

Political Actor or Ritualized Other? Mapping Ethnic Attitudes in the Russian Blogosphere by Methods of Big Data Research

Svetlana S. Bodrunova

National Research University Higher School of Economics, St. Petersburg, Russia, visual@jf.spbu.ru
School of Journalism and Mass Communication, St. Petersburg State University

Sergey I. Nikolenko

National Research University Higher School of Economics, St. Petersburg, Russia
Steklov Mathematical Institute at St. Petersburg

Sergei Koltcov

National Research University Higher School of Economics, St. Petersburg, Russia

Olessia Koltsova

National Research University Higher School of Economics, St. Petersburg, Russia

Abstract

Qualitative studies, such as sociological research, opinion analysis, or media studies, can benefit greatly from automated topic mining provided by topic models such as LDA. However, examples of qualitative studies that employ topic modelling as a tool are currently few and far between. In this work, we identify two important problems along the way to using topic models in qualitative studies: lack of a good quality metric that closely matches human judgement in understanding topics and the need to indicate specific sub-topics that a specific qualitative study may be most interested in mining. For the first problem, we propose a new quality metric, tf-idf coherence, that reflects human judgement more accurately than regular coherence, and conduct an experiment to verify this claim. For the second problem, we propose an interval semi-supervised LDA approach, or ISLDA, in which certain predefined sets of keywords (that define the topics researchers are interested in) are restricted to specific intervals of topic assignments. We also present a case study on a Russian Livejournal dataset aimed at ethnicity discourse analysis.

In Russia of 2006–2013, grassroots political activity and its radicalization have both risen rapidly. Urban ethnic conflicts with Caucasian-origin resettlers and emigrants from ex-Soviet states, including Moscow of 2010–2013, have created new research agendas focused on mapping and prognosis of current ethnic attitudes. The project aims at mapping current ethnic attitudes in the Russian-language Livejournal community by big data research methods, Livejournal selected for its ‘platform for Runet elite’ image. To interpret the topics, framing theory is used: ‘politicized vs. ritualized ethnicities’ semi-automated frame analysis and manual discourse analysis are deployed, altogether forming a mixed quality-assessment method for LDA topics. Factors influencing ‘politicization/ritualization’ are suggested.

Keywords

Latent Dirichlet Allocation; topic quality; topic modelling; LDA extensions, Russia, ethnic discourse, blogosphere

1. Introduction

Over recent years, topic modelling techniques based on probabilistic latent semantic analysis (pLSA) [1] and latent Dirichlet allocation (LDA) [2, 3] have seen growing use for various applications, including direct applications to unsupervised analysis with many extensions to additional information and structure that may be available in specific settings (see Section 2.2 for a brief review), supervised versions to be used as text classifiers [4], analysing topic evolution in a set of documents over time [5, 6], image recognition and classification [7, 8] and others. In essence, topic modelling is an unsupervised model that learns the set of underlying topics (in terms of word distributions) for a set of documents and each document’s affinities to these topics (see Section 2 for a more detailed introduction).

Recently, a novel field of applications for topic modelling began to emerge. In this field, topic modelling assists qualitative and quantitative research over user-generated texts coming from the blogosphere or social networks. Such research may include sociological studies, opinion analysis for marketing purposes, discourse analysis in media studies and so on. By studying the set of topics learnt from, say, a dataset of entries from popular blogs over some period of time, it may become possible to find out what the users are talking about, identify underlying topical trends and follow them through time, identify documents that are most relevant to a specific topic. Ideally, a researcher may be able to draw important conclusions exclusively from the set of word distributions for topics, and only then read documents with high affinities for interesting topics for a more in-depth study. There exists some recent research that brings topic

modelling into this picture, but mostly as a means of statistical analysis in order to detect biased words that may define media bias [9] or ideological discourse [10] rather than as direct help for qualitative studies.

However, there still exist important obstacles along this path. One obstacle is that in order to find “interesting” topics, it would be desirable to have a reasonable metric that automatically ranks learned topics in an order that closely matches their potential to be interesting for a researcher. Human judgement represents the ground truth for the types of studies mentioned above. A good topic quality metric that closely matches human judgement would help filter uninteresting topics in a single topic model and also compare different topic models (with different parameters, different number of topics etc.). Researchers need criteria to choose the best topics and, accordingly, the best model that provides the best topics. As we demonstrate in Section 3, existing metrics such as coherence are less than perfect in this regard, and further research towards topic quality metrics is warranted.

Another obstacle is that topic modelling learns a set of topics that attempts to describe the entire dataset as a whole, while a specific study may be interested in specific subtopics, points of interest where the study requires a more in-depth view. For instance, the case study we present in Section 4 deals with ethnical discourse; while a researcher can identify the most important keywords related to ethnic groups and nations, it is hard to come up with a comprehensive list of such keywords, and documents relevant for discourse analysis may not contain them, so it would not suffice to simply filter the documents by keyword occurrence. On the other hand, it is still desirable to guide the topic model to pay the most attention to topics identified with these specific keywords. In this work, we deal with these two obstacles, extending and augmenting the LDA model.

The paper is organized as follows. In Section 2, we introduce the basic LDA model and briefly survey its most important extensions. Then we proceed to present a modification of the basic LDA model that assigns a predefined number of topics to points of interest represented by specific keywords; this results in a deeper analysis of these topics and answers the second obstacle mentioned above. In Section 3, we introduce a novel topic quality metric that more closely corresponds to human judgement than existing ones; we also conduct an experiment with human experts that supports this claim. In Section 4, we present a case study of the kind mentioned above: a discourse analysis study based on a dataset of Russian Livejournal texts aimed at the analysis of ethnical discourse in the Russian blogosphere.

2. LDA and Semi-Supervised LDA

2.1. Latent Dirichlet Allocation

The basic latent Dirichlet allocation (LDA) model [2, 3] is depicted on Fig. 1a. In this model, a collection of D documents is assumed to contain T topics expressed with W different words. Each document $d \in D$ is modeled as a discrete distribution $\theta(d)$ over the set of topics: $p(z_w = j) = \theta^{(d)}$, where z is a discrete variable that defines the topic for each word $w \in d$. Each topic, in turn, corresponds to a multinomial distribution over the words, $p(w|z_w = j) = \phi_w^{(j)}$. The model also introduces Dirichlet priors α for the distribution over documents (topic vectors) θ , $\theta \sim \text{Dir}(\alpha)$, and β for the distribution over the topical word distributions, $\phi \sim \text{Dir}(\beta)$. The inference problem in LDA is to find hidden topic variables z , a vector spanning all instances of all words in the dataset. There are two approaches to inference in the LDA model: variational approximations and MCMC sampling which in this case is convenient to frame as Gibbs sampling. In this work, we use Gibbs sampling because it generalizes easily to semi-supervised LDA considered below. In the LDA model, Gibbs sampling after easy transformations [3] reduces to the so-called collapsed Gibbs sampling, where z_w are iteratively resampled with distributions

$$p(z_w = t | \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) \propto q(z_w, t, \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) = \frac{n_{-w,t}^{(d)} + \alpha}{\sum_{t' \in T} (n_{-w,t'}^{(d)} + \alpha)} \frac{n_{-w,t}^{(w)} + \beta}{\sum_{w' \in W} (n_{-w,t}^{(w')} + \beta)}, \quad (1)$$

where $n_{-w,t}^{(d)}$ is the number of times topic t occurs in document d and $n_{-w,t}^{(w)}$ is the number of times word w has been generated by topic t , not counting the current value z_w . The basic LDA model has been used in numerous applications, including studies that survey scientific literature [3] and attempt to mine how topics change in time [11, 12, 13].

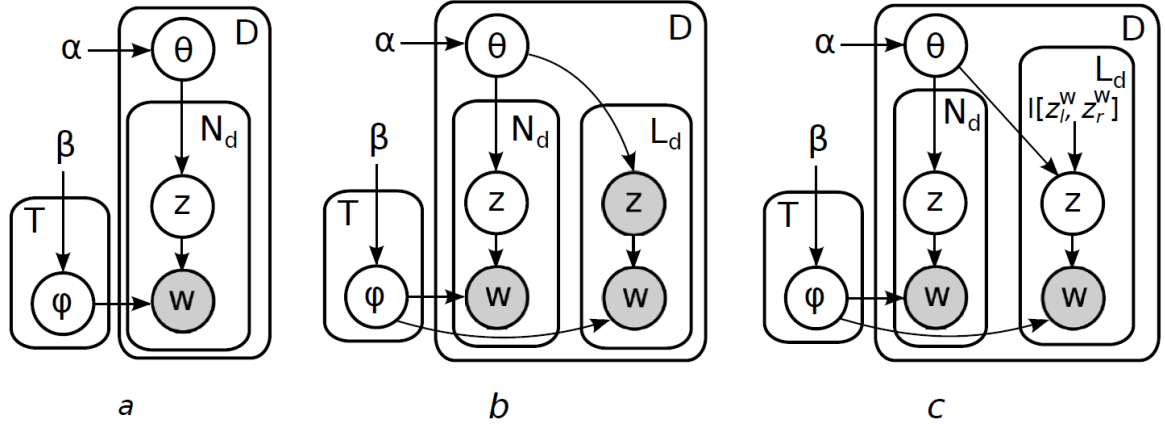


Figure 1. Probabilistic models: (a) LDA; (b) semi-supervised LDA; (c) interval semi-supervised LDA.

2.2. Related work: LDA extensions

Over the last decade, after LDA was introduced in [2], the basic LDA model has been subject to many extensions. Each of these extensions presents either a variational or a Gibbs sampling algorithm for a model that builds upon LDA to incorporate some additional information or additional presumed dependencies. Before proceeding to our specific extension introduced in the rest of this section, we give a brief survey of the most important existing extensions.

One large class of extensions deals with imposing new structure on the set of topics that are independent and uncorrelated in the base LDA model. Correlated topic models (CTM) avoid this unrealistic assumption, admitting that some topics are closer to each other and share words with each other; CTMs use logistic normal distribution instead of Dirichlet to model correlations between topics [14]. Markov topic models use Markov random fields to model the interactions between topics in different parts of the dataset (different text corpora), connecting a number of different hyperparameters β_i in a Markov random field that lets one subject these hyperparameters to a wide class of prior constraints [15]. Syntactic topic models introduce syntactic constraints inside a document that replace the bag-of-words model with syntactic parse trees [16]. Relational topic models construct a hierarchical model that reflects the structure of a document network as a graph [17], while Spatial LDA extends the LDA approach to image recognition by imposing spatial structure on “visual words” [7].

Another class of extensions takes into account additional information that may be available together with the documents and may reveal additional insights into the topical structure. Many such extensions deal with time; for instance, the Topics over Time model applies when documents have timestamps of their creation (e.g., news articles or blog posts); it represents the time when topics arise in continuous time with a beta distribution [5]. Dynamic topic models represent the temporal evolution of topics through the evolution of their hyperparameters α and β , either with a state-based discrete model [18] or with a Brownian motion in continuous time [19]. Online topic detection with a temporal component (based on tensor factorization) has been applied to topic mining in continuous streams [20]. Other extensions deal with other kinds of labels attached to documents. For example, supervised LDA assigns each document with an additional response variable that can be observed; this variable depends on the distribution of topics in the document and can represent, e.g., user response in a recommender system [21]. DiscLDA assumes that each document is assigned with a categorical label and attempts to utilize LDA for mining topic classes related to this classification problem [22]. The Author-Topic model incorporates information about the author of a document, assuming that texts from the same author will be more likely to concentrate on the same topics and will be more likely to share common words [23, 24]. Yet other extensions improve upon the bag-of-words assumption, modelling correlations between words and individual sentences [25]. Finally, a lot of work has been done on nonparametric LDA variants based on Dirichlet processes that we will not go into in this paper; for the most important nonparametric approaches to LDA see [26, 27, 28, 29, 30] and references therein.

The extension that appears to be closest to the one proposed in this work is the *Topic-in-Set knowledge* model and its extension with Dirichlet forest priors [31, 32]. In [32], words are assigned with “z-labels”; a z-label represents the topic this specific word should fall into; in the remainder of this section, we build upon and extend this model.

2.3. LDA and Tikhonov Regularizers

Generally speaking, LDA is an optimization problem whose solution is not unique and/or unstable; practical experiments (see, e.g., [33]) show that LDA inference results produced by Gibbs sampling may change significantly depending on the generated random numbers and certain topics may go in or out of the resulting set of topics. One general approach to handling ill-posed problems has been developed in the works of Tikhonov [34, 35, 36]; the basic idea is to introduce additional constraints (regularizers) in the problem, thus narrowing down the set of solutions. Recent works of Vorontsov [37] apply this approach to topic modelling, introducing additive regularization of topic models (ARTM) as an alternative to the Bayesian approach. This method is free of superfluous probabilistic assumptions, does not require Dirichlet distributions, and lets one use regularizers that do not have a clear probabilistic interpretation. The idea is that certain topics can be characterized by a kernel of words that represent terms in a certain subject domain that will have high probability in this topic and low probabilities in other topics. Thus, the collection of word kernels for various topics whose probabilities do not change during topic modelling serves as a regularizer. However, this idea has so far been implemented only for the PLSA model and has been used to improve interpretability of the topics. In this work, the proposed LDA extension is, in a sense, a kernel of keywords for a specific topic. For instance, in our case study (Section 4) a subset of topics is defined by a group of words related to a specific ethnicity, and the purpose is to find other words that often characterize these ethnicities and ultimately study documents with ethnic-related content. In essence, we propose a form of a regularizer that serves as a “crystallization point” for the topics being extracted.

2.4. Semi-Supervised LDA

In real life text mining applications, it often happens that the entire dataset D deals with a large number of different unrelated topics, while the researcher is actually interested only in a small subset of these topics. In this case, a direct application of the LDA model has important disadvantages. Relevant topics may have too small a presence in the dataset to be detected directly, and one would need a very large number of topics to capture them in an unsupervised fashion. For a large number of topics, however, the LDA model often has too many local maxima, giving unstable results with many degenerate topics. To find relevant subsets of topics in the dataset, we propose to use a semi-supervised approach to LDA, fixing the values of z for certain key words related to the topics in question; similar approaches have been considered in [31, 32]. The resulting graphical model is shown on Fig. 1b. For words $w \in W_s$ from a predefined set W_s , the values of z are known and remain fixed to $\tilde{z}_w$ throughout the Gibbs sampling process:

$$p(z_w = t | \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) \propto \begin{cases} [t = \tilde{z}_w], & w \in W_s, \\ q(z_w, t, \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta), & \text{otherwise.} \end{cases} \quad (2)$$

Otherwise, the Gibbs sampler works as in the basic LDA model; this yields an efficient inference algorithm.

2.5. Interval Semi-Supervised LDA

One disadvantage of semi-supervised LDA approach is that it assigns only a single topic to each set of keywords, while many qualitative studies have specific reasons to reveal subtopics related to the same set. One often has to discern between different contexts: for instance, in our case it is important to separate politically neutral posts about Ukrainian resorts from texts dealing with tensions between East and West Ukraine or Russian–Ukrainian relations. Drawing them all together with semi-supervised LDA would have undesirable consequences: some “Ukrainian” topics would be cut off from the single supervised topic and left without Ukrainian keywords because it is more likely for the model to cut off a few words even if they fit well than bring together two very different sets of words under a single topic. The two political topics on Ukraine would most likely stick together, while resorts would be lost in a more general topic of travel to different countries. Therefore, we propose to map each set of key words to several topics; it is convenient to choose a contiguous interval, hence interval semi-supervised LDA (ISLDA). Each key word $w \in W_s$ is thus mapped to an interval $[z_l^w, z_r^w]$, and the probability distribution is restricted to that interval; the graphical model is shown on Fig. 1c, where $I[z_l^w, z_r^w]$ denotes the indicator function: $I[z_l^w, z_r^w] = 1$ iff $z \in [z_l^w, z_r^w]$. In the Gibbs sampling algorithm, we simply need to set the probabilities of all topics outside $[z_l^w, z_r^w]$ to zero and renormalize the distribution inside:

$$p(z_w = t | \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta) \propto \begin{cases} I[z_l^w, z_r^w](z) \frac{q(z_w, t, \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta)}{\sum_{z_l^w \leq t' \leq z_r^w} q(z_w, t', \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta)}, & w \in W_s, \\ q(z_w, t, \mathbf{z}_{-w}, \mathbf{w}, \alpha, \beta), & \text{otherwise.} \end{cases} \quad (3)$$

Interval semi-supervised LDA is able to account for several sets of keywords representing different topics of interest: one simply has to assign them disjoint intervals of topics. For instance, in our case study described in Section 4 we chose four different sets of keywords and assigned them to four different intervals of topics, all in the same model.

3. LDA quality estimation: beyond coherence

3.1. The coherence metric and its problems

Topic models such as LDA produce a set of topics characterized by distributions over words; however, they do not by themselves produce any characteristic features that might help a researcher identify the most useful topics, i.e., choose a subset of topics that are best suitable for human interpretation. The problem of finding a metric that characterizes such interpretability has been a subject of study for some years now. There is a difference between evaluating the entire solution (set of topics) and evaluating individual topics to filter out the junk. For the entire solution, researchers usually either look at perplexity [38] on the original dataset or measure the predictive power of a model by computing the log probability of a held-out set of documents [39]; an important recent study advocates a Bayesian approach based on posterior predictive checking over [40].

As for individual topics, which is what we are interested in now, recent studies [41, 42] agree that topic coherence is a good candidate. For a topic t characterized by its set of top words W_t , coherence is defined as

$$c(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{d(w_1, w_2) + \epsilon}{d(w_1)}, \quad (4)$$

where $d(w_i)$ is the number of documents that contain w_i , $d(w_i, w_j)$ is the number of documents where w_i and w_j cooccur, and ϵ is a smoothing count usually set to either 1 or 0.01. Coherence and word cooccurrence statistics in general have been used for initialization of LDA parameters [43]. There also exist other, less widely known approaches; e.g., in [44] the authors attempt to quantify properties of word distributions that may indicate a junk/insignificant topic and construct a ranking scheme. In our studies, however, we have found the coherence metric to be a less than perfect guide. It has been able to consistently identify bad topics (i.e., topics with poor coherence are indeed bad topics) but has not performed well at the positive end of the spectrum. We identify two main reasons for it. First, many topics that have good coherence are composed of common words that do not represent any topic of discourse per se. A representative example is shown in Table 1, where we list the top words of top ten topics w.r.t. coherence in one of our experiments. Obviously, these common words do indeed co-occur often, but as a result, only two of the top ten topics can be readily identified ((8) is about history, and (10) is about Russian law). Thus, the first problem is that coherence does not distinguish between high frequency words and informative words that define topics. Second, in the blogosphere many topics stem from copies, reposts, and discussions of a single text. All of them either directly copy or extensively cite the original text, so words that appear in this text both turn out to be top words in the corresponding topic and have very good coherence with each other, especially when these words are relatively rare and do not often occur in other topics. This is a characteristic feature of user-generated web datasets rather than a general disadvantage of the coherence metric, but it still needs to be addressed. Reposts of the same message, due to their extreme similarity, tend to be torn from other relevant texts that would have been assigned to the same topic had the reposts been filtered out beforehand. This deprives the resulting topic of potentially rich connotations and makes it less informative.

Table 1. Topics with top coherence scores: top words (translated) from top ten topics with respect to coherence in one of our experiments.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
say	just	issue	just	author	life	have	century	just	right
tell	stay	solution	instance	fact	know	instance	appear	know	law
know	solve	problem	example	article	just	image	beginning	say	Russian
just	problem	situation	often	say	live	example	history	nothing	state
need	moment	side	say	issue	see	system	well	known	general
nothing	know	relation	have	write	say	follow	end	see	federation

3.2. Tf-idf coherence

To alleviate the drawbacks shown in the previous subsection, we propose a modification of the basic coherence metric that takes into account informative content of the topics. We have found that a simple tf-idf weighting scheme works very well in this regard: instead of counting the number of [co]occurrences we substitute tf-idf scores. Namely, we define the tf-idf coherence metric as

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} \log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d) \text{tf-idf}(w_2, d) + \epsilon}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)}, \quad (4)$$

where the tf-idf metric is computed with augmented frequency,

$$\text{tf-idf}(w, d) = \text{tf}(w, d) \text{idf}(w, d) = \left(\frac{1}{2} + \frac{f(w, d)}{\max_{w' \in d} f(w', d)} \right) \log \frac{|D|}{|\{d \in D: w \in d\}|}, \quad (5)$$

where $f(w, d)$ is the number of occurrences of term w in document d . Intuitively, we skew the metric towards topics with high tf-idf scores in top words, since the numerator of the coherence fraction has quadratic dependence on the tf-idf scores and the denominator only linear.

Table 2 shows the ten top topics in the same experiment, but ranked with respect to their tf-idf coherence. It is clear that these topics are much more informative, uncovering both important current events that fell in the experiment’s period (terrorist attack in Boston, Pope Benedict’s resignation, Cyprus default etc.) and all-time favourites (recipes, World War II). The second row shows the topics’ ranks with respect to regular coherence; it is clear that more general topics score higher in regular coherence, while interesting and important topics related to current events are lost in the middle of the pack. We advise to use tf-idf coherence to rank topics in projects where topic mining and further human interpretation are of essence.

3.3. Experimental comparison

In order to make a fair comparison between different LDA topic quality estimation metrics, we have performed an experimental evaluation of topic quality based on lists of top words. For each topic, we asked the subjects (among them media studies experts) two binary questions.

- (1) Do you understand why the words in this topic have been united together, do you see obvious semantic criteria that unite the words in this topic?
- (2) If you have answered “yes” to the first question: can you identify specific issues/events that documents in this topic might address?

Note that in this experiment, we are more interested in the subjective assessment of topic quality and whether a topic is readily identified by human subjects than objective parameters that other metrics usually concentrate on. Therefore, in

Table 2. Topics with top tf-idf coherence scores: top words (translated) from top ten topics w.r.t. tf-idf coherence in the same experiment as Table 1. The second row shows a topic’s rank with respect to regular coherence. Word^{adj} denotes a word in adjectival form (in Russian, adjectives are more often different from their corresponding nouns than in English).

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(64)	(58)	(42)	(63)	(75)	(48)	(34)	(61)	(28)	(25)
ter. attack	pope	church	butter	Korea	add	Cyprus	Syria	military	war
explosion	Vatican	orthodox	sugar	North	butter	bank	Syrian	army	German ^{adj}
Boston	Roman	temple	add	DPRK	onion	Russian	AI	service	Germany
terrorist	church	cleric	flour	South	meat	Cypriot	country	general	German
brother	cardinal	faith	dough	Kim	pepper	Euro	Muslim	officer	Hitler
police	catolic	church ^{adj}	recipe	Korean	minute	financial	fighter	mil. force	Soviet
Boston ^{adj}	Benedict	patriarch	egg	Nuclear	dish	money	Arab	defense	world

this experiment we asked the subjects direct questions about understanding topics rather than used subjects as oracles to judge objective topic qualities like the experiments conducted in [41]. At the same time, since we are not interested in subjective judgment of a particular individual, each set of topics was offered to 6–8 experts for independent assessment. In total, we assessed four different datasets with ten different experts.

To evaluate different quality metrics, we have computed, for each subject and each metric, the area-under-curve (AUC) measure [45]. AUC is a popular quality metric for classifiers that produce ranking results; by definition it represents the probability that for a uniformly selected pair consisting of a positive and a negative example the classifier ranks the positive one higher. Thus, the optimal AUC is 1 (all positive examples come before negative ones), the worst possible AUC is 0, and a random classifier would get an AUC of 0.5. AUC can be easily computed as

$$AUC = \frac{\sum_{i \in \text{pos}} r_i - n_{\text{pos}}(n_{\text{pos}} + 1)/2}{n_{\text{pos}}n_{\text{neg}}}, \quad (6)$$

where n_{pos} and n_{neg} are the number of positive and negative examples in the dataset, and $\sum_{\text{pos}} r_i$ is the sum of the ranks of all positive examples [46]. We sort the topics by each metric and use AUC to measure how close to the top positive answers to the questions above (“relevant” topics) turn up. Table 3 presents the results. It is clear that tf-idf coherence wins over “vanilla” coherence in all experiments with significantly higher AUC for all test subjects without exception. Therefore, we can conclude that in practice, tf-idf coherence is preferable over regular coherence as a quality metric for LDA topics aimed at further qualitative analysis. We also verified the reliability and repeatability of our results in two ways. First, apart from the two binary questions, we asked each subject to briefly (in 1–4 words) identify each topic if possible; this was done in order to verify that different subjects understand topics in a similar way. This qualitative test has shown that all subjects understood most topics in a very similar way. Second, we computed average Hamming distances between the subject’s responses for each dataset divided by the number of topics (response vector length); they are also shown in Table 3. These distances indicate a good match between the individual subjects’ opinions: two vectors of answers differ, on average, only in 10–20% of positions, with the second question, which is more subjective than the first, naturally leading to more difference in opinion.

4. Mining Ethnic Discourse in the Russian blogosphere

We have applied the method outlined above as well as in our earlier research [47] to a sociological project intended to study the discourse on ethnic groups and nations in the Russian blogosphere. The project seeks to fulfil the two aims: 1) to discover a methodology (either automated or mixed) to dig up ethnic discourse (or any other topic- or issue-oriented discourse) from text collections of the natural ‘oral-written’ language [48]; 2) to map the ethnic discourse in the Russian Internet (exemplified by *Livejournal* blog platform) along two axes: ‘politicized’/‘ritualized’ [49, 50, 51] as one of the most persistent dualities in interpreting the nature of ethnicities and ‘dominant’/‘subordinate’ as the sign of post-imperial [52] and ‘internal colonization’ [53] mentality. Combination of the two axes and geography will create interpretable groups of ethnicities with shared imposed attitudes.

Table 3. Experimental comparison between LDA topic quality metrics. All experiments were conducted on Livejournal blog posts from top bloggers over different periods of time. The columns show the time period, number of topics in the experiment, and coherence, tf-idf coherence, and average unit Hamming distance between answer vectors for different users; these metrics are computed for question 1 and question 2 in the study.

Dataset	Topics	Question 1			Question 2		
		AUC, coherence	AUC, tf-idf coherence	Hamming distance	AUC, coherence	AUC, tf-idf coherence	Hamming distance
March 2012	100	0.66	0.74	0.15	0.59	0.65	0.24
March 2013	200	0.72	0.76	0.19	0.67	0.73	0.24
April 2012	100	0.66	0.74	0.10	0.59	0.65	0.22
September 2012	200	0.67	0.73	0.14	0.65	0.70	0.25

4.1. Case study: ethnic attitudes in Russian blogs as a research object

Ethnic attitudes have long been a huge area of research that may be conventionally divided into essentialist and constructivist fields [54]. Among other issues, links between intergroup relations and ethnic perceptions have been thoroughly studied; public attitudes towards ethnic groups are believed to be a factor influencing the spill-over potential in an ethnic conflict [55].

In Russia of 2000-2010s, along with the rise of social activism and radical movements [56], ethnic tensions have become a widespread phenomenon resulting in violent conflicts in Kondopoga (2006) and Moscow (2010, 2013). Public attitudes towards ethnic groups are believed to be a factor influencing the spill-over potential in an ethnic conflict [55].

Ethnicity is the concept that has long been contextualized with the help of similar concepts. Several research lines put 'ethnicity' into the context of 'ritualized' sides of life (appearance, traditions, customs, social habits, or everyday behavior patterns). Though being criticized as primordialist in the Academe, they remain significant ethnic markers in popular mind. Within another line of contextualization, the concept of 'nation' is, in contrast, often built upon linkages between ethnic community and territorial polity (e.g. in case of nation states or ethnic regions within them), which politicizes our understanding of ethnicity [57: 5-7] and allows for interpreting ethnicities as political social imaginaries [58] – collective actors on world arenas pursuing unified political and economic goals. This interpretation is also believed to be important for an ordinary person [59]. The 'ritual' and the 'political' create two poles of strangeness (culturally differing communities / geopolitical antagonists). Several studies have shown that the 'political – ritual(ized)' axis exists in self-perception of the European nations [50] and in perception of out-groups; for the latter, several variables are believed to be influencing attitude formation. Among those, experience of ethnic conflict is crucial for forming politicized attitudes, while for the 'ritualized' ones experience of co-living and neighboring (from individual collaboration to nation-wide competition) is a natural fostering factor [60]. This paradigm was challenged by a range of researchers, first and foremost by Stewart Hall [61]. Thus, it is interesting to know whether the 'political – ritualized' axis exists within the public discourse *en masse*; today, big data methods allow testing it.

Ethnic attitudes are believed to be represented online as well as offline. With the growth of Internet in 1990s, hopes arose that the negative ethnic attitudes online would be gradually diminishing; but soon it was clear enough that negative perceptions of ethnic communities persist in online communication [62]. Being, in general, an under-researched area until recently [63], today, ethnic Internet studies occupy growing space and pose the questions on how ethnic attitudes are expressed, whether traditional fostering factors work online, or what the role of a given platform in forming ethnic attitudes is. By far, as Daniels claims [63], people tend to reproduce pre-existing ethnic attitudes online; thus, studies on Runet data could be, at least to some extent, generalized to the Russian-speaking community offline. The duality described above as the axis of ethnic perceptions has been documented for the USSR, starting from the classic works by Yulian Bromley [49]. Western researchers also spotted this axis in development of 'nationhood' policy in the Soviet Union [51]. Thus, it is interesting to know to what extent the axis is seen in the online discourse of the Russian-speaking population 25 years after the fall of the USSR and what factors shape it now.

Soviet policies have left a controversial imprint on ethnic attitudes in Russia. On the one hand, ethnic diversity and equality was perceived as an indicator of 'Soviet democracy'; thus, 'peoples' friendship' rhetoric was adopted, and ethnic cultures supported. On the other hand, actual policing included policies aimed at assimilation of smaller ethnic groups and the formation of the Soviet quasi-national identity. Today, this contradiction reinvents itself in the 'post-imperial vs. ethnic nation' building in the Russian modernization and transition studies [52] [64, 65]. Being a fundamentally fragmented society, Russia has post-Soviet migrant and North Caucasian population as one of 'four Russias' – a distinct social milieu neither absorbed nor accepted as 'inherently Russian' by the urban post-industrial, post-Soviet industrial, and rural 'Russias' [66]. This milieu is in rapid growth, especially in absence of a sound migration policy that could combine freedom of movement with crime control and efficient integration measures. The recent decade has been marked by resonant conflicts between residents and diasporas on the local level including those in Kondopoga (2006) and in Moscow (2010, 2013), all involving the growing migrant population in urban and suburban areas, especially Central Asians and North Caucasians both perceived as immigrants. In Russian ethnic studies, all the traditions of contextualization of ethnicity are represented. However, often studies of ethnic attitudes use the 'binary' approach that studies perception of one nation by another one. There are attempts to map ethnic attitudes within a territory or a social stratum [67, 68, 69]. Research of attitudes towards multiple ethnic groups is usually based on the data from polls. Thus, we will be trying to map the ethnic attitudes in the Russian blogosphere, as blogs provide enough material to do that: the texts are long enough, and the speech online is described in literature as 'oral-written' [48], that is, spontaneous, dialogue-oriented, and non-regulated in comparison, e.g., to media speech.

4.2. Case study: application of methodology

As already told above, ISLDA differs from LDA, as it uses keywords as ‘chrystallizing agents’, thus combining keyword search with LDA algorithm; as the keywords (in our case, ethnic names) being assigned several times within one dataset, ISLDA is expected to exhaust the possible topics related to one keyword. ISLDA is also expected to work much better than naive keyword search, as 1) keyword search retrieves many “false negative” and “false positive” texts; 2) keyword search does not allow one to quickly define different contexts surrounding various ethnic groups and discern between these contexts; 3) modelling also digs out texts that do not contain the assigned keywords but may be crucial for understanding of the topic. Thus we have used ethnonyms, including pejoratives (e.g., Azeri), nation names (e.g., Azerbaijanis), and nation state names (e.g. Azerbaijanian) as keywords for semi-supervised LDA rather than keyword search. To choose the exact set of ethnonyms, we calculated their frequencies in the whole corpus represented by the Russian Livejournal posts. Contrary to our assumptions, Asian ex-Soviet nations that are connected by the media to the most serious problems such as crime, unemployment growth, illegal migration, “islamization” of Russia etc., were not the most mentioned. The top of the list was occupied by Americans, Ukrainians, Germans, and Jews. Top ten ethnic groups / nations were selected for running the ISLDA algorithm. For qualitative tests of ISLDA topic mining quality in comparison to LDA, four ethnicities were chosen according to their importance for the final qualitative goal. The Ukrainians were chosen due to their ethnic, geographical, and historical proximity to the Russian public, which could potentially lead to a bigger number of issues discussed international the blogosphere about this ethnicity. Tajiks were chosen as the representatives of a big migrant community potentially linked in public mind to social trouble. Georgians were chosen, as the 2008 armed conflict between South Ossetia and Georgia deeply affected and polarized public opinion in all the three communities (Russian, Ossetian, and Georgian). French were chosen due to the multiple contexts (from Napoleonic wars to cuisine and fashion) to which they may belong in the Russian mind. To test the suggested methodology, the following steps were conducted (see Table 4).

4.3. Case study: LDA vs. ISLDA

In this section, we present the results of automated and qualitative comparison between LDA and ISLDA. In this research, we focus on Livejournal, the blogging platform that, for Russia, has in early 2000s become the most significant one in terms of the public sphere [70], in contrast to other popular platforms de-politicized and overwhelmed with trivial content. The basic sample is the full collection of posts of top 2,000 Livejournal Russian-language bloggers (listed in the Livejournal ‘social capital’ rating) of February to May 2013 (11 weeks, 363,580 texts, 1,072,283 stems; the dictionary, after cleaning stopwords and low frequency words, contained 192,614 words with about 53.5 million total instances of use). This period is ‘calm’, as no ethnic-related events (except for Easter) were salient in the public agenda.

We have performed experiments with different numbers of topics (50, 100, 200, and 400) for both regular LDA and IS LDA; 50 to 200 were made on a smaller test dataset, while 400 topics were run on actual dataset. To compare the results, we looked at the word collections automatically assigned to the topics in LDA- and ISLDA-processed datasets. In order to assess the quality of the topics more accurately, it may be necessary to analyse the underlying document collection assigned to the relevant topics by hand; here, we rather evaluate the visibility of topics for a researcher on the first stage of LDA assessment. Comparing regular LDA results for 100 and 400 topics, it is clear that ethnic topics need to be dug up at 400 rather than 100 topics. The share of ethnic topics was approximately the same: 9 out of 100 (9%) and 34 out of 400 (8.5%), but in terms of quality, the first iteration gives “too thick” topics like Great Patriotic war,

Table 4. Stages of the project ‘Mapping Ethnic Discourse in the Russian blogosphere’

Project stage	LDA	ISLDA
Dataset collection	V	V
Topic modeling with stability measures applied	V	V
Automated similarity analysis (Jaccard and Kullback-Leibler measures applied)	V	V
Top words assessment	V	V
Comparing the quality of LDA and ISLDA by keyword probability and tf-idf coherence	V	V
Selection and indexing of ethnic-related topics	V	Work in progress
Selection of text collection for manual coding	V	Work in progress
Elaboration of codebook (test coding, Kappa tests, correction of codebook)	V	Work in progress
Manual coding of posts	Work in progress	Work in progress
Analysis of coding results	Work in progress	Work in progress

Muslim, CEE countries, “big chess play” (great world powers and their roles in local conflicts), Russian vs. Western values, US/UK celebrities and East in travel (Japan, India, China and Korea). The 400-topic LDA iteration looks much more informative, providing topics of three kinds: event-oriented (e.g., death of Kim Jong-Il or boycotting Russian TV channel NTV in Lithuania), current affairs oriented (e.g., armed conflicts in Libya and Syria or protests in Kazakh city Zhanaozen), and long-term, or ‘chtonic’, topics. The latter may be divided into “neutral” descriptions of country/historic realities (Japan, China, British Commonwealth countries, ancient Indians etc.), long-term conflict topics (e.g., the Arab-Israeli conflict, Serb-Albanian conflict and the Kosovo problem), and two types of “problematized” topics: internal problems of a given country/nation (e.g., the U.S.) and “Russia vs. another country/region” topics (Poland, Chechnya, Ukraine). There are several topics of particular interest for the ethnic case study: a topic on Tajiks, two opposing topics on Russian nationalism (“patriotic” and “negative”), and a Tatar topic. Several ethnic groups, e.g., Americans, Germans, Russians, and Arabs, were subject of more than one topic.

To assess the ISLDA results, the 100-topic test dataset and the 400-topic actual dataset were examined. For the former, four ethnicities were chosen according to their importance for the final qualitative goal. The Ukrainians were chosen due to their ethnic, geographical, and historical proximity to the Russian public, which could potentially lead to a bigger number of issues discussed international the blogosphere about this ethnicity. Tajiks were chosen as the representatives of a big migrant community potentially linked in public mind to social trouble. Georgians were chosen, as the 2008 armed conflict between South Ossetia and Georgia deeply affected and polarized public opinion in all the three communities (Russian, Ossetian, and Georgian). French were chosen due to the multiple contexts (from Napoleonic wars to cuisine and fashion) to which they may belong in the Russian mind.

With 100 topics, ISLDA performed better than LDA in most cases. For ex-Soviet ethnic groups (Tajik and Georgian), one of two pre-assigned topics clearly showed a problematized context. For Tajiks, it was illegal migration: we also saw writers from opposing opinion camps (Belkovsky, Kholmogorov, Krylov) and vocabulary characteristic of opinion media texts. For Georgians, the context of the Georgian-Ossetian conflict of 2008 clearly showed up, enriched by current events like election issues in South Ossetia. ISLDA found new important topics related to the chosen semi-supervised subjects. Table 5 shows topics from our runs with 100 and 400 topics related to Ukraine. With 100 topics, ISLDA distinguishes a Ukrainian nationalist topic (very important for our study) that was lost on LDA.

With 400 topics, LDA finds virtually the same topics that are found via assigning 100 topics, while ISLDA finds three new important topics: scandals related to Russian natural gas transmitted through Ukraine, a topic devoted to Crimea, and again the nationalist topic (this time with a Western Ukrainian spin).

In our main dataset, the results by far seem a bit less promising than expected but still have their advantages. Our question was whether ISLDA indeed could help gather the topics relevant for the keywords – or the same topics could be found in vanilla LDA output by searching for this word. Our results indicate that this is not the case and that ISLDA indeed helps find more concentrated and more readily interpretable topics. To illustrate this, Fig. 2 shows the probability of a chosen keyword in the topics ordered by this probability for several characteristic examples, namely (of 22 most frequent ethnicities that were used for ISLDA tests) American, German, European, Tatar, and Georgian.

The graphs indicate that the top topic is nearly always more concentrated and better captures the keyword in ISLDA than LDA. Note also how ISLDA topic probability often drops to nearly zero even before three topics have been found; this indicates that ISLDA helps uncover the “true” content of the dataset with regard to the keywords and concentrate it in a few topics (otherwise it would always take up all available topics). It also means that the number of topics per set of keywords is not overly important, and it suffices to set a “large enough” number as a parameter.

To doublecheck the automated results, we assessed the content of the topics in the main dataset (top words lists) for four ethnicities: Germans and Europeans as those who have more than one topic found by ISLDA, whereas Tatar and Georgians who had only one topic each related to the ethnic naming. What we saw was that ISLDA still was not overwhelmingly better in digging up event-oriented topics in case of multiple topics available in LDA: in the very similar cases of American, German, and Tatar, LDA performed a bit better for top2 and top3 topics for German and Tatar, while for American it was vice versa. But ISLDA was definitely better in separating one topic from another: thus, European got into the context of pirate tales and history of the European civilization, and the topics were separated well. In case of one ‘real’ (most relevant, the one that matters) topic searched for in a dataset, as for European and Georgian, ISLDA performs better in terms of keyword probability, thus, presumably, extracting the topic with higher accuracy.

For further sociological studies directed at specific issues, we recommend to combine LDA with ISLDA, the latter equipped with the number of preassigned topics (interval sizes) chosen a priori larger than the possible number of relevant topics: in our experiments, we saw that extra slots are simply filled up with some unrelated topics and do not deteriorate the quality of relevant topics. Looking for stable discourses might be suited by LDA with the number of topics no smaller than several hundred; but for separating contexts and better digging for a single topic (e.g. for an event-oriented or issue-oriented one) one should employ ISLDA technique.

In terms of the topic quality metrics discussed in Section 3, both average coherence and tf-idf coherence are slightly higher in ISLDA than in LDA. In our experiment, we have obtained average coherence of -873.95 for LDA and -870.25 for ISLDA and tf-idf coherence of -537.14 for LDA and -516.79 for ISLDA. This case study also supports the claim that tf-idf coherence is better for human judgement than regular coherence: readily interpretable topics consistently appear higher in the list of topics ranked by tf-idf coherence than by regular coherence.

One final question is whether ISLDA indeed helps to gather the topics relevant for the keywords or the same topics could be found in vanilla LDA output by searching for this word. Our results indicate that this is not the case and ISLDA indeed helps find more concentrated and more readily interpretable topics. To illustrate this, Figure 2 shows the probability of a chosen keyword in the topics ordered by this probability for several characteristic examples. The graphs indicate that the top topic is nearly always more concentrated and better captures the keyword in ISLDA than LDA. Note also how ISLDA topic probability often drops to nearly zero even before three topics have been found; this indicates that ISLDA helps uncover the “true” content of the dataset with regard to the keywords and concentrate it in a few topics (otherwise it would always take up all available topics). It also means that the number of topics per set of keywords is not overly important, and it suffices to set a “large enough” number as a parameter.

4.4. Case study: word frequency analysis and test coding of LDA topics

We have obtained two types of results that cross-validate each other: word frequency analysis and test coding of the topics discovered by regular LDA. By September 3, 2014 we will have full mapping results for LDA runs, which should allow for mapping ethnic attitudes expressed in the dataset and suggesting underlying factors for these attitudes.

Word frequency analysis. Word frequency statistics available at this stage of the research is quite informative. 166 ethnonyms and quasi-ethnonyms were discovered in the word list of the 5,000 most frequent stems of our text collection

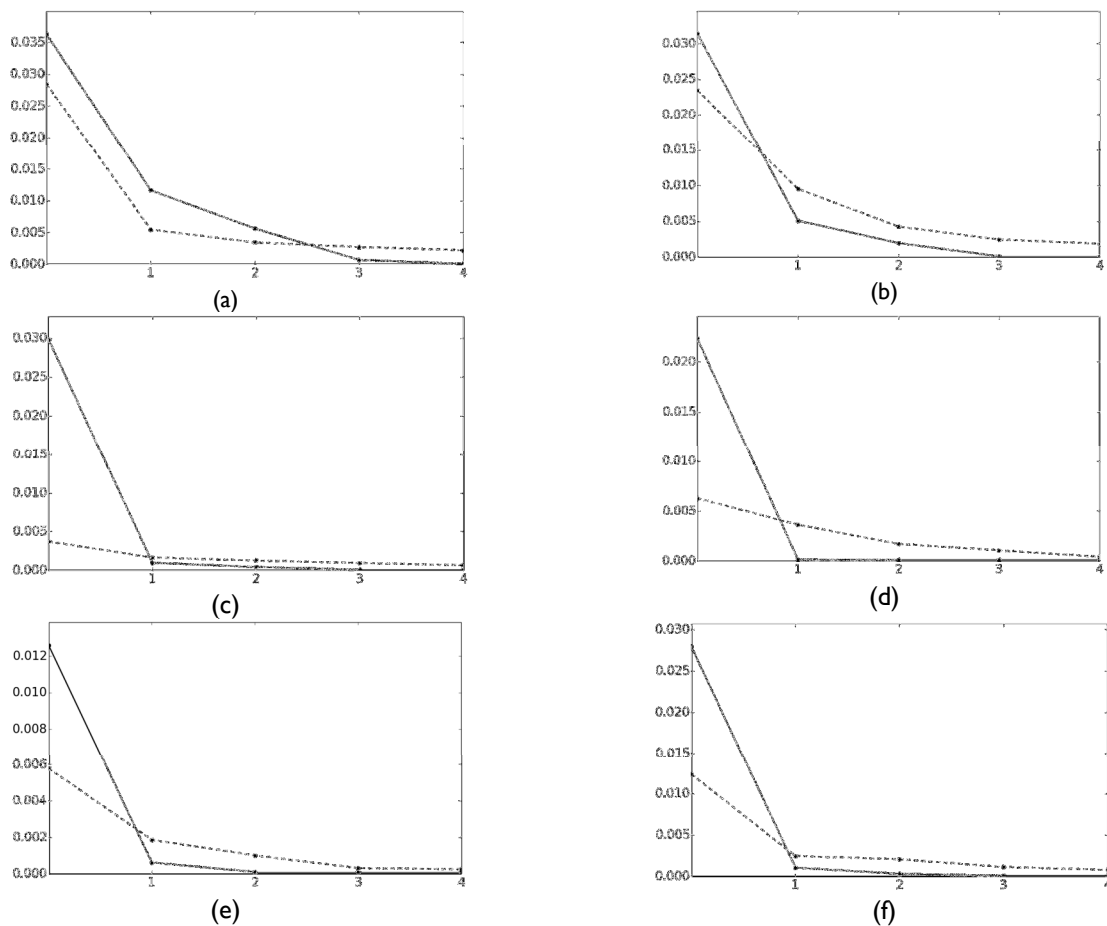


Figure 2. Keyword probabilities for keywords in topics ranked according to these probabilities. ISLDA, solid lines; LDA, dashed lines. Ethnicities/keywords: (a) American; (b) German; (c) European; (d) Tatar; (e) Georgian; (f) English.

Table 5. A comparison of LDA topics related to Ukraine: LDA, 100 topics; ISLDA, 100 topics; LDA, 400 topics; ISLDA, 400 topics.

Model	Word	Prob	Word	Prob	Word	Prob	Word	Prob
LDA, 100 topics	Ukraine	0.043	Ukraine	0.049				
	Ukrainian	0.029	Ukrainian	0.017				
	Polish	0.012	Timoshenko	0.015				
	Belorussian	0.011	Yanukovich	0.015				
	Poland	0.011	Victor	0.012				
	Belarus	0.010	president	0.012				
ISLDA, 100 topics	Russian	0.125	Ukraine	0.056				
	Russia	0.023	Ukrainian	0.022				
	nation	0.022	president	0.014				
	language	0.013	Timoshenko	0.011				
	national	0.012	Yanukovich	0.010				
	Ukraine	0.009	Victor	0.009				
LDA, 400 topics	Ukraine	0.098	Ukraine	0.054	dragon	0.026		
	Ukrainian	0.068	Timoshenko	0.019	Kiev	0.022		
	Belorussian	0.020	Yanukovich	0.018	Bali	0.012		
	Belarus	0.018	Ukrainian	0.016	house	0.010		
	Kiev	0.018	president	0.015	place	0.006		
	Kievan	0.012	Victor	0.013	work	0.006		
ISLDA, 400 topics	Ukraine	0.065	Ukraine	0.062	Ukrainian	0.040	Crimea	0.046
	gas	0.030	Timoshenko	0.023	Ukraine	0.036	Crimean	0.015
	Europe	0.026	Ukrainian	0.022	Polish	0.021	Sevastopol	0.015
	Russia	0.019	Yanukovich	0.018	Poland	0.017	Simferopol	0.008
	Ukrainian	0.018	Kiev	0.015	year	0.009	Yalta	0.008
	Belorussian	0.018	Victor	0.014	L'vov	0.006	source	0.007
	Belarus	0.017	president	0.013	Western	0.005	Orjonikidze	0.005
	European	0.015	party	0.013	Cossack	0.005	sea	0.005

via manual search; their word frequency was contrasted to word frequency dictionaries of the Russian language. For 58 of the 166 discovered ethnonyms, it was possible to check whether their use goes beyond average. We learnt that 48 of 58 ethnicities go above average; thus, ethnic-oriented discourse is quite intense and stable in the blogosphere. There are, indeed, ethnic groups that are discussed up to 5.5 times over average. The most frequently mentioned ethnicities in our dataset are (top 30, in the descending order): American, Jew, German, Chinese, Englishman, Frenchman, Japanese, Cossack, Ukrainian, European, Chechen, Polish, Tatar, Slav, Arab, Turk, American Indian, Negro¹, Greek, Italian, Finn, Briton, Tajik, Caucasian (pejorative), Spaniard, Georgian², Swede, Israeli, Czech, Maya.

If we apply stability measures and analyze 5 topic modeling runs rather than 1, there are 23 ethnic groups that are discussed more than on average *and* form stable topics. Among those, 5 groups may be traced, with no dominance of in-Russian and post-Soviet ethnic groups, neither by quantity of topics nor by instances of use of ethnonyms. These ethnonyms are the following: Briton, Tajik, Ukrainian, European, Chinese, Caucasian (pejorative), Uzbek, Spanish, American Indian, Korean, American, Pole, Japanese, Palestinian, Jew, Israeli, Arab, Czech, Cossack, Englishman, Chechen, Frenchman, Tatar.

Contrary to media discourse, those most discussed are not South-Asian or North Caucasian ethnicities but nations that play key roles on global (Americans, Germans) or regional scale (Ukrainians, Arabs). In-Russian nations are practically (and strikingly) absent from the list, with only 3 of 30 that could belong to Russia. Of those 3, Cossacks share Ukrainian and Russian identity, Caucasians live both in and outside the official Russian border, and Tatars form a distinct polity (Republic of Tatarstan) within the federal structure of Russia. This, probably, speaks of an inherently outside-oriented ethnic discourse, with blurred borders of the country in its post-Soviet public mind.

As top words analysis shows, there are very few ethnic groups that form stable topics themselves; the overwhelming majority of topics are multi-ethnic. One of the interesting findings (though quite expectable) is that macroregions play the key role in shaping public imaginary on ethnicities, as the level of discussion in >50% of the 'ethnic' topics is macroregional in how the ethnonyms combine within the topics. But what is somewhat less self-evident is, again, how the borders of these macroregions are shaped. E.g., along with quite conventional Western Europe, Scandinavia, and East Asia, there are also several 'axes' like post-Soviet Muslim republics – 'militarized' Middle East – North Caucasus or Belarus – Ukraine – Crimea.

As stated above, we discovered three types of ethnicity-related topics: 'chronical', issue-based and event-based. In the 33 topics, the proportion was 24:6:3 (with one topic non-defined). This is counter-intuitive, as event-oriented topics

in other web 2.0 media tend to occupy more space. Stable ethnic-related topics are detectable in both LDA and ISLDA. For LDA, of 2,000 topics, 291 (15%) were ethnicity-related. Stability of ethnic discourse does not differ much from the average. Thus, Livejournal is, obviously, not very much concerned of ethnic topics; but they constitute stable discussion milieus. By far, the 'political – ritual' axis is very evident in the content. Top words showed that proportion of political, 'ritual', both, other, and non-definable topics was 16:5:5:4:3. Thus, 79% of the topics seem to be located within this axis; interpretative reading will show if this is true. The post-empirical nature of the discourse is also clearly evident. Of 166 ethnicities found, in-Russian are just 39 (28% of the 139 indigenous nationalities registered by the 2010 census), and most small nationalities are absent. While this is explainable, more surprising is that, counter-intuitively, of the 23 stably discussed nations, European, distant (Chinese, Japanese, American, American Indian), Middle East & North African, in-Russian, and post-Soviet nations are 8:4:4:3:3, the discourse being oriented to past and current armed conflicts, international politics, immigrants, and border issues rather than to internal nationalities and their agendas.

Test coding of LDA topics and its results. While genuine stabilization of LDA is still waiting for its development, here we propose a semi-automated methodology of detecting the most stable topics from multiple runs of the algorithm. All thresholds in this methodology are conventional but derived from empirical observations. At first, after five runs of the algorithm have been performed, all 2,000 topics were compared pairwise with Jaccard and Kullback-Leibler measures. The most similar topics from different runs were combined into topic sets if their similarity was detected by both measures or if the Jaccard coefficient was not less than 0,150. After that, if at least one topic of a set contained words possibly related to ethnicity (e.g. ethnonyms, country names, traditional subjects etc.) among its top ten most probable words, the entire set became the subject of further scrutiny. All topics within it were checked manually to contain ethnonyms with probability of no less than 0,002, thus producing 291 topics within 154 sets. Finally, 33 topics were left in the analysis based on the *index of their 'real repeatability'*. To construct the index, we manually defined if each set constituted a single 'real' theme or a number of related 'real' themes (e.g. for the French, French cuisine and Napoleon wars). Then we divided the number of the ethnic-related topics in the set by the number of the 'real' themes in it; e.g. for a set of 5 topics defined as similar by stability measures, there were 2 'real' topics – thus, the index for this set was 2,5. There were 35 topics that acquired the index of 2,5 and higher (of max 5), but one of them was repost-based and one was based on very few texts; these were filtered out, and thus 33 sets remained for coding and reading.

Test coding was performed for 9 of these 33 topics; the rest is still under scrutiny. In the 260 posts of the first 9 topics that were test-coded (10 posts were missing, which meant that some of the 9 topics were created by less than 30 posts). 72 posts contained no ethnonyms; in the 188 remaining posts, 460 instances of use of ethnonyms (that is, 460 units of analysis) were designated, comprising 93 nations. The 24 most often mentioned nations are depicted on Fig. 3; of them, only 6 (25%) may be to some extent allocated as in-Russian (marked orange on Fig. 3). Of those 460 instances, 113 were just mentionings (and thus were not further coded), and 347 instances were more or less substantial and possible to code. Thus, the topics tend to be multi-ethnic indeed. On the other hand, though, dominant ethnicities tend to show up; the coders were able to detect a dominant ethnonym in 5 of 9 topics, even if none of the topics was really mono-ethnic; these ethnicities were Americans, Ukrainians, Soviet³, Syrian, and Venezuelan. This, as well as analysis of the 13 nations mentioned over 10 times, shows that the most often mentioned nations may be divided into 'contextual' (traced in many topics) and 'topical' (met in one topic only). The most 'contextual' of the 13 nations are Jews and English (7 topics of 9), American and French (5 of 9), and Russians, Chechens, Ukrainians, and Germans (4 of 9), while Syrians and Venezuelans were met, understandably, in the topic dedicated to the Syrian armed conflict and the one about Hugo Chavez's death. This makes us think that ethnic discourse in the blogosphere in some cases may form a 'universal context' for other topics.

As to the 'political vs. ritual' axis, it has emerged even in much bigger extent, and the 'political' pole is overwhelming. For now, we only measured the substantially discussed instances. Of 347 instances, the context in which they were discussed (that is, the main orientation of the post) the proportion of political/economic/social-issue/cultural-'ritual'/mixed of these/other was 323:1:4:9:8:6, thus bringing all the contexts except the political one (which occupied 93% of the discussion context) practically beyond significance.

The 'dominant - subordinate' axis is more diverse. The 'victim'/'neutral'/'ambivalent'/'mainly aggressor' ratio is 86:48:45:166; thus, discourse is polarized in these terms. Some nations are shown mainly as victims (e.g., Belarussians), some ambivalently (e.g., Jews or Ukrainians), and some mainly as aggressors (e.g., Americans or Germans).

The posts were also analyzed for presence of Russians in the text. This was coded in a special way: 0 was for absence, 1 – for Russians as 'россияне' ('inhabitants of Russia'), 2 for Russian as 'русские' (ethnic Russian), 3 for markers of pan-Slavonic, Old Russian, or pagan Russian naming (like 'русич', 'рус'). The proportion in definable 525 instances is 316:158:49:3; for the 347 substantial instances, it is similar: 214:83:47:2 (with 2 flaws in coding) – that is, ethnic Russians are a bit more discussed within the substantial ethnic-oriented discourse. This proportion is quite expectable. The most interesting group is the 49 instances thus somehow connected to Russian ethnos in the text; ethnonyms most co-mentioned with Russians are Americans (10), Chechens (6), Englishmen / Britons / Anglo-Saxons (6), Germans (5), and Soviet people (4); other 14 co-mentioned nations are quite diverse, from a small nation of Uyгур to historical Khazars.

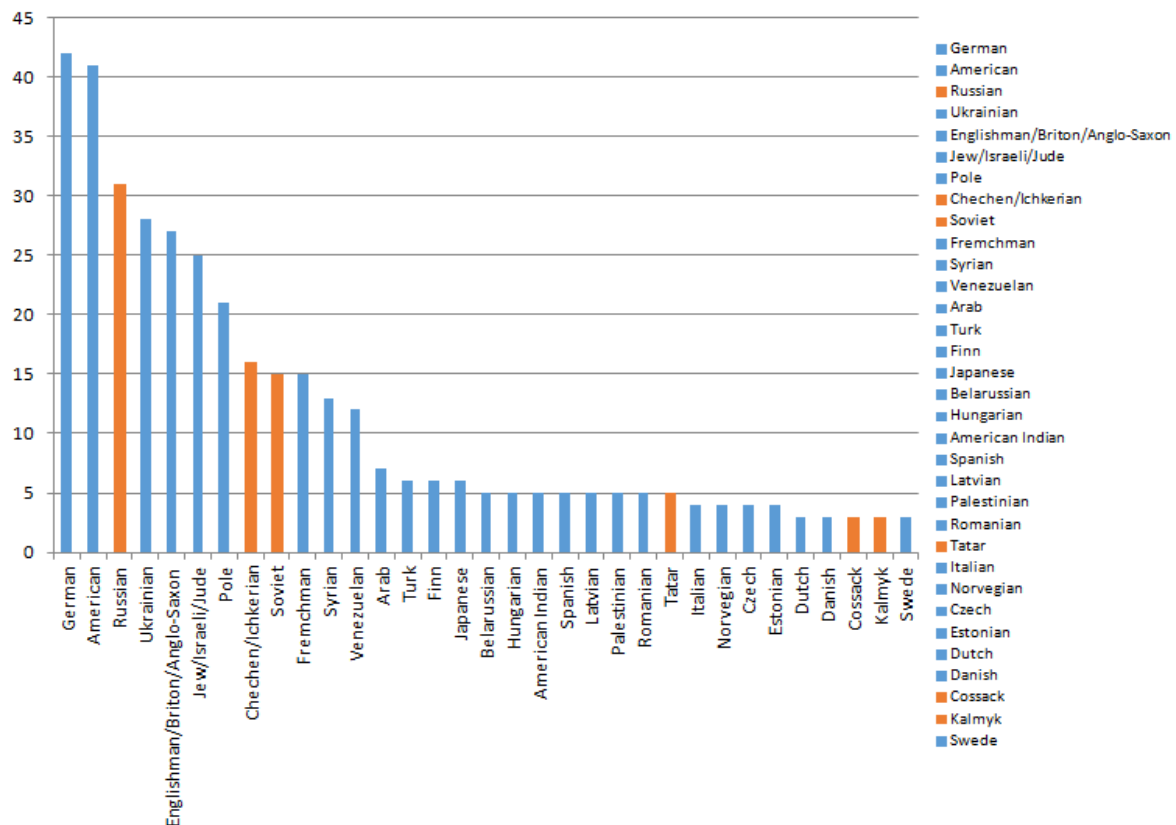


Figure 3. 24 ethnic groups most often mentioned in the dataset.

5. Conclusion

In this work, we have introduced two new ideas that are important for topic modelling in qualitative studies. First, we have presented the Interval Semi-Supervised LDA model (ISLDA) as a tool for a more detailed analysis of a specific set of topics inside a larger dataset and have showed an inference algorithm for this model based on collapsed Gibbs sampling. With this tool, we have described a case study in ethnical discourse analysis on a dataset comprised of the Russian Livejournal blogs. We show that topics relevant to the subject of study do indeed improve in the ISLDA analysis and recommend ISLDA for further use in sociological studies of the blogosphere. Second, we have presented a new topic quality metric, tf-idf coherence that improves upon regular coherence in predicting human judgement about topic quality. We have supported this claim with an experiment designed to extract this judgement from human subjects.

Research in quality metrics seems a promising direction for further work; in particular, one interesting question for further study is how to best extend the tf-idf coherence metric (which at present deals with topics) so that it can serve as a good measure for the overall quality of a specific topic model. In this way, the new metric may help find optimal sets of parameters for the LDA model (number of topics, α , β), which is always a nontrivial problem in specific applications.

The approach we introduced helps in separating topics and better formulation of those topics which are discussed as issues or events, but not ‘chronically’; for the latter, regular LDA could be equally suitable.

As to substantial results of mapping of ethnic attitudes, we have arrived at the following conclusions.

1. The Russian blogosphere allows for research on ethnicity and provides a lot of interpretable material, as circa 15% of it is dedicated to ethnic-related discussion, and most of the ethnonyms are discussed more intensely than in average speech, which could mean that the ethnic debate on Livejournal is more or less heated. Ethnic-related topics are stable enough by today’s automated measures as well as by manual check.

2. Ethnic-related topics may be divided into ‘chronic’, issue-oriented and event-oriented. Surprisingly, ‘chronic’ topics overwhelm the Livejournal top bloggers’ agendas, which may be explained by the fact that the platform experiences more and more a repost in-flow of texts written outside it.

3. Test mapping shows that ethnic discourse in the Russian blogs is oriented outside Russia and shows clear signs of post-imperialism, as it is oriented to macroregions, geopolitical competitors, and border issues. Historic background seems to be a crucial factor shaping the expressed attitudes; some ethnicities play a ‘framework’ role for the discourse. In over 9 cases of 10, ethnicities are discussed within political context, making other contexts practically irrelevant. Mapping on the axis of dominance/subordination shows that polarization on this vector is significant, and some ethnicities are portrayed as ‘eternal victims’ (like Belarussians) and some as ‘eternal enemies’ (Germans or Americans).

Funding

This work was done at the Laboratory for Internet Studies, National Research University Higher School of Economics (NRU HSE), Russia, and supported by the Basic Research Program of NRU HSE.

References

- [1] Hoffmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 2001; 42(1): 177–196 .
- [2] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003; 3(4–5): 993–1022.
- [3] Griffiths T and Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences* 2004; 101 (Suppl. 1): 528–533.
- [4] Quercia D, Askham H and Crowcroft J. TweetLDA: supervised topic classification and link prediction in twitter. In: Contractor NS, Uzzi B, Macy MW and Nejd W (ed.) *Proceedings of the ACM Web Science Conference 2012*. New York: ACM, 2012, pp. 247–250.
- [5] Wang X and McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2006, pp. 424–433.
- [6] Gohr A, Hinneburg A, Schult R and Spiliopoulou M. Topic evolution in a stream of documents. In: *SIAM International Conference on Data Mining (SDM09)*. New York: SIAM, pp. 859–872.
- [7] Wang X and Grimson E. Spatial Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems* 2007; 20.
- [8] Pan CC and Mitra P. Event detection with spatial latent Dirichlet allocation. In: Newton G (ed.) *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital libraries (JCDL'11)*. New York: ACM, 2011, 349–358.
- [9] Fortuna B, Galleguillos C and Cristianini N. Detection of bias in media outlets with statistical learning methods. In: Srivastava AN and Sahami M (eds.) *Text Mining – Classification, Clustering, and Applications*. London: Chapman & Hall, 2009, pp. 27–50.
- [10] Lin WH, Xing EP and Hauptmann AG. A joint topic and perspective model for ideological discourse. In: Daelemans W, Goethals B and Morik K (eds.) *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Lecture Notes in Computer Science vol. 5112. Berlin: Springer, 2008, pp. 17–32.
- [11] Wu Q, Zhang C, Hong Q, and Chen L. Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science* 2014; 00: 1–00 (accessed August 2014).
- [12] Wu Q, Zhang C and An X. Topic segmentation model based on ATNLDA and co-occurrence theory and its application in stem cell field. *Journal of Information Science* 2013; 39(3): 319–332.
- [13] He Q, Chen B, Pei J, Qiu B, Mitra P and Giles L. Detecting topic evolution in scientific literature: how can citations help? In: Cheung D and Song IY (eds.) *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM'09)*. New York: ACM, 2009, pp. 957–966.
- [14] Blei DM and Lafferty JD. Correlated topic models. *Advances in Neural Information Processing Systems* 2006; 18.
- [15] Li SZ. *Markov Random Field Modelling in Image Analysis*. *Advances in Pattern Recognition*. Springer, Berlin Heidelberg, (2009).
- [16] Boyd-Graber JL and Blei DM. Syntactic topic models. In: Koller D, Schuurmans D, Bengio, Y and Bottou L (eds.) *Proceedings of the 2008 NIPS Conference*. Red Hook, NY: Curran Associates Inc., 2008, pp. 185–192.
- [17] Chang J and Blei DM. Hierarchical relational models for document networks. *Annals of Applied Statistics* 2010; 4(1): 124–150.
- [18] Blei DM and Lafferty JD. Dynamic topic models. In: Cohen W and Moore A (eds.) *Proceedings of the 23rd International Conference on Machine Learning*. New York: ACM, 2006, pp. 113–120.
- [19] Wang C, Blei DM and Heckerman D. Continuous time dynamic topic models. In: McAllester DA and Myllymäki P (eds.) *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*. Arlington, VA: AUAI Press, 2008, pp. 579–586.
- [20] Guo X, Xiang Y, Chen Q, Huang Z and Hao Y. LDA-based online topic detection using tensor factorization. *Journal of Information Science* 2013; 39(4): 459–469.
- [21] Blei DM and McAuliffe JD. Supervised topic models. *Advances in Neural Information Processing Systems* 2007; 22.
- [22] Lacoste-Julien S, Sha F and Jordan MI. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems* 2008; 20.

- [23] Rosen-Zvi M, Griffiths T, Steyvers M and Smyth P. The Author-Topic model for authors and documents. In: Chickering DM and Halpern JY (eds.) *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Arlington, VA: AUAI Press, 2004, pp. 487–494.
- [24] Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P and Steyvers M. Learning Author-Topic models from text corpora. *ACM Transactions on Information Systems* 2010; 28(1): 1–38.
- [25] Bagheri A, Saraee M and de Jong F. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science* 2014; 0:1-00 (accessed August 2014).
- [26] Teh YW, Jordan MI, Beal MJ and Blei DM. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 2004; 101(476): 1566–1581.
- [27] Blei DM, Jordan MI, Griffiths TL and Tennenbaum JB. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems* 2003; 16.
- [28] Teh YW, Jordan MI, Beal MJ and Blei DM. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems* 2004; 17: 1385–1392.
- [29] Williamson S, Wang C, Heller KA and Blei DM. The IBP compound Dirichlet process and its application to focused topic modelling. In: Fürnkranz J and Joachims T (eds.) *Proceedings of the 27th International Conference on Machine Learning*. New York: ACM, 2010, pp. 1151–1158.
- [30] Chen X, Zhou M and Carin L. The contextual focused topic model. In: Yang Q, Agarwal D and Pei J (eds.) *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2012, pp. 96–104.
- [31] Andrzejewski D, Zhu X and Craven M. Incorporating domain knowledge into topic modelling via Dirichlet forest priors. In: Danyluk AP, Bottou L and Littman ML (eds.) *Proceedings of the 26th Annual International Conference on Machine Learning*. New York: ACM, 2009, pp. 25–32.
- [32] Andrzejewski D and Zhu X. Latent Dirichlet allocation with topic-in-set knowledge. In: Wang QI, Duh K and Lin D (eds.) *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2009, pp. 43–48.
- [33] Koltcov S, Koltsova O and Nikolenko S. Latent dirichlet allocation: stability and applications to studies of user-generated content. In: Menczer F, Hendler J and Dutton W (eds.) *Proceedings of the 2014 ACM conference on Web science*. New York: ACM, 2014, pp. 161–165.
- [34] Tikhonov A and Arsenin V. *Solution of Ill-posed Problems*. Washington: Winston & Sons, 1995.
- [35] Tikhonov A, Goncharsky A, Stepanov V and Yagola A. *Numerical Methods for the Solution of Ill-Posed Problems*. Dordrecht: Kluwer, 1995.
- [36] Tikhonov A, Leonov A and Yagola A. *Nonlinear Ill-Posed Problems*. London: Chapman and Hall, 1998.
- [37] Vorontsov K. Additive regularization for topic models of text collections. *Doklady Mathematics* 2014; 89(3): 301–304.
- [38] Asuncion A, Welling M, Smyth, P. and Teh YW On smoothing and inference for topic models. In: Bilmes J and Ng AY (eds.) *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. Arlington, VA: AUAI Press, 2000, pp. 27–34.
- [39] Wallach HM, Murray I, Salakhutdinov R and Mimno D. Evaluation methods for topic models. In: Danyluk AP, Bottou L and Littman ML (eds.) *Proceedings of the 26th Annual International Conference on Machine Learning*. New York: ACM, 2009, pp. 1105–1112.
- [40] Mimno D and Blei D. Bayesian checking for topic models. In: Merlo P (ed.) *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2011, pp. 227–237.
- [41] Chang J, Boyd-Graber J, Gerrish, S, Wang C and Blei DM. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* 2009; 20: 288–296.
- [42] Mimno D, Wallach HM, Talley, E, Leenders M and McCallum A. Optimizing semantic coherence in topic models. In: Merlo P (ed.) *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, 2011, pp. 262–272.
- [43] Rathore AS and Roy D. Performance of LDA and DCT models. *Journal of Information Science* 2014; 40(3): 281–292.
- [44] AlSumait L, Barbar D, Gentle J and Domeniconi C. Topic significance ranking of LDA generative models. In: Buntine WL, Grobelnik M, Mladenic D and Shawe-Taylor J (eds.) *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Lecture Notes in Computer Science vol. 5781. Berlin: Springer, 2009, pp. 67–82.
- [45] Ling CX, Huang J and Zhang HM. AUC: a statistically consistent and more discriminating measure than accuracy. In: Gottlob G and Walsh T (eds.) *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 2003, pp. 519–526.
- [46] Hand D J. and Till R J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 2001; 45: 171–186.
- [47] Bodrunova S, Koltsov S, Koltsova O, Nikolenko SI and Shimorina A. Interval Semi-Supervised LDA: Classifying Needles in a Haystack. In: A. Gelbukh (ed.) *Proceedings of the 12th Mexican International Conference on Artificial Intelligence (MICAI'2013)*. Lecture Notes in Artificial Intelligence vol. 8265. Berlin: Springer, 2013, pp. 265–274.

- [48] Lutovinova OV. Internet as a new 'oral-written' communication system. *Bulletin of Russian State Pedagogical University n.a. Herzen* 2008; 71: 58-65. (in Russian)
- [49] Bromley YV. *Ethnos and ethnography*. Moscow: Nauka, 1973.
- [50] Krejči Y, Velimsky V. *Ethnic and Political Nations*. London: Croom Helm, 1981.
- [51] Brubaker R. Nationhood and the National Question in the Soviet Union and Post-Soviet Eurasia: An Institutional Account. *Theory and Society* 1994; 23(1): 47-78.
- [52] Payin EA. *Between empire and nation: modernist project and its traditionalist alternative in Russian national policy*. Moscow: Liberal Mission Foundation, 2003.
- [53] Etkind A. *Internal Colonization*. Oxford: Oxford University Press, 2013.
- [54] Hutchinson J., Smith AD. *Ethnicity*. Oxford: Oxford University Press, 1996.
- [55] LeVine RA, Campbell DT. *Ethnocentrism: Theories of Conflict, Ethnic Attitudes, and Group Behavior*. London: John Wiley & Sons, 1972.
- [56] Bodrunova SS and Litvinenko AA. (2013b) New media and the political protest: the formation of a public counter-sphere in Russia of 2008-12 // *Russia's Changing Economic and Political Regimes: Putin Years and Beyond* / ed. by A. Makarychev, A. Mommen. London: Routledge. P. 29-66.
- [57] Eriksen TH. *Ethnicity and Nationalism: Anthropological Perspectives*. London: Pluto Press, 1993.
- [58] Anderson B. *Imagined communities: reflection on the origin and spread of nationalism*. London: Verso, 1983.
- [59] Prazauskas AA. *Ethnos and Politics: A Reader*. Moscow: URAO, 2000. (in Russian)
- [60] Drobizheva LM. Identity and ethnic attitudes of Russians in their own and alternative ethnic environment. *Socis (Social Research)* 2010; 12: 49-58. (in Russian)
- [61] Hall S. *Representation: Cultural Representations and Signifying Practices*. London: SAGE Publications, 1997.
- [62] Nakamura L. *Cybertypes: Race, Ethnicity, and Identity on the Internet*. London: Routledge, 2002.
- [63] Daniels J. Race and racism in Internet Studies: A review and critique. *New Media & Society* 2013; 15(5): 695-719.
- [64] Gavrov SN. *Modernization of Russia: Post-imperial transit*. Lincoln: Nebraska 4-H Clubs: Historical Materials and Publications, 2009.
- [65] Mukomel V and Payin E. The Causes and Demographic-Social Consequences of Interethnic and Regional Conflicts in the post-Soviet Union. In: Demko GJ, Ioffe G, Zayonchkovskaya Z (eds.) *Population Under Duress: The Geodemography of Post-Soviet Russia*. Boulder: Westview Press, 1999, pp. 177-183.
- [66] Zubarevich N. Four Russias. *Vedomosti* 2011; December 30. Available online at: http://www.vedomosti.ru/opinion/news/1467059/chetyre_rossii.html. (in Russian)
- [67] Ryzhakova SI, Zavyalova MV. *Ethnic images and stereotypes: upon materials of comparative research among youth of Latvia and Lithuania*. Moscow: RAS Institute of Ethnology and Anthropology, 2008. (in Russian)
- [68] Osadchaya GI. Values, aims and stereotypes of ethnic behavior of young Muscovites. *Social Policy and Sociology* 2011; 11: 86-105. (in Russian)
- [69] Huhlaev OE. Ethnonational attitudes of today's Russian youth. *Voprosy psikhologii* 2011; 1:46-57. (in Russian)
- [70] Gorny E. Understanding the Real Impact of Russian Blogs. *Russian Analytical Digest* 2009; 69: 8-11.