A full-cycle methodology for news topic modeling and user feedback research

Sergei Koltsov[1][0000-0002-2932-2746], Sergei Pashakhin[1][0000-0003-0361-2064] and Sofia Dokuka[2][1111-2222-3333-4444]

[1] National Research University Higher School of Economics, St. Petersburg 190008, Russia
[2] Institute of Education, National Research University Higher School of Economics, Moscow, 101000, Russia
skoltsov@hse.ru

Abstract. Online social networks (OSNs) play an increasingly important role in news dissemination and consumption, attracting such traditional media outlets as TV channels with growing online audiences. Online news streams require appropriate instruments for analysis. One of such tools is topic modeling (TM). However, TM has a set of limitations (the problem of topic number choice and the algorithm instability, among others) that must be addressed specifically for the task of sociological online news analysis. In this paper, we propose a full-cycle methodology for such study: from choosing the optimal topic number to the extraction of stable topics and analysis of TM results. We illustrate it with an analysis of online news stream of 164,426 messages formed by twelve national TV channels during a one-year period in a leading Russian OSN. We show that our method can easily reveal associations between news topics and user feedback, including sharing behavior. Additionally, we show how uneven distribution of document quantities and lengths over classes (TV channels) could affect TM results.

Keywords: topic modeling, text mining, TV news, news consumptions, online social networks, social media

1 Introduction

Social media play an increasingly important role in information spread within and across societies. In particular, younger generations of news consumers increasingly access them through social media news streams rather than through traditional media channels. As a result, social media aggregate digital traces of both news content and audience feedback that are matched together. This gives media professionals and social scientists a unique possibility to directly establish relations between content features of news and audience reactions to them in a way never possible before. However, the research community still lacks methodological routines that would allow social scientists to carry out full-cycle studies without inventing and testing new algorithms or data mining techniques.

In this paper, we develop a full-cycle approach for such a research by proposing a system of methodological steps arranged into a sequence. All elements of the system were tested in previous research, but here we show how to unite them, use them thoughtfully with attention to algorithmic limitations and how to interpret the outputs sociologically. We further apply our approach to the task of assessing audience feedback to the topics in a stream of Russian-language news in a Russian social networking site.

The first part of a research aimed at relating news content to audience feedback is to determine the features of the content. As the scope of our work is limited to news texts, further on we address texts only. They may differ in a number of aspects, such as source, time of issuing, length, genre, but above all – topics. The latter, unlike most other news features, are particularly hard to determine when news are too numerous to read. Topic modeling as a group of algorithms has been used for this [1]. However, TM is difficult to understand, its features and behavior are underresearched, and its limitations are not widely known and even less resolved. One of the most straightforward problems of TM that researchers face at the very beginning of its use is the problem of the "right" number of topics that has to be somehow set by a user. Due to these problems, despite some progress in the recent years, TM is still not widely adopted by social scientists and media practitioners. Meanwhile, news topic is a most important factor among others that can offer useful explanations of audience feedback. In this work we show how to integrate this method in a social research pipeline.

The second part of such research is to measure audience behavior which, in case of social networking sites, is limited to their technical functionality. News consumption is usually measured through the number of visits (clicks), unique visitors and the time spent on a news item page, however, this information is most often missing from the open access. What is usually present is user feedback embodied in user likes, comments, and sharing actions. Meanings that stand behind these types of actions are different. Likes are most often used to express approval, solidarity, or at least satisfaction with the news item features, such as its newsworthiness or ability to entertain. Comments, on the other hand, are signs of high involvement of a commenter into the issue, which does not necessarily mean agreement but usually indicates issue importance and controversy [2]. Moreover, long threads of comments often conceal heated and even polarized discussions in which commenters oppose each other rather than news authors and which use the news itself just as a starting point for discussing issues that are related to the news topic but are not identical to it. Finally, sharing occurs when users think a news item has a practical utility [3] or demands involvement of other audience members.

While sharing is of ultimate importance for media practitioners as it increases media audiences and advertising revenues, all three types of feedback are meaningful for broader social science. In this paper, we outline how these types of feedback may be interpreted in relation to topics they refer to.

The rest of the paper is organized as follows. In the next section we introduce information needed before the start of the proposed research cycle: we give a brief description of topic modeling as a method and review the entropic approach to the choice of optimal topic number. Then, in the third section, we describe our data set used to illustrate our methodological pipeline. Sections four and five introduce the suggested sequence of data analysis, while describing various problems and some rules of thumb on how to overcome them. In particular, section 4 describes the procedure we used to choose the optimal number of topics, extract stable topics and for topic labeling. Section five describes how information on topics was matched with user feedback and what interpretations were obtained. We conclude with a brief summary of the proposed approach.

2 Before the start: thermodynamic and entropic approaches to finding the optimal number of topics

Before describing the suggested research pipeline we give a most general explanation of this method, then paying more attention to its key problem – choice of the number of topics – and the solution that we offer as a part of the proposed research cycle.

As an extended version of cluster analysis, topic modeling (TM) is a family of mathematical algorithms that allows simultaneous fuzzy co-clustering of both objects and features, namely texts (documents, d) and terms (words, n-grams or other text properties, w). Mathematical foundations of topic modeling are described elsewhere [1]. TM input data is the term-document matrix where cells are frequencies of terms in documents, and its output consists of two matrices: term-topic (e.g. word-topic) matrix ϕ_{wt} and document-topic matrix θ_{td}, where cells are probabilities of either terms or documents in topics. The sums of probabilities of all topics in a document and of all words in a topic are equal to one, however, the sums of probabilities of all documents in a topic and of all topics assigned to a word may take any values. The latter two sums can be interpreted as the salience of a topic in the collection and the importance of a word in the collection, respectively.

Topics are viewed as latent variables whose distribution is unknown. Thus, to build a topic model means to solve the reverse task of finding an array of latent topics T or an array of one-dimensional conditional distributions $p(w|t) = \varphi(w, t)$ for each topic t constituting matrix ϕ_{wt} as well as an array of one-dimensional distributions $p(t|d) = \theta(t, d)$ for each document d using observable variables d and w.

One of the convenient methods of distribution restoring is Gibbs sampling employed by Steyvers and Griffiths who based their approach of the Potts model [4, 5]. This version of topic modeling is used here as it has shown better suitability for finding the optimal topic number with the thermodynamic approach that we also adopt in this paper as a part of the suggested research pipeline.

The problem of optimal topic number choice remains largely unsolved. However, topic modeling literature suggests several approaches to the problem. For instance, Cao et al. [6], based on the ideas of cluster analysis, propose to look at a topic as a semantic cluster (a set of terms) which makes it possible to compute its intracluster distance. They propose to use the cosine similarity measure as the function to be minimized. Thus, the optimal topic number would be at the minimum of average cosine similarity measure computed between all pairs of topics in a given solution. Another approach is based on Kullback-Leibler (KL) divergence [7]. The authors propose to look for the minimum KL across solutions for different topic numbers as follows. First, matrices ϕ_{wt} and θ_{td} are decomposed using SVD; then, in a pairwise manner KL divergences for vectors of singular values are computed. The optimal topic number corresponds to the situation when both matrices contain the same number of singular values. The described approaches have several limitations. First, it is unclear how the minimum of the proposed functions is related to the principle of entropy maximization – a standard approach to account for "information usefulness" in information theory. Second, additional SVD transformation or KL computation hinder application of such approaches to big data. For example, in the second approach, the authors compute KL for a document collection of less than 2,500 texts. These approaches do not consider the influence of initial distribution on the results of TM as was shown in [8]. Finally, a well-known approach to automatically finding an optimal number of topics is the hierarchical Dirichlet process (HDP) [9]. It constructs a hierarchy of topics in the form of a tree whose depth must be predefined by a user. The problem of the number of topics is thus transformed into a problem of the levels of the tree, but not truly resolved [10]. Here, we do not consider this algorithm because the thermodynamic approach developed further below demands additional investigation to be adapted for HDP.

In this work, to find the optimal number of topics, we follow an approach based on finding the minimum of free energy or the minimum of the Rényi entropy [11]. This approach assumes that it is possible to view a collection of documents and words as a mesoscopic informational statistical system (a complex system). This allows to formulate and compute Gibbs-Shannon entropy (S), as well as internal energy (E) and the Helmholtz free energy, of such a mesoscopic system:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\frac{\sum_{t=1}^{T}\sum_{n=1}^{N} P_{nt}}{T}\right) - T \cdot \ln\left(\frac{N_{k1}}{N \cdot T}\right) \qquad (1),$$

where $S(T) = -\ln\left(\frac{\sum_{t=1}^{T}\sum_{n=1}^{N} P_{nt}}{T}\right)$ is an internal energy, $\ln\left(\frac{N_{k1}}{N \cdot T}\right)$ is the Gibbs-Shannon entropy and Λ_F is a free energy. Following this, the number of topics (or clusters) is the informational system temperature; this parameter should be set by a user. The principle to guide the user's choice that is proposed in this approach is the search of the minimum of non-extensive entropy of the system.

Since the information measure is the entropy with opposite sign, the maximum entropy corresponds to the minimum of information. Thus, it is possible to reduce the search of the optimal topic number to the search of the minimum of the Rényi entropy expressed through free energy using escort distribution [12]:

$$S_{q=1/T}^{R} = \frac{F}{T-1},\ q = \frac{1}{T} \qquad (2).$$

Here, the $q = \frac{1}{T}$ value is viewed as a formal parameter (the number of topics/clusters) which is possible to change during a computational experiment. Thus, the search for the optimal topic number is reduced to varying topic/cluster number in TM and the search for the minimum Rényi entropy for each topic solution. It is necessary to note, that the presented approach assumes that the divergence of entropy could be achieved with $q = 1$. This means that the information of topic solution for just one topic is equal to zero. On the other hand, with $T \to \infty$ we have uniformly distributed probabilities of terms over topics which also corresponds to the maximum entropy or the minimum of information.

3 Selecting data

Following the general goal of the type of studies for which our pipeline is proposed, here we narrow this goal to the task of finding the most liked, most commented and most shared topics in the news stream generated by the leading Russian TV channels in the most popular Russian social networking site – VKontakte (VK). Television, until recently, was the most popular media channel in Russia and hence the object of the major concern for the government. It is through television (especially *Channel 1*) that the government has been used to disseminate most of its messages to the population. With the massive outflow of younger audiences to the Internet, however, the government has made special efforts to channel its controlled content via social media and to direct social media users to TV websites that host video content. Therefore, it is important to understand what content is thus channeled to online audiences and how those audiences react to it.

To account for this, we collected the data from the VK pages of eight leading state TV channels including *Russia Today* (*RT*) in Russian, one government news agency and three media outlets of varying degree of independence: oppositional online TV *Dozhd, Echo Moskvy* radio as "permitted opposition" and *RBC* business channel that attempts to be neutral (for full list see table 4). Other national media did not show any substantial presence in VK.

The data collection proceeded as follows. First, we collected news texts posted by the 12 chosen media channels on social network *VK* during the entire year of 2017. The resulting dataset consists of 164,426 Russian language posts and 185,029 unique words. Then, for each post we collected the following metadata: 1) number of likes for posts; 2) number of shares; 3) number of comments; 4) number of likes for comments; 5) date and time of publishing; 6) post URL. The dataset thus contains the total number of 111,626 comments, 527,147 likes and 121,073 shares. Finally, the news texts were cleared and lemmatized with MyStem lemmatizer, with stop-words being removed.

4 Finding the optimal topic number

To describe the topical structure of the news collection, we began with choosing the optimal number of topics. For this, we ran models with varying number of topics in the range $T = [2;\ 400]$ in the increments of two. During each iteration, we calculated the Gibbs-Shannon entropy, internal energy, free energy as defined in (1) and the Rényi entropy as defined in (2). Given the known instability problem of TM algorithm [13], at each iteration we performed three runs of our model with identical starting conditions and then averaged the values of all four aforementioned parameters. Fig.1 presents the curve of the Rényi entropy and shows that there are two minima corresponding to the maxima of information. The first minimum lies at 12 topics and the second one is at 146. Thus, we use two topic models, corresponding to the maximum information, for further analysis, and illustrate how to choose between them.

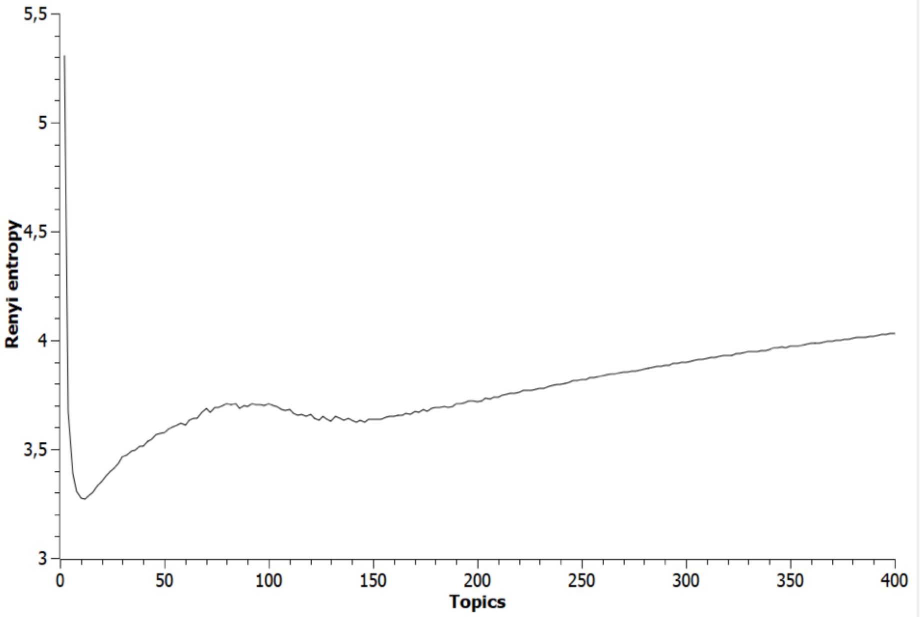

Fig. 1. The Rényi entropy of a topic model as a function of topic number.

4.1 Extracting stable topics

In topic modeling, there is a chance that a solution will contain topics impossible to reproduce even with the same data and algorithm parameters. For end users it means that they cannot make reliable judgments about topic composition of collections of their interest. At the moment, for unreproducible topics it is impossible to say whether their instability is explained by the inability of the algorithm to detect them or they are just algorithmic artifacts. It is thus not recommended to make any substantial conclusions based on the analysis of such topics. However, the topics that do get reproduced in each solution, can be analyzed and compared to each other, however, keeping in mind that the set of stable topics may be not a complete list of all the topics occuring in a given collection.

To select stable topics, we suggest to run TM several times with the same parameters, but not less than three, which is what we did here. Then, for each topic, we ranked words by probability and performed a pairwise comparison of each topic from every solution with the normalized Kulbak-Leibler measure [13]. Finally, we selected similar topics at the level of KL ≥ 90%. The use of such threshold for TM is justified in [14]. We applied the proposed algorithm to the two topic solutions (12 topics and 146 topics).

We found only five stable topics in the first solution, and 86 the second solution. For further analysis, we used only these stable topics.

4.2 Labeling stable topics

To interpret topics, it is necessary to label them, and a problem of choosing between the runs of the same solution arises. However, if only stable topics are taken into analysis, this choice can be made at random since topics that get reproduced in three and more runs, are virtually identical. Therefore, in this study we picked a random run, ranked matrices ϕ_{wt} and θ_{td} by probabilities and asked three independent coders to assign a label for each stable topic based on interpretation of its top words and documents. For labeling, we used up to 500 top words and documents (but usually much fewer, around 20). A standard problem with labeling TM results is a difficulty to calculate intercoder agreement since labels are usually open-end expressions. However, in the case of news, and especially in the subset of stable topics, topics are usually well pronounced and therefore labeled unanimously except a few “trash” topics that are hard to interpret. The judgment on whether similar labels are the same or different is easy to made, but it can be made by the consensus between coders, if necessary. Then intercoder agreement can be calculated which is what we did. Namely, we assessed it with Krippendorff's alpha for multiple raters. The coders achieved excellent overall agreement with alpha = 1 for five stable topics (12 topics solution) and 0.88 for 86 stable topics (146 topics solution).

5 Matching topics with user feedback data and interpreting results

5.1 Choosing the best solution

The global entropy minimum most often occurs at a relatively small number of topics (between 10 and 20), even for large collections with dozens of thousands of texts. It usually corresponds to solutions that yield very general topics. These solutions are suitable for getting a quick and most general understanding of a collection’s content if it is completely unknown. Oftentimes, the global minimum may indicate solutions with topics that are not most general, but are most lexically dissimilar thus covering only the most visible “top” of the collection’s topical structure (see table 1). The next best minimum, however, usually points at solutions that are of the highest analytical usefulness for researchers (compare tables 1 and 2). Subject-area expertise choice based on interpretability and analytic utility is an approach between solutions suggested by the founders of topic modeling Blei and Lafferty [15], however, prior finding of global and local minima helps to reduce the number of alternatives to just two or three. All this suggests that a TM user should not blindly follow the global entropy minimum. Instead, the better practice would be to examine all entropy minima for relevance to the research goals before committing to the analysis. Further on, we use a 146-topic solution.

Table 1. Likes, reposts, comments and comments likes distributions for stable topics of the 12 topics solution (estimated on ~500 most probable documents with no missing metadata).

Stable topic	N likes	N reposts	N comments	N comments likes
Russian sport achievements	106234	3800	16134	23328
Money & Russian markets analysis	61584	9032	17180	22952
Russian culture	71830	5623	6756	14630
Science news	110953	8594	24526	25605
American movie culture	93225	5537	20403	32286

5.2 Assessing relation of user feedback to topics

As each text contains all topics in different proportions, and this proportions decrease rapidly from the top text in a given topic to its bottom text, it is not easy to match topics to user feedback. Users attach their likes or comments to texts, not topics, whose representation in text may be high or low. One way suggested in [16] is to multiply the number of likes by the probability of a given topic in a given text and then to summarize the obtained values across all texts. A limitation of this approach is that the long tail of low probabilities may in the end outweigh the influence of the small number of texts in which a topic is best pronounced. Another approach is to use only top N texts for calculating the "likability" or "sharebility" of a topic. A limitation of this approach is that it ignores the degrees to which each text belongs to a given topic. And a third approach is to combine both. Here, we use the second approach as an example selecting 500 top texts. The average probability of all topics in these texts is 0.27 which is 38 times higher than random (1/146). This allows us to ignore differences in probabilities among those 500 texts.

Table 2 shows top ten stable topics from the 146-topic solution ranked by the numbers of likes, reposts, comments and likes to comments. Top topics shared across all types of feedback are topics of *Mixture of controversial events* as well as coverage of *Russian Orthodox church* and various Christian celebrations. *Mixture of controversial events* unites a number of most resonant scandals that burst out during the year, so it is well understandable why it is the leader in all aspects of feedback.

Table 2. Top 10 stable topics from the 146 topics solution ranked by the number of likes, reposts, comments and likes to comments (estimated on ~500 most probable documents with no missing metadata).

Rank	By likes	By reposts	By comments	By likes in comments
1	Mixture of controversial events	Mixture of controversial events	Mixture of controversial events	Mixture of controversial events

2	Russian sport achievements	WW2 commemoration	2018 Russian presidential campaign	2018 Russian presidential campaign
3	WW2 commemoration	Navalny-Usmanov controversy	'Matilda' movie controversy	Russian athletes doping controversy
4	Sport: hockey	Rulemaking	Russian opinion polls	Finance, pension fund and the Finance Ministry
5	Sport: figure skating	Russian opinion polls	Street actions & protests (international)	Russian opinion polls
6	Food & recipes	Finance, pension fund and the Finance Ministry	Finance, pension fund and the Finance Ministry	'Matilda' movie controversy
7	Russian Orthodox Church	FSB and counterterrorist activities	WW2 history related events (international)	Street actions & protests (international)
8	FSB and counterterrorist activities	Russian Orthodox Church	Russian Orthodox Church	Putin & his addresses
9	Syria & Russia	Food & recipes	Russian athletes doping controversy	Russian Orthodox Church
10	Russian navy	Astronomy & NASA news	Ukraine & separatist proto-states	FSB and counterterrorist activities

While most liked topics are related to *Russian sport achievements* (including hockey and figure skating) as well as Russian military campaign in Syria (*Syria & Russia*) and other military advances, the most shared topics are about corruption (*Navalny-Usmanov controversy*), *Rulemaking* and Russian finance news (*Finance, pension fund and the Finance Ministry*). Most commented topics include *Ukraine & separatist proto-states* as well as *Russian athletes doping controversy* and *Street actions and Protests*. It thus can be seen that high “likability” is most likely to be generated by topics that can produce and maintain national pride. High sharing levels are likely to be produced by social problems and some practically useful topics (recipes), and high numbers of comments correspond to the largest number of sharp conflicts. Large numbers of likes to comments, however, are caused by topics other than those that produce the highest likability of news themselves. As said in the introduction, large number of comments usually indicate hot, often polarized discussions (which is confirmed by the nature of the detected topics that generate them), and thus likes born in the course of such discussions indicate solidarity with the parties of the discussion rather than satisfaction with news.

From this, it is clear that likes to comments should not be used together with likes to news items as the indicators of audience's satisfaction with news content.

5.3 Assessing topical compositions of collection subsets and feedback received by them: topicality of TV channels

The discussed above results reflect only the cumulative effect of all 12 TV channels combined. However, both topic composition and user feedback related to different topics could significantly vary by channel. For instance, a given topic might be most represented in only two channels, but one of them might still gain more likes than the other. To account for this, we, first, calculated the proportion of texts from each channel among 500 top texts of all stable topics. Second, we calculated the proportion of likes and shares received by the texts of each given channel from among all likes and shares received by 500 top texts of each stable topic, respectively. This approach may also be easily transformed into the approach offered in [16] by multiplying the number of likes and shares by the probability of a given topic in a text.

Table 3. Contribution of TV channels into topic salience, "likability" and "sharebility"

Stable topic	Channel weight in a topic, %	Channel likes, %	Channel reposts, %
Mixture of controversial events	RIA News – 98%	RIA News – 99.54%	RIA News – 99.80%
Russian sport achievements	Russia Today – 41.29%	Russia Today – 31.90%	Russia Today – 26.75%
	Russia-24 – 17.43%	Russia-24 – 8.76%	Russia-24 – 10.96%
	RIA News – 16.6%	RIA News – 47.03%	RIA News – 42.19%
	NTV – 12.45%	NTV – 2.86%	NTV – 4.61%
Syria & Russia	RIA News – 33.68%	RIA News – 54.46%	RIA News – 44.63%
	Russia Today – 30.58%	Russia Today – 33.42%	Russia Today – 36.60%
	Russia-24 – 13.00%	Russia-24 – 4.93%	Russia-24 – 7.05%
	NTV – 10.95%	NTV – 2.40%	NTV – 4.36%
Russian athletes doping controversy	Russia Today – 30.50%	Russia Today – 24.66%	Russia Today – 23.42%
	RBC – 17.3%	RBC – 17.13%	RBC – 20.15%
	RIA News – 15.87%	RIA News – 37.75%	RIA News – 30.52%
	NTV – 15.67%	NTV – 5.83%	NTV – 7.10%
Russian Orthodox Church	NTV – 21.49%	NTV – 5.93%	NTV – 8.37%
	RIA News – 20.66%	RIA News – 51.86%	RIA News – 39.25%

	Russia-24 – 16.53% Dozhd – 10.95%	Russia-24 – 7.89% Dozhd – 2.03%	Russia-24 – 8.93% Dozhd – 2.61%
'Matilda' movie controversy	Dozhd – 21.44% Russia Today – 19.17% RBC – 18.76% RIA News – 15.00%	Dozhd – 13.88% Russia Today – 16.18% RBC – 27.73% RIA News – 34.85%	Dozhd – 17.02% Russia Today – 14.80% RBC – 34.26% RIA News – 23.89%
Street actions & protests (international)	Dozhd – 25.67% RBC – 22.77% RIA News – 20.08% NTV – 13.04%	Dozhd – 12.19% RBC – 40.83% RIA News – 33.49% NTV – 2.08%	Dozhd – 15.38% RBC – 47.90% RIA News – 19.33% NTV – 2.89%
Putin & his addresses	NTV – 21.85% Russia Today – 21.65% RBC – 14.64% RIA News – 12.57%	NTV – 5.66% Russia Today – 22.88% RBC – 13.54% RIA News – 48.68%	NTV – 8.73% Russia Today – 19.08% RBC – 23.44% RIA News – 35.62%
FSB and counter-terrorist activities	NTV – 23.81% RIA News – 20.08% Russia-24 – 13.25% Russia Today – 12.42%	NTV – 6.78% RIA News – 57.47% Russia-24 – 7.45% Russia Today – 12.33%	NTV – 9.05% RIA News – 40.94% Russia-24 – 10.62% Russia Today – 14.10%

From table 3 we can see that the dominant role in topic composition as well as in total topic likability and sharebility belongs to just four TV channels. Namely, state-controlled channels *Russia Today*, *RIA News, Russia 24* and *NTV* dominate the landscape and produce thematically similar content that forms a hegemonic discourse [17]. Conventionally neutral *RBC* closely follows the leaders by its presence in the most liked and shared topics. The only oppositional TV channel, *Dozhd* is also present, and it predictably emerges within some of the most conflict topics *Street Actions & Protests* and *Matilda movie controversy*. The movie was heavily criticized by the officials, banned from display and then reinstated.

While difference between the mainstream and the oppositional agendas in quite predictable, the apparent domination of just a few channels in our topic model deserves special consideration. It persists across all stable topics when measured both as the proportion of texts and as salience. This effect can be explained via examination of the document and word distributions in the data. The largest influence on TM results is exerted by sources with (1) relatively large number of messages, and (2) with relatively longer texts. As can be seen from table 4, each channel on average posts 200-300 word long messages. Because the four dominant sources have more documents and words, they accumulate higher topic probabilities (see table 4).

This effect thus should be taken into consideration by researchers. One way to deal with it is to obtain more balanced samples. However, as sampling may distort true topical structures, another strategy can be chosen. Topic saliences, as suggested elsewhere [16], may be calculated as the sums of probabilities of each topic over all texts, or, as we would suggest, over N most probable texts in each topic. When aggregated by source (e.g. TV channel), topics, on average, will tend to show higher saliences in overrepresented channels. Therefore, to compare topic saliences across channels, they can be normalized based on the distribution of text quantities and lengths among channels.

Table 4. General distribution of posts, likes, reposts and subscribers in dataset

Chanel	N of messages	N of comments	Posts likes	N of reposts	N of subscribers
Russia Today	96440	7491681	13405188	869569	1083472
RIA News	28947	3562463	13024009	706123	2149674
NTV	28298	466976	1583383	160728	313140
Russia-24	22300	553084	742878	1597518	142129
Dozhd	19300	557997	1214643	108189	425127
Channel 5	18097	137680	500796	53522	90607
Russia-1	17640	127940	820069	113462	78492
RBC	10800	178627	2111679	236643	631525
Channel 1	5700	346299	3758479	322667	1740665
Mir-24	5500	4412	63197	11237	24700
TVC (News)	4400	2994	29811	4053	7895
Russia-Culture	2900	6363	138730	25336	35709

6 Conclusion

In this paper, we presented a full-cycle methodology for news topic modeling and user feedback research. This methodology offers a series of steps to overcome various latent limitations of topic modeling and, above all, mitigates the problem of topic number choice. Our solution to this problem is based on the search for the minimum of the Rényi entropy. Furthermore, we formulated an approach for stable topic extraction based on the normalized Kullback-Leibler divergence. Additionally, we illustrated the proposed research pipeline with an analysis of a one-year online stream formed by 12 national TV channels and broadcasted online via VK social network (12 and 146 topics models). We demonstrated that audience feedback varies depending on news topics and showed how this and other effects can be captured with our approach. Positive topics like athletic achievements receive more likes and are discussed significantly less; more discussed topics, on the other hand, are more related to problems and conflicts and are generally receive less likes. Finally, using metadata of online news items, we showed the overrepresentation effect of the leading TV channels in topic models. This finding opens a new question about data normalization in topic modeling.

Acknowledgement

The reported study was funded by RFBR according to the research project № 18-011-00997 A.

References

1. Daud A, Li J, Zhou L, Muhammad F (2010) Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China 4:280–301. doi: 10.1007/s11704-009-0062-y
2. Ziegele M, Breiner T, Quiring O (2014) An Exploratory Analysis of Discussion Factors in User Comments on News Items: What Creates Interactivity in Online News Discussions? Journal of Communication 64:1111–1138. doi: 10.1111/jcom.12123
3. Bobkowski PS (2015) Sharing the News: Effects of Informational Utility and Opinion Leadership on Online News Sharing. Journalism & Mass Communication Quarterly 92:320–345. doi: 10.1177/1077699015573194
4. Landau DP, Binder R (2009) A guide to Monte Carlo simulations in statistical physics. Cambridge University Press.
5. Griffiths TL, Steyvers M (2004) Finding scientific topics. PNAS 101:5228–5235. doi: 10.1073/pnas.0307752101
6. Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive LDA model selection. Neurocomputing 72:1775–1781. doi: https://doi.org/10.1016/j.neucom.2008.06.011
7. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN (2010) On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V (eds) Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 391–402
8. Roberts ME, Stewart BM, Tingley D (2016) Navigating the Local Modes of Big Data: The Case of Topic Models. In: Alvarez RM (ed) Computational Social Science. Cambridge University Press, Cambridge, pp 51–97
9. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101: 1566–1581
10. Blei DM, Griffiths TL, Jordan MI, Tenebaum JB (2004) Hierarchical Topic Models and the Nested Chinese Restaurant Process. Advances in Neural Information Processing Systems 16: 17—24, doi: 10.1016/0169-023X(89)90004-9
11. Koltcov SN (2017) A thermodynamic approach to selecting a number of clusters based on topic modeling. Technical Physics Letters 43:584–586. doi: 10.1134/S1063785017060207.
12. Beck C, Schlögl F (1993) Thermodynamics of Chaotic Systems. Cambridge University Press.

13. Koltcov S, Nikolenko SI, Koltsova O, et al (2016) Stable Topic Modeling with Local Density Regularization. In: Bagnoli F, Satsiou A, Stavrakakis I, et al. (eds) Internet Science. Springer International Publishing, Cham, pp 176–188
14. Koltcov S, Koltsova O, Nikolenko S (2014) Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content. In: Proceedings of the 2014 ACM Conference on Web Science. ACM, Bloomington, Indiana, USA, pp 161–165
15. Blei DM, Lafferty JD (2009) Topic models. In: Text Mining: Classification, Clustering, and Applications. CRC press, p 71—94.
16. Nagornyy O and Koltsova O (2017) Mining Media Topics Perceived as Social Problems by Online Audiences: Use of a Data Mining Approach in Sociology. Higher School of Economics Research Paper No. WP BRP 74/SOC/2017. Available at SSRN: https://ssrn.com/abstract=2968359 or http://dx.doi.org/10.2139/ssrn.2968359
17. Prozorov S (2005) Russian conservatism in the Putin presidency: The dispersion of a hegemonic discourse. Journal of Political Ideologies 10:121–143. doi: 10.1080/13569310500097224