Adaptation of Static and Contextualized Topic Modeling Techniques to Hidden Community Detection

Ivan Mamaev and Olga Mitrofanova

1 Introduction

Nowadays the Internet can be considered as a set of web pages, which becomes larger as more people use it every day. The most widespread domain of the Internet is social networks. Their main feature is increasing the number of users' content. Users can publish posts on different topics, repost the texts of other users, and react to all of them. Such interactions can be analyzed from the point of view of hidden communities.

A hidden community is defined as an elusive group of people that cannot be detected without the application of specific approaches. Scholars usually propose three main strategies: graph-based methods, clustering methods, and hybrid ones. Recent papers show that hybrid procedures can be applied to Russian datasets [11, 12], and the basis of the hybrid approach becomes topic modeling that is a way of creating a semantic model of a text collection that describes transition from a set of documents and their words to a set of topics that characterize the content of documents.

The issue of a plethora of the algorithms becomes pivotal as the algorithm one is going to deal with strongly depends on a corpus. Therefore, a blow-by-blow comparison of methods as regards their performance is required. In the current study, we are going to focus on the dataset of 2021 Russian LiveJournal posts, which was developed in course of our research. Although LiveJournal is not so popular among

I. Mamaev (✉)
Baltic State Technical University "Voenmeh" named after D.F. Ustinov, St. Petersburg, Russia

Saint Petersburg State University, St. Petersburg, Russia
e-mail: mamaev_id@voenmeh.ru; i.mamaev@spbu.ru

O. Mitrofanova
Saint Petersburg State University, St. Petersburg, Russia
e-mail: o.mitrofanova@spbu.ru

R. Bolgov et al. (eds.), *Digital Geography*, Springer Geography,
https://doi.org/10.1007/978-3-031-50609-3_7

users as other social networks in Russia (VK, Instagram, etc.), it has some reasons to be analyzed. First of all, its inner structure is not as difficult as the structure of other sites like Facebook or Instagram so it can be web-scraped without using any complex methods. Moreover, studies [3, 7] prove deep interest of researchers to LiveJournal linguistic data; thus, this social network is of current importance among researchers. It is also worth mentioning that the Russian segment of LiveJournal is characterized by a specific set of communication practices that distinguish it from other social networks. Posts on this social network contain more textual information, which is necessary for conducting experiments in the field of topic modeling, while posts on other social networks tend to use a lot of audio content and video content; as a consequence, textual information is likely to be far from being presented in full. Resultant findings are presented and discussed to guide the choice of topic modeling approaches, especially in terms of detecting hidden communities.

2 Main Topic Modeling Approaches and Related Works

Nowadays the techniques of topic modeling are divided into two main groups: algebraic and probabilistic. In probabilistic models, we distinguish static topic models and contextualized hybrid topic models. Among the algebraic text models, the most common are the standard Vector Space Model (VSM), Latent Semantic Analysis/Indexing (LSA/LSI), Non-negative Matrix Factorization (NMF), etc. As for the static probabilistic models, one can use probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), its multimodal extensions like Author-Topic Modeling (ATM), etc. [10]. Finally, recent years have proved that pre-trained language models like BERT or ELMo can improve the quality and content of corpora processing and topic modeling. It is also worth mentioning that contextualized embeddings allow one to describe topical structure of documents in a corpus more consistently.

In contemporary linguistics, topic modeling is widely used for social network texts. In [6], the perception of social problems by readers was investigated, and the corpora of regional Russian news on social networks were used for the current experiment. With the help of LDA, the main topics, which characterized the news, were formed. The authors also assessed the importance of the topics with the help of news comments. Despite the different degrees of sentiment expressed by users, the authors concluded that a large degree of polarization of opinions led to the actualization of issues.

The paper [14] focuses on the development of the corpus of Pikabu Russian posts. The authors ran several experiments including standard LDA and ATM extended with topic label assignment based on hashtag distribution. The experiments allowed them to obtain groups of authors with similar interests. Actually, similar interests showed semantic similarity of authors that was considered as a basis for generating a model of hidden communities.

In [17], the authors set a goal to evaluate the methods of aggregating semantic features in the gender classification of texts of Russian users on social networks, a corpus of Facebook posts being collected. They used three models: LDA, ATM, and distributive semantic clustering (DSC), with ATM showing the best results. Political topics were found to be prevalent among male users.

LSA is also used for web-texts. In [1], the authors argue that customers often search for product reviews to be sure if a product is worth buying. These reviews most often contain emotional vocabulary that can influence the quality of a review and a purchase decision. The authors believe that reviews with emotional words that indicate confidence will have a positive effect on the overall rating of a review. In this study, LSA is used to measure the emotional content of reviews.

When discussing contextualized topic models, we should mention that there is not a unified algorithm for contextualized topic modeling as some of them may be a combination of LDA and distributed embedding models (e.g., LDA2Vec), and others do not include any probabilistic topic model as a core algorithm and generate or predict topics by triggering a sequence of dimensionality reduction techniques over contextualized vectors. For instance, in [18], authors propose a novel topic-informed BERT-based architecture for semantic similarity detection. They show that the proposed model improves performance over strong neural baselines across a variety of English datasets. It is observed that the addition of topics to BERT helps to resolve domain-specific cases. In [16], BERT topic modeling is also applied to a corpus of English micro-blogs. A so-called T-BERT framework is proposed to show the enhanced performance by using both latent topics and BERT embeddings. The experiments are conducted on 42,000 datasets. The empirical results allow the authors to state that the model improves the resultant performance when one adds topics to BERT. Moreover, the authors classifies the resultant topics in terms of sentiment analysis, and the accuracy rate is about 90.81% with the proposed approach.

Meanwhile the comparison of topic modeling approaches in foreign studies is a pivotal issue; in Russian studies, the problem of comparing topic modeling algorithms is not covered in such detail, especially in terms of detecting hidden communities. Thus, our study aims to fill in the gap in this area of computational linguistics.

3 Experimental Design

3.1 Developing a Corpus of LiveJournal Russian Posts

The corpus of LiveJournal Russian posts includes texts that were downloaded from LiveJournal social network. The current textual collection was created with the help of Python 3.7 programming language, as well as beautiful soup and requests libraries. The following rules were followed:

- To obtain interpretable topics, it was necessary to include texts which length was equal to or more than 200 symbols.
- To show the current relations among users in the model of hidden communities and track topical trends, we decided to choose posts that were published no earlier than 01.01.2020.
- The authors are not friends in the LiveJournal social network.
- The authors, who published images instead of textual information, were not taken into account in course of corpus development.

The step of filtering friends was the most pivotal one. If two users were friends on social networks, they could not be united by latent links and form a model of a hidden community. This situation conflicted with the term of hidden communities, and the links between such users were considered to be obvious. To omit those users, the id number of each user was checked iteratively in the list of friends of other users. If it was in the list of friends, we did not include the users in the resultant model.

The next step implied corpus preprocessing. The stanza library was chosen for this purpose as it allows us to create a non-stop pipeline in a single code environment. We also used a stop-list during lemmatization to check each token; the stop-list includes prepositions, conjunctions, particles, interjections, symbols of various alphabets, obscene vocabulary, abbreviations, etc. The stop-list is based on a Frequency Dictionary of Contemporary Russian by Olga N. Lyashevskaya and Serge A. Sharoff, as well as words and expressions that were included after checking topic models of the first preliminary procedures: expressions of laughter like xa (hah), graphical representations of emoticons, etc. The total number of stop words is more than 1400. As a result of the above procedures, the size of the final corpus turned out to be about 600,000 tokens, and the initial number of users was 125. On average, each of the users published about 100 posts within 2020 and 2021, and the average length of each text was 35 words.

3.2 Author-Topic Models

Author-topic model (ATM) [19] represents a multimodal procedure, being an extension of a classical LDA technique. In ATM, topic generation is based on an expanded set of parameters, namely, term-document distribution, term-topic distribution, document-topic distribution, and author-topic distribution. The major difference of ATM from LDA consists in putting forward text authorship as a target object of investigation.

The basic generative ATM procedure works as follows: We select an author for each lexical unit from the set of an authors' document. The next step is to select a topic from the distribution of topics corresponding to the author.

Finally, we choose a lexical unit from the distribution of words corresponding to the topic. The graphical model of ATM is presented in Fig. 1.

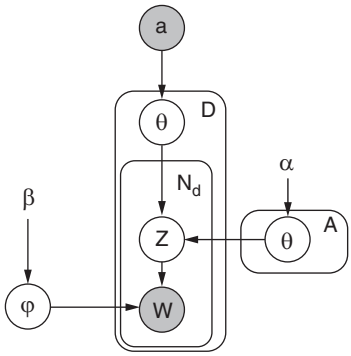

Fig. 1 The representation of ATM

One of the publicly available implementations of ATM can be imported from the gensim library. Its usage is similar to the usage of the standard LDA technique. The choice of this ATM is determined by the successful application of LDA models on different genres of Russian texts [11, 14, 17]. In this library, it is possible to preselect the best model with the highest topic coherence parameter. To do this, we used a cyclic iteration of parameters such as random_state that sets the state of the random number generator inside the author-topic model (from 1 to 7 with a step that is equal to 1) and num_topics which is needed to create the required number of topics covering all text documents in LiveJournal corpus (from 5 to 35 with a step that is equal to 5). As a result, the following parameters turned out to be the most optimal: random_state = 2, num_topics = 15, topic coherence = −2.06. When deriving the resulting topics for each author, we also introduced an additional condition: if a topic occupies less than 10% of all topics of each author, we do not take it into account.

Examples of the ATM output are represented in Table 1. As we also used additional filters, some insignificant topics were not included into resulting topic sets. The topic index corresponds to a conventional index instead of a random number assigned to ATM output. As we focus mainly on author-topic distribution, the resultant sets of topics per authors may overlap, and this peculiarity is exemplified by the repetition of topics marked with *, #, @ indices. The observed overlap is explained by the fact that topics are distributed among all the authors in the corpus with the assignment of the percentage coverage of the topic in each author's subcorpus.

3.3 Contextualized Topic Models

In the NLP domain, standard topic modeling procedures allow one to extract meaningful topical sets of words from both structured and unstructured texts [2]. Unfortunately, such models do not take context into account, therefore, some semantic features of topical sets are unlikely to be mentioned. Nowadays, pretrained language models like BERT fill in this gap, they are used in numerous NLP

Table 1 ATM topics

User	Topic index	Percentage of coverage (%)	ATM lemmata
boyskaut0	+	26.6	бензин, тема, письмо, мама, собственный, подруга, счастье, телефон, нож, учить (gasoline, topic, letter, mother, own, girlfriend, happiness, phone, knife, teach)
	*	14.4	друг, страна, город, два, день, дом, место, история, оказаться, первый (friend, country, city, two, day, house, place, history, turn out, first)
	#	11.04	человек, день, хороший, новый, время, сегодня, первый, вопрос, жизнь, ребёнок (person, day, good, new, time, today, first, question, life, child)
	@	12.16	президент, владимир, выборы, экономика, пожаловать, альманах, глава, студент, сша (president, Vladimir, elections, economics, welcome, almanac, head, student, United States)
sisj	*	13.86	друг, страна, город, два, день, дом, место, история, оказаться, первый (friend, country, city, two, day, house, place, history, turn out, first)
	^	11.97	книга, россия, присоединяться, высота, цена, топливо, радиус, джет, поршневой, новый (book, Russia, join, height, price, fuel, radius, jet, piston, new)
	~	14.3	фото, рубль, июль, результат, поставить, роман, официальный, спутник, бали, рождение (photo, ruble, july, result, put, novel, official, satellite, Bali, birth)
	#	25.91	человек, день, хороший, новый, время, сегодня, первый, вопрос, жизнь, ребёнок (person, day, good, new, time, today, first, question, life, child)
md_ prokhorov	#	27.63	человек, день, хороший, новый, время, сегодня, первый, вопрос, жизнь, ребёнок (person, day, good, new, time, today, first, question, life, child)
	@	21.62	президент, владимир, выборы, экономика, пожаловать, альманах, глава, студент, сша (president, Vladimir, elections, economics, welcome, almanac, head, student, United States)

application, and topic modeling is not an exception. One of such implementations is BERTopic that is an approach that uses transformers and c-TF-IDF to create dense clusters for interpretable topics; it allows keeping important words in the topic descriptions [4]. The algorithm consists of three stages: creating document embeddings, predicting semantic clusters, and printing topic representation from clusters. The c-TF-IDF compares the importance of lexical units to a specific cluster and reveals the most significant lexical units in a topic. It is calculated according to Eq. (1).

$$\mathrm{c\ \ TF\ \ IDF} = \frac{f_i}{\mathrm{wd}_i} \times \log \frac{m}{\sum_j^n f_j}, \tag{1}$$

The frequency in the formula for each word *f* is extracted from each particular cluster *i* and then divided by the total number of lexical units wd of a cluster *i*. It is a way of normalizing the frequency of words in each cluster. Then the number of clusters *m* is divided by the total frequency of the word *f* across all the clusters.

After generating the c-TF-IDF representations, a user obtains a set of lemmata that describe a collection of documents. Of course, it does not mean that the collection of words describes a coherent topic. To improve the coherence of words, the author of the library uses Maximal Marginal Relevance to find the most coherent words without having too much overlap among the words themselves. This action results in getting rid of words that do not contribute to a particular topic.

The graphical representation of BERTopic architecture is presented in Fig. 2.

First of all, we need to tune the model. The posts are distributed among all the authors so that it would be possible to assign particular topics to each author. As we manually include the authorship as a pivotal parameter for the model, the basic BERTopic models are transformed into contextualized quasi-author-topic models. In the ATM algorithm, the assignment of words to a topic and to an author is automatic, so it is important to find out in which model topics will be most fully described from a semantic point of view: in an automatic ATM model or in an automated contextualized quasi-author-topic models. For both standard LDA and contextualized BERTopic model, we choose the topic size that is equal to 10 lemmata. As this number of lemmata is usually set as a standard one in a lot of topic modeling techniques, we decided to use the following filter: if the number of lemmata is less than 10, a topic will not be included in the resultant model and will not be seen in the output. The next step is to choose the language. As the Russian model is absent in the current library, we choose the multilingual BERT model as it is also trained on some Russian data. Some of the results are presented in Table 2 that includes full lists of topics for each user. As we use additional filters, some preliminary topics were not printed in the output; that is why some topical indices are not mentioned in Table 2.

Since we described the system of filters for printing topic models like the set of the optimal topic size, as a result, some authors with little textual data at their pages got no topic assignments. Such users are aleksa_i_sandra, avryabov, etc.

Thus, we did not assign topics to 25 authors out of 125. This author set reduction should not be considered as a data loss, as BERTopic applied to corpora is aimed at dimensionality reduction and structural generalization which is manifested in

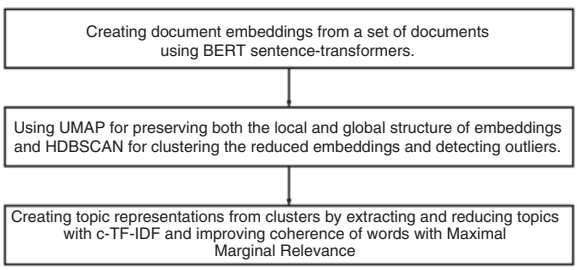

Fig. 2 BERTopic architecture

Table 2 BERTopic topics

User	Topic index	BERT lemmata
boyskaut0	boyskaut0_1	польша, прокатиться, сегодня, рейс, день, вернуться, закончиться, утро, последний, посетить (Poland, ride, today, flight, day, return, end, morning, last, visit)
	boyskaut0_2	осудить, порок, мизулина, борьба, правило, работа, делать, суд, уволить, приказ (condemn, vice, mizulina, fight, rule, work, do, court, dismiss, order)
	boyskaut0_3	кудрин, ретивый, начальник, баба, капитан, опоздать, калининград, захлестнуть, есстесный [естественный], довлатов (kudrin, zealous, boss, woman, captain, be late, Kaliningrad, overwhelm, natural, dovlatov)
	boyskaut0_4	лежать, греф, человека, герман, папка, папенбург, обратить, обозначаться, цыган, адыгейский (lie, gref, human, german, folder, Papenburg, turn, be designated, gypsy, adyghe)
	boyskaut0_5	море, вода, рыбалка, орлан, метр, пароход, огромный, последний, пейзаж, паук (sea, water, fishing, eagle, meter, steamer, huge, last, landscape, spider)
	boyskaut0_6	импортозамещение, тема, брот, непутевой, страна, замена, метаться, гражданин, лакомиться, заправка (import substitution, topic, bread, bad, country, replacement, rush about, citizen, regale, refueling)
	boyskaut0_7	шувалов, неудержимый, путин, владимир, впасть, гнев, файл, пространство, постсоветский, перевод (shuvalov, unstoppable, putin, Vladimir, fall, anger, file, space, post-soviet, translation)
	boyskaut0_8	кошуба, клубника, фестиваль, цель, стол, рынок, резерв, прикупить, предпразничный, праздничный (koshuba, strawberry, festival, goal, table, market, reserve, buy, pre-holiday, celebratory)
	boyskaut0_9	диск, компьютер, собственный, педофил, догадаться, осмотреть, файл, техника, трофей, автомобиль (disk, computer, own, pedophile, guess, inspect, file, technique, trophy, car)
	boyskaut0_10	понятный, навальный, качество, товар, разница, разниться, продовольственный, проветриться, проверить, насколько (easy to understand, bulk, quality, product, difference, vary, food, air, check how much)

(continued)

Table 2 (continued)

User	Topic index	BERT lemmata
sisj	sisj_0	мир, новый, суд, ученый, работа, делать, исследование, день, вода, лес (peace, new, court, scientist, work, do, research, day, water, forest)
	sisj_1	часть, часто, финальный, финал, слово, связывать, освободить, второй, алкоголь, ждать (part, often, final, ending, word, bind, release, second, alcohol, wait)
	sisj_2	январь, февраль, декабрь, наступать, рождество, дорогой, праздновать, праздник, женщина, друг (january, february, december, advance, christmas, dear, celebrate, holiday, woman, friend)
	sisj_3	март, приурочить, полноценный, шествие, грядущий, ближайший, образец, собрать, открытие, десяток (march, time, full-fledged, procession, upcoming, nearest, sample, collect, opening, dozen)
md_prokhorov	md_prokhorov_1	партия, выборы, сегодня, гражданский, власть, дело, последний, время, день, человек (party, election, today, civic, power, matter, last, time, day, man)
	md_prokhorov_2	привет, благодарить, отличный, здравый, дорогой, довольный, гораздо, якиманка, друг, болотный (hello, thank, excellent, healthy, dear, satisfied, much, yakiman, friend, bolotny)

ranking authors as regards significance of their impact to the corpus. Unlike BERTopic, in ATM we did not have any loss of the authors, each of them received at least one topic, but the resulting model seems to be redundant due to topic overlaps.

4 Evaluation

First, before quantitative evaluation of experimental results on formal grounds, we will try to perform qualitative analysis of paradigmatic relations between the topics [5] obtained by means of ATM and BERTopic. Our framework used in qualitative evaluation is inspired by wordnet-thesauri architecture, which reflects variety of hierarchic relations in the lexicon established within and between synonymic sets (synsets), for example, synonymy, antonymy, hyponymy, meronymy, conversion, troponymy, etc. In the given case study, we chose major paradigmatic relations detected in our data. This approach allows us to say if the obtained topics have any common linguistic features.

For instance, judging from Tables 1 and 2, we see that the topics партия, выборы, сегодня, гражданский, власть, дело, последний, время, день, человек (party, election, today, civic, power, matter, last, time, day, man) and президент, владимир,

выборы, экономика, пожаловать, альманах, глава, студент, сша (president, vladimir, elections, economics, welcome, almanac, head, student, united states) are in relation of relative synonymy. The BERT topic deals with the problem of internal politics, while the ATM topic describes foreign politics, although they are both in the semantic field of politics.

Among other paradigmatic relations, we can also discuss antonymic relations. Among the main topics of the user yuripasholok, the BERTopic algorithm highlighted the following one: танк, военный, время, техника, война, являться, машина, история, порой, технический (tank, military, time, technique, war, be, machine, history, sometimes, technical). The ATM topic is действие, особо, ситуация, приводить, куча, происходить, государство, истребитель, политика, режим (action, especially, situation, lead, heap, occur, state, fighter, politics, mode). They are both in relation ground forces/air forces as the second set contains the истребитель (fighter) word. It is worth noting that the topic of the air force can also refer to a general political topic. As a result, these two sets can be in relation of the hyponym and hypernym because military forces are an integral part of politics.

Finally, antonymic topics are clearly observed in the profile of diak_kuraev: церковь, патриарх, православный, русский, андрей, история, россия, монастырь, храм, день (church, patriarch, orthodox, russian, andrey, history, russia, monastery, temple, day) (BERTopic) and путин, россия, русский, мотор, дождь, народ, чиновник, умереть, завтра, вирус (putin, russia, russian, motor, rain, people, official, die, tomorrow, virus) (ATM). These two sets are in relation to church and state. The problem of their relationship has always been one of the most important in the history of Russia.

Turning to the formal parameters of the assessment, it is important to say that for probabilistic topic models, the u-mass topic coherence metric is most often used.

In general, the issue of automatic evaluation of the quality of topic models has always been acute. In practice, in addition to the u-mass topic coherence, NLP researchers also resort to other metrics. For example, you can use the normalized pointwise mutual information (NPMI) coherence. It has been found to correlate best with human judgment in most experiments for English corpora [9]. However, this process is more time consuming and more often used for larger corpora, so in our experiment we decided to use the u-mass metric. The closer the value of the metric to zero is, the more coherent and stable the resulting topical sets are. In Sect. 3, we mentioned that, when looping over for this dataset, we got a topic coherence that is approximately equal to -2, which really showed that the final author-topic models had the right to exist. Unfortunately, in libraries for contextualized topic modeling algorithms that are based on pre-trained BERT language models, we cannot apply the same metric. Nonetheless, we should note that in BERTopic, there is such a parameter as topic similarity, which shows the heterogeneity of topics for each of the authors of the corpus. Figure 3 shows the proximity table for the topics of the zelenyislon user. On average, the topic similarity parameter is less than 0.6, which really indicates the heterogeneity of topics and, as a result, the absence of the need to remove duplicate topics, which characterizes the stability of the resulting sets.

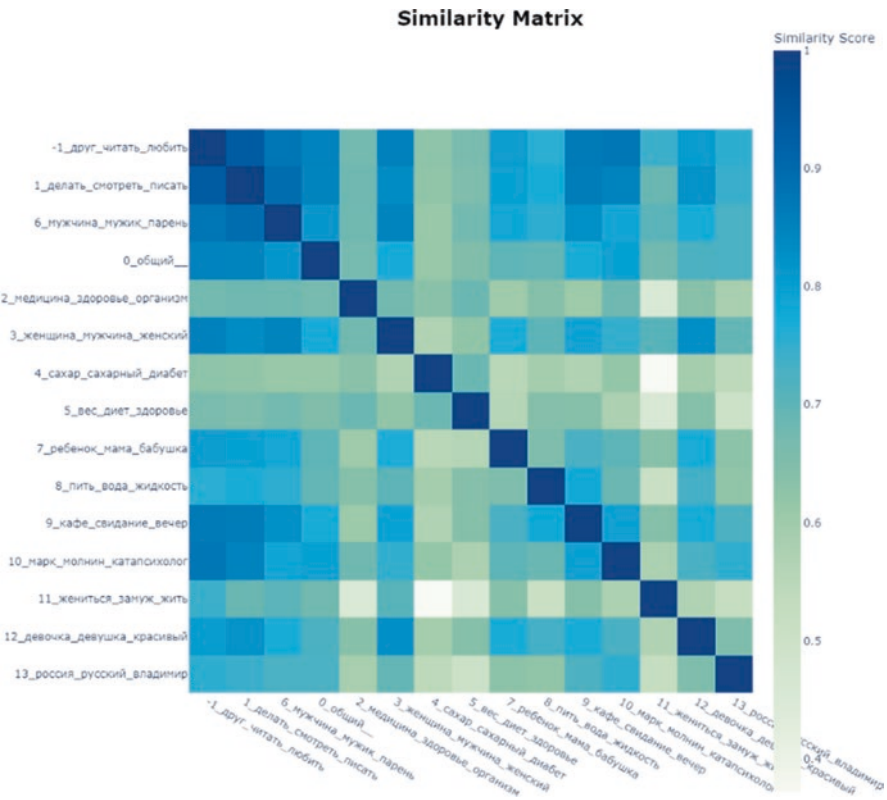

Fig. 3 The topic similarity of the *zelenyislon* user

5 Conclusion

Based on the linguistic analysis and description of the models obtained, we can conclude that despite some shortcomings in each of the models, both of them can be used to build a model of hidden communities. Depending on the choice of a topic modeling algorithm, various techniques may be used. Nevertheless, from the point of view of traditional linguistics, it seems important that the resulting sets, in their paradigmatic properties, should be closer to co-hyponyms than to hypernyms, since we will be able to observe specific ideas that are reflected in text collections. Therefore, contextualized models may have an advantage over probabilistic ones.

When working with topics obtained with the help of BERT, it is necessary to simplify manual unification of topics. The procedure for automatic topic labeling will reduce the dimension of a set of 10 lemmata to a lexical unit, which will approximately cover the idea of an original text. Since the source texts were taken from the web, the labels have to be obtained from the vector model of a corpus, which is based on web texts [8, 15]. It is important to note that, unfortunately, the

speed at which large open-access representative web corpora appear is too slow for the resulting labels to perfectly match topic models. Therefore, another way might be connected with obtaining labels from search engines [13]. To do this, the topic lemmata can be turned into a search query. Labels will be used for creating a graph of users that will be a resultant model of hidden communities. As for ATM topics, the step of automatic topic labeling is unnecessary since some author-topic models are the same with the same set of lemmata. We can use topical sets for creating a graph.

Further research will be aimed at developing procedures for automatic topic labeling and testing them, creating a model of hidden communities for the LiveJournal social network and its linguistic description, as well as using text corpora collected from other Russian segments of social networks.

References

1. Ahmad, S.N., Laroche, M.: How do expressed emotions affect the helpfulness of a product review? Evidence from reviews using latent semantic analysis. Int. J. Electron. Commer. **20**(1), 76–111 (2015)
2. Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E.: Cross-lingual contextualized topic models with zero-shot learning. arXiv preprint arXiv:2004.07737. (2020)
3. Danilova, V., Popova, S., Karpova, V.: A pipeline for graph-based monitoring of the changes in the information space of Russian social media during the lockdown. arXiv preprint arXiv:2110.13626. (2021)
4. Grootendorst, M.: BERTopic: leveraging BERT and c-TF-IDF to create easily interpretable topics (2020). https://doi.org/10.5281/zenodo.4381785
5. Koltsov, S.N., Koltsova, O.J., Mitrofanova, O.A., Shimorina, A.S.: Interpretation of semantic relations in the texts of the Russian LiveJournal segment based on LDA topic model. In: Proceedings of the XVII All-Russia Joint Conference "Internet and Modern Society" IMS-2014, Saint-Petersburg, 19–20 November 2014
6. Koltsova, O., Nagorny, O.: Redefining media agendas: topic problematization in online reader comments. Media Commun. **7**(3), 145–156 (2019)
7. Koltsova, O., Alexeeva, S., Pashakhin, S., Koltsov, S.: PolSentiLex: sentiment detection in socio-political discussions on Russian social media. In: Filchenkov, A., Kauttonen, J., Pivovarova, L. (eds.) Artificial Intelligence and Natural Language. AINL 2020 Communications in Computer and Information Science, 1292, pp. 1–16. Springer, Cham (2020)
8. Kriukova, A., Erofeeva, A., Mitrofanova, O., Sukharev, K.: Explicit semantic analysis as a means for topic labelling. In: Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings, pp. 167–177. Springer, Cham (2018)
9. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539. Association for Computational Linguistics, Gothenburg (2014)
10. Li, A., Daud, J., Zhou, L., Muhammad, F.: Knowledge discovery through directed probabilistic topic models: a survey. Proc. Front. Comput. Sci. China. **4**, 280–301 (2010)
11. Mamaev, I., Mitrofanova, O.: Automatic detection of hidden communities in the texts of Russian social network corpus. In: Filchenkov, A., Kauttonen, J., Pivovarova, L. (eds.)

Artificial Intelligence and Natural Language. AINL 2020 Communications in Computer and Information Science, 1292, pp. 17–33. Springer, Cham (2020)
12. Mamaev, I., Mitrofanova, O.: Hidden communities in the Russian social network corpus: a comparative study of detection methods. In: CMCL, pp. 69–78 (2020) https://ceur-ws.org/Vol-2780/
13. Mitrofanova, O.A., Mirzagitova, A.: Automatic assignment of labels in topic modeling for Russian corpora. In: Botinis, A. (ed.) Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics: ExLing 2016 Proceedings, pp. 115–118. Saint Petersburg State University, St. Petersburg (2016)
14. Mitrofanova, O., Sampetova, V., Mamaev, I., Moskvina, A., Sukharev, K.: Topic modelling of the Russian corpus of Pikabu posts: author-topic distribution and topic labelling. In: IMS, pp. 101–116 (2020) https://ceur-ws.org/Vol-2813/
15. Mitrofanova, O., Kriukova, A., Shulginov, V., Shulginov, V.: E-hypertext media topic model with automatic label assignment. In: Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings Communications in Computer and Information Science, 1357, pp. 102–114. Springer Nature, Cham (2021)
16. Palani, S., Rajagopal, P., Pancholi, S.: T-BERT-model for sentiment analysis of micro-blogs integrating topic model and BERT. arXiv preprint arXiv:2106.01097. (2021) https://arxiv.org/abs/2106.01097
17. Panicheva, P., Mirzagitova, A., Ledovaya, Y.: Semantic feature aggregation for gender identification in Russian Facebook. In: Conference on Artificial Intelligence and Natural Language, pp. 3–15. Springer, Cham (2017)
18. Peinelt, N., Nguyen, D., Liakata, M.: tBERT: topic models and BERT joining forces for semantic similarity detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7047–7055. Association for Computational Linguistics, Gothenburg (2020)
19. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. arXiv preprint arXiv:1207.4169. (2012) https://arxiv.org/abs/1207.4169