Topic Modeling of Literary Texts Using LDA: on the Influence of Linguistic Preprocessing on Model Interpretability

Tatiana Sherstinova, Anna Moskvina, Margarita Kirina, Irina Zavyalova,
Asya Karysheva, Evgenia Kolpashchikova, Polina Maksimenko, Alena Moskalenko
National Research University Higher School of Economics, Saint Petersburg
Saint Petersburg, Russia
{tsherstinova, admoskvina, mkirina}@hse.ru
{izavyalova, askarysheva, eokolpaschikova, pimaksimenko, eyumoskalenko}@edu.hse.ru

***Abstract*—The article describes the results of the research, the purpose of which was to evaluate the influence of linguistic preprocessing on the interpretability of topic models for literary texts. The study was carried out as part of a large project aimed to obtain topic models of Russian short stories written in the first three decades of the 20th century and divided into three successive historical periods: 1) the period of the beginning of the century before the First World War (1900-1913), 2) the time of acute social cataclysms, wars and revolutions (World War I, the February and October revolutions, and the Civil War) (1914-1922), and 3) the early Soviet period (1923-1930). The material of the study was 3 samples of different sizes for each period, containing 100, 500 and 1000 short stories each. Preprocessing included lemmatization using spaCy library and four POS-filtering options: 1) nouns only, 2) nouns and verbs, 3) nouns, adjectives, adverbs, verbs, and 4) no filtering. Using the latent Dirichlet allocation (LDA), 36 topic models were built (9 models for each preprocessing option). The research showed that in case of literary texts topic models built without any POS filters are the most interpretable. The study made it possible to obtain information about topic diversity of Russian short stories, to assess their expert interpretability, and to offer some recommendations for optimizing topic modeling, which are to be used in the development of artificial intelligence systems that process large volumes of literary texts.**

I. Introduction

Topic modeling is a machine learning method used to categorize large unstructured text data. At present, the construction of topic models for a corpus or a collection of text documents is an actively developing area of text processing [1], [2], [3], [4], [5]. Originally topic modeling methods were developed for processing non-fiction (scientific, technical, news related, etc.) documents. However, in recent years there have been more and more cases when these methods were used for clustering fiction [6], [7], [8], [9], [10], [11], [12], [13], [14]. Moreover, it seems that topic modeling can become an effective way of solving the current tasks of modern humanities, e.g. building a topic model of national literature [15], broadening the possibilities of digital humanities [16], [17], [18], [19] and 'distance reading' methods [20], [21], as well as the developing the artificial intelligence systems that process large volumes of literary texts in any language [15].

This research continues a series of studies dedicated to the research of linguistic and topic diversity of Russian short stories in the first three decades of the XX century [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], and aims to build topic models for three consecutive historical periods, each of these corresponding with important historical events:

- the 1st period (1900-1913) refers to the beginning of the 20th century before World War I,
- the 2nd period (1914-1922) is the era of social cataclysms, wars, and revolutions (World War I, February and October Revolution, the Civil War),
- the 3rd period (1923-1930) refers to the beginning of the Soviet Union's formation.

It is assumed that the historical background of the era, during which literary texts were written, will be somehow reflected in fiction, influencing not only the language, but also the content and topics of these texts [24], [25].

It is known that the quality of topic modeling results is affected not only by the algorithm behind the model, but also by the data preprocessing. The purpose of this research is to evaluate the influence of linguistic preprocessing on quality and interpretability of resulting topics: first of all, POS filtering of the data after lemmatization. An important result of this research is the assessment of the topic model in terms of its content interpretation, i.e. the degree of semantic correlation of words combined into one topic.

II. Data and Methods

A. Data

In comparison with the earlier research based on topic modeling of Russian short stories [23], [24], [25], [26], [31], [32], [33], in this study, we significantly expanded the sample size — up to 1000 short stories from every period. Moreover, it seemed to us appropriate to see how the results of topic

modeling change with a consistent expansion of the research sample — for 100, 500 and 1000 texts.

Hence, 9 samples were prepared for conducting the research: 3 samples of 100, 500 and 1000 short stories for 3 sequential periods — pre-war period (1900–1913), war-revolutionary one (1914–1922), and early period of Soviet Union (1923–1930). This was done to estimate the change of "topic diversity" in the dynamics. Since the ultimate goal of the research is to model the national literary process [34], [15], when forming the sample, the main emphasis was placed on ensuring representation of texts for the maximum number of Russian writers who worked in the genre of the story [22], [27]. The list of these writers includes not only the famous authors but also relatively unknown writers and even the virtually forgotten ones.

The main sources for forming the samples were two open-access literary resources. These involve: 1) Lib.ru — the library of Maxim Mashkov [35], one of the oldest and most representative resources of Russian literature, and 2) the Corpus of Russian Short Stories 1900–1930 that has been developed by the authors [36]. The latter contains a large volume of rare texts digitized specially for the Corpus. However, these two text collections turned out not to be sufficient for the tasks set. We had to use other open-access Internet resources (in particular, [37], [38], [39]) to complete the sample for the third period — the early stage of the Soviet Union.

While forming the samples, two tasks were set: 1) ensuring a relatively even distribution of texts by the year of writing or the year of their first publication, and 2) ensuring the maximum representativeness of different authors in both the whole period and each individual year. Stories were selected randomly, regardless of their themes/motifs and their content. The samples did not include relatively large stories (more than 10000 word usages) and very short ones (less than 200 words). The texts selected were analyzed as a whole, regardless of their size. The samples of a smaller size were subsets of the larger samples (i.e., texts from a sample of 100 stories are also included in samples of 500 and 1000 stories).

B. Data Preprocessing

The modeling we used employs a bag-of-words approach, so we had to go through some steps before providing input for the model training. While preprocessing, punctuation, digits, and function words were removed from the texts, the words were lemmatized using the ru_core_news_sm model of the spaCy library [41].

Pilot experiments have shown that proper names are an issue to consider: first names and patronymics were quite common in topics, though they do not carry any meaningful information. In a corpus consisting of heterogeneous stories, the same name, as a rule, functions as a highly polysemantic word: in each story, the name Ivan would refer to some separate character, and the absolute frequency within the whole corpus is not significant. Therefore, in the final preprocessing, we also removed all proper names from selected texts, using for that the morphological analysis functionality in spaCy.

In some studies, only nouns are used to build a model, since they often carry more specific semantics. Exploring fiction texts, we've decided to compare four possible cases. For all 9 samples, preprocessing was carried out with different POS filtering options (i.e. when having a certain filter only the words of the parts of speech indicated in the filter remained in texts and, respectively, were lemmatized):

1) Without any filtering options (F-0).
2) With filters "Noun", "Verb" and "Adjective" (F-NVA).
3) With filters "Noun" and "Verb" (F-NV).
4) With the only filter "Noun" (F-N).

Too rare and too frequent words — those found in less than 6 documents and no more than in 80% of the corpus — were removed from all texts. The left border prevents the lexical preferences of only one separate author to influence the topics and was based on the principals of corpus content selection, that is not the issue of this article. The upper border was chosen as an intuition provided by pilot experiments and should prevent the considering lexical items inherent to the whole corpus. We also used custom stop-list for removing some frequent words.

Sample sizes obtained are presented in Table I.

TABLE I. SAMPLE SIZES OBTAINED FOR DIFFERENT POS-FILTRATION OPTIONS

Samples	Before lemmatization	POS-Filtering			
		F-0	F-NVA	F-NV	F-N
1-100	369113	219110	183334	149012	83790
1-500	1820446	1069285	888781	725271	400235
1-1000	3564493	2099711	1746658	1421549	786676
2-100	321693	191239	159612	130718	72767
2-500	1355994	812613	677527	554989	311631
2-1000	2702207	1614738	1343028	1098107	615165
3-100	302682	187889	157891	131564	75444
3-500	1516696	922588	770669	636302	366807
3-1000	2541865	1543123	1288382	1062122	613462

Another common stage of data preparation for the bag-of-words approach is determining frequency collocations (bigrams) and combining them into one token. Then the model dictionary may consist not only of unigrams but also of multiword expressions. To select collocations and combine them into one unit, we used the Phrases class in gensim library. Consecutive words are considered a collocation and are combined if function exceeds the threshold value (the approach is based on the works [42] and [43]). We used the default function threshold and set a constraint that a pair of words must occur at least 4 times (the value was chosen based on the insight from the pilot experiments) in the corpus to be combined into a collocation. Surprisingly, we have not found many bigrams within the topics obtained (though there were some, e.g. '*starshiy ofitser*' — 'senior officer'). Probably, it can be explained by specifics of literary texts or by the relatively small size of the samples. In any case, it leaves room for future experiments.

C. Latent Dirichlet allocation (LDA) algorithm

Latent Dirichlet Allocation, or LDA, is a multinomial generative probabilistic model [44], [45]. In accordance with the LDA model, the corpus of texts is considered as a mixture of topics. Thus, each document in the collection is described by a set of hidden semantic structures — topics, which in turn are composed of words that contribute to it with a certain degree of probability. A document can have an unlimited number of topics, so that documents are considered to be described by a family of topic distributions [5]. One topic characterizes the document not only with a certain degree of probability but is also generated by a larger or smaller part of the document. The number of topics to be found in the text collection is predetermined by the user, on the one hand, or calculated statistically, on the other, and is equal to *k* [46].

One of the advantages of LDA is that the model does not require any preliminary training. As a result, it is possible to obtain generally well-interpretable data, regardless of the length of the texts. However, on big data, the model can be retrained, therefore, the results may not be reproducible. Another disadvantage is the amount of time that the algorithm requires to process the data. In addition, the number of topics must, as already mentioned, be set in advance — if there are too few of them, then the output will be too general topics, but if, on the contrary, the number turned out to be too large, then the topics will intersect with each other and in this case, it will be difficult to distinguish them [47].

D. Topic model evaluation

Regarding topic model evaluation, two aspects are mainly considered: first, how interpretable the resulting topics are and, second, how well these topics describe the collection of documents in question. On the one hand, the assessment can be proceeded using a number of statistical measures, and on the other hand, by involving experts. Among statistical measures for assessing the quality of topicc models, the most popular are perplexity and coherence. Among the experts, when the assessment of the consistency of the model and its interpretability is based on the judgments of people, the method of intrusions — word intrusion and topic intrusion — can be noted.

Perplexity is used to evaluate how well a topic model fits test data. In other words, it evaluates how different and evenly distributed words are in texts [48]. It is believed that the less perplexity, the better the model. However, it is noted that the interpretation of this measure is difficult, since, in fact, it determines the quality of the matrix decomposition, and not the quality of the topics themselves and their interpretability. Another measure, coherence, on the contrary, is considered more appropriate for model evaluation, since it more often than others correlates with expert assessments [49]. The coherence measure allows determining how often words that occur together in one text fall into the same topic. The larger the coherence value, the better. However, these measures do not always correctly determine the quality of the model and indicate their interpretability [50].

In this regard, another important technique is the expert evaluation of the topic model. As a rule, several people are involved in this procedure, who are offered a list of criteria by which topics should be evaluated for their interpretability — how they correlate with some, in their opinion, thematic category. Another interesting experimental technique aimed at searching for “extra” words (topics) artificially added to the topic by an expert and called the word (topic) intrusion tasks described in [50]. If this word (topic) was correctly defined by a person, then this indicated a higher coherence of the model, and vice versa. And finally, the last of the main expert methods is the comparison of the results of automatic and expert annotations. This approach to assessing the quality of a model, which allows one to determine how well topics describe the distribution of topics in a corpus, also seems promising [31]. The methodology for assessing the interpretability of a topic is presented below in Section IV.

III. Results of Topic Modeling

Table II presents summary statistics about the trained topic models for 9 samples with four different preprocessing conditions, including the optimal number of topics, perplexity and coherence for each model.

TABLE II. The optimal number of topics, perplexity and coherence for different preprocessing

Sample	Number of Topics	Perplexity	Coherence
F-0			
1-100	20	-8.251630597	0.3297
1-500	40	-9.220208614	0.2884
1-1000	15	-9.057504691	0.3605
2-100	35	-8.385480029	0.3256
2-500	40	-9.238728492	0.2846
2-1000	15	-8.986393197	0.3402
3-100	25	-8.130335837	0.3183
3-500	45	-9.351973386	0.2983
3-1000	30	-9.088551366	0.3453
F-NVA			
1-100	35	-8.169626314	0.3076
1-500	30	-8.984397623	0.2887
1-1000	15	-8.991972464	0.3421
2-100	20	-8.001491291	0.3027
2-500	35	-8.828669303	0.2785
2-1000	30	-8.929520329	0.3085
3-100	40	-8.065844991	0.2685
3-500	30	-8.917253343	0.2744
3-1000	30	-8.949371349	0.3452
F-NV			
1-100	25	-7.908964388	0.3212
1-500	30	-8.699320364	0.2802
1-1000	20	-8.747728317	0.3549
2-100	25	-7.785691599	0.2934
2-500	40	-8.589043185	0.2621
2-1000	15	-8.679811428	0.3084
3-100	35	-7.782870956	0.3065
3-500	35	-8.692940262	0.2936
3-1000	15	-8.75306189	0.3382
F-N			
1-100	30	-7.194386812	0.3423
1-500	20	-7.824747288	0.316
1-1000	15	-7.884775195	0.3735
2-100	10	-6.804736668	0.3084
2-500	10	-7.691362341	0.2999
2-1000	30	-7.924730806	0.368
3-100	15	-6.883887804	0.345
3-500	35	-8.124425314	0.3188
3-1000	25	-7.951963042	0.3446

TABLE III. 15 TOPICS FOR 1-1000 SAMPLE (1900-1913, 1000 SHORT STORIES), NO POS-FILTERING

Topic #	Topic content
0	'doktor' (doctor), 0.0206, 0.0206, 'professor' (professor), 0.0090, 'ninochka'* (ninochka), 0.0077, 'babushka' (grandmother) 0.0074, 'rebyonok' (child), 0.0060, 'muzh' (husband), 0.0043, 'vagon' (train car), 0.0032, 'komnata' (room), 0.0032, 'zhenshchina' (woman), 0.0028, 'ujti' (to go away), 0.0028
1	'djakon' (deacon), 0.0155, 'kupets' (merchant), 0.0082, 'starosta' (headman), 0.0078, 'aktyor' (actor), 0.0067, 'pjesa' (a play), 0.0055, 'gospodin' (gentleman), 0.0054, 'gorozhanin' (citizen), 0.0052, 'gulyanije' (festivities), 0.0046, 'rytsar' (knight), 0.0046, 'parashka'* (parashka), 0.0040
2	'grob' (coffin), 0.0040, 'sklep' (crypt), 0.0035, 'igra' (game), 0.0018, 'strashniy' (scary), 0.0018, 'pozhar' (fire), 0.0015, 'pokoynik' (deadman), 0.0015, 'strakh' (fear), 0.0013, 'uzhas' (horror), 0.0013, 'ispytat' (to try), 0.0012, 'noch' (night), 0.0012
3	'noch' (night), 0.0007, 'poyti' (to go), 0.0006, 'stoyat' (to stand), 0.0006, 'lyubit' (to love), 0.0005, 'okno' (window), 0.0005, 'beliy' (white), 0.0005, 'otets' (father), 0.0005, 'golos' (voice), 0.0004, 'tyomniy' (dark), 0.0004, 'sidet' (sit), 0.0004
4	'denga' (money), 0.0055, 'komnata' (room), 0.0046, 'zhena' (wife), 0.0044, 'rubl' (rouble), 0.0043, 'sprosit' (ask), 0.0039, 'kvartira' (flat), 0.0039, 'chas' (hour), 0.0035, 'poyti' (to go), 0.0031, 'dom' (house), 0.0031, 'sluzhba' (service), 0.0029
5	'lyubit' (to love), 0.0030, 'zhena' (wife), 0.0028, 'rebyonok' (child), 0.0028, 'sidet' (sit), 0.0026, 'dom' (house), 0.0026, 'muzh' (husband), 0.0025, 'sprosit' (ask), 0.0024, 'komnata' (room), 0.0024, 'otets' (father), 0.0024, 'minuta' (minute), 0.0023
6	'zemlya' (earth), 0.0036, 'poyti' (to go), 0.0031, 'otets' (father), 0.0028, 'stoyat' (to stand), 0.0026, 'starik' (old man), 0.0026, 'stariy' (old), 0.0026, 'dom' (house), 0.0024, 'noch' (night), 0.0024, 'sidet' (sit), 0.0023, 'les' (forest), 0.0023
7	'zhulik' (rogue), 0.0074, 'poyti' (to go), 0.0005, 'zemlya' (earth), 0.0005, 'golos' (voice), 0.0004, 'serdtse' (heart), 0.0004, 'golov' (heads), 0.0003, 'otvetit' (answer), 0.0003, 'stoyat' (to stand), 0.0003, 'storona' (side), 0.0003, 'lyubit' (to love), 0.0003
8	'matros' (sailor), 0.0132, 'starshiy_ofitser' (senior_officer), 0.0105, 'doktor' (doctor), 0.0080, 'blagorodije' (honor), 0.0080, 'kapitan' (captain), 0.0077, 'botsman' (boatswain), 0.0076, 'kayuta' (cabin), 0.0054, 'tyurma' (jail), 0.0048, 'vasheskobrodie' (your honor), 0.0048, 'smotritel' (caretaker), 0.0047
9	'kapitan' (captain), 0.0560, 'admiral' (admiral), 0.0144, 'professor' (professor), 0.0091, 'tort' (cake), 0.0064, 'svetlost' (grace), 0.0056, 'knyaz' (prince), 0.0045, 'grafinya' (countess), 0.0039, 'dobriy' (kind), 0.0038, 'bak' (tank), 0.0034, 'starshiy_ofitser' (senior_officer), 0.0029
10	'poyti' (to go), 0.0003, 'otets' (father), 0.0003, 'sprosit' (ask), 0.0003, 'golos' (voice), 0.0003, 'lyubit' (to love), 0.0003, 'zhena' (wife), 0.0003, 'zemlya' (earth), 0.0003, 'dusha' (soul), 0.0002, 'sidet' (sit), 0.0002, 'stoyat' (to stand), 0.0002
11	'vagon' (train car), 0.0123, 'poyezd' (train), 0.0121, 'stantsiya' (station), 0.0120, 'polkovnik' (colonel), 0.0093, 'parovoz' (locomotive), 0.0080, 'platforma' (platform), 0.0060, 'mashinist' (driver), 0.0056, 'poruchik' (lieutenant), 0.0040, 'roza' (rose), 0.0036, 'telegrafist' (telegrapher), 0.0035
12	'soldat' (soldier), 0.0327, 'ofitser' (officer), 0.0177, 'poruchik' (lieutenant), 0.0093, 'rota' (company), 0.0088, 'pulya' (bullet), 0.0081, 'yaponets' (japanese), 0.0072, 'polk' (regiment), 0.0061, 'vystrel' (shot), 0.0056, 'raneniy' (wounded), 0.0054, 'blagorodije' (honor), 0.0054
13	'arestant' (prisoner), 0.0229, 'tyurma' (jail), 0.0120, 'nadziratel' (warden), 0.0083, 'negr' (black person), 0.0074, 'kamera' (cell), 0.0061, 'kedr' (cedar), 0.0051, 'amerikanskiy' (american), 0.0048, 'amerika' (america), 0.0043, 'bratik' (little brother), 0.0042, 'khishchnik' (predator), 0.0042
14	'lyubit' (to love), 0.0034, 'dusha' (soul), 0.0033, 'komnata' (room), 0.0029, 'lybov' (love), 0.0028, 'noch' (night), 0.0028, 'mysl' (thought), 0.0028, 'sprosit' (ask), 0.0028, 'zhenshchina' (woman), 0.0028, 'stranniy' (strange), 0.0027, 'serdtse' (heart), 0.0026

Each topic is a list of ten words sorted in descending order of the probability that a word belongs to this topic. The absolute values of such probabilities vary from topic to topic, meaning some words are less probable to their topic even if they have the same rank within the corresponding topics. An example of a topic model obtained for the 1-1000 sample without the use of POS filters with the number of topics equal to 15 is presented in Table III. A total of 36 such models were obtained during the experiment.

(*) It is worth mentioning that in Table III one of the models includes two proper names 'ninochka' and 'parashka' although preprocessing was meant to remove all personal names. This fact can be explained by the diminutive form of the presented words, which were not included in the dictionary.

At the next stage of the study, the task was to assess the expert interpretability of these models.

IV. Expert Assessment of Topic Interpretability

The assessment of the interpretability of the constructed models can be carried out in various ways, some of which are described above in Section II (D). For our experiment, we decided to use the expert assessment method.

Three philologists with both literary and linguistic backgrounds were invited as experts, who had to independently assess interpretability of all topics obtained for all models on a binary scale (*interpretable* vs *uninterpretable*), and then the degree of consistency of responses for each model was analyzed.

The experts were instructed to consider the topics interpretable if the absolute majority of the words forming the topic relate to the same semantic field or to adjacent semantic fields, or if it was possible to construct a plausible story or a fragment of a story, based on the words of the topic. Along with that, it was allowed to have one to three "extra" words that did not fit into the general context if their probability in the topic was not high enough, implying that the higher the probability of an irrelevant word is, the less the topic can be considered interpretable. One expert tended to consider the topics consisting mostly of synonymous words, for example, 'stariy' (old), 'starik' (old man) and 'ded' (grandfather) as too narrow for the literary collection theme and thus not interpretable.

Here is an example of a topic that was considered interpretable by the first expert (Exp1). It contained the words 'grob' (coffin), 'sklep' (crypt), 'igra' (game), 'strashnyj' (scary), 'pozhar' (fire), 'pokoynik' (deadman), 'strakh' (fear), 'uzhas' (horror), 'ispytat' (to try), 'noch' (night) (this topic was taken from the variant, which was lemmatized without any POS filters, sample 1-1000 (a thousand stories from the first time period, topic #2, see Table III). It is obvious that here we have *a theme of death*, perhaps a *funeral* (however, in this case the presence of the word 'noch' (night) is poorly explained, except for the metaphorical meaning). It can also be a *theme of murder* or *death because of gas suffocation in a fire*. However, the word 'igra' (game) does not fit the series; of course, it is possible to imagine a situation (or a metaphor), in context of which this word will be more than appropriate, but this

requires some imposition of meanings that are unlikely to be found in the original stories. It can be added that 'pozhar' (fire) partially narrows the topic, but it cannot be said that it categorically contradicts it. Thus, there was one word in this topic that does not seem to be a logical part of it. Since there is only one word, it seems possible not to take its position into account when sorting by decreasing probability of hitting (furthermore, in absolute terms, this probability is 0.002, which is relatively small, in two neighboring topics, for example, the probability of the third word is about 0.007) and in general consider the topic as interpretable. This topic was given the conditional title "*death*", and it was considered interpretable by all three experts.

Nevertheless, in some cases, there were discrepancies between experts in terms of how many "extra" words can be allowed in an interpretable model. The degree of rigidity of the experts' assessment was different.

For example, topic #14, sample 2-1000, F-N was considered to be interpretable by two experts (Exp1 and Exp2). It contains the following set of words: 'yolka' (Christmas tree), 'palets' (finger), 'pianist' (pianist), 'rebonok' (child), 'yolochka' (Christmas tree), 'aktrisa' (actress), 'devochka' (girl), 'sochel'nik' (Christmas Eve), 'mamochka' (mommy), 'balet' (ballet). Exp2 evaluated this topic as interpretable, since using these words, it is possible to construct one small narrative about a family that watches a festive ballet shortly before Christmas. But the third expert (Exp3) was prone to disagree with this conclusion, because from his point of view there could not be more than 1-2 "extra" words in the topic.

At the same time, the second expert allowed that even if up to four words in a topic do not correspond to the semantic field of the rest of the words in it, the topic can still be considered interpretable. For example, for a sample of 3-500 F-NVA, topic #12 was formed. The topic includes the following words: 'devushka' (girl), 'byt' (everyday life), 'mat' (obscene words), 'rebyata' (guys), 'moda' (fashion), 'sovetskiy' (Soviet), 'styd' (shame), 'guba' (lip), 'diskussiya' (discussion), 'vopros' (question). This topic was recognized by Exp2 as interpretable; though the name given by him to this topic — '*Everyday life and obscene words*' – is not exhaustive for the combination of all the words in the topic. Thus, the words 'fashion', 'lip' and 'discussion' seem to go beyond a domestic conflict situation.

In addition, here is an example of a topic that was not considered to be a topic by the first expert. It contained the words 'noch" (night), 'poyti' (go), 'stoyat" (stand), 'lyubit" (love), 'okno' (window), 'belyy' (white), 'otets' (father), 'golos' (voice), 'tiomnyy' (dark), 'sidet' (sit) (topic #3, sample 1-1000, F-0, see Table III). Despite the fact that during the initial data preprocessing, the upper part of the frequency words list was removed, a certain set of words, that occur in many topics, has emerged. Most of them seems necessary when building a narrative — such words can include, for example, 'go' and others not represented in this topic, verbs of movement, color designations, etc. The concentration in one topic of such words, though not meaningless, but rather neutral leads to a noticeable decrease in both the probabilities of including even for the first word (only 0.0007) as well as the interpretability of the whole topic. Undoubtedly, a situation where all these words are more or less appropriate can be imagined, but, according to expert 1, the interpretable topic should not require rampant imagination.

Thus, in the experiment conducted, the degree of rigidity of the assessment among experts turned out to be different. Because of this, we do not have many topics that are considered to be interpretable by all three experts. But this in no way affects the solution of the problem assigned in this study, since all 36 models were described by the same experts working in parallel and independently.

V. The Results of Expert Assessment

Table IV presents the results of the assessment of the interpretability for the obtained models by three different experts.

TABLE IV. Comparative interpretability of models by three independent experts (Exp1, Exp2, Exp3)

Sample	Number of topics	Number of interpretable topics			Share of interpretable topics, %		
		Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
F-0							
1-100	20	5	5	6	25	25	30
1-500	40	13	7	12	33	18	30
1-1000	15	9	6	8	60	40	53
2-100	35	9	5	9	26	14	26
2-500	40	4	7	4	10	18	10
2-1000	15	4	2	5	27	13	30
3-100	25	11	4	8	44	16	32
3-500	45	9	3	8	20	07	20
3-1000	30	5	3	8	17	10	30
F-NVA							
1-100	35	7	5	4	20	14	11
1-500	30	4	4	3	13	13	10
1-1000	15	6	3	2	40	20	13
2-100	20	7	3	4	35	15	20
2-500	35	10	3	5	29	09	14
2-1000	30	15	8	8	50	27	27
3-100	40	10	1	5	25	03	13
3-500	30	6	2	3	20	07	10
3-1000	30	11	2	6	37	07	20
F-NV							
1-100	25	12	4	4	48	16	16
1-500	30	8	2	5	27	07	17
1-1000	20	11	4	7	55	20	35
2-100	25	6	0	4	24	00	16
2-500	40	13	4	4	33	10	10
2-1000	15	7	1	3	47	07	20
3-100	35	12	2	6	34	06	17
3-500	35	11	4	7	31	11	20
3-1000	15	5	3	3	33	20	20
F-N							
1-100	30	9	4	8	30	13	27
1-500	20	6	5	4	30	25	20
1-1000	15	4	4	4	27	27	27
2-100	10	0	2	1	00	20	10
2-500	10	1	1	2	10	10	20
2-1000	30	12	9	13	40	30	43
3-100	15	3	3	4	20	20	27
3-500	35	5	5	9	14	14	26
3-1000	25	7	4	8	28	16	32

Table IV shows that the maximum number of interpretable topics was found in the largest samples containing 1000 texts

of the first (1900-1913) and second (1914-1922) periods. At the same time, in a sample of the beginning of the century (1-1000), the best results are shown by a model built without POS filtering and a model that includes only nouns and verbs. For the war and revolution period (sample 2-1000), we have the best preprocessing results with the F-NVA filter (nouns, verbs, adjectives) and with the F-NV filter (nouns and verbs), although the absolute and relative indicators are somewhat lower here, than for the texts that were written in the beginning of the century.

TABLE V. PERCENTAGE OF UNAMBIGUOUSLY INTERPRETABLE TOPICS FOR EACH OF THE MODELS

Samples	Preprocessing			
	F-0	F-NVA	F-NV	F-N
1-100	20,0	8,6	8,0	0,0
1-500	7,5	3,3	6,7	10,0
1-1000	33,3	13,3	20,0	6,7
2-100	8,6	5,0	0,0	0,0
2-500	5,0	5,7	2,5	10,0
2-1000	13,3	16,7	0,0	20,0
3-100	8,0	2,5	5,7	0,0
3-500	4,4	0,0	11,4	8,6
3-1000	3,3	6,7	13,3	0,0
Mean, %	11,5	6,9	7,5	6,1

We can get a more objective picture from Table V, containing the percentage of unambiguously interpretable topics for each model, that is, the proportion of topics for each model that were determined to be interpretable by all three experts. In this table different shades of gray highlight the best values in terms of expert opinions' consistency. It can be seen that the maximum interpretability can be found in 1-1000 sample of the first period without POS filtering. In this case, a third of all topics turned out to be quite interpretable. Preprocessing with zero POS filtering occurred to be the best for all three samples of 100 stories (1-100, 2-100, 3-100), although the shares of interpretable topics for the 2nd and 3rd periods are significantly lower than those of the 1st period. In general, it is the models built without POS processing that show the best average value (11.5%) of "interpretability".

The second rank is shown by models built exclusively on nouns and verbs (F-NV) preprocessing. It was this variant of preprocessing that showed the best (although not very convincing) results for large samples of the 3rd period (3-500 and 3-1000). We see the smallest share of interpretable topics among topics formed by nouns alone (FN filtering), although this particular preprocessing option showed the best result for samples 1-500, 2-500 and 2-1000.

It should be noted that the seven models did not reveal any interpretable topic common to all three experts. Most of these topics are for FN preprocessing on small samples (1-100, 2-100, 3-100) and for 3-1000.

Thus, based on our material, it was not possible to obtain an unambiguous conclusion that one or another preprocessing option is significantly superior to all the others. But one can argue about some advantage of building models without any POS filters (F-0), which on average gives performance that is 1.5-2 times better compared to other preprocessing options.

In addition, based on the results obtained, it can be assumed that interpretability of a topic model trained on a collection of texts depends not only on the preprocessing options considered, but also on the lexical and stylistic homogeneity of the given text collection. Thus, the prose of the beginning of the 20^{th} century is quite homogeneous (traditional) in style. The second (military-revolutionary) period is more diverse, combining both pre-revolutionary and revolutionary prose. Finally, the prose of the third, early Soviet period, is distinguished by the maximum stylistic diversity [51], possibly this fact leads to a greater number of poorly interpretable topics. It can also be assumed that for the construction of interpretable models of this period, the volume of 1000 texts is not sufficiently representative. However, more research is needed to test this hypothesis.

VI. CONCLUSION

The study allowed to evaluate how different options of text preprocessing affect the quality of topic interpretability obtained in the result of expert assessment. The results showed that when building topic models using the LDA algorithm, among the tested preprocessing options, there was no one that leads to optimal results for all text samples. However, in general, a better result is shown by topic modeling without any POS filters. Therefore, this particular option can be recommended in cases where it is not possible to conduct a parallel study, as described in this article.

Another conclusion of our research is the suggestion that the interpretability of the topic model built on a collection of texts depends not only on the considered preprocessing options, but also on lexical and stylistic homogeneity of the collection of texts under consideration. This may explain the small number of interpretable models obtained for the texts of the early Soviet period. If this is true, then the volume of 1000 texts is not sufficient enough to build interpretable models of Russian short prose of 1923-1930 and should be expanded.

The study also showed that for large text collections of literary texts, it is not enough to use ready-made dictionaries of stop words or proper names. If the resulting topic models contain proper names, one should expand the stop words list and rebuild the model. In the case of the Russian language of short prose, it is worth considering the word formation of diminutive forms from proper names.

To assess interpretability of topics by comparing the opinions of experts, we can recommend the use of more formalized instructions for the latter. This will reduce the arbitrariness of certain evaluation factors (in our case, for example, how many semantically "extra words" it is permissible to have in a topic or what can be considered as a plausible literary theme). Obviously, some "softening" to the requirements of peer review can lead to more "interpretable topics" when compared with the results of this work (e.g., if all experts agree that the topic can be considered interpretable even if it contains 3 or 4 "extra" words).

Further, a specificity of the research was the fact that experts, when assessing the interpretability of a topic, took into account only the set of words that form this topic and their

probabilities, but as is usually the case with topic modeling, they did not refer to the original literary texts. It seems that the continuation of research should be seen precisely in the combination of formal methods and the analysis of literary works themselves [15], [31], although it is obvious that the expansion of the research material (in our case, up to 3000 texts) makes a detailed expert analysis of the entire text collection virtually impossible. But it is this approach that can serve as the basis for constructing an optimal system of hybrid annotation of literary texts.

ACKNOWLEDGMENT

The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2021 (grant # 21-04-053 'Artificial Intelligence Methods in Literature and Language Studies').

REFERENCES

[1] D. A. McFarland, D. Ramage, J. Chuang, J. Heer, C. D. Manning, D. Jurafsky, "Differentiating language usage through topic models", *Poetics*, vol. 41(6), 2013, pp. 607–625.

[2] D. Greene, J. P. Cross, "Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis", *in Proc. of the ACM Web Science Conference (WebSci'15)*, 2015.

[3] S. Nikolenko, S. Koltcov, O. Koltsova, "Topic modelling for qualitative studies", *Journal of Information Science*, vol. 43(1), 2017, pp. 88–102.

[4] P. Panicheva, O. Litvinova, T. Litvinova, "Author Clustering with and Without Topical Features", *in Proc. of the 21st Int. Conf. Speech and Computer (SPECOM 2019)*, LNAI 11658, 2019, pp. 348–358.

[5] O. A. Mitrofanova, "Topic modelling of special texts based on LDA algorithm", *in Proc. of XLII International Philological Conference. Selected* works, 2014, pp. 220–233.

[6] L. M. Rhody, "Topic Modelling and Figurative Language', *Journal of Digital Humanities*, vol. 2(1), 2012.

[7] O. A. Mitrofanova, A. G. Sedova, "Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose)", *in Proc. of the International Conference IMS-2017*, 2017, pp. 175–180.

[8] O. A. Mitrofanova, "Topic modeling of A.N. Afanasjev's Russian Fairytales", *in Proc. of the International Conf. Corpus Linguistics*, 2015.

[9] O. A. Mitrofanova, "Issledovanie strukturnoj organizacii hudozhestvennogo proizvedenija s pomoshhju tematicheskogo modelirovanija (opyt raboty s tekstom romana «Master i Margarita» M.A. Bulgakova)" [Analysis of Fiction Text Structure by means of Topic Modelling: Case Study of "Master and Margarita" Novel by M.A. Bulgakov], *in Proc. of the International Conf. in Corpus Linguistics*, 2019, pp. 387–394.

[10] M. L. Jockers, D. Mimno, "Significant themes in 19th-century literature", *Poetics*, vol. 41(6), 2013, pp. 750–769.

[11] B. Navarro-Colorado, "On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry", *Frontiers in Digital Humanities*, vol. 5, 2018.

[12] C. Schöch, "Topic modeling genre: an exploration of French classical and enlightenment drama", arXiv preprint, 2021.

[13] I. Uglanova, E. Gius, "The Order of Things. A Study on Topic Modelling of Literary Texts", *CHR 2020: Workshop on computational Humanities Research*, Nov. 18–20, 2020.

[14] N. Z. Da, "The computational case against computational literary studies", *Critical Inquiry*, vol. 45(3), 2019, pp. 601–639.

[15] T. Sherstinova, A. Moskvina, M. Kirina, "Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian", *in 29th Conference of Open Innovations Association (FRUCT)*, 2021, pp. 313–323.

[16] M.L. Jockers, *Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities)*, University of Illinois Press, 2013.

[17] A. M. Ronchi, *eCulture. Cultural Content in the Digital Age*, Springer, Berlin, Heidelberg, 2009.

[18] S. Fish, *Mind Your P's and B's: The Digital Humanities and Interpretation*, New York Times, 2012.

[19] L. Manovich, *Cultural Analytics*, The MIT Press, 2020.

[20] F. Moretti, *Distant Reading*, London: Verso, 2013.

[21] F. Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*, London: Verso, 2007.

[22] G. Martynenko, T. Sherstinova, "Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century", *in CEUR Workshop Proceedings*, vol. 2552, 2020, pp. 105–120.

[23] E. Zamiraylova, O. Mitrofanova, "Dynamic topic modelling of Russian fiction prose of the first third of the 20th century by means of non-negative matrix factorization*", in R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings*, vol. 2552, 2020, pp. 321–339.

[24] T. G. Skrebtsova, "Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories", *in CompLing, CEUR Workshop Proceedings*, vol. 2813, 2021, pp. 265–276.

[25] T. Sherstinova, T. Skrebtsova, "Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930", *in CompLing, CEUR Workshop Proceedings*, 2020.

[26] T. G. Skrebtsova, "Struktura narrativa v russkom rasskaze nachala XX veka" [Narrative structure of the Russian short story in the early XX century], *in Proceedings of the International Conference 'Corpus Linguistics-2019'*, 2019, pp. 426–431.

[27] G. Ya. Martynenko, T. Yu. Sherstinova, T. I. Popova, A. G. Melnik, E. V. Zamirajlova, "On the principles of creation of corpus of Russian short stories of the first third of the 20th century", *in Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics 'TEL 2018'*, 2018, pp. 180–197.

[28] A. M. Lavrentiev, T. Yu. Sherstinova, A. M. Chepovskiy, B. Pincemin, "Using TXM platform for research on language changes over time: The dynamics of vocabulary and punctuation in Russian Literary Texts". *in Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya*, 70, 2021, pp. 69–89.

[29] V. Zarembo, T. Sherstinova, "The dynamics of extensive text variables in Russian short stories", *in CEUR Workshop Proceedings*, 2780, 2020, pp. 102–114.

[30] T. G. Skrebtsova, A. O. Grebennikov, T. Yu. Sherstinova, "The Dynamics of Vocabulary in Russian Prose (Based on Frequency Dictionaries of the Corpus of Russian Short Stories 1900-1930)", *in Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2021-June(20), 2021, pp. 646–659.

[31] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina, "Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction", *in Martínez-Villaseñor L., Herrera-Alcántara O., Ponce H., Castro-Espinoza F.A. (eds.) Advances in Computational Intelligence. MICAI 2020. Lecture Notes in Computer Science*, vol. 12469, pp. 134–152. Springer, Cham (2020).

[32] E. Gryaznova, M. Kirina, "Defining Kinds of Violence in Russian Short Stories of 1900–1930: A Case of Topic Modelling with LDA and PCA", *in Proceedings of the International Conference IMS-2021*, 2021, in print.

[33] T. Sherstinova, M. Kirina, "Normalization Issues in Digital Literary Studies: Spelling, Literary Themes and Biographical Description of Writers", *in Communications in Computer and Information Science*, 1503 CCIS, 2022, pp. 332–346.

[34] Yu. N. Tynyanov, *Arkhaisty i novatory* [Archaists and Innovators], Priboi Publ., Leningrad, 1929.

[35] Lib.ru: "Classics" (Maxim Moshkov's Library), Web: http://az.lib.ru/ (date of access: 15.02.2021).

[36] The Corpus of Russian Short Stories, Web: https://russian-short-stories.ru/ (date of access: 05.03.2022).

[37] Digital library of books Aldebaran, Web: https://aldebaran.ru (date of access: 15.02.2022).

[38] Digital library Litmir, Web: https://www.litmir.me (date of access: 15.02.2022).

[39] Digital Library of IMLI RAS, Web: http://biblio.imli.ru (date of access: 15.02.2022).

[40] Sovlit project, Web: http://www.ruthenia.ru/sovlit/.

[41] SpaCy models for Russian, Web: https://spacy.io/models/ru.

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and

[43] their compositionality", *in Proceedings of the 26th International Conference on Neural Information Processing Systems*, vol. 2, 2018.
[44] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction", *in Proceedings of the Biennial GSCL Conference*, 2009.
[45] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3(4–5), 2003.
[46] D. M. Blei, "Probabilistic topic models", *Communications of the ACM*, vol. 55(4), 2012, pp. 77–84.
[47] P. Kherwa, P. Bansal, "Topic modeling: a comprehensive review", *EAI Endorsed transactions on scalable information systems*, vol. 7(24), 2020.
[48] R. Albalawi, T. H. Yeap, M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis", *in Frontiers in Artificial Intelligence*, vol. 3, 2020.
[49] K.V. Vorontsov, A. I. Frej, M. A. Apishev, A. A. Potapenko, "Tematicheskoe modelirovanie v BigARTM: teoriya, algoritmy, prilozheniya" [Topic modeling in BigARTM: theoty, algorithms, applications], 2015, Web: http://www.machinelearning.ru/wiki/images/b/bc/Voron-2015-BigARTM.pdf
[50] M. Röder, A. Both, A. Hinneburg, "Exploring the Space of Topic Coherence Measures", *in Proceedings of the 8th International Conference on Web Search and Data Mining*, 2015.
[51] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, D. M. Blei "Reading tea leaves: How humans interpret topic models", *Advances in neural information processing systems*, 2009, p. 288–296.
[52] G. Ya. Martynenko, Stilizovannyye sintaksicheskiye triady v russkom rasskaze pervoy treti XX veka [Stylized syntactic triads in a Russian short story of the first third of the 20th century], *in Proc. of the Int. Conf. 'Corpus Linguistics – 2019'*, 2019, pp. 395–404.