Article

JIS

Journal of Information Science
1–15

DOI: 10.1177/0165551515617393
jis.sagepub.com
SAGE

Topic modelling for qualitative studies

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

Sergei Koltcov
National Research University Higher School of Economics, Russia

Olessia Koltsova
National Research University Higher School of Economics, Russia

Abstract

Qualitative studies, such as sociological research, opinion analysis and media studies, can benefit greatly from automated topic mining provided by topic models such as latent Dirichlet allocation (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 subtopics 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 approach (ISLDA) where certain predefined sets of keywords (that define the topics researchers are interested in) are restricted to specific intervals of topic assignments. Our experiments show that ISLDA is better for topic extraction than LDA in terms of tf-idf coherence, number of topics identified to predefined keywords and topic stability. We also present a case study on a Russian LiveJournal dataset aimed at ethnicity discourse analysis.

Keywords

Latent Dirichlet allocation; LDA extensions; topic modelling; topic quality

1. Introduction

Over recent years, topic modelling techniques based on probabilistic latent semantic analysis [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], analysis of 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 has begun 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 users are talking about, identify underlying topical trends and follow them through time, as well as identifying the most relevant documents for a specific topic. Ideally, a researcher may be able to draw conclusions from word distributions for topics and only then read documents with high affinities for specific topics for a more in-depth study.

Corresponding author:
Sergey I. Nikolenko, National Research University Higher School of Economics, ul. Soyuza Pechatnikov, d. 16, 190008 St. Petersburg, Russia.
Email: snikolenko@gmail.com

Recent research brings topic modelling into this picture mostly as statistical analysis to detect words that may define media bias [9] or ideological discourse [10] rather than as direct help for qualitative studies.

One obstacle for the use of LDA in qualitative studies 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 describe the 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 5 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. Furthermore, keyword search does not allow one to quickly define different contexts surrounding various ethnic groups (or other objects) and distinguish between these contexts. On the other hand, one still wants to guide the topic model to topics identified with these specific keywords. In this work, we deal with these two obstacles, extending and augmenting the LDA model.

Specifically, we suggest a methodology for qualitative studies based on topic modelling that is both able to narrow the search down to specific topics defined by keywords and better evaluate topic quality with a metric that better reflects human judgement. Our novel contributions in this work are twofold. First, we propose and develop an interval semi-supervised modification of the LDA model (ISLDA) where topic values for a set of predefined keywords are fixed. Second, we design a novel topic quality metric, tf-idf coherence, and show that it corresponds to human interpretability better than regular topic coherence. The paper is organized as follows. In Section 2, we introduce the basic LDA model and briefly survey its most important extensions. Then we present a modification of LDA 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. Section 4 presents our experiments with human experts to evaluate both LDA and tf-idf coherence, and in Section 5 we present a qualitative case study: 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 LDA model [2, 3] is shown in Figure 1a. A collection of D documents is assumed to contain T topics expressed with W different words. Each document $d \in D$ of length N_d is modelled as a discrete distribution $\theta(d)$ over the set of topics: $p\left(z_j = t\right) = \theta_t^{(d)}$, where z is a discrete variable that defines the topic for each word instance $j \in d$. Each topic, in turn, corresponds to a multinomial distribution over the words, $p\left(w|z_j = t\right) = \phi_w^{(t)}$. Dirichlet priors α are assigned to the distribution of topic vectors θ, $\theta \sim Dir(\alpha)$, and β for the distributions of words in topics, $\phi \sim Dir(\beta)$. On Figure 1a, the outer plate spans documents and the inner plate spans word instances in each document (so the w node denotes the observed word at the current instance, and the z node denotes its topic). The inference problem in LDA is to find hidden topic variables $\boldsymbol{z}$, 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 it is convenient to frame as Gibbs sampling. In this work, we use Gibbs sampling because it generalizes easily to the semi-supervised LDA considered below. After easy transformations [3], Gibbs sampling reduces to the so-called collapsed Gibbs sampling, where z_j are iteratively resampled with distributions

$$p\left(z_j = t|\boldsymbol{z}_{-j}, \boldsymbol{j}, \alpha, \beta\right) \propto q\left(z_j, t, \boldsymbol{z}_{-j}, \boldsymbol{j}, \alpha, \beta\right) = \frac{n_{-j,t}^{(d)} + \alpha}{\sum_{t' \in T}\left(n_{-j,t'}^{(d)} + \alpha\right)} \frac{n_{-j,t}^{(w)} + \beta}{\sum_{w' \in W}\left(n_{-j,t}^{(w')} + \beta\right)} \tag{1}$$

where $\boldsymbol{j}$ is the vector of word instances, $n_{-j,t}^{(d)}$ is the number of times topic t occurs in document d and $n_{-j,t}^{(w)}$ is the number of times word w has been generated by topic t, not counting its current instance z_j. LDA has had numerous applications, including studies that survey scientific literature [3] and mine how topics change in time [11–13].

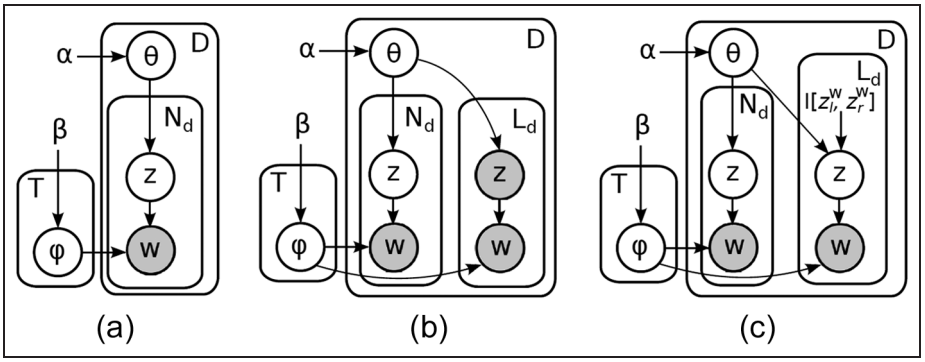

Figure 1. Probabilistic models: (a) LDA; (b) semi-supervised LDA; and (c) interval semi-supervised LDA.

2.2. Related work: LDA extensions

Over the last decade, after LDA was introduced by Blei et al. [2], the basic LDA model has been the subject of many extensions. Each of them 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, which 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, where i spans the corpora, 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. Many such extensions deal with time; for instance, the Topics over Time model represents the time when topics arise in a corpus of time-stamped texts 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 instance, 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, for example, user response in a recommender system [21]. 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 [22, 23]. Yet other extensions improve upon the bag-of-words assumption, modelling correlations between words and individual sentences [24]. Finally, a lot of work has been done on non-parametric LDA variants based on Dirichlet processes that we will not go into in this paper; for the most important non-parametric approaches to LDA see previous studies [25–29] and references therein.

There exist a number of LDA extensions that appear to be similar to the one proposed in this work; however, none of them serve exactly the same purpose. First, in the *Topic-in-Set knowledge* model and its extension with Dirichlet forest priors [30, 31], words are assigned with 'z-labels'. Second, a line of similar models has been proposed in relation to aspect-based sentiment analysis, with a goal to make aspect and opinion words better interpretable. In Wayne et al. [32], a hybrid MaxEnt-LDA model was proposed for joint modelling of aspects and opinions, in particular, to identify aspect-specific opinion words. In the work by Lu et al. [33], a multilevel tree-like model has been proposed that allows a user to specify *seed words* for specific aspects in a review and then infer aspects related to these seed words and corresponding opinions. Specifically, seed words are used to capture specific aspects of restaurant and hotel reviews (which was the dataset in Lu et al. [33]). To capture this, the seed words are assigned with an asymmetric beta prior $Dir(\beta + C_w)$, which makes it *a priori* more probable for the word w to occur in this topic with equivalent sample size C_w. The same idea has

also been used in recent works on LDA extensions to sentiment analysis [34, 35]: sentiment words from a predefined dictionary are assigned asymmetric beta priors that reflect prior knowledge and make specific sentiment labels for those words more likely. The work by Jagarlamudi et al. [36] addresses the same goal as our proposed approach, but from the opposite direction. In Jagarlamudi et al. [36], seed words are learned automatically from texts labelled into classes; words with the most discriminative power are extracted and used as seed words. The LDA modification in Jagarlamudi et al. [36] assumes that documents are generated from a mixture of regular topics and seed topics that produce only seed words, which are also subsequently trained. Thus, the model of Jagarlamudi et al. [36] uses already labelled texts as input and produces seed words, while in our case the labelled texts are unknown and are precisely what we are ultimately looking for. In Xiang et al. [37], a bootstrapping approach is used to expand a set of seed words (offensive words on Twitter in this case) for the purposes of document classification (detecting offensive tweets).

In this work, we propose an LDA extension, interval semi-supervised LDA (see Section 2.4), which is simpler than the modifications above: instead of tweaking the prior we simply project the sampling probability onto a given subset of topics. At the same time, we provide extensive experimental data, including human experiments, that shows that semi-supervised LDA manages to significantly improve human interpretability and coherence metrics for the resulting topics. Different from Lu et al. [33] where single-labelled sentences are used for quality tests, we deal with multi-topic texts, including those devoted to multiple ethnic groups, and our goal is not to break them into single-aspect sentences with *a priori* known aspects. On the contrary, ethnicity-related connotations that can be regarded as loose equivalents of aspects is what we are trying to learn (likewise, our ethnic groups can be regarded as equivalents of objects – restaurants or hotels – so the whole problem is, in a way, inverse compared with Lu et al. [33]). We therefore cannot apply such quality metrics as precision or recall used in Lu et al. [33] and instead rely on the experimental design proposed in Chang et al. [38].

2.3. 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 'junk' topics (in particular, topics that cannot be readily interpreted by human experts, see also Section 4.2). To find relevant subsets of topics in the dataset, we propose using 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 Andrzejewski and co-workers [30, 31]. The resulting graphical model is shown in Figure 1b (on Figure 1b and c, L_d is the number of key words in document d). 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:

$$p\left(z_j = t | \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right) \propto \begin{cases} 1, & w_j \in W_s \text{ and } t = \tilde{z}_{w_j}, \\ 0, & w_j \in W_s \text{ and } t \neq \tilde{z}_{w_j}, \\ q\left(z_j, t, \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right), & \text{otherwise.} \end{cases} \tag{2}$$

Otherwise, the Gibbs sampler works as in the basic LDA model; this yields an efficient inference algorithm that does not incur additional computational costs.

2.4. Interval semi-supervised LDA

One disadvantage of the semi-supervised LDA approach is that it assigns only a single topic to each set of keywords. However, many qualitative studies have specific reasons to reveal subtopics related to the same set. However, in many qualitative studies, including the case study presented in Section 4, the researcher has to discern between different contexts, if any, in which an ethnic group or a nation is mentioned: for instance, it is important to separate politically neutral posts about Ukrainian resorts from texts dealing with tensions between Eastern and Western Ukrainians or with Russian–Ukrainian relations. Drawing them all together with the semi-supervised LDA model 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. Thus, the two mentioned 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 $\left[z_l^w, z_r^w\right]$, and the probability distribution is restricted to that interval; the graphical model it shown on Figure 1c, where $I\left[z_l^w, z_r^w\right]$ denotes the indicator function: $I\left[z_l^w, z_r^w\right] = 1$ iff $z \in \left[z_l^w, z_r^w\right]$. In Gibbs sampling, we simply need to set the probabilities of all topics outside $\left[z_l^w, z_r^w\right]$ to zero and renormalize the distribution ($w = w_j$ is the word at instance j):

$$p\left(z_j = t | \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right) \propto \begin{cases} I\left[z_l^w, z_r^w\right](z) \dfrac{q\left(z_j, t, \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right)}{\sum_{z_l^w \le t' \le z_r^w} q\left(z_j, t', \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right)}, & w \in W_s, \\ q\left(z_j, t, \boldsymbol{z}_{-\boldsymbol{j}}, \boldsymbol{j}, \alpha, \beta\right), & \text{otherwise.} \end{cases} \tag{3}$$

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. Quality estimation: coherence and tf-idf coherence

3.1. The coherence metric and its problems

To evaluate our LDA extension, a reliable quality measure is needed, which, as we have found, is an unresolved problem in LDA literature. 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, that is, 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 [39] on the original dataset or measure the predictive power of a model by computing the log probability of a held-out set of documents that are shown to work not very well [40]; an important recent study advocates a Bayesian approach based on posterior predictive checking over [41].

As for individual topics, which is what we are interested in now, recent studies [38, 42] agree that topic coherence is a good candidate. In fact, to the best of our knowledge it is nearly the only candidate. This is somewhat surprising given the strong demand for quality assessment from LDA end users. The only alternative to coherence we have found is given by AlSumait et al. [43], where topic quality is defined as the difference between word distribution in a given topic and 'junk' distribution. However, unlike Mimno et al. [42], AlSumait et al. [43] has no human experiments, so we have limited our comparison to coherence as the baseline model. In Mimno et al. [42], for a topic t characterized by a set of top words W_t (either a fixed number of top words or all words whose probabilities exceed a predefined threshold), coherence is defined as

$$c(t, W_t) = \sum_{w_1 \neq w_2,\, w_1, w_2 \in W_t} \log \frac{\mathrm{d}(w_1, w_2) + \varepsilon}{\mathrm{d}(w_1)}, \tag{4}$$

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 [44]. 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. There are two main reasons for this. First, many topics that have good coherence are composed of common words that do not represent any topic of discourse *per se*. Table 1 lists the top words of the top 10 topics w.r.t. coherence in one of our experiments (dataset and experimental design are described in Secion 4.1). These common words do indeed co-occur often, but as a result, only two of the top 10 topics can be readily identified (topic (8) is about history, topic (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, either directly copying or extensively citing the original, so words 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, so in this work we do not address this second challenge, concentrating on the first one.

Table 1. Topics with top coherence scores: top words (translated) from top 10 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

Table 2. Topics with top tf-idf coherence scores: top words (translated) from top 10 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)
terrorist 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	al	service	Germany
terrorist	church	cleric	flour	South	meat	Cypriot	country	general	German
brother	cardinal	faith	dough	Kim	pepper	Euro	Muslim	officer	Hitler
police	catholic	church^{adj}	recipe	Korean	minute	financial	fighter	military force	Soviet
Boston^{adj}	Benedict	patriarch	egg	nuclear	dish	money	Arab	defense	World

3.2. tf-idf coherence

To alleviate the first drawback shown above, we propose tf-idf coherence, a rather straightforward modification of the basic coherence metric that accounts for the informative content of the topics. We have found that a tf-idf weighting scheme works very well and significantly outperforms standard coherence. Our idea is to introduce tf-idf scores [45] instead of the number of [co]occurrences. The tf-idf score of a word, as defined by Salton and Buckley [45], is its weighted frequency that is calculated so as to privilege the words that not only occur frequently in a given text, but also occur rarely in other texts of a corpus; thus, a coherence metric with tf-idf scores penalizes co-occurrence of common words that have low discriminative power. Namely, we define the tf-idf coherence metric as

$$c_{\text{tf}-\text{idf}}(t, W_t) = \sum_{w_1 \neq w_2,\, w_1, w_2 \in W_t} \log \frac{\sum_{d: w_1, w_2 \in d} \text{tf} - \text{idf}(w_1, d)\text{tf-idf}(w_2, d) + \varepsilon}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)} \tag{5}$$

where W_t is the set of (top) words in topic t, and the tf-idf metric is computed with augmented frequency

$$\text{tf-idf}(w, d) = \text{tf}(w, d)\text{idf}(w) = \left(\frac{1}{2} + \frac{f(w, d)}{\max\limits_{w' \in d} f(w', d)}\right) \log \frac{|D|}{|\{d' \in D : w \in d'\}|}, \tag{6}$$

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 top 10 topics in the same experiment but ranked with respect to tf-idf coherence. It is clear that these topics are 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; 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.

4. Experimental results

4.1. Dataset and LDA parameters

We now proceed to the description of our experimental dataset and LDA parameters chosen. We used the collection of Russian-language blogs because Russian social media is the object of our long-term research interest and a constant source for our sociological studies. All datasets consisted of LiveJournal posts written by 2000 top Russian bloggers over a given period of time; top bloggers were chosen from publicly available LiveJournal rankings, with weekly downloads based on the current rating. The dataset used for ISLDA tests and in the case study on ethnicity consisted of four months of LiveJournal posts written by 2000 top Russian bloggers and embraced the period from January to April 2013. The dataset contains 235,407 documents, and the dictionary of 700,000 tokens was compressed by cleaning stopwords and words that occur fewer than three times to 192,614 words with about 53.5 million total instances. The four one-month datasets from March 2012 (57,000 documents, 80,000 words), April 2012 (62,000 documents, 90,000 words), September 2012 (54,000 documents, 83,000 words) and March 2013 (103,000 documents, 110,000 words) were used in experiments with quality metrics. Raw data for the experiments in this work, including these datasets, has been made available at http://linis.hse.ru/rnf-topic-modelling-for-qualitative-studies. We used the Russian language lemmatizer MyStem, our own extensive list of stopwords developed over previous studies, and our own LDA implementation based on Gibbs sampling that was extended to ISLDA. In tokenization, we discarded punctuation and converted all words to lowercase before lemmatization. All experiments were run on an Intel Core i5 desktop with 64 Gb RAM.

Literature analysis shows that the values as well as methods of optimization for LDA hyperparameters α and β depend on the final research goals and on quality measures used; suggested values vary greatly [39, 46, 47]. We have found no works testing different values of α and β against human judgement about the resulting topics. Since our main goal is to evaluate our quality measure and our LDA extension, we have chosen to fix hyperparameters throughout all experiments at the values most commonly used (e.g. in Lu et al. [33]), namely $\alpha = 0.1$ and $\beta = 0.1$.

For similar reasons, the number of topics is also a problematic issue; human judgement is missing from the attempts to estimate it [3]. In some of our earlier works [48] we too used a non-human-based approach based on jump theory [49], but our experience with sociological interpretation of the results has convinced us that the 'good' number of topics depends on the level of 'resolution' a social scientist desires to obtain. For collections of the size we use here, our sociologists have tended to choose values between 100 and 400 topics, and this is the range we work with in this paper. For each experiment further below we specify the number of topics used. Models were trained for 200 Gibbs sampling iterations, the number determined experimentally in our earlier studies [50]. In Koltcov et al. [50], using similar data from the Russian LiveJournal, we show that after approximately 150 iterations both the word ratio and the document ratio virtually stop changing (this is our burn-in period), while after 50 more iterations both curves just turn into a straight line (and this is our number of subsequent iterations); to get the parameters we average every 5th sample after the burn-in. Finally, the problem of LDA stability is hardly addressed at all, while the differences between different runs are dramatic [50]. We therefore use from two to five runs in each experiment, the exact number being specified further below.

4.2. Experimental comparison of coherence and tf-idf coherence

To evaluate the two quality metrics, coherence and tf-idf coherence, we have compared them against human judgement. For each topic, we asked the subjects 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?

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; negative answers to both questions would indicate a junk topic. Therefore, in this experiment we asked the subjects direct questions about understanding topics rather than used subjects as oracles to judge objective topic qualities, as has been done in the experiments conducted in Chang et al. [38] and as we will do in Section 3.4. At the same time, since we are not interested in subjective judgements of a particular individual, each set of topics was offered to six to eight assessors for independent assessment, with some assessors evaluating all datasets and others only one or two, about 300 topics per assessor on average. In total, we evaluated four different datasets corresponding to four different LDA runs (two runs of 100 topics and two runs of 200 topics) with 10 different assessors, obtaining 2900 data points (topic evaluations) covering 600 topics.

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 questions 1 and 2 in the study

Dataset	Topics	Question 1			Question 2		
Time period	Number of topics	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

All assessors in this experiment were educated (university diploma or higher), but with different specializations. We mixed experienced assessors and newcomers and did a pilot coding, after which assessors discussed discrepancies and clarified assessment criteria.

After that, for each subject and each metric, we computed the area-under-curve (AUC) measure [51]. 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

$$\mathrm{AUC} = \frac{\sum_{i \in \mathrm{pos}} r_i - n_{\mathrm{pos}}(n_{\mathrm{pos}} + 1)/2}{n_{\mathrm{pos}} n_{\mathrm{neg}}}, \tag{7}$$

where n_{pos} and n_{neg} are the number of positive and negative examples in the dataset, and $\sum_{pos} r_i$ is the sum of the ranks of all positive examples [52]. 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 one to four words) identify each topic if possible; this was done as a sanity check to make sure that different subjects understood topics in a similar way. The results were checked by hand, and no major discrepancies were found. However, we drew all subsequent conclusions from the second test in which 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 4. These distances indicate a good match between the individual subjects' opinions: two vectors of answers differ, on average, only in 10–25% of positions, with the second question, more subjective than the first, leading to more difference in opinion. This was the main reason why we asked several independent individuals to assess the topics in this experiment's design.

4.3. Experimental comparison of LDA and ISLDA

In order to make a fair comparison between LDA and ISLDA, we have run several experiments complementing the tf-idf coherence metric. First, we performed an experimental evaluation of topic quality and the general interpretability of the models. To account for reproducibility, in all experiments we used five runs of ISLDA and five runs of LDA with different (random) initial approximations. We used 200 topics for both LDA and ISLDA. ISLDA had 22 different keywords and three topics per keyword (66 topics in total). Since the project was intended to study ethnicity-related public opinion, the keywords were ethnicity/nationality names in Russian. Nineteen were the most frequently mentioned ethnonyms in the corpus, and three more were chosen by experts in ethnic studies from among ethnonyms where one could expect an interesting discourse (and which also scored relatively high in frequency): 'American', 'German', 'Chinese', 'Jewish', 'French', 'English', 'Japanese', 'Ukrainian', 'European', 'Cossack', 'Polish', 'Slavic', 'Finnish', 'Chechen', 'Arabian', 'Greek', 'Tatar', 'Spanish', 'Italian', 'Czech', 'Georgian' and 'Swedish'. In a few cases, we used several similar keywords if several forms of the same word appear in Russian. Our experiments were standard word intrusion and topic

Table 4. Experimental results: word intrusion and topic intrusion experiments

Experiment	Model	Correct answers	Total answers	Ratio
Word intrusion	LDA	1064	1523	0.699
	ISLDA	1091	1604	0.680
	ISLDA, key topics	534	638	0.837
Topic intrusion	LDA	1595	1810	0.881
	ISLDA	1662	1895	0.877
	ISLDA, key topics	609	640	0.952

intrusion experiments as described in Chang et al. [38]. In the word intrusion experiment, the subjects were presented with a list of five words four of which were top words in a topic, and the fifth was an intruder from a different topic. In the topic intrusion experiment, the subjects were presented with a document (a blog post) and several topics, with one intruder topic and the rest actual topics that are expressed in the document; we selected at most four 'correct' topics per document but also applied a threshold so that if not enough topics were expressed in the document, the low-probability nearly random topics would not appear as correct. Word and topic intruders were randomly selected from words not found in the topics and from topics not found in the text, respectively, similar to Chang et al. [38]. Experiments were performed by 11 assessors who, like in the previous experiment, were educated individuals with widely varying backgrounds. A web interface was created for this experiment, tested and improved after test usage by the assessors.

Results of these experiments are shown in Table 4. In both experiments, a common pattern emerged: ISLDA in general is not that much better in the word and topic intrusion experiments, but when we restrict our attention to the topics that had been assigned with keywords, the result turns out to be much better than both for the rest of ISLDA topics and the regular LDA topics. This supports our claim that semi-supervised LDA produces good, readily interpretable topics on the subjects of interest (expressed with keywords).

However, the next question is whether ISLDA indeed helps to gather more topics relevant for the keywords and whether these topics are more concentrated, or the same number and concentration of topics could be found in vanilla LDA output by searching for this word. Our results indicate that the latter is not the case and ISLDA indeed helps find more concentrated and more numerous topics. To illustrate this, Figure 2 shows the probability $\phi_w^{(t)}$ of top words 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: extra topics will simply be used for something different.

We proceed with addressing the problem of LDA stability: it is a known problem that LDA may be unstable across different runs even on the same dataset with the same parameters [50], and if ISLDA happened to be even less stable than LDA it would devalue any gain in coherence. Thus, we have studied how well keyword-related topics hold across different runs. Table 5 shows the number of topics identified with a keyword (checked by hand) and the average number of runs out of five that these topics have appeared in for both LDA and ISLDA. Two topics from different runs were considered matching if they were nearest neighbours of each other in terms of both Kullback–Leibler divergence and Jaccard similarity on the set of top 100 words, and Jaccard similarity was at least 0.2. Results indicate that ISLDA is no less and often more stable than regular LDA; in general, ISLDA does not lose stable keyword-related topics and sometimes finds new ones. Finally, we have applied our tf-idf coherence quality measure for LDA and ISLDA topics (see Table 5); again, ISLDA on average outperforms LDA in terms of tf-idf coherence, a metric that we have already linked with human interpretability in our experiments above.

5. Mining ethnical discourse

5.1. Case study: project description

To evaluate ISLDA on real-life tasks, we have applied both LDA and ISLDA to a sociological project intended to study the discourse on ethnic groups and nations in the Russian blogosphere. The project seeks to find out how the issue of ethnicity is perceived by bloggers and, in particular, answer the following questions:

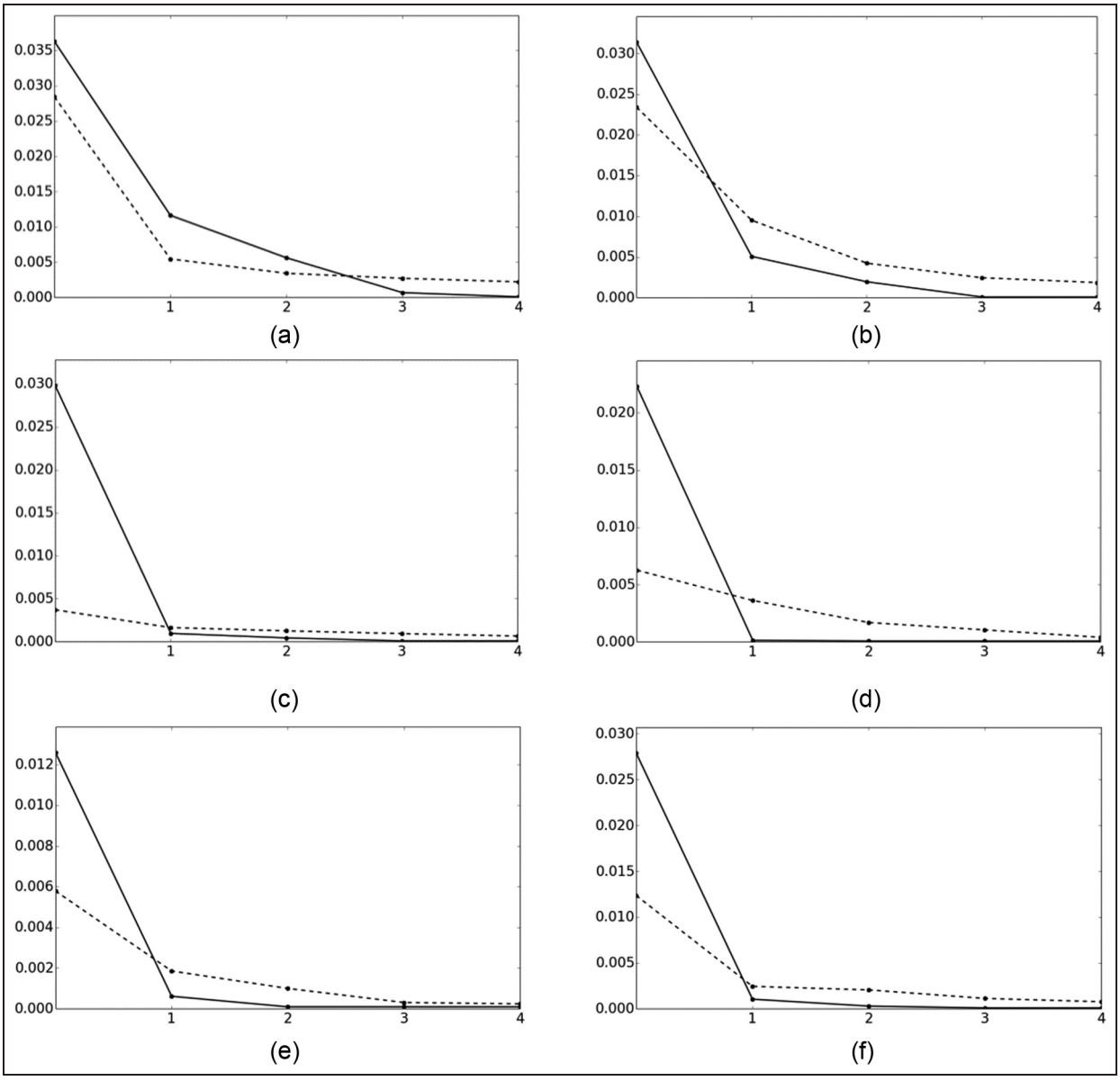

Figure 2. Keyword probabilities for keywords in topics ranked according to these probabilities; horizontal axis shows keyword ranks from 0 (top) to 4; vertical axis, probabilities. ISLDA, solid lines; LDA, dashed lines. Ethnicities/keywords: (a) American; (b) German; (c) European; (d) Tatar; (e) Georgian; (f) English.

Table 5. Experimental comparison between LDA and ISLDA in terms of number of topics, topic stability and tf-idf coherence

	Number of topics identified		Average number of similar topics (out of 5)		Average tf-idf coherence	
Keyword	LDA	ISLDA	LDA	ISLDA	LDA	ISLDA
American	2	3	4	5	−323.9	−306.2
German	1	1	5	5	−395.3	−229.4
Jewish	1	2	5	5	−250.5	−253.9
Chinese	1	1	5	5	−459.0	−410.3
Chechen	2	2	3.5	4	−258.8	−212.4

(1) Which ethnic or national groups are associated with unproblematic contexts (such as national cuisine, music and other cultural issues), and which are related to social tensions or problems (such as crime, unemployment, ethnical or international conflict)?
(2) Which groups are perceived as alien and dangerous?
(3) What are related topics with which all issues of ethnicity are closely connected (migration, religion, international relations, leisure travel)?
(4) What ethnic groups and nations are covered more than others?

For each interval (set) of topics we have used a set of related keywords of different types: 'ethnonyms' (e.g. Azeri), nation names (e.g. Azerbaijanis), and nation state names (e.g. Azerbaijan). To choose ethnonyms, we calculated their frequencies in the whole corpus, while the full list of ethnonyms had been compiled by our ethnicity studies expert from the Russian Census 2010 data and from other relevant sources. 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. Top of the list was occupied by Americans, Ukrainians, Germans and Jews (note that this was before the Ukrainian crisis). The ten most often occurring ethnic groups/nations were selected for running the ISLDA algorithm. For qualitative expert assessment of ISLDA topic mining quality in comparison to LDA, four ethnicities were chosen according to the final sociological goal. The Ukrainians were chosen owing to their ethnic, geographical and historical proximity to the Russian public, which could potentially lead to more issues being discussed. Tajiks were chosen as the representatives of a big migrant community potentially linked in the public mind to social trouble. Georgians were chosen because the 2008 armed conflict between South Ossetia and Georgia deeply affected and polarized public opinion in all three communities (Russian, Ossetian and Georgian). The French were chosen owing to the multiple contexts (from Napoleonic wars to cuisine and fashion) to which they may belong in the Russian mind.

5.2. Case study: qualitative results

In this section, we present qualitative results of the case study and compare LDA with ISLDA. In this case study we have performed experiments with different numbers of topics (100, 200 and 400) for both regular LDA and Interval Semi-Supervised LDA. To compare the results, an expert in ethnic studies looked at the top word collections automatically assigned to the topics in LDA- and ISLDA-processed datasets. In order to accurately assess topic quality, one may have to analyse the underlying documents assigned to this topic by hand; here, we rather evaluate the visibility of topics for a researcher on the first stage of LDA assessment. A topic was considered 'ethnic' by the expert if it met the requirements of topic interpretability as described in Section 4.2 and contained an ethnonym with other top words providing a meaningful social context (i.e. if top words did not look unrelated to the key word). To analyse the content of topics sociologically, the expert read 20 texts with the highest probabilities in each 'ethnic' topic.

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 broad' topics like Great Patriotic war, Islam, 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 merged in a single topic). The 400 topic LDA run looks much more informative, providing topics of two kinds. The first is event-oriented and is represented either by 'spot news' (e.g. the death of Kim Jong-Il or the boycotting of a Russian TV channel NTV in Lithuania), or by 'continuing news'/current affairs comments (e.g. armed conflicts in Libya and Syria or protests in Kazakh city Zhanaozen). The second kind includes long-tem topics, such as 'neutral' descriptions of country/historic realities (Japan, China, British Commonwealth countries, ancient Indians etc., each in a separate topic), long-term conflict topics (e.g. the Arab–Israeli conflict, the 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).

LDA has found several topics of special interest for the ethnic study: a topic on Tajiks, two opposing topics on Russian nationalism ('patriotic' and 'anti-nationalist') and a Tatar topic. Several ethnic groups, for example, Americans, Germans, Russians and Arabs, were the subject of more than one topic.

In ISLDA results, the 100 topic modelling covered the same ethnic topics as regular LDA, but Ukrainian ethnonyms produced a new result as discussed below. A 400 topic ISLDA gave a result much more informative than regular LDA. 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 (as had been expected); we also saw writers from opposing opinion camps (Krylov, Belkovsky, Kholmogorov) and vocabulary characteristic of opinion media texts. For Georgians, the context of

Table 6. Comparison of Ukraine-related topics: LDA, 100 topics; ISLDA, 100 topics; LDA, 400 topics; ISLDA, 400 topics

Model	Word	Probability	Word	Probability	Word	Probability	Word	Probability
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

the Georgian–Ossetian conflict of 2008 clearly showed up, enriched by current events, including election issues in South Ossetia. In general, we found that ISLDA finds new important topics related to the chosen semi-supervised subjects. Table 6 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 while ISLDA finds three new important topics: scandals on Russian natural gas transmitted through Ukraine, a topic on Crimea and the nationalist topic again (this time with a Western Ukrainian spin). ISLDA produces more informative topics on other specified ethnical subjects as well.

For further sociological studies with specific issues in mind, we recommend ISLDA with relatively large number of topics and the number of preassigned topics (interval sizes) chosen *a priori* to be larger than the possible number of relevant topics: in our experiments, we saw that extra slots are simply filled up with unrelated topics and do not interfere with relevant topics. In terms of topic quality metrics, both average coherence and tf-idf coherence are slightly higher in ISLDA than in LDA, with 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 when ranked by tf-idf coherence than by regular coherence.

6. Conclusion

We have introduced two new ideas for topic modelling in qualitative studies. First, we have presented ISLDA as a tool for a more detailed analysis of a specific set of topics in a larger dataset, with an inference algorithm based on Gibbs sampling. We have shown that topics relevant to the subject of study do improve in the ISLDA analysis, so we recommend using ISLDA in sociological studies that aim at mining specific topics, especially at learning unknown contexts (aspects) related to pre-determined objects. At the same time, it is less obvious that ISLDA is preferable for detecting overall topic structure of text corpora. Second, we have presented a new topic quality metric, tf-idf coherence, which improves upon regular coherence in predicting human judgement about topic interpretability.

We have supported this claim with experiments designed to extract this judgement from human subjects. These experiments have shown that, compared with regular coherence, tf-idf coherence performs much better in the task of privileging issue-oriented topics over coherent, but meaningless topics composed of frequently cooccuring common words. With these two tools, we conducted a study on a dataset of Russian LiveJournal blogs, which showed that ISLDA outperforms LDA in terms of tf-idf topic coherence, stability across training runs (instability is a known problem for LDA models), and relevant topics recognized by a human expert. Raw data for the experiments in this work has been made available at http://linis.hse.ru/rnf-topic-modelling-for-qualitative-studies.

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 or run. 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 non-trivial problem in specific applications. Another interesting direction would be to unite the approaches proposed in this work with other extensions designed to improve topic interpretability, such as the ones proposed in Newman et al. [53]; we believe that the benefits of our approach and other regularizers for topic models may be cumulative.

Funding

This work was done at the Laboratory for Internet Studies, National Research University Higher School of Economics (NRU HSE), Russia. It was supported by the Russian Research Foundation grant no. 15-18-00091.

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. In: *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 Nejdl W (eda) *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; 40(5): 611–620.
[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*. Berlin: Springer, 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, 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] 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.
[23] 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.
[24] 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; 40(5): 621–636.
[25] Teh YW, Jordan MI, Beal MJ and Blei DM. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 2004; 101(476): 1566–1581.
[26] 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, 17–24.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] Wayne XZ, Jing J, Hongfei Y and Xiaoming L. In: Hang L and Luis M (eds), *Proceedings of the 2010 conference on empirical methods in natural language processing*. Stroudsburg, PA: Association for Computational Linguistics, 2010, pp. 56–65.
[33] Lu B, Ott M, Cardie C and Tsou BK. Multi-aspect sentiment analysis with topic models. In: Spiliopoulou M, Wang H, Cook DJ, Pei J, Wang W, Zaiane OR and Wu X (eds), *Proceedings of the 11th IEEE international conference on data mining workshops*. Washington, DC: IEEE Computer Society, 2011, pp. 81–88.
[34] Lin C, He Y, Everson R and Ruger S. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering* 2012; 24: 1134–1145.
[35] Jo Y and Oh A. Aspect and sentiment unification model for online review analysis. In: *Proceedings of the Fourth ACM international conference on web search and data mining*. New York, NY, USA: ACM, 2011, pp. 815–824.
[36] Jagarlamudi J, Daume H and Udupa R. Incorporating lexical priors into topic models. In: *Proceedings of the 13th conference of the European Chapter of the Association for Computational Linguistics*. Avignon: Association for Computational Linguistics, 2012, pp. 204–213.
[37] Xiang G, Fan B, Wang L, Hong JI and Rose CP. Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus. In: *Proceedings of the 21st ACM international conference on information and knowledge management*. Maui, HI: ACM, 2012, pp. 1980–1984.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[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] 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.

[44] Rathore AS and Roy D. Performance of LDA and DCT models. *Journal of Information Science* 2014; 40(3): 281–292.
[45] Salton G and Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 1988; 24(5): 513–523.
[46] Tang J, Meng Z, Nguyen X, Mei Q and Zhang M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: *Proceedings of the 31st international conference on machine learning*, Beijing, 2014. JMLR, W&CP Vol. 32.
[47] Wallach HM. Structured topic models for language. Thesis submitted for the degree of Doctor of Philosophy, University of Cambridge, 2008.
[48] Koltsova O and Koltcov S. Mapping the public agenda with topic modeling: The case of the Russian LiveJournal. *Policy & Internet* 2013; 5(2): 207–227.
[49] Sugar C and James G. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association* 2003; 98: 750–63.
[50] 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.
[51] 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.
[52] Hand DJ and Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* 2001; 45: 171–186.
[53] Newman D, Bonilla EV and Buntine W. Improving topic coherence with regularized topic models. *Advances in Neural Information Processing Systems* 2011; 24: 496–504.