse», «city», «collection», «fairy tale», «retelling», «translation»)
4	«Ребенок», «планета», «написать», «стихотворная_форма», «рассказывать», «сказка», «русский», «приключение» («child», «planet», «to write», «poetic form», «to tell», «fairy tale», «Russian», «adventure»)

5	«Рассказ», «ребенок», «поучительный», «чтение», «церковный», «любить», «Богородица», «учебник» («story», «child», «instructive», «reading», «church», «to love», «Mother of God», «textbook»)
6+	
1	«Школьный_возраст», «рассказ», «младший_школьник», «детский_писатель», «ребенок», «известный», «сборник», «повесть» («school age», «story», «younger school student», «children's writer», «child», «famous», «collection», «story»)
2	«Сказка», «волшебный», «сказочный», «серия», «японский», «украинский», «индийский», «арабский» («tale», «magic», «fairy tale», «series», «Japanese», «Ukrainian», «Indian», «Arabic»)
3	«Упражнение», «язык», «закрепление», «брошюра», «английский», «самоучитель», «лингвистический», «испанский» («exercise», «language», «fastening», «brochure», «English», «checkbook», «linguistic», «Spanish»)
4	«Ребенок», «планета», «написать», «стихотворная_форма», «рассказывать», «сказка», «русский», «приключение» («child», «planet», «to write», «poetic form», «to tell», «fairy tale», «Russian», «adventure»)
5	«Стихотворная_форма», «планета», «рассказывать», «растение», «алфавит», «иллюстрация», «стихотворение» («Poetic form», «planet», «to tell», «plant», «alphabet», «illustration», «poem»)
12+	
1	«Школьный_возраст», «рассказ», «младший_школьник», «детский_писатель», «ребенок», «известный», «сборник», «повесть» («school age», «story», «younger school student», «children's writer», «child», «famous», «collection», «story»)
2	«Сказка», «волшебный», «сказочный», «серия», «японский», «украинский», «индийский», «арабский» («tale», «magic», «fairy tale», «series», «Japanese», «Ukrainian», «Indian», «Arabic»)
3	«Упражнение», «язык», «закрепление», «брошюра», «английский», «самоучитель», «лингвистический», «испанский» («exercise», «language», «consolidation», «brochure», «English», «checkbook», «linguistic», «Spanish»)
4	«Ребенок», «планета», «написать», «стихотворная_форма», «рассказывать», «сказка», «русский», «приключение» («child», «planet», «to write», «poetic form», «to tell», «fairy tale», «Russian», «adventure»)
5	«Стихотворная_форма», «планета», «рассказывать», «растение», «алфавит», «иллюстрация», «стихотворение» («Poetic form», «planet», «to tell», «plant», «alphabet», «illustration», «poem»)
16+	

1	«Жадный», «погон», «замучить», «милость», «заметить» » счастье», «свадебный», «глотать» («greedy», «epaulet», «to torture», «mercy», «to notice» «happiness», «wedding», «to swallow»)
2	Фронт», «мужчина», «трагедия», «гипотетический», «догнать», «былой», «предшественник», «достижение» («forefront», «man», «tragedy», «hypothetical», «to catch up», «past», «predecessor», «achievement»)
3	«Метод», «французский», «упрощение», «повторяемость», «заучивание», «уникальность», «текст», «лексический» («method», «French», «simplification», «repeatability», «memorization», «uniqueness», «text», «lexical»)
4	«Детский», «ребенок», «отношение», «педагог», «искусство», «воспитание», «зависимость», «женщина» («children's», «child», «relation», «teacher», «art», «upbringing», «addiction», «woman»)
5	«Книга», «прошлое», «человек», «любовь», «горький», «терять», «пройти», «мир» («book», «past», «man», «love», «bitter», «to lose», «to pass», «peace»)
18+	
1	«Чувство», «отношение», «новелла», «представлять», «реальность», «интрига», «удовольствие», «фантазия» («feeling», «relation», «short story», «to imagine», «reality», «intrigue», «pleasure», «fantasy»)
2	«Зодиак», «гороскоп», «светить», «знак», «поведение», «тип», «предопределять», «сексуальный» («Zodiac», «horoscope», «to shine», «sign», «behavior», «type», «to determine», «sexual»)
3	«Книга», «прошлое», «человек», «любовь», «горький», «терять», «пройти», «мир» («book», «past», «man», «love», «bitter», «to lose», «to pass», «peace»)
4	«История», «друг», «жанр», «смешной», «личность», «диалог», «герой», «весёлый» («history», «friend», «genre», «funny», «personality», «dialogue», «hero», «funny»)
5	«Знание», «отличие», «мужчина», «женщина», «мир», «здоровье», «действие», «тело» («knowledge», «difference», «man», «woman», «world», «health», «action», «body»)

The books for children under 6 years old (0+) may contain episodic unnaturalistic images justified by the genre or descriptions of physical or psychological violence, provided that the victim is compassionate and happy ending. The prevailing topics in the 0+ category are fairy tales of the world (topic 1), developing children's benefits (2), poems about the world around us (3-4) and Christian literary works for children (5).

In the books for children over the age of 6 (6+), non-naturalistic images or descriptions of human diseases, accidents, catastrophes or violent death without demonstrating their consequences are permissible. The books for children over 12 (12+) may contain scenes of violence or murder, descriptions of illnesses, disasters, but without details. Alcohol, tobacco and drug use may be present,

but should be condemned. A schematic description of the hugs and kisses of men and women may be present. These two categories are described by similar topics in our topic model. These are short stories and tales for primary and secondary school age (topic 1), fairy tales of the world (2), study guides (3) and poems for children (4-5).

The books for children over 16 (16+) may contain scenes of illnesses, disasters without detailed descriptions. Violence, alcohol and drug use can be described, but should be condemned. Rough words may be present, with the exception of swear words. Scenes of sexual relations cannot be described with anatomical details. In our example, this category is represented by military (topics 1-2) and human condition (5) fiction, teaching aids (3), psychological and pedagogical literature (4).

A book should be marked with the 18+ label if the book contains a naturalistic description of illnesses, disasters, non-condemned drug and alcohol use, naturalistic scenes of sexual relations, non-traditional relationships, obscene language, scenes that encourage suicide. In our topic model, love stories (topic 1), horoscopes (2), human condition fiction (3-4) and possibly relationship psychology books (5) were detected in this category.

3.2 Most Common Topics for Categories

Table 3 shows most of the common topics discovered using the three sigma rule. These topics are most widely represented in their category. The column «Main topic» shows the percentage of texts for which this topic is the main, i.e. it has the largest proportion in the topic distribution.

Table 3: The most typical topics for categories.

Category	Keywords	Main topic
0+	«tale», «magic», «fairy tale», «series», «Japanese», «Ukrainian», «Indian», «Arabic»	7,55%
16+	«greedy», «epaulet», «to torture», «mercy», «to notice» «happiness», «wedding», «to swallow»	1,55%
16+	«forefront», «man», «tragedy», «hypothetical», «to catch up», «past», «predecessor», «achievement»	1,53%
18+	«feeling», «relation», «short story», «to imagine», «reality», «intrigue», «pleasure», «fantasy»	1,78%
0+	«story», «fairy tale», «notebook», «small», «lesson», «manual», «intelligence», «series»	5,66%
6+	«school age», «story», «younger school student»,	2,45%
12+	«children's writer», «child», «famous», «collection», «story»	2,48%

3.3 Age-specific Topics

In this subsection, we provide the topics that are typical mainly for one age rating category using the Dixon's Q-test. The tabulated Q_{crit} value is equal to 0.642 for confidence level 90% and $m = 5$.

We noticed that documents in category 0+ are largely mono-thematic. At the same time documents of other categories are usually mixtures of topics. Therefore, our topic model has many specific topics for texts from the 0+ category. In Table 4, we present the list of the most common age-specific topics in our data set.

As it would be logical to assume, age-specific topics generally relate to children's books, as well as to specific literature from the 18+ category (in our case, literature on business and success).

Table 4. Age-specific categories.

Category	Keywords	Q_{crit}
18+	«success», «activity», «man», «business», «city», «collection», «european», «necessary»	0,94
0+	«story», «fairy tale», «notebook», «small», «lesson», «manual», «intelligence», «series»	0,93
0+	«fairy tale», «plunge», «fairytale atmosphere», «emotion», «character», «interesting», «exciting», «impression»	0,85
0+	«child», «planet», «to write», «poetic form», «to tell», «fairy tale», «Russian», «adventure»	0,78
0+	«tale», «magic», «fairy tale», «series», «Japanese», «Ukrainian», «Indian», «Arabic»	0,76
0+	«malachite box», «tale», «storyteller», «humor», «funny», «forest», «hero», «capable» [1]	0,75

4 Conclusion

In this paper, we empirically analyzed the topics of texts assigned to different age rating categories. We introduced the distribution of topics for age categories and the list of the most common topics for categories and age-specific topics. These list of topics were obtained using statistical methods. Our analysis confirmed the existing differences between the categories and demonstrated that topic models can be a good source of features for age rating identification. In our future work, we will try to develop a machine learning classifier for automatically determining the text age rating.

[1] This topic is probably related to «The Malachite Box» («Малахитовая шкатулка»). This is a book of fairy tales and folk tales of the Ural region of Russia compiled by Pavel Bazhov and published from 1936 to 1945. It is written in contemporary language and blends elements of everyday life with fantastic creatures of mountains and forests. This book significantly popularized the folklore of the Urals [8].

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I. Latent dirichlet allocation. In: Journal of machine Learning research. Vol. 3(Jan). Pp. 993-1022 (2003).
2. Dixon, W. J.: Processing data for outliers. Biometrics **1**(9), 74–89 (1953). https://doi.org/10.2307/3001634
3. 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) [Federal'nyj zakon ot 29.12.2010 N 436-FZ (red. ot 01.05.2019) «O zashchite detej ot informacii, prichinyayushchej vred ih zdorov'yu i razvitiyu» (s izm. i dop., vstup. v silu s 29.10.2019).], http://www.consultant.ru/document/cons_doc_LAW_108808/. Last accessed 7 Apr 2020.
4. Glazkova, A., Kruzhinov, V., Sokova, Z.: Dynamic Topic Models for Retrospective Event Detection: A Study on Soviet Opposition-Leaning Media. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 145-154, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37334-4_13
5. Gong, H., You, F., Guan, X., Cao, Y., Lai, S.: Application of LDA Topic Model in E-Mail Subject Classification. In: 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City. Atlantis Press (2018). https://doi.org/10.2991/tlicsc-18.2018.24
6. How are age-based gaming ratings set?, https://www.kaspersky.com/blog/gaming-age-ratings/11647/. Last accessed 7 Apr 2020.
7. Hu X.: News hotspots detection and tracking based on LDA topic model. In: 2016 International Conference on Progress in Informatics and Computing (PIC). IEEE, pp. 248-252 (2016). https://doi.org/10.1109/pic.2016.7949504
8. Ilyasova, R. S.: Dialectal lexis of P. P. Bazov's narrations «Malachite casket». Letters of the Chechen State University **3**(11), 103-107 (2018).
9. Manning, C.D.; Raghavan, P.; Schutze, H.: Scoring, term weighting, and the vector space model. Introduction to Information Retrieval. p. 100 (2008). https://doi.org/10.1017/CBO9780511809071.007.
10. Mitrofanova, O., Sedova, A.: Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose). In: Proceedings of the International Conference IMS-2017, pp. 175-180 (2017). https://doi.org/10.1145/3143699.3143734
11. Raschka, S.: A questionable practice: Dixon's Q test for outlier identification, https://sebastianraschka.com/Articles/2014_dixon_test.html. Last accessed 13 Apr 2020. https://doi.org/10.13140/2.1.3000.0004
12. Rehurek, R., Sojka, P.: Gensim—statistical semantics in python, Retrieved from genism. org. (2011).
13. Natasha - high quality compact solution for extracting named entities from news articles in Russian, https://natasha.github.io/ner/. Last accessed 26 Jul 2020.
14. Loper, E., Bird, S.: NLTK: the natural language toolkit, arXiv preprint cs/0205028 (2002).
15. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 29-46, Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12580-03.
16. Wang, H., Wang, J., Zhang, Y., Wang, M., Mao, C.: Optimization of Topic Recognition Model for News Texts Based on LDA. Journal of Digital Information Management **5**(17), 257 (2019). https://doi.org/10.6025/jdim/2019/17/5/257-269

17. Yang, S. and Zhang, H.: Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. In: Int. J. Comput. Inf. Eng, 12, pp.525-529 (2018).
18. Zhao, F., Zhu, Y., Jin, H., Yang, L. T.: A personalized hashtag recommendation approach using LDA-based topic model in microblog environment //Future Generation Computer Systems **65**, 196–206 (2016). https://doi.org/10.1016/j.future.2015.10.012
19. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European conference on information retrieval, pp. 338-349. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-53_4