CHAPTER 23

Topic Modeling in Russia: Current Approaches and Issues in Methodology

Svetlana S. Bodrunova

23.1 Introduction

23.1.1 Topic Modeling as a Scientific Method

Topic modeling is a method of probabilistic clustering of textual documents mostly used for large text collections. It finds itself on the crossroads of probabilistic and predictive text classification, natural language processing methodologies, semantic analysis, and discourse studies. In this chapter, we look at how the teams involving Russian-speaking scholars have enhanced the topic modeling algorithms, tested their efficiency, and employed them for interpretation of real-world datasets, including those from today's social media—either in Russian only or for the Russian cases in comparison with those in other languages.

For many scholars, topic models are about latent semantic analysis, or LSA (Steyvers and Griffiths 2007), but, algorithmically, LSA appears to be only one option of topic modeling; a large variety of algorithmic approaches and extensions to them have been suggested within the last two decades (Blei and Lafferty 2009; Korshunov and Gomzin 2012). The main goal of using any topic modeling algorithm is to detect the so-called topics in a text collection. In communication terms, a topic is a theme around which the discussion is evolving; but, in topic modeling, topics express themselves via collections of words and/or documents that the modeling algorithm considers similar and/or related to each other.

S. S. Bodrunova (✉)
St Petersburg State University, St Petersburg, Russia

D. Gritsenko et al. (eds.), *The Palgrave Handbook of Digital Russia Studies*, https://doi.org/10.1007/978-3-030-42855-6_23

The basis for the texts to be related to each other is word co-occurrence. The method implies that texts belonging to one topic may be described as those where particular words stay close to each other or, at all, can be found. This understanding leads to the probabilistic iterative process which sees the dataset as a "bag of words" where the word order and syntactic links between the words are ignored. It defines, by multiple iterations, which words most probably stand together in which documents. Computationally, a topic is a discrete (multinomial) probability distribution over terms in a given vocabulary (Mcauliffe and Blei 2008, 121). Thus, each document belongs to each topic with some probability (often negligibly small), but some texts belong to some topics with much higher probabilities, with an arbitrary threshold for where to cut the "long tail" of nonrelevant texts. The results of the modeling are represented in two matrices: the word-topic one (the probabilities of particular words to belong to a topic) and the topic-document one (the probabilities of a topic to be found in a particular document); but the end-users usually assess the top words (the words with the highest probability for the topics) and the most probable texts in the topics. For the end-user, a topic is a collection (cluster) of texts that belong with high enough probability to one theme slot and are expected to be linked by topicality of their content.

The quality of modeling—that is, how well the topics are separated from each other, how many texts they involve above the relevance threshold, and how interpretable they are—may be measured by the metrics of topic interpretability, coherence, robustness, et cetera. The baseline for topic quality assessment is human coders' interpretation, but a lot of automated metrics of quality have been developed to make the topic quality assessment quicker and easier.

Of the Bayesian bag-of-words algorithms, the one based on Dirichlet distributions and called latent Dirichlet allocation (LDA) (Blei et al. 2003) is, undoubtedly, the most developed today. Along with it, for various types of data, several other algorithms have matured and also gained important extensions that allow the scholars to intervene and change the parameters of the algorithm. In terms of allowed intervention, topic modeling may be unsupervised, supervised (Mcauliffe and Blei 2008), semi-supervised (Bodrunova et al. 2013), or weakly supervised (Lin et al. 2011). Alternative promising approaches to topic detection mostly try to preserve the semantics that stem from word order and grammatical relations between words, like the approaches based on Markov chains (Gruber et al. 2007) or n-grams (Wang et al. 2007).

Topic modeling has advantages quite attractive for scholars, as well as shortcomings inherent for the method. Among the latter, there are the principal instability of clustering results (i.e. different runs resulting in a slightly different shape of topics) and an impossibility of a priori definition of the optimal number of topic slots for getting the most robust and interpretable topics. Due to this, multiple runs are practiced, with the varying number of slots for topics—usually, 50 to 400, depending on the nature of the dataset. Another inherent problem is dependence of the results upon the length of texts in the dataset: the longer the texts, the more material there is for an algorithm to analyze;

thus, the topics formed of shorter texts are more vulnerable to non-interpretability. Another set of complications lies in malformation of topics from the human viewpoint; for example, among "bad" topics, there are topics dominated by general words, mixed and "chained" topics, or those where one theme splits to several topics (Boyd-Graber et al. 2014, 235–37).

Technical issues about topic modeling are, first, its relatively low feasibility, as the data for topic modeling, especially the real-world datasets, demand several steps of preprocessing (including stemming, lemmatization, and cutting out stop-words) and then either human interpretation or automated quality assessment plus reading by coders; second, it is the dependence on available software and hardware, as collection and processing of large datasets demands a lot of resources.

But, despite the aforementioned discrepancies, topic modeling remains attractive to the scholars, as it has several key (even if arguable) advantages. The first one is that, in comparison with naïve keyword search, the topics unite the texts that might belong to a discussion subtheme but do not contain the keyword, thus enriching our understanding of how people discuss the theme and what it is linked to. The second advantage is that topic modeling may be easily combined with other methods and can serve as a processing tool for other computational goals, including dataset dimensionality reduction. Topic modeling has already proven to be efficient "for a wide range of research-oriented tasks, including multi-document summarization, word sense discrimination, sentiment analysis, machine translation, information retrieval, discourse analysis, and image labeling" (Boyd-Graber et al. 2014, 227).

The third advantage is that the method is believed to be language-independent (given that the language is not hieroglyphic): it means that the algorithms work with words as independent units of analysis, and this approach is suitable for any language. However, today, this assumption is questioned. Topic modeling *per primo* was created for analytical languages such as English, and synthetic languages including Russian, where a role of inflexions for transferring meanings is high, experience additional complications in word preprocessing. Thus, 12 possible case forms of a noun in singular/plural need to be distinguished from numerous forms of the same-root verb in singular/plural in three tenses; for modeling, both the noun and the verb need to "collapse" into singular-nominative (for nouns) or indefinite (for verbs) forms. Moreover, contextual linkages between words arranged, for example, with the help of diminutives, may be lost in stemming.

An overwhelming multitude of descriptions of topic modeling in general, with their advantages and shortcomings (Boyd-Graber et al. 2014; in Russian, Korshunov and Gomzin 2012), as well as particular algorithms, may be found elsewhere (for more detailed example of the procedures of topic modeling applied to a Russian language, see Chaps. 24 and 25). Here, we will focus on how the scholars who deal with the Russian-language datasets develop the topic modeling methods tackling the issues stated above, including

topic quality assessment, and interpret the public discussions in Russia with the help of topic models.

23.1.2 Topic Modeling for the Russian Language

To our best knowledge, there has so far been no extensive review of how topic modeling has developed for the Russian language. This gap exists despite the fact that Russian-oriented topic modeling studies appear to be one of the most developed beyond the English-language realm, outnumbering German, French, and Spanish in terms of methodological suggestions and cases of application. Also, topic modeling for Russian is considered the most developed among the highly inflected languages like Slavonic ones. Contributions by the scholars working with the Russian-language datasets have become internationally recognized.

To make our review more systematic, we will divide the works into groups. For Russian, topic modeling studies may be divided into *methodological* (that develop, compare, and extend models as well as evaluate their quality), *applied* (that apply topic modeling to extract the meanings from datasets), and *relational* (that relate topic modeling results to other features of the datasets or external factors). Of course, in the case of a rapidly developing method like topic modeling, nearly all the works that use it become methodological, as the method is used in a particular variation which needs to be chosen, grounded, and often reworked or extended. But still we see this distinction as fruitful to structure the results that have been achieved by the scholars. Also, a separate group of works focuses on topic quality assessment. We will also mention topic modeling for short texts like tweets, as, first, modeling for Twitter occupies a separate arena in international topic modeling studies and, second, it has also started to be developed in Russia (for more, see Chap. 30).

The chapter is, thus, organized as follows. In Sect. 23.2, we provide an overview of the methodological papers; here, we summarize the main directions of development of topic modeling for Russian and the main issues that the researchers work upon, including modeling for short texts. In Sect. 23.3, we review the works that deal with topic quality assessment. In Sect. 23.4, we focus on both Russian- and English-language papers about meaning extraction; here, we review the papers that link topic models to other text features, research methods, and contextual knowledge. In particular, we will look at how topic models are used in a wider context of aspect extraction and sentiment analysis. In concluding remarks, we indicate the potential research gaps and the prospects for future studies.

23.2 Methodological Studies of Topic Modeling for the Russian Language

23.2.1 Model-oriented Works: LDA and pLSA

In the recent decades, there have been several groups within Russia who have been focusing on various topic modeling algorithms.

Thus, in a sequence of influential works, Koltsova and colleagues have been developing LDA (Koltcov et al. 2014) and a range of extensions and improvements to it. What this group has tried to tackle, with the help of Russian-language datasets from LiveJournal and VKontakte, were the dataset-level and the topic-level issues.

On the level of dataset, the group has dealt with instability of the results of modeling, nonexhaustive LDA results, the quality of sampling and optimization of the number of topics; we will now review the group's achievements in the stated order.

Thus, the topics that appear in two runs of the model are, logically, more stably present in the dataset than those that appear only once and may be occasional. Based on Kullback-Leibler divergence for topic models, the authors have introduced the normalized Kullback-Leibler topic similarity metric (NKLS) for multiple runs (Koltcov et al. 2014). They have used NKLS and also the Jaccard topic similarity metric (Bodrunova et al. 2017) to assess the stability of topics. They have also introduced several LDA extensions to make the results more stable: among others, one is granulated LDA (Koltcov et al. 2016a) similar to the idea of using *n*-grams (Batura and Strekalova 2018; Sedova and Mitrofanova 2017a), and another is LDA with local density regularization (Koltcov et al. 2016b).

Doing topic modeling in search for a particular result (say, the public opinion on a particular event or issue), a researcher cannot be sure that the topics (s)he finds in the modeling results represent the full picture of the public discourse. Thus, the group has introduced interval semi-supervised LDA (ISLDA) that links naïve keyword search with probabilistic clustering by attaching word labels to topic slots, thus making the algorithm "crystallize" the topics around keywords (Bodrunova et al. 2013). By attaching the same keyword to several topic slots, a researcher can exhaust the respective theme in the dataset, at the same time getting the topics "thin" enough to see multiple aspects of the discussion (Koltcov et al. 2017).

As to sampling, it is the core procedure of the method that defines in which order the words are sampled (metaphorically, "taken out of the bag of words") to be probabilistically put together. Most researchers use Gibbs sampling for LDA (Blei et al. 2003), while expectation maximization (EM-algorithm; Mashechkin et al. 2013) and Expectation-Propagation algorithm can also be used (Minka and Lafferty 2002). After introducing the granulated LDA, Koltcov et al. (2016c) have also suggested an optimization for Gibbs sampling for granulated data.

And, last but not least, selecting the optimal number of clusters was tackled. The number of topics is crucial for the results, and, in unsupervised models, it is the only parameter set by the researcher. Usually, multiple runs with varying number of topics are necessary to choose the number closer to optimal, and automation of selection of the number of topics is a separate scientific task. Using the maximum entropy principle, Koltcov et al. (2018) have suggested applying Rényi and Tsallis entropies to find the optimum number of topics. Other groups of scholars have suggested using text representations by dense vectors and sentence embeddings for the same purpose (Krasnov and Sen 2019; Bodrunova et al. 2020).

Topic-level discrepancies of the method were less a focus of attention for this research group, but, in most of their works, they describe the coding experience and the problems of topic interpretability. Thus, they show that human interpretability is linked to the writing style of the authors of the texts in the dataset, as well as to the number of topics, and that the focus of the topic ("war" vs. "Israeli-Palestinian conflict") matters much for qualitative studies (Koltsova and Koltcov 2013). For dealing with specifically Russian-related issues like the synthetic structure of the language, the group has successfully used pre-developed decisions on lemmatization and have involved contextual interpretations in their works described below, successfully linking the use of topic modeling to qualitative studies of social media and beyond (Koltcov et al. 2017). The group has developed its own software TopicMiner and has worked mostly with texts from the Russian LiveJournal, VK.com, and other social media datasets.

Similarly, the works by Vorontsov and colleagues (e.g. Vorontsov et al. 2015a, 2015b; Vorontsov and Potapenko 2015) have been influential in exploring probabilistic LSA (pLSA) and its modifications based on non-Bayesian regularization. PLSA differs from LDA, as parameters of discrete distributions are estimated via likelihood maximization, with nonnegativity and normality constraints, while LDA uses Dirichlet distribution and additional parameters that help reduce overfitting (Potapenko and Vorontsov 2013, 784). In particular, Vorontsov and colleagues have shown that robust pLSA performs better than LDA for certain tasks; they have also suggested a generalized learning algorithm for probabilistic topic models (PTM), arguing that the currently used algorithms of topic modeling may all be viewed as specific cases of such an algorithm but with differing sets of algorithmic features like regularization, sampling, update frequency, sparsing, and robustness (Potapenko and Vorontsov 2013, 784).

Within this logic, and also advocating for avoidance of unnecessary probabilistic assumptions in natural language processing (Vorontsov and Potapenko 2015, 304), the group has developed ARTM—a non-Bayesian additive regularization of topic models. The authors have argued that, mathematically, "[l]earning a topic model from a document collection is an ill-posed problem of approximate stochastic matrix factorization" and that "[m]any requirements for a topic model can be more naturally formalized in terms of optimization criteria rather than prior distributions. Regularizers may have no probabilistic

interpretation at all" (Vorontsov and Potapenko 2015, 304). ARTM as a regularization framework that integrates many potential regularizers for topic modeling parameters, as the authors have shown. The authors' claim of high efficiency of their approach, as well as of BigARTM, an open-source library for additive-regularized topic models (Vorontsov et al. 2015a, b), remains unchallenged (Kochedykov et al. 2017). Later, the group has developed TransARTM based on hyper-graph multimodal modeling for "transactional data" where transactions are interactions between network nodes, for example, users on social networks (Zharikov et al. 2018) and have suggested an ARTM improvement by relying on segmental structure of texts (Skachkov and Vorontsov 2018).

Also, this group of scholars has tested two algorithms for the Russian-language short texts, namely biterm topic modeling (BTM) and word network topic model (WNTM) (Kochedykov et al. 2017, 191). These algorithms were also tested against LDA for short texts including tweets (see below) and user queries (Völske et al. 2015).

Despite their varying algorithmic preferences, the research groups led by Koltsova and Vorontsov have collaborated on additive and regularized topic models (Apishev et al. 2016a, b). Also, Vorontsov and colleagues have published important methodological and review papers in Russian, including one on regularization, robustness, and sparsity of probabilistic topic models (Vorontsov and Potapenko 2012).

The similarity between these groups of scholars lies in their focus. First, they both develop the methodologies on the level of dataset, and the level of word in a text corpus mostly remains their secondary concern. This, it seems, stems from the fact that, second, they both treat Russian as "language as such"—just as English is used in topic modeling, often without discussing inherent linguistic or contextual limitations of analytical/inflective languages. This has its advantages, as the language is not treated as "local," and thus the scholars avoid the "colonial" relations between more universal English and more localized other languages. Also, the authors' contributions can be easily applied to other languages. But, at the same time, they, to some extent, overlook the word-level of topic modeling, being, of course, well aware of the achievements of Russian computer linguists in developing opinion mining for Russian.

23.2.2 Computer-linguistic Approaches to Topic Modeling

The latter efforts have, for decades, been concentrating in several groups vaguely linked to each other via the conference on computational linguistics and intellectual technologies called "Dialogue" dedicated to, inter alia, sentiment analysis and aspect detection (for details, see dialog-21.ru). For years, in the conference proceedings and individual papers, the notion of topicality and topic detection has been developing on the level of word semantics and lexical relations. Semantic proximity, ambiguity of meaning, inflections and their impact upon word semantics, sentiment, and other features of lexical units have

been the focus of attention of this sparse "school" or, rather, array of research groups.

Here, we find an understanding of goals of topic modeling that differs from that in the previously described studies. Topic modeling is seen here as a tool for resolution of grapheme-, word-, or fragment-level tasks, such as, for example, relevance detection for automatic text annotation (Mashechkin et al. 2013), automatic content filtration and genre detection (Voronov and Vorontsov 2015), or aspect-based (Rubtsova and Koshelnikov 2015) and non-aspect-based sentiment analysis (Koltsova et al. 2016a; Tutubalina and Nikolenko 2015). Such an approach shifts the very notion of what a topic is: thus, already as early as in 2000, Loukachevitch and Dobrov (2000) noted that topics may be viewed as semantically linked chains of words, thus stating the necessity for a topic to preserve both the grammatical and semantic relations between lexical units. Loukachevitch, Dobrov and their colleagues who, for over two decades, have been dealing with both hard and fuzzy classification methods for the Russian language have developed the notions of "thematic knots" and "thematic text representation" based not on co-occurrence but on semantic relatedness of words in documents (Loukachevitch and Dobrov 2009; for more, also see Chap. 18).

In accordance to this, within computational-linguistic approaches, topic modeling is often used for the tasks that deal with the level of a lexical unit, and not always with great success in comparison with other methods of computational linguistics. Thus, one recent work by Davydova (2019) unites LSA-base modeling with the use of contextual vectors for the task of disambiguation and differentiation of meaning. It successfully unites LSA with word-vector logic to detect thematic relevance of lexemes. In other works (see, e.g., Lopukhin and Lopukhina 2016; Lopukhin et al. 2017), though, it was argued that, for lexical disambiguation, word2vec approaches were more efficient than LDA and other topic modeling approaches based on bag-of-words logic, as topic modeling works on the level of document/dataset.

Thus, the two approaches to developing topic models—the method-oriented one and the computational-linguistic one—seem to be moving forward but without being interconnected, not integrating each other's achievements into research practice, even despite co-publications and collaboration. There is an evident lack of works that would both develop the topic modeling algorithms *and* have in mind the peculiarities of the Russian language. Despite the evident necessity of integration of the two logics, it is rarely found also for other inflective languages; we see this logic explicitly employed by only one group working in Slovenian (see, e.g., Maučec et al. 2004, and later works). Beside this, several works by computer linguists have suggested decisions for the Russian language, including adding automated labeling to Russian-language topics (Mirzagitova and Mitrofanova 2016) and showing the possibility of domain term extraction by topic modeling (Bolshakova et al. 2013). Automatic topic labeling by a single word or phrase is expected to ease topic interpretation; working upon it continued in the recent years by

comparing quality of two labeling algorithms, namely the vector-based Explicit Semantic Analysis (ESA) and graph-based method, with the former one preferred by the authors (Kriukova et al. 2018).

23.2.3 Topic Modeling for the Russian Twitter

Unlike for longer texts, short-text modeling for Russian is also done within comparative international context. For instance, there are at least three methodological works that explore topic modeling for the Russian Twitter (Mimno et al. 2009; Sridhar 2015; Gutiérrez et al. 2016; for more, see Chap. 30) while developing multilingual modeling tools. The first two do not discuss individual results for any single language, and the third only observes one difference in description of sports between Russian- and English-language Twitter. Similarly, only a small handful of works applies topic modeling to Russian Twitter to detect substantial meanings or discussion features. Thus, one work (Chew and Turnley 2017) has shown the divergence between Russian- and English-language "master narratives" on Russian cyber-operations.

The works by Bodrunova and colleagues appear to be the only continuous effort (since 2013) to combine topic modeling for Twitter with various other instruments of automated text analysis, also in comparison with other languages (Bodrunova et al. 2019a, c). Thus, we have tested three topic modeling algorithms, namely unsupervised LDA, WNTM, and BTM (Blekanov et al. 2018), and have shown that BTM works best, as measured by normalized PMI and Umass (see below). We have also applied BTM to detect the dynamics of topicality in conflictual discussions (Smoliarova et al. 2018) and have demonstrated that the saliency of topics in time may help detect pivotal points in mediated discussions. Experiments with datasets on Twitter in three languages, including Russian (Smoliarova et al. 2018, Bodrunova et al. 2019a), show that sentiment of tweets is linked to topicality: thus, more interpretable topics are more sentiment-loaded, in particular negativity-loaded (Bodrunova et al. 2019a). Another study (Bodrunova et al. 2019b) has shown that topic interpretability may be linked to topic robustness and topic saliency.

23.3 Quality Assessment and Interpretability of the Russian-language Topics

All around the world, a vast array of works on topic modeling is dedicated to finding and testing the metrics of its quality. Arguably, these metrics may be divided into those assessing the overall quality of the modeling and those of the topic level. Here, we will review the contribution by the Russian scholars to topic modeling quality studies.

One of the first metrics that were used to assess the modeling itself was *perplexity*—a predictive metric of how well the current distribution matrices

predict the results for new samples. Perplexity has been assessed by Koltcov et al. (2014); they have shown that it is unclear how to use it for qualitative studies in topic modeling, due to inability to establish how dictionary-dependent perplexity is linked to human interpretability of topics. Instead, to measure the quality of modeling, the group has introduced word and document ratios that allow drastically cutting the dictionary of the dataset for computation and suggested a new metric for topic stability measurement. The idea of this metric is that "good" topics are both human-interpretable and stable in multiple runs. As we mentioned above, Koltsova and colleagues have introduced normalized Kullback-Leibler divergence-based metric of topic similarity (NKLS) that allows for detecting similar and stable topics.

They have also improved another traditional metric such as term frequency–inverse document frequency (*tf-idf*). Tf-idf calculates values for each word in a document through an inverse proportion: frequency of the word in a particular document against the percentage of documents the word appears in—which gives a hint on how relevant a given word is in a given document. Tf-idf values allow for calculating the tf-idf coherence metric, to see whether the topics are composed of the words highly relevant for them (Koltcov et al. 2017).

Coherence as a measure of topic quality is one of the basic metrics suggested in early years of topic modeling, but later, other automated metrics were introduced. An extensive study of nine automated metrics juxtaposed to the human-coding baseline was performed by Nikolenko (2016). The author has looked at several classes of metrics, including coherence, pairwise pointwise mutual information (PMI), and metrics elaborated by the author based on distributed word representations where each word is represented as a vector in a semantic space (word2vec approach). The author shows that normalized PMI (NPMI) suggested in the paper outperforms PMI as well as other conventional metrics like tf-idf, but vector-based metrics work even better than NPMI. But the question remains whether both NPMI and word2vec metrics work well for short texts, as there is evidence that NPMI marks the topics as good while they remain low-interpretable for human coders (Bodrunova et al. 2019b). For automated topic assessment versus human interpretability, an important attempt to introduce a quality metric has recently been made. Mavrin et al. (2018) have introduced a new interpretability score for top words, based both on assessing the word probability against an external dataset of frequently used words and on pairing the words and assessing the pairs' coherence. In parallel, Alekseev et al. (2018) have suggested intra-text coherence as a metric to improve interpretability, fairly arguing that topic coherence and interpretability cannot stand for each other, due to a very small percentage of text volume covered by the topic's top words. Another work has discussed metrics based both on linguistic and probabilistic similarity for hierarchical topic modeling, a special sort of topic modeling (Belyy et al. 2018).

But none of these works has primarily focused on the causes in human (non-)interpretability of the topics, mostly seeing human coding as a baseline—perhaps because, for longer texts, when interpretability was at stake, the

models performed well enough. Thus, Koltsova and Koltcov (2013) have shown that, for long texts like LiveJournal posts, circa two-thirds of the topics are interpretable after LDA has been applied. They have also identified three types of uninterpretable topics: "language" (other than Russian), "style" (writing styles, including offensive language), and "noise" (uninterpretable texts/combinations of texts) (Koltsova and Koltcov 2013, 218). In our pilot studies, though, we have seen that topics for Twitter are less interpretable, with only up to 40–45% identified as such in all the three languages (Bodrunova et al. 2019a, b); thus, it is not the nature of Russian alone that seems to be causing lower topic interpretability in the case of Russian Twitter. Also, we examined the features of top words and found that their negative sentiment could actually raise topic interpretability (Bodrunova et al. 2019a).

23.4 Use of Topic Modeling for Content Interpretation

In this part of our chapter, we provide a short overview of how the topic models have been applied to social and language studies. A detailed review, though, would demand a separate chapter, as many findings by scholars working with the Russian data are illuminating enough; here, we will only indicate the examples of content-exploring research aiming to demonstrate the variety of possible applications of topic modeling for today's social science. Also, many works have already been discussed above, and, here, we will only mark the major findings.

The works exploring content may be divided into "purely applicational" and "relational." The former apply the methods to generate findings relevant for social science; the latter relate such findings to other phenomena or research methods. Also, content-exploring research has scrutinized both social media and text collections beyond them.

In social media studies, topic modeling was first employed to map the agenda of the Russian LiveJournal (Koltsova and Koltcov 2013), finding that the topical structure of posts of the top 2000 Russian LiveJournal authors was quite stable across time and, thus, challenging the notion of dissipative social media agendas. Later, this structural finding was amplified by analyzing the structure of co-commenting communities in LiveJournal (Koltsova et al. 2016b) which showed that the role of individual authors and active commentators was higher than that of topics for the stability of commenting structure.

The two major themes explored via topic modeling have been politics and ethnicity. Thus, Koltsova and Shcherbak (2015) have shown how the bias in political LiveJournal posts correlated with the ratings of the leading parties and presidential candidates in the 2011–2012 election campaigns in Russia. This chapter is an example of combining topic modeling as a dataset reduction instrument with manual coding and descriptive statistics performed for the reduced dataset. Also, Smoliarova and colleagues (2018) have shown that

assessment of topic saliency (i.e. which topics stick out and when) may help detect pivotal moments in development of conflictual political discussions online.

Other important works add to media effects theory, including agenda setting and media framing. They, inter alia, demonstrated how agendas on the Ukrainian conflict were gradually diverging on the Russian and Ukrainian TV, thus coming from different framing to building differing agendas (Koltsova and Pashakhin 2017), and that the agendas in news and user comments on Russian regional news portals diverge (Koltsova and Nagornyy 2019). Later, a full-cycle methodology was suggested for co-analysis of news topicality and user feedback (Koltsov et al. 2018). Another group of scholars has also applied LDA to analysis of newspaper coverage on climate change in 2000–2014 (Boussalis et al. 2016) identifying national-level and newspaper-level factors influencing the volume and framing of the coverage.

In a separate line of research, the scholars have explored ethnic content of the Russian social media (Apishev et al. 2016b; Nagornyy 2018a), including detection of most hated ethnicities (Bodrunova et al. 2017), as well as user ethnicity and gender versus attitudes toward ethnic groups (Nagornyy 2018b). Here, topic modeling has produced results unavailable by means of surveys or field research. It has been shown that Americans (outside Russia) and Caucasian nations (inside Russia) provoke the most negative discussion; also, a clear division of attitudes in the Ukrainians-related topics had shown up in LiveJournal much before the Ukrainian conflict started.

Last but not least, beyond the social networking realm, LDA has been applied to Russian and English prose with the aim of facilitating translation of fiction (Sedova and Mitrofanova 2017b) and to a corpus of musicological texts, with the purpose of automated defining syntagmatic and paradigmatic relations between terms (Mitrofanova 2015). In the former work, the authors have added bigrams to the LDA algorithm to detect the differences in various translations of novels. The paper shows high differences in topical structure between English and Russian versions of novels but shows that this diversity may be used for lexical and topical comparison of prose translations. The latter paper is of descriptive nature and was conducted to show that automated text clustering provides the results that are in line with expert knowledge on musicology.

23.5 Conclusion

Among highly inflected languages, Russian is today the most researched upon in terms of topics models and their applications. The scholars working with Russian-language data have successfully employed the existing methods and have suggested both their universally applicable modifications and new quality metrics. Significant results going much beyond the modeling methodology have been achieved in analysis of social structures of online communication, agenda setting and framing, ethnic studies, and political factors of user discussions.

At the same time, we have identified a gap between method-oriented works that develop topic modeling for Russian as "language as such" and the math-linguistic approach that is Russian-oriented but often sees topic modeling as a secondary, not very useful tool for aspect extraction. Also, there is already a slight "method fatigue" among the researchers who have, to a large extent, reached the limits of the method and are willing to combine it with other methods for resolving tasks in social science. Topic modeling suits well for mapping subthemes inside a stable corpus of documents or understanding the configuration of a particular subtheme beyond the naïve search; it fits a bit less for regular monitoring or precise classification of highly noisy data from social media. There is also lack of studies of human interpretability of Russian-language topics and the factors behind it. In future, we need more discussion on how the properties of Russian influence the modeling results, how text semantics may be used to enhance topic extraction, and whether topic modeling may be used to monitor the dynamics of the discussions. Also, within practically all Slavic languages, no attempts have so far been made to use topic detection in image studies; all these fields are open for rigorous research.

Acknowledgments This chapter is supported by presidential grants of the Russian Federation for young Doctors of science, grant MD-6259.2018.6.

References

Alekseev, Vasily A., Vladmir G. Bulatov, and Konstantin V. Vorontsov. 2018. Intra-Text Coherence as a Measure of Topic Models' Interpretability. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2018*, 1–13.

Apishev, Murat, Sergei Koltcov, Olessia Koltsova, Sergei Nikolenko, and Konstantin Vorontsov. 2016a. Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In *Proceedings of Mexican International Conference on Artificial Intelligence (MICAI)*, 169–184. Cham: Springer.

———. 2016b. Mining Ethnic Content Online with Additively Regularized Topic Models. *Computacion y Sistemas* 20 (3): 387–403.

Batura, Tatyana, and Svetlana Strekalova. 2018. Podhod k postroeniů rasširennyh tematičeskih modelej tekstov na russkom âzyke [An Approach to Constructing Extended Topic Models in the Russian Language]. *Bulletin of Novosibirsk State University, Information Technologies Series* 16 (2): 5–18.

Belyy, Anton, Maria Seleznova, Aleksei Sholokhov, and Konstantin Vorontsov. 2018. Quality Evaluation and Improvement for Hierarchical Topic Modeling. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2018*, 110–123.

Blei, David M., and John D. Lafferty. 2009. Topic Models. In *Text Mining: Classification, Clustering, and Applications*, ed. Ashok Srivastava and Mehran Sahami, 101–124. Chapman and Hall/CRC.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blekanov, Ivan, Nikita Tarasov, and Alexey Maksimov. 2018. Topic Modeling of Conflict Ad Hoc Discussions in Social Networks. In *Proceedings of the 3rd International Conference on Applications in Information Technology*, 122–126. ACM.

Bodrunova, Svetlana, Sergei Koltsov, Olessia Koltsova, Sergei Nikolenko, and Anastasia Shimorina. 2013. Interval Semi-Supervised LDA: Classifying Needles in a Haystack. In *Proceedings of the Mexican International Conference on Artificial Intelligence*, 265–274. Berlin – Heidelberg: Springer.

Bodrunova, Svetlana S., Olessia Koltsova, Sergei Koltcov, and Sergei Nikolenko. 2017. Who's Bad? Attitudes Toward Resettlers from the Post-Soviet South Versus Other Nations in the Russian Blogosphere. *International Journal of Communication* 11: 3242–3264.

Bodrunova, Svetlana S., Ivan Blekanov, and Mikhail Kukarkin. 2019a. Topics in the Russian Twitter and Relations Between Their Interpretability and Sentiment. In *Proceedings of the IEEE International Workshop on Sentiment Analysis and Mining of Social Networks (SAMSN)*, 549–554. IEEE.

———. 2019b. Topic Modelling for Twitter Discussions: Model Selection and Quality Assessment. *Proceedings of the 6th SWS International Scientific Conference on Social Sciences* 6 (5): 207–214. Sofia: STEF92 Technology.

Bodrunova, Svetlana S., Ivan Blekanov, Anna Smoliarova, and Anna Litvinenko. 2019c. Beyond Left and Right: Real-World Political Polarization in Twitter Discussions on Inter-Ethnic Conflicts. *Media and Communication* 7 (3): 119–132.

Bodrunova, S. S., Orekhov, A. V., Blekanov, I. S., Lyudkevich, N. S., & Tarasov, N. A. (2020). Topic Detection Based on Sentence Embeddings and Agglomerative Clustering with Markov Moment. *Future Internet* 12 (9): 144–160.

Bolshakova, Elena, Natalia Loukachevitch, and Michael Nokel. 2013. Topic Models Can Improve Domain Term Extraction. In *Proceedings of European Conference on Information Retrieval*, 684–687. Berlin – Heidelberg: Springer.

Boussalis, Constantine, Travis G. Coan, and Marianna Poberezhskaya. 2016. Measuring and Modeling Russian Newspaper Coverage of Climate Change. *Global Environmental Change* 41: 99–110.

Boyd-Graber, Jordan, David Mimno, and David Newman. 2014. Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In *Handbook of Mixed Membership Models and Their Applications*, ed. Edoardo M. Airoldi et al., 225–255. Chapman and Hall.

Chew, Peter A., and Jessica G. Turnley. 2017. Understanding Russian Information Operations Using Unsupervised Multilingual Topic Modeling. In *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 102–107. Cham: Springer.

Davydova, Yulia. 2019. Defining Thematic Relevance of Messages in the Task of Online Social Networks Monitoring in Providing Information-Psychological Security. *International Journal of Open Information Technologies* 7 (4): 11–18.

Gruber, Amit, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden Topic Markov Models. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (PMLR)* 2: 163–170.

Gutiérrez, Elkin D., Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting Cross-Cultural Differences Using a Multilingual Topic Model. *Transactions of Association for Computer Linguistics* 4: 47–60.

Kochedykov, Denis, Murat Apishev, Lev Golitsyn, and Konstantin Vorontsov. 2017. Fast and Modular Regularized Topic Modelling. In *Proceedings of 21st Conference of Open Innovations Association (FRUCT)*, 182–193. IEEE.

Koltcov, Sergei. 2018. Application of Rényi and Tsallis Entropies to Topic Modeling Optimization. *Physica A: Statistical Mechanics and Its Applications* 512: 1192–1204.

Koltcov, Sergei, Olessia Koltsova, and Sergei Nikolenko. 2014. Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated Content. In *Proceedings of ACM Conference on Web Science*, 161–165. ACM.

Koltcov, Sergei, Sergei Nikolenko, Olessia Koltsova, and Svetlana Bodrunova. 2016a. Stable Topic Modeling for Web Science: Granulated LDA. In *Proceedings of 8th ACM Conference on Web Science*, 342–343. ACM.

Koltcov, Sergei, Sergei Nikolenko, Olessia Koltsova, Vladimir Filippov, and Svetlana Bodrunova. 2016b. Stable Topic Modeling with Local Density Regularization. In *Proceedings of International Conference on Internet Science (INSCI)*, 176–188. Cham: Springer.

Koltcov, Sergei N., Sergei I. Nikolenko, and Elena Y. Koltsova. 2016c. Gibbs Sampler Optimization for Analysis of a Granulated Medium. *Technical Physics Letters* 42 (8): 837–839.

Koltcov, Sergei N., Sergei I. Nikolenko, and Olessia Koltsova. 2017. Topic Modelling for Qualitative Studies. *Journal of Information Science* 43 (1): 88–102.

Koltsov, Sergei, Sergei Pashakhin, and Sofia Dokuka. 2018. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research. In *Proceedings of the International Conference on Social Informatics (SocInfo)*, 308–321. Cham: Springer.

Koltsova, Olessia, and Sergei Koltcov. 2013. Mapping the Public Agenda with Topic Modeling: The Case of the Russian Livejournal. *Policy and Internet* 5 (2): 207–227.

Koltsova, Olessia, and Oleg Nagornyy. 2019. Redefining Media Agendas: Topic Problematization in Online Reader Comments. *Media and Communication* 7 (3): 145–156.

Koltsova, Olessia, and Sergei Pashakhin. 2017. Agenda Divergence in a Developing Conflict: Quantitative Evidence from Ukrainian and Russian TV Newsfeeds. *Media, War and Conflict*. https://doi.org/10.1177/1750635219829876.

Koltsova, Olessia, and Andrey Shcherbak. 2015. 'LiveJournal Libra!': The Political Blogosphere and Voting Preferences in Russia in 2011–2012. *New Media & Society* 17 (10): 1715–1732.

Koltsova, Olessia, Svetlana Alexeeva, and Sergei Kolcov. 2016a. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016*, 277–287.

Koltsova, Olessia, Sergei Koltcov, and Sergei Nikolenko. 2016b. Communities of Co-Commenting in the Russian LiveJournal and Their Topical Coherence. *Internet Research* 26 (3): 710–732.

Korshunov, Anton, and Andrey Gomzin. 2012. Tematičeskoe modelirovanie tekstov na russkom âzyke [Topic Modeling of Texts in the Russian Language]. *Proceedings of the Institute of Systemic Programming of the Russian Academy of Science* 23: 215–243.

Krasnov, Fedor, and Anastasiia Sen. 2019. The Number of Topics Optimization: Clustering Approach. *Machine Learning and Knowledge Extraction* 1 (1): 416–426.

Kriukova, Anna, Aliia Erofeeva, Olga Mitrofanova, and Kirill Sukharev. 2018. Explicit Semantic Analysis as a Means for Topic Labelling. In *Proceedings of the 7th International Conference on Artificial Intelligence and Natural Language Processing (AINL)*, 110–118. Cham: Springer.

Lin, Chenghua, Yulan He, Richard Everson, and Stefan Ruger. 2011. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering* 24 (6): 1134–1145.

Lopukhin, Konstantin, and Anastasia Lopukhina. 2016. Word Sense Disambiguation for Russian Verbs Using Semantic Vectors and Dictionary Entries. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016*, 393–405.

Lopukhin, Konstantin, Boris Iomdin, and Anastasia Lopukhina. 2017. Word Sense Induction for Russian: Deep Study and Comparison with Dictionaries. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016*, 121–134.

Loukachevitch, Natalia V., and Boris V. Dobrov. 2000. Issledovanie tematičeskoj struktury teksta na osnove bol'šogo lingvističeskogo resursa [Studying Topical Structure of Text with the Help of a Large Linguistic Dataset]. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016*, 252–258.

———. 2009. Avtomatičeskoe annotirovanie novostnyh klasterov na osnove tematičeskogo predstavleniâ [Automated Annotation of News Clusters Based on Topical Representation]. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2009*, 299–305.

Mashechkin, Igor, Mikhail Petrovsky, and Dmitry Tsarev. 2013. Metod vyčisleniâ relevantnosti fragmentov teksta na osnove tematičeskih modelej v zadače avtomatičeskogo annotirovaniâ [Methods of Relevance Calculation of Textual Fragments for Automated Annotation Based on Topic Models]. *Vyčislitel'nye metody i programmirovanie [Computational Methods and Programming]* 14 (1): 91–102.

Maučec, Miriam S., Zdravko Kačič, and Bogomir Horvat. 2004. Modelling Highly Inflected Languages. *Information Sciences* 166 (1–4): 249–269.

Mavrin, Andrey, Andrey Filchenkov, and Sergei Koltcov. 2018. Four Keys to Topic Interpretability in Topic Modeling. In *Proceedings of AINL Conference*, 117–129. Cham: Springer.

Mcauliffe, Jon D., and David M. Blei. 2008. Supervised Topic Models. *Advances in Neural Information Processing Systems* 20: 121–128. Neural Information Processing Systems Foundation.

Mimno, David, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual Topic Models. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* 2: 880–889. ACL.

Minka, Thomas, and John Lafferty. 2002. Expectation-Propagation for the Generative Aspect Model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, 352–359. San Francisco: Morgan Kaufmann Publishers.

Mirzagitova, Aliya, and Olga Mitrofanova. 2016. Automatic Assignment of Labels in Topic Modelling for Russian Corpora. In *Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics (ExLing)*, 115–118. ExLing.

Mitrofanova, Olga. 2015. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In *Proceedings of the International Workshop on Language, Music, and Computing*, 69–76. Cham: Springer.

Nagornyy, Oleg. 2018a. Topics of Ethnic Discussions in Russian Social Media. In *Proceedings of the International Conference on Digital Transformation and Global Society*, 83–94. Cham: Springer.

———. 2018b. User Ethnicity and Gender as Predictors of Attitudes to Ethnic Groups in Social Media Texts. In *Proceedings of International Conference on Internet Science (INSCI)*, 33–41. Cham: Springer.

Nikolenko, Sergei. 2016. Topic Quality Metrics Based on Distributed Word Representations. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Developments in Information Retrieval*, 1029–1032. ACM.

Potapenko, Anna, and Konstantin Vorontsov. 2013. Robust PLSA Performs Better than LDA. In *Proceedings of the European Conference on Information Retrieval*, 784–787. Berlin – Heidelberg: Springer.

Rubtsova, Yuliya, and Sergey Koshelnikov. 2015. Aspect Extraction from Reviews Using Conditional Random Fields. In *International Conference on Knowledge Engineering and the Semantic Web*, 158–167. Cham: Springer.

Sedova, Anastasia, and Olga Mitrofanova. 2017a. Tematičeskoe modelirovanie russkoâzyčnyh tekstov s oporoj na lemmy i leksičeskie konstrukcii [Topic Modeling of Russian Texts Based on Lemmata and Lexical Constructions]. *Komp'ûternaâ lingvistika i vyčislitel'nye ontologii [Computer Linguistics and Computational Ontologies]* 1: 132–144.

———. 2017b. Topic Modelling in Parallel and Comparable Fiction Texts (the Case Study of English and Russian Prose). In *Proceedings of the International Conference on Internet and Modern Society (IMS)*, 175–180. ACM.

Skachkov, Nikolay, and Konstantin Vorontsov. 2018. Improving Topic Models with Segmental Structure of Texts. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2018*, 652–661.

Smoliarova, Anna S., Svetlana S. Bodrunova, Alexander V. Yakunin, Ivan Blekanov, and Alexey Maksimov. 2018. Detecting Pivotal Points in Social Conflicts via Topic Modeling of Twitter Content. In *Proceedings of the International Conference on Internet Science*, 61–71. Cham: Springer.

Sridhar, Vivek K.R. 2015. Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 192–200. Association for Computational Linguistics.

Steyvers, Mark, and Tom Griffiths. 2007. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, ed. Thomas K. Landauer et al., 427–448. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Tutubalina, Elena, and Sergei Nikolenko. 2015. Inferring Sentiment-Based Priors in Topic Models. In *Proceedings of the Mexican International Conference on Artificial Intelligence (MICAI)*, 92–104. Cham: Springer.

Völske, Michael, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. 2015. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 1571–1580. ACM.

Voronov, Sergei O., and Konstantin V. Vorontsov. 2015. Avtomatičeskaâ fil'traciâ russkoâzyčnogo naučnogo kontenta metodami mašinnogo obučeniâ i tematičeskogo modelirovaniâ [Automated Filtration of Russian-Language Academic Content by Machine Learning and Topic Modeling]. *Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016*. dialog-21.ru/media/1143/voronovsovorontsovkv.pdf.

Vorontsov, Konstantin, and Anna Potapenko. 2012. Regulârizaciâ, robastnost' i razrežennost' veroâtnostnyh tematičeskih modelej [Regularizaton, Robustness, and Sparsity of Probabilistic Topic Models]. *Komp'ûternye issledovaniâ i modelirovanie [Computer Research and Modeling]* 4 (4): 693–706.

———. 2015. Additive Regularization of Topic Models. *Machine Learning* 101 (1–3): 303–323.

Vorontsov, Konstantin, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. 2015a. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, 29–37. ACM.

Vorontsov, Konstantin, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015b. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In *Proceedings of International Conference on Analysis of Images, Social Networks and Texts*, 370–381. Cham: Springer.

Wang, Xuerui, Andrew McCallum, and Xing Wei. 2007. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, 697–702. IEEE.

Zharikov, Ilya, Murat Apishev, and Konstantin Vorontsov. 2018. Gipergrafovye mnogomodal'nye veroâtnostnye tematičeskie modeli tranzakcionnyh dannyh [Hypergraph Multimodal Probabilistic Topic Models of Transactional Data]. In *Proceedings of the Conference "Intellektualizaciâ obrabotki informacii"*, 148–149.