DOI: 10.14515/monitoring.2021.6.2044

V. V. Danilova, S. V. Popova, V. M. Karpova

A PIPELINE FOR GRAPH-BASED MONITORING OF THE CHANGES IN THE INFORMATION SPACE OF RUSSIAN SOCIAL MEDIA DURING THE LOCKDOWN

For citation:

Danilova V. V., Popova S. V., Karpova V. M. (2021) A Pipeline for Graph-Based Monitoring of the Changes in the Information Space of Russian Social Media during the Lockdown. *Monitoring of Public Opinion: Economic and Social Changes.* No. 6. P. 254–295. https://doi.org/10.14515/monitoring.2021.6.2044.

Правильная ссылка на статью:

Данилова В. В., Попова С. В., Карпова В. М. Пайплайн для графического мониторинга динамики информационного пространства российских социальных сетей в период карантина // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 6. С. 254—295. https://doi.org/10.14515/monitoring.2021.6.2044. (In Eng.)

DOI: 10.14515/monitoring.2021.6.2044

A PIPELINE FOR GRAPH-BASED MONITORING OF THE CHANGES IN THE INFORMATION SPACE OF RUSSIAN SOCIAL MEDIA DURING THE LOCKDOWN

Vera V. DANILOVA[1] — Cand. Sci. (Applied Linguistics), Senior Research Fellow, International Laboratory for Social Media Computational Studies
E-MAIL: maolve@gmail.com
https://orcid.org/0000-0003-0868-522X

Svetlana V. POPOVA[2,3] — Senior Lecturer, Faculty of Applied Mathematics and Control Processes; PhD candidate
E-MAIL: spbu.svp@gmail.com
https://orcid.org/0000-0001-5827-984X

Vera M. KARPOVA[4] — Cand. Sci. (Sociology), Senior Lecturer, Faculty of Sociology
E-MAIL: Wmkarpova@yandex.ru
https://orcid.org/0000-0003-2560-6140

[1] Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia
[2] Saint Petersburg State University, Saint Petersburg, Russia
[3] Technological University Dublin, Dublin, Ireland
[4] Moscow State University, Moscow, Russia

Abstract. With the COVID-19 outbreak and the subsequent lockdown, social media became a vital communication tool. The sudden outburst of online activity influenced information spread and consumption patterns. It increases the relevance of studying the dynamics of

ПАЙПЛАЙН ДЛЯ ГРАФИЧЕСКОГО МОНИТОРИНГА ДИНАМИКИ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА РОССИЙСКИХ СОЦИАЛЬНЫХ СЕТЕЙ В ПЕРИОД КАРАНТИНА

ДАНИЛОВА Вера Владимировна — кандидат филологических наук, старший научный сотрудник, Международная лаборатория математических методов исследования социальных сетей, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Москва, Россия
E-MAIL: maolve@gmail.com
https://orcid.org/0000-0003-0868-522X

ПОПОВА Светлана Владимировна — старший преподаватель, факультет прикладной математики — процессов управления, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия; аспирант, Технологический университет Дублина, Дублин, Ирландия
E-MAIL: spbu.svp@gmail.com
https://orcid.org/0000-0001-5827-984X

КАРПОВА Вера Михайловна — кандидат социологических наук, старший преподаватель, социологический факультет, Московский государственный университет, Москва, Россия
E-MAIL: Wmkarpova@yandex.ru
https://orcid.org/0000-0003-2560-6140

Аннотация. Со вспышкой COVID-19 и последующей всеобщей изоляцией социальные сети стали жизненно важным инструментом коммуникации. Внезапный всплеск онлайн-активности повлиял на распространение информации и модели ее потребления. Это

DOI: 10.14515/monitoring.2021.6.2044

social networks and developing data processing pipelines that allow a comprehensive analysis of social media data in the temporal dimension. This paper scopes the weekly dynamics of the information space represented by Russian social media (Twitter and LiveJournal) during a critical period (massive COVID-19 outbreak and first governmental measures). The approach is twofold: 1) build the time series of topic similarity indicators by identifying COVID-related topics in each week and measuring user contribution to the topic space, and 2) cluster user activity and display user-topic relationships on graphs in a dashboard application. The paper describes the development of the pipeline, explains the choices made and provides a case study of the adaptation to virus control measures. The results confirm that social processes and behavior in response to pandemic-triggered changes can be successfully traced in social media. Moreover, the adaptation trends revealed by psychological and sociological studies are reflected in our data and can be explored using the proposed method.

Keywords: COVID-19, topic modeling, topic dynamics, time series, user activity monitoring, LDA, feature elaboration, content analysis, social graphs

повышает актуальность изучения динамики различных тем в социальных сетях и разработки пайплайнов для обработки данных, позволяющих проводить всесторонний анализ информации из социальных сетей во временном измерении. В статье рассматривается понедельная динамика информационного пространства, представленного российским сегментом социальных сетей (русскоязычные Twitter и LiveJournal), в критический период (массовая вспышка COVID-19 и первые правительственные меры). Авторский подход состоит из двух частей: 1) построение временных рядов индикаторов сходства тем путем выявления тех из них, которые связаны с COVID-19, за каждую неделю и измерение вклада пользователей в тематическое пространство, 2) кластеризация активности пользователей и отображение взаимосвязей между ними и темами на дашборде. В статье описывается разработка пайплайна, объясняются принятые решения и приводится тематическое исследование адаптации к мерам борьбы с коронавирусом. Результаты подтверждают, что социальные процессы и общественный ответ на изменения, вызванные пандемией, можно успешно отслеживать в социальных сетях. Более того, тенденции адаптации, выявленные психологическими и социологическими исследованиями, отражены в наших данных, а следовательно, могут изучаться с помощью предлагаемого метода.

Ключевые слова: COVID-19, тематическое моделирование, тематическая динамика, временные ряды, мониторинг активности пользователей, LDA, разработка признаков, анализ контента, социальные графы

Introduction

The 2020 lockdown resulting from the outbreak of COVID-19 led to significant and sudden changes in people's lives and attitudes. In this period, social media (SM) has become a vital source of connection between people and an indispensable tool employed by governments, universities, organizations and others for information production, spread, exchange and consumption. At the same time, the informational overload went ahead feeding the online infodemic including over-interpreting, personal views, rumors, fake news, propaganda, and misinformation that affected users' well-being and behavior and threatened the effectiveness of public health measures[1] [Cinelli et al., 2020]. Being the main producer of the image of pandemic-related processes and events, SM undoubtedly plays a key role in their perceptions and the potential consequences thereof [Tsao et al., 2021; Al-Dmour et al., 2020].

Researchers in different fields recurred to SM data to help the prevention and treatment procedures, as well as to explore, understand and predict the changes caused by the onset of COVID-19 and associated events [Chakraborty et al., 2020; Tsao et al., 2021]. Text analysis tools are applied in this context, among others, to study and compare user activity across platforms, model the information and infodemic spread, find sources that are susceptible to misinformation [Cinelli et al., 2020], find correlations between the increase of new infection cases and public attention peaks [Hou, Hou, Cai, 2021], detection and prediction of outbreaks [Jordan et al., 2019].

The goal of the present research is the study of the impact of the pandemic on the online information environment in the Russian-language SM. It complements previous research on the analysis of changes in online information space around COVID-19 conducted for (mainly English-language) SM platforms including Twitter, Instagram, YouTube [Tsao et al., 2021; Cinelli et al., 2020]. The relevance of this study derives from its key aims to understand the issues faced by people in the online environment where information is growing and spreading uncontrollably. The investigation of information spread is particularly important during critical moments, self-isolation, and lockdown in this case study.

This article describes collection, processing and visualization of data reflecting dynamic topic-user relationships for the exploration of the changes in the Russian information space from the social science perspective. The study collects Russian-language data from two SM platforms: the LiveJournal (LJ) community hosting and Twitter microblogging service. The study period starts several days before the first Russian President's Address to the Nation on the coronavirus pandemic (March 25, 2020) and covers the major lockdown (until June 1, 2020).

On the one hand, we study the process of discussion formation and evolution in SM during the first-ever similar unusual period: self-isolation regime was introduced in almost all regions; a non-working month was announced; social, economic and a part of political activity (the vote on the amendments to the Constitution of Russian Federation) stopped and moved online. The sudden and drastic changes undoubtedly

[1] Managing the COVID-19 Infodemic: Promoting Healthy Behaviors and Mitigating the Harm from Misinformation and Disinformation (2020) *World Health Organization*. 23 September. URL: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation (accessed: 01.12.2021).

DOI: 10.14515/monitoring.2021.6.2044

influenced the information spread in SM. The milestone events that took place in that period including the Addresses to the Nation by the President, the self-isolation decrees, and the restrictive measures introduced by Moscow authorities allow tracing the response of the online community. On the other hand, this study complements and expands similar research based on other online platforms providing new data to analyze information spread patterns in SM.

In the COVID-19-related studies of SM in the previous work (overviewed in the Related Work section), the use of topic modeling (mostly Latent Dirichlet Allocation (LDA)-based, see [Blei, Ng, Jordan, 2003]) is mostly limited to getting a general view of themes in the dataset. A few works explore topic dynamics within a certain period. However, they do not consider the changes that take place in the user layer. The user dimension allows representing the size of a topic in a given time unit using not only the ratio of topic-related texts but also the ratio of contributors. Different ways of grouping the contributors (by activity type, by dispersion) provide new insights into the study of topic dynamics and cross-network comparison of content generation behavior. Also, to our knowledge, there is a lack of studies of public attitudes in Russian SM during the lockdown.

To trace topic dynamics in connection with the user dimension and explore the changes in content generation behavior related to COVID-19 in SM within the study period, the following pipeline is used (see fig. 1). The week is adopted as the time unit. One branch is responsible for retrieving document topics, constructing the topic hierarchy and weekly topic representations, and the other clusters account activity patterns across the weeks. The topic modeling (TM) branch solves the following main tasks:

(1) for document *d* from the subset D_{wn} published during week *w* in network *n*, identify the prevalent topic[2] *t* from the subset of LDA-produced topics T_{wn} related to COVID-19 and active during *w* in *n*,

(2) trace similar topics across all T_{wn} subsets across time units in each *n*,

(3) consolidate similar topics into larger themes and build the set *T* of unique topics across *w* and *n* to facilitate cross-network topic dynamics comparison.

This branch produces 1) time series of topic dynamics indicators and 2) data for graphs. Latent Dirichlet Allocation (LDA) with Gibbs sampling [Mimno et al., 2011] is used as the base tool for topic identification. The influence of parameter tuning on the resulting topics is explored, and the optimal settings are determined based on comparative analysis and literature overview.

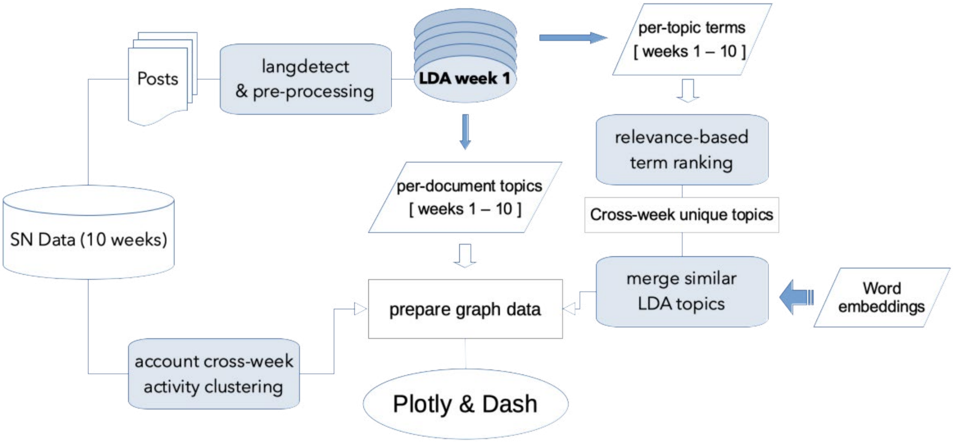
The other branch clusters all contributors by activity pattern. The graph data preparation step uses the output of both branches and additionally groups users by dispersion (by the number of prevalent topics in their posts). The resulting dashboard application allows producing graphs for each week, network and activity cluster by changing the corresponding settings.

The analysis based on this tool explores patterns, if any, in the dynamics of content generation behavior across the weeks based on the obtained graphs and time series of topic dynamics. The collected data is expected to provide insights on whether the changes in the information space for both platforms are similar, what patterns are

[2] Sets of words or phrases that tend to co-occur in texts and are associated with a certain subject, event or knowledge area.

shared, or, if not, what are their distinctive features. Moreover, the gathered data is aimed to provide a better understanding of the framing of discussions around topics related to COVID-19 lockdown.

Fig. 1. Graph generation workflow



The case study (based on the Twitter data) aims to trace personal adaptation to the sudden changes and restrictive measures in SM and compare the observed pattern with the results of the available psychological studies of the adaptation during the lockdown.

Related Work

The present literature overview summarizes, on the one hand, text mining and, specifically, topic modeling techniques that were used for SM processing in the COVID-19-related research, and, on the other hand, studies that explored the impact of the pandemic on the population.

Text mining of SM in COVID-19-related research

The 2019—2020 events resulted in an outburst of COVID-19-related research. An extensive bibliometric analysis of officially published and indexed reports on COVID-19 collected on October 14, 2020 [Wang, Tian, 2021] shows a large number of contributions in healthcare, biology, medicine, and epidemiology. Fewer contributions are reported in transmission (disease transmission route), psychology, and even fewer in other research directions including social impact and social science. The applications of SM data to COVID-19-related research are particularly discussed in an article by Tsao et al. [2021]. This survey summarizes 2,405 peer-reviewed studies for the period from November 2019 to November 2020. It reports the use of SM mining mainly to investigate problems in psychology (mental health assessment), healthcare (detection and prediction of infection cases, evaluation of health information in prevention education videos), and social science (monitoring of public attitudes, analysis of government responses to pandemic). Also, another important research issue concerns the mechanisms of infodemics spread detection on SM. Twitter and Sina Weibo (a Chinese microblogging website) are the most frequently studied platforms [ibidem].

TM techniques (mainly LDA) are used to analyze public attitudes and concerns in English-language SM [Boon-Itt, Skunkan, 2020; Jelodar et al., 2020; Kurten, Beullens, 2021; Xue et al., 2020; Wicke, Bolognesi, 2020]. Some works combine LDA with sentiment analysis tools to measure public sentiment towards particular topics [Das, Dutta, 2020: 163; Abd-Alrazaq et al., 2020]. Although in studies by Boon-Itt and Skunkan [2020] as well as Kurten and Beullens [2021], LDA is used to obtain a general view of the discussion topics within the studied period, topic dynamics is not covered. Mainly manual analysis of per-topic per-word distributions is performed. Boon-Itt and Skunkan [2020] group LDA topics into larger themes by means of qualitative content analysis. These works employ sentiment analysis tools separately from TM, therefore, user sentiments towards the topics cannot be traced. Xue et al. [2020] apply LDA on a large dataset of English-language tweets to trace public discourse on family violence. As in Boon-Itt and Skunkan [2020], similar topics (fifty in total) are manually grouped into nine larger themes followed by a detailed interpretation. Wicke and Bolognesi [2020] explore both less and more granular LDA topics in the dataset by setting the number of *k* to 4 and 16, respectively. The topics are interpreted and summarized based on the manual analysis of lexical units. They are further used to check and analyze the presence of figurative frames, such as WAR, and the conventional metaphor DISEASE TREATMENT IS WAR in COVID-19-related themes on Twitter in March and April 2020. The above studies do not consider the temporal and user dimensions.

Medford et al. [2020] consider topics and sentiments separately, however, they are further grouped with respect to the percentage of top retweeted tweets to show the prevalent sentiment and top-3 topics in each group. In research by Das and Dutta [2020], two LDA models are built for two parts of the dataset previously classified by sentiment using the R-based software package “sentiment” and the NRC Emotion Lexicon. Word cloud visualization of twenty topics from each sentiment group is made to provide insights into the contexts that are characteristic of each group. The resulting topic coverage trends for both sentiments turned out to be very similar. Abd-Alrazaq et al. [2020] first perform TM of tweets followed by sentiment analysis (Python textblob library) and interaction rate calculation for each of the topics based on the number of retweets, likes, and followers. The number of topics that represent people’s concerns in Twitter is selected manually based on the LDA output and n-gram clouds examination.

Some of the overviewed studies conducted time series analysis. Das and Dutta [2020], Boon-Itt and Skunkan [2020] explore the dynamics of sentiment indicators. In Das and Dutta [2020: 158], daily dynamics of sentiment in tweets tagged with \#IndiaLockdown and \#IndiafightsCorona are analyzed for the period from March 22 to April 21, 2020. Here, the daily sentiment is measured using both the whole corpus and individual tweets. Boon-Itt and Skunkan [2020] analyze the changes in the number of retweets and likes in a corpus of 107,990 English tweets related to COVID-19 between December 13, 2019, and March 9, 2020. The peaks are aligned with the events that were reported in the news media in the corresponding period. Kurten and Beullens [2021] investigates daily changes in the number of English, Dutch, and French tweets and retweets posted in Belgium within the period from February 25 to March 30, 2020 and associates the peaks with the events that took place in this country in the corresponding period.

The work by Gozzi et al. [2020] is closer to ours in that it explores topic dynamics. The base 64 topics are extracted from a corpus of news outlets published from February 7 to May 15, 2020. Temporal changes in the attention of Reddit users to the 64 news topics are measured by tracing the corresponding topics in Reddit comments using LDA for the period from February 15 to May 15, 2020. The authors notice that the view of Reddit topics is limited since only news-triggered discussions are covered. Medford et al. [2020] visualize LDA-based topic dynamics from January 14 to January 27, 2020 using a t-distributed Stochastic Neighbor Embedding (t-SNE) graph.

Following the mentioned researchers, we use LDA for topic modeling in our task. Expanding the overviewed works that predominantly process English texts and focus on the dynamics of keywords, retweets, and sentiment [Das, Dutta, 2020; Boon-Itt, Skunkan, 2020], our research, being closer to Medford et al. [2020] and Gozzi et al. [2020], collects features to explore the temporal changes in two dimensions of topics and users in Russian SM. It allows exploring the content generation behavior within the critical period of strict lockdown and self-isolation and performing the cross-network comparison.

Research on the impact of the COVID-19 pandemic on the population

Nowadays, there is a growing body of research that studies the social and psychological effects of the COVID-19 pandemic on the population and people's adaptability to a new way of life [Xiong et al., 2020; Chu et al., 2020]. Many psychological studies conducted during the pandemic, and primarily during the lockdown, explore the changes in the psychological state during the confinement starting from the establishment of lockdown measures and in 1—2 months. The results show that within the first 1—2 weeks anxiety levels rise sharply and then over the next 4—6 weeks they gradually decrease to return to their original level [Daly, Robinson, 2021: 606].

In general, the psychological and social consequences of the COVID-19 pandemic and restrictive measures introduced around the world are subject to further research. However, the first large-scale literature surveys show that the consequences for well-being and mental health may be quite severe. So, in their overview Clemente-Suárez et al. [2020] mention such manifestations as anxiety, panic attacks, depression, signs of PTSD, and even suicidality. These negative consequences are aggravated by social isolation, a forced decrease in physical activity and social contacts, as well as grief after the death of friends, spouses, or relatives.

As noted in the study by Ruggieri et al. [2021], one of the effective ways of coping with the experience of restrictions caused by the lockdown is online communication. The results of some studies confirmed that online communication and comparing one's own experiences of social isolation with other people reduces negative psychological consequences of quarantine [ibidem]. Thus, the analysis of SM posts can be a valid way to examine the coping strategies with respect to the effects of the COVID-19 pandemic and lockdown measures since this analysis is based on the product of these experiences. This type of interaction in the absence of live communication generated

DOI: 10.14515/monitoring.2021.6.2044

a collective experience of social upheavals associated with the imposed restrictions and alarming news.

There is a large-scale study currently being conducted in the US[3] on a similar topic (coping during COVID-19). One of its methodological tasks is to combine the greatest number of qualitative methods and approaches to deal with the great diversity of changes observed in the public sphere during the pandemic and lockdown. However, many of the difficulties faced by this kind of research during textual data processing can be solved automatically. For instance, one of the important parameters is the frequency of polls. In the text processing, the analysis can be carried out without interruption and the duration of analysis stages can be set to any value from one day to one week or even a month. Longitudinal surveys are particularly challenging because they are conducted multiple times. In SM mining, the total duration of the study will be primarily limited by computational performance and the researcher's decisions.

Another methodological advantage of this study is the generation of a large number of questionnaires for different topics, which enables tracing specific short-term (local) topics that are discussed for one to three weeks and then dropped. Local topics represent issues that are relevant for the audience during a short time period. In this setting it is impossible to launch a survey with a tailored questionnaire or interview, however, we can perform text mining and identify/trace the corresponding topics.

The alignment of SM content with the main news feed for the corresponding period shows, on the one hand, that there are news stories that were definitely mirrored in SM (according to our examination of the first three weeks, the peak of discussions occurs 2—3 days after the event). On the other hand, if some news is not reflected in the SM, it means that it did not cause a significant response from the population.

Another important issue is the spread of false news (fake news) and different manifestations of conspiracy theory [Tagliabue, Galassi, Mariani, 2020] that negatively affect the audience by increasing the levels of anxiety and stress. The identification of news content on SM contributes to the detection of those fake news that gets the greatest response from the audience and must be blocked in the first place.

Methodology of Feature Generation

This section discusses the approaches to the generation of topic- and user-specific features. It starts with a brief description of LDA as the base feature extraction technique explaining the choice of parameter settings and tools.

Latent Dirichlet Allocation

This study employs a conventional topic modelling scheme based on Latent Dirichlet Allocation (LDA) [Blei, Ng, Jordan, 2003]. As indicated by Hagen [2018], properly trained and evaluated LDA-based topic models are a powerful tool for content analysis in social science that helps discover themes overlooked by human coders and are less bias-prone. LDA is a particularly popular parametric approach that models documents as mixtures of topics and topics as mixtures of words (probabilistic distributions over words).

[3] Rusch T., Han Y., Liang D., Hopkins A., Lawrence C., Maoz U., Paul L. K., Stanley, D. A. (2021) COVID-Dynamic: A Large-Scale Multifaceted Longitudinal Study of Socioemotional and Behavioral Change across the Pandemic. *PsyArXiv Preprints*. URL: https://psyarxiv.com/75eyx/ (accessed: 21.12.2021).

An overview of recent papers reporting the application of two widely used LDA-based packages, namely Java-based Mallet[4] [Zhou, Awasthi, Cardinal, 2020; Fang, Partovi, 2021; Cho, Park, Song, 2020] and Python-based Gensim[5] [Porter, 2018; Kastrati, Kurti, Imran, 2020; Riesener et al., 2019], as well as extensive experiments with both packages on the 1st-week data (LJ) followed by the analysis of LDAvis [Sievert, Shirley, 2014] output, we settled on the use of the Mallet package [Mimno et al., 2011]. Ebeid and Arango[6] compare both tools and point out that both have their strengths and weaknesses. Mallet's underlying approach relies on Gibbs sampling, which has well-known implications for the runtime complexity [Jelodar et al., 2020: 2736] because the training process requires keeping the entire dataset in memory. On the other hand, as shown by Zhou, Awasthi and Cardinal [2021], Mallet performs better than Gensim from the perspective of the coherence value. Roughly, coherence reflects the degree of mutual support between subsets (word sets) within each topic in a topic model. "C_v" coherence used in this paper combines the indirect cosine measure with the NPMI (Normalized Pointwise Mutual Information) and the boolean sliding window and it is reported to be the best measure in terms of runtime and correlation to human ratings [Röder, Both, Hinneburg, 2015].

Dataset pre-processing for LDA includes lemmatization with PyMystem3[7], Russian "stopwords" removal using NLTK[8] and bag-of-words representation using Gensim libraries.

The best LDA setup is found as follows. Following Wallach, Mimno and Mccallum [2009] we use asymmetric alpha (prior for topic proportions within documents), which combined with symmetric beta (prior for word weights in topic distributions) proved to enhance the quality of topic models. In Mallet, alpha can be optimized for each *N* iterations using the "optimize_interval" parameter equal to *N*. In the field, it was observed that although frequent optimization increases coherence, it influences topic quality due to the growing prevalence of topics with small coverage (topics that are present in few documents)[9]. Since we aim to capture the most prominent topics, less frequent optimization is given a priority. The optimal number of topics for each week is identified by maximizing the value of "c_v" coherence over the following parameters: topic number in the interval from 2 to 50 and the "optimize_interval" in (10, 50, 100, 500, 1000). The plots of the corresponding coherence values are examined. A wider interval is avoided because, as observed in the study by Porter [2018], selecting too many topics leads to overfitting. Also, a low number of topics ensures the explicability and efficient analysis of each topic. We consider the best number of topics as corresponding to the best "c_v" coherence value as it is done in a number of works including Zhou, Awasthi and Cardinal [2021] as well as Fang and Partovi [2021]. Also, in case

[4] Mallet: Machine Learning for LanguagE Toolkit. *GitHub*. URL: http://mallet.cs.umass.edu/ (accessed: 14.03.2021).

[5] GENSIM: Topic modelling for humans. *RaRe Consulting*. URL: https://radimrehurek.com/gensim/ (accessed: 5.05.2021).

[6] Ebeid I., Arango J. (2016) Mallet vs GenSim: Topic Modelling for 20 News Groups Report. Fayetteville, AR: University of Arkansas.

[7] Pymystem3 0.2.2: Python wrapper for the Yandex MyStem 3.1 morpholocial analyzer of the Russian language. *PyPI: The Python Package Index*. URL: https://pypi.org/project/pymystem3/ (accessed: 20.03.21).

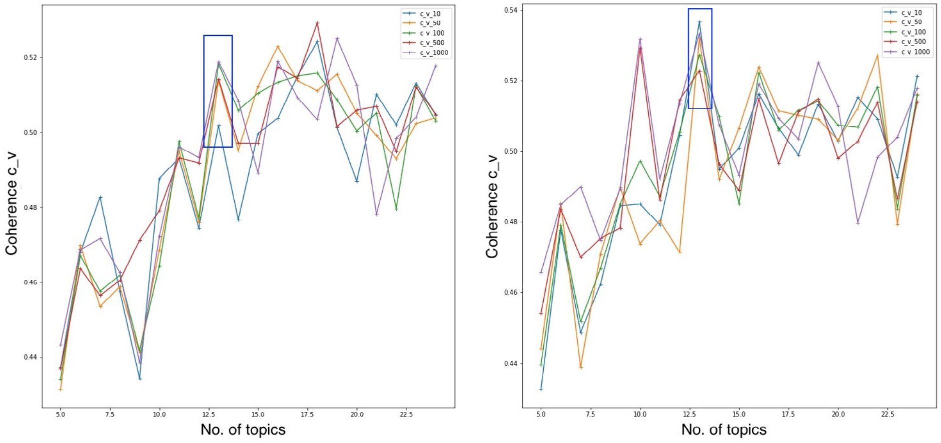
[8] Natural Language Toolkit. *NLTK*. URL: https://www.nltk.org/ (accessed: 5.05.2021).

[9] Topic Modeling with MALLET: Hyperparameter Optimization. *The Dragonfly's Gaze*. URL: https://dragonfly.hypotheses.org/1051 (accessed: 20.04.2021).

a coherence peak that includes the highest coherence value is shared by the LDA runs with different “optimize_interval” values, it is considered to indicate the best number of topics. Ebeid and Arango[10], Fang and Partovi [2021] exploit the same idea by comparing the results with varying seeds. Since coherence tends to grow with the increase of the number of topics [Hasan et al., 2021: 350], the maximums (particularly those shared by most “optimize_interval” values) at the beginning of the interval are given a priority. Additionally, we calculate coherence values in the same way for an extended reference corpus to compare the maximums. In this study, we identified the optimum number of topics for the first week using this method as equal to 13. Figure 2 shows coherence peaks for the first week for the number of topics in the range 5—25 (with and without an external corpus for comparison). The same procedure is conducted for the subsequent weeks. The resulting maximums are 13 or close to 13 in most weeks (9, 10, 12, 14, 16). We decided to adopt 13 as the optimum topic number for all weeks to explore the topics that come to prominence in each week and facilitate the examination of changes in the topic space. Also, we built two LDA models for the whole period to compare the unique topics, which is discussed further in the Topic Hierarchy subsection.

Following Porter [2018], we use the relevance metric with λ = 0.6 as proposed in the original paper by Sievert and Shirley [2014] to enhance the quality of topic interpretation and grouping. Relevance allows taking into account both topic-specific term frequency and exclusivity under a given topic. λ denotes the term’s probability weight relative to the term’s lift, the ratio of a topic-specific term’s probability to its marginal probability across the corpus [ibidem].

Fig. 2. Week 1 coherences for “optimize_interval” in (10, 50, 100, 500, 1000) with an external corpus (left) and without it (right)



[10] Ebeid I., Arango J. (2016) Mallet vs GenSim: Topic Modelling for 20 News Groups Report. Fayetteville, AR: University of Arkansas.

In the standard output, the corpus common terms dominate representation rankings of multiple topics in the model. In this work, the main function of relevance-based ranking is to enhance the comparison of topic representations across the weeks.

Topic Dynamics Features

Topic dynamics features are defined in the present study as the quantitative and qualitative representations of each COVID-19 related theme that are specific to each of the time units within the considered period. The generation of topic dynamics features for both networks is achieved by:

(1) pre-processing of textual data,

(2) LDA parameters selection,

(3) construction of topic models for a predefined number of topics for each week,

(4) creation of a set S_u of unique topics,

(5) consolidation of similar topics into larger themes to facilitate cross-week analysis and comparison of SM,

(6) iterative similarity calculation between each topic in S_u and each topic in a given week's topic set to identify the changes in the representations of S_u topics across the weeks,

(7) calculation of document and user-related statistics.

In the following, we describe data sources, pre-processing, and feature construction.

Data Sources

Publicly available data from the Russian version of LiveJournal (LJ) and Russian-language Twitter tagged with the word “coronavirus” for the period from March 22 to June 1, 2020 that covers the strict lockdown in Russia.

LJ is the largest online community in the Russian-language Internet that hosts the majority of the Russian top blogs[11]. According to the news media reports[12], in 2019 its audience was around twelve million people. Despite not being the largest in terms of the number of authors and messages, it is considered to be a strongly connected blogging community with a rather constant audience. As reported elsewhere, the audience is mostly male, politically oriented, and 35+.

The Russian Twitter is reported to have the most active audience as compared to other SM used in Russia with an average of 47,1 messages per each of 690 thousand active authors[13]. The audience is also predominantly male (66 %) and 35+ (60 %)[14].

Textual Data Pre-processing

The basic dataset pre-processing encompasses the following steps:

[11] About LiveJournal. *LiveJournal*. URL: https://www.livejournal.com/about/ (accessed: 10.03.2021).

[12] Afanasyeva A. (2019) Rambler will re-release LiveJournal. *Kommersant*. January 18. URL: https://www.kommersant.ru/doc/3855808 (accessed: 10.03.2021). (In Rus.)

[13] Chernyi V. (2020) Social Media in Russia: Figures and Trends, Autumn 2020. *Brand-Analytics*. November 30. URL: https://br-analytics.ru/blog/social-media-russia-2020/ (accessed: 10.03.2021). (In Russ.)

[14] 10 facts from Twitter statistics worth knowing about in 2020 (2020) *LPGENERATOR*. January 21. URL: https://lpgenerator.ru/blog/2020/01/21/10-faktov-iz-statistiki-twitter-o-kotoryh-stoit-znat-v-2020-godu/ (accessed: 10.03.2021). (In Russ.)

(1) for each sentence, the language is detected using Langdetect[15], non-Russian sentences are removed,

(2) empty/image posts are excluded (for LJ, empty full texts are substituted with titles, if any),

(3) URLs are removed,

(4) for each post, lemmatization is performed with pymystem3,

(5) punctuation and parts of speech except for semantically loaded parts of speech, such as verb, noun, adjective, adverb, adverbial pronoun, interjection, numeral adjective, and compounds are excluded (numeral, particle, conjunction, preposition, substantive pronoun). This is done to enhance the topic modeling performance by preserving the most semantically loaded words,

(6) standard Russian "stopwords" are removed using NLTK.

Next, the dataset is divided into 10 weeks:
week 1 ("2020-03-22—2020-03-28"),
week 2 ("2020-03-29—2020-04-04"),
week 3 ("2020-04-05—2020-04-11"),
week 4 ("2020-04-12—2020-04-18"),
week 5 ("2020-04-19—2020-04-25"),
week 6 ("2020-04-26—2020-05-02"),
week 7 ("2020-05-03—2020-05-09"),
week 8 ("2020-05-10—2020-05-16"),
week 9 ("2020-05-17—2020-05-23"),
week 10 ("2020-05-24—2020-06-01").

Relative word frequency and text length

To enhance topic modelling results we performed additional pre-processing of the datasets (LJ and Twitter). For each network, we identified the outliers in terms of relative word frequency that are persistent across all weeks. We explored the influence of these outliers, as well as too short and too long texts (1st and 5th quantile of length in each week) on the LDA output for the first week (in LJ). The words with significantly higher frequencies (in terms of the distance from the upper border of the main group of closely spaced frequencies) are expected to dominate the top of per-topic word distributions and multiple topics and will skew the inter-topic distances in the LDAvis visualization. The same LDA settings are used for the experiments.

Relative Term Frequency. The relative frequency of terms is calculated to find possible "stopwords" (outliers). Per-week word frequency estimation takes into account the per-document term count and the number of documents in each week. The equation is as follows:

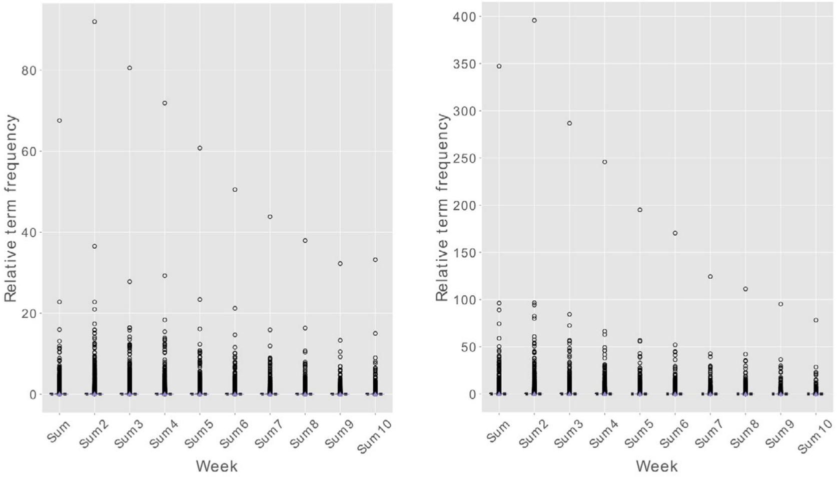
$$f_w = 1/T * \sum_{1}^{d} n_d / N_d \,,$$

[15] Langdetect 1.0.9: Language detection library ported from Google's language-detection. *PyPI: The Python Package Index*. URL: https://pypi.org/project/langdetect/ (accessed: 20.04.2021).

where *n* is the number of term occurrences in document *d*, *N* is the length of *d* and *T* is the size of the corresponding week's dataset (see fig. 3). The analysis of the frequency distribution shows that, for both LJ and Twitter, the term "coronavirus" has much higher frequency across all weeks. In LJ, the word "person" is another outlier. The removal of outliers increased topic coherence (from 0.46 to 0.47) and improved topic distances in the LDAvis output, which was to be expected, since, in this case, they do not help distinguish between the topics.

Document length. When we consider LJ posts, the influence of document length on the topic modeling output is particularly important, because the lengths vary on average from 1.1 to 2914.5 words across the weeks in our dataset. The model assumes that documents are mainly mono-thematic, therefore it attempts to assign as few topics as possible to each text. Too long texts are likely to dominate word occurrence statistics and create too general topics.

Fig. 3. Relative term frequency across weeks in LJ (left) and Twitter (right)



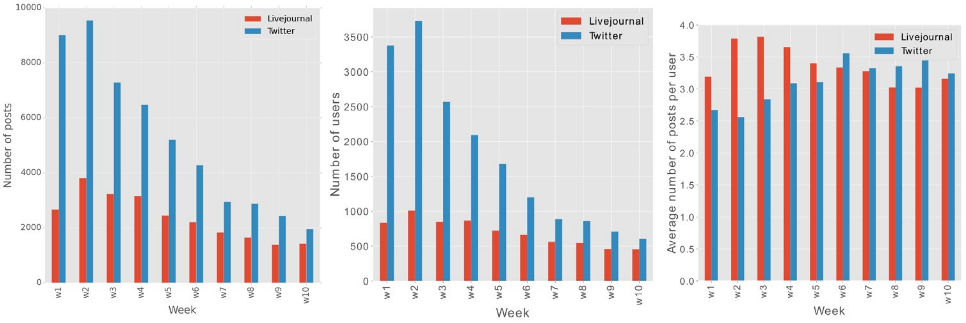
Too short texts tend to be ambiguous and belong to a variety of contexts, which may skew per-topic document distributions and deteriorate topic modeling quality. The length influence is examined by calculating weekly and mean values of the first and fifth length quantiles and comparing coherence values and LDAvis outputs for the first week's full dataset and its pruned version (without first and fifth quantiles).

The exclusion of the longest and shortest texts provides an insignificant increase in coherence from 0.4704 to 0.4705. However, the fact that it changed the inter-topic distances and topic contents (a new topic appeared, per-topic word distributions changed in the visualization) led to the decision to exclude too short/long texts from the dataset for further experiments. In the Twitter case, one-word texts are excluded.

The total number of posts in the final pre-processed version of the LJ dataset is 23 925, the pre-processed Twitter corpus includes 66 616 tweets. The dynamics of the number of posts in both networks across the weeks is shown in figure 4 (left).

DOI: 10.14515/monitoring.2021.6.2044

Fig. 4. The dynamics of the number of posts (left), number of users (middle), mean number of posts per user (right) across weeks in LJ (blue bars) and Twitter (orange bars)



Construction of the unique topics set

For each week and platform, per-topic term distributions are ranked by relevance and the list of all different topics (base list) is built as follows. First, week 1 LJ topics are assigned to the base list. Next, a pairwise comparison with the topics from weeks 2—10 is performed. If the intersection between the lists of top-50 relevant topic terms in a pair is equal or exceeds 30 %, the topics are considered similar. The topic whose intersection with the base list topic is below 30 % is appended to the base list. In this way, an iterative comparison and base list extension is accomplished. The resulting base lists include 42 topics (LJ) and 94 topics (Twitter).

Topic hierarchy construction

The construction of sets of unique topics for LJ and Twitter resulted in an overall number of 136 unique topics for both networks. On the one hand, it is to be expected that Twitter discussions are more dynamic, and the topics are more diverse due to a large number of authors with their characteristic vocabularies. In the LJ community, groups of contributors develop and maintain a more “constant” set of topics over the weeks. On the other hand, manual examination of topic content showed that within Twitter there are quite a number of topics that can be grouped into larger themes. Groups of short tweets cover only certain aspects of the same theme and LDA “sees” them as individual topics. Also, Twitter per-topic term distributions ranked by relevance differ from LJ ones in that, in Twitter, only top-15 terms clearly define the topic, while in LJ all 50 terms are semantically connected. Moreover, we observed that most unique topics are similar across networks and represent parts of broader concepts.

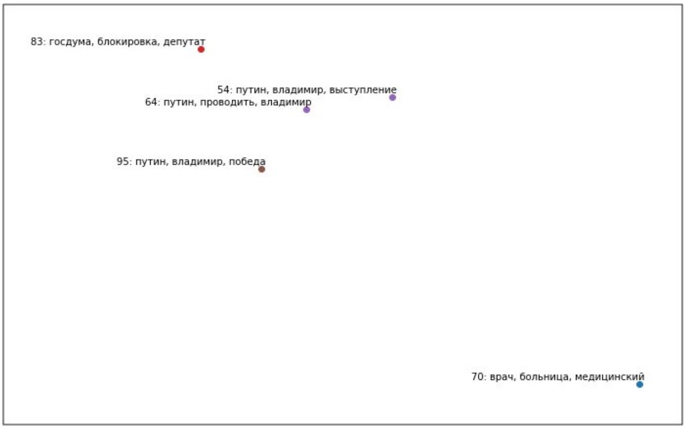
To obtain a topic hierarchy and facilitate cross-network comparison, we tried both automatic (semantic clustering based on a specially trained skipgram model[16]) and manual topic consolidation. In this work, we settled on the use of the manual gold standard version since the semantic clustering approach needs additional improvements. The manual clustering is made by two experts and assisted

[16] Models.word2vec — Word2vec embeddings. *GENSIM: Topic Modeliing for Humans*. URL: https://radimrehurek.com/gensim/models/word2vec.html (accessed: 17.07.2021).

DOI: 10.14515/monitoring.2021.6.2044

by a sklearn t-SNE[17] visualization of BERT[18] embeddings of the unique set topics' representations.

Fig. 5. An example t-SNE plot of the distribution of LDA topics from the unique topics set



The t-SNE algorithm uses dimension reduction to produce 2D scatter plots (see fig. 5) of the distribution of high-dimensional objects, such as word or document embeddings. Most dissimilar objects are separated by larger distances.

For example, in figure 5, the topics related to governmental authorities (83: State Duma, blocking, deputy; 54: Putin, Vladimir, speech; 64: Putin, hold a speech, Vladimir; 95: Putin, Vladimir, Victory) are in the left upper corner and in the lower right corner, there is an LDA topic that belongs to the theme "Hospitalization" (70: physician, hospital, medical). The resulting eighteen large themes together with the logic of their formation (high-relevance terms that define each theme) are given in Appendix A.

Time series of topic-specific indicators

Similar topics are traced by performing an iterative comparison of the unique topics set with the LDA topics from each week in the corresponding datasets. In case there are > 1 topics similar to the unique set topic, all of them are included in the time series of cross-week topic similarities. The resulting datasets for LJ and Twitter contain, for each topic *t*, a time series s_t of topic similarity indicators representing the evolution across the study period. For a given topic, the indicators include median intersection relevance (if several similar topics are found, it helps to identify the closest one), the ratio of topic-specific texts to all week texts, the ratio of topic contributors to the total number of unique users in a given week, ratio of one-topic contributors to the total number of unique users in a given topic, ratio of one-topic contributors to the total number of one-topic users, as well as the shifts in term composition.

[17] Sklearn.manifold.TSNE. *Scikit-Learn: Machine Learning in Python*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (accessed: 20.07.2021).

[18] BERT. *Pre-Training Embedding*. URL: http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html (accessed: 20.07.2021).

Generation of User-specific Features for Graph Visualization

The account data is grouped in two ways. On the one hand, the cross-week account activity data (number of texts per week) is clustered using k-means to explore activity patterns. Particularly, we test the assumption that there are users sharing similar activity patterns (in per-week per-user number of posts) within the study period. The following groups are expected:

(1) users whose period of maximum activity lasts one or two weeks, which may reflect their response to some impacting news or events, or purposeful information spread in case of a large number of messages,

(2) users who contribute uniformly or almost uniformly during the study period (the cluster of the most active users, these are often news accounts);

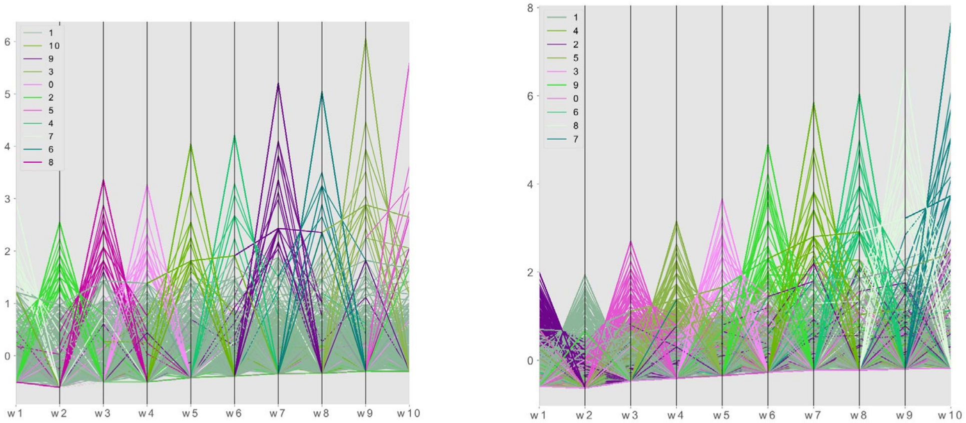
(3) users whose interest in COVID-19-related topics gradually fades from the first to the tenth week.

On the other hand, we group users by dispersion (participation in *n* topics) to explore the per-topic contribution of the corresponding groups during a given week.

Account Activity Clustering

The total number of users that contributed during the study period is 14 129 (Twitter) and 2 470 (LJ). The general trends in weekly dynamics of the number of users and the mean number of posts per user are depicted in figure 3 (middle and right), respectively. In both networks, the activity decreases, in Twitter, a more drastic change is observed, while the mean number of messages per user lies between two and four, which is most likely due to a large number of users with very few messages per week in both networks.

Fig. 6. User activity clusters: 11 clusters in LJ (left) and 10 clusters in Twitter (right)



The clustering is performed using k-means from the Python-based sklearn library[19]. The matrix of per-user per-week messages is normalized by rows (by the overall per-us-

[19] Sklearn.cluster.KMeans. *Scikit — Learn. Machine Learning in Python.* URL: https://scikitlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html (accessed: 20.04.2021).

er contribution) and by columns using StandardScaler from sklearn. The number of clusters is selected using the popular elbow curve method [Marutho et al., 2018] based on the within-cluster sum of square distances. Based on the experiment with nine different random seed values and the number of clusters in range (2, 15) the optimal number of clusters is determined as equal to eleven for LJ and ten for Twitter. The percentage of users per cluster in LJ and Twitter is shown in table 1. The clusters are visualized in figure 6. The colors correspond to cluster numbers that are identical to those shown in table 1.

Both in LJ and Twitter, there are groups whose largest contributions are made in one of the weeks with none or almost no activity before and after the corresponding week. Also, the largest groups (cluster three in LJ and zero in Twitter) are represented, among others, by the users who were the most active (or uniformly inactive) throughout the whole period.

Table 1. ***The percentage of users per cluster in LJ and Twitter***

№	0	1	2	3	4	5	6	7	8	9	10
Users %, LJ	4.42	7.13	4.59	36.55	3.59	4.01	6.84	3.46	9.18	7.47	12.77
Users %, Twitter	31.04	7.13	21.24	13.06	4.78	2.98	11.77	3.47	2.59	1.94	—

In Twitter, this group also includes users with the contribution maximum at week 2 followed by a slight decrease. In LJ, the general trend in this group is the same, however, there is an independent cluster 1 (see fig.6) with the maximum at week 2 preceded and followed by zero or almost zero activity.

The examination of user activity and attention shows consistent and predictable dynamics: a gradual decline in user activity after the second week. This trend can be due to the fact that it was in week 2 from March 29 to April 5 when the introduction of the self-isolation regime from March 30 was announced as well as the start of a non-working month from March 30. In week 2, the response is the highest and it gradually fades as people are getting used to new circumstances and start creating the appropriate behavioral patterns.

Account Grouping by Dispersion

We tested three approaches to summarize user data given the topic distribution per user. Two previous graph versions displayed, for each week, full graphs of users and all different dispersion groups, respectively. By dispersion, we mean the length of topic distribution per account. The main function of full user graphs was to display each user's "interest" in a given topic (ratio of his topic-specific texts to his total contribution) and the ratio of his per-topic publications to the per-week per-topic contribution of all users. When exploring this graph type, we noticed that, surprisingly, certain news accounts turned out to be the most "interested" in small-sized specific topics within the theme "Entertainment and Leisure", according to LDA per-document topic prevalence. We then weighed user dispersions by the use of *n* top relevant terms in a given account's documents. With *n* = 5 the mentioned news accounts lost their positions in

DOI: 10.14515/monitoring.2021.6.2044

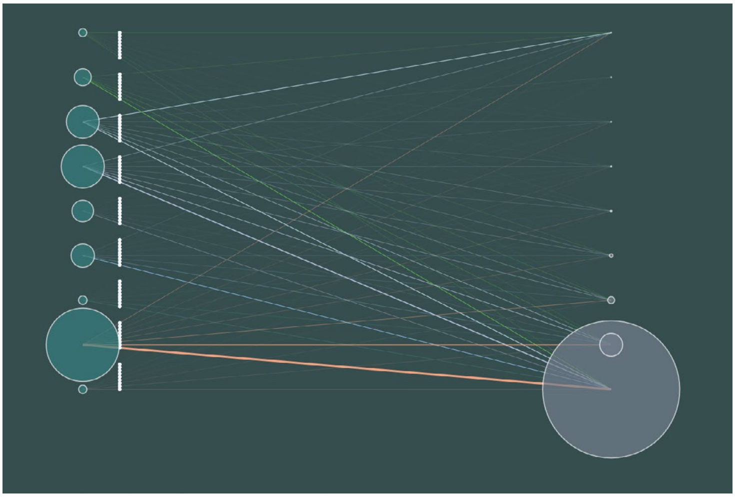
the "interest" ratings since their posts did not contain top-5 relevant words. We keep working on this graph type to expand the functionalities of the current application version.

The graph of dispersion groups visualized all types of topic combinations per account in a given week where the accounts with the same topic combinations were grouped together into one node. We decided to further summarize the data since the previous two graph types are large and difficult to interpret when performing a cross-week and cross-network analysis. We settled on summarizing the dispersion data since it represents the audience focus properties. It is achieved by gathering all users belonging to a certain dispersion type (e. g., one-topic users) in one node (n-topic group).

Graph Visualization in Dash

To visualize the connections between n-topic user groups (user dispersion groups) and the corresponding topics in each week and cluster for each network (LJ and Twitter) we built a Dash[20] application based on the graphs created using networkxs[21] and Plotly[22]. The app allows switching between the following parameters: week (from the first to the tenth week), cluster (the main cluster and the one with the peak at a given week), and network (LJ or Twitter). For each week, the graph is a bidirected graph where the vertices aligned on the left are topics and the vertices on the right are user groups (see fig. 7). The information appears when the user hovers the pointer over the nodes/edges.

Fig. 7. A sample graph (Twitter, week 1, main cluster) of connections between user dispersion groups (right) and topics (left) for each week



[20] Dash Python User Guide. *Plotly*. URL: https://dash.plotly.com/ (accessed: 15.05.2021).

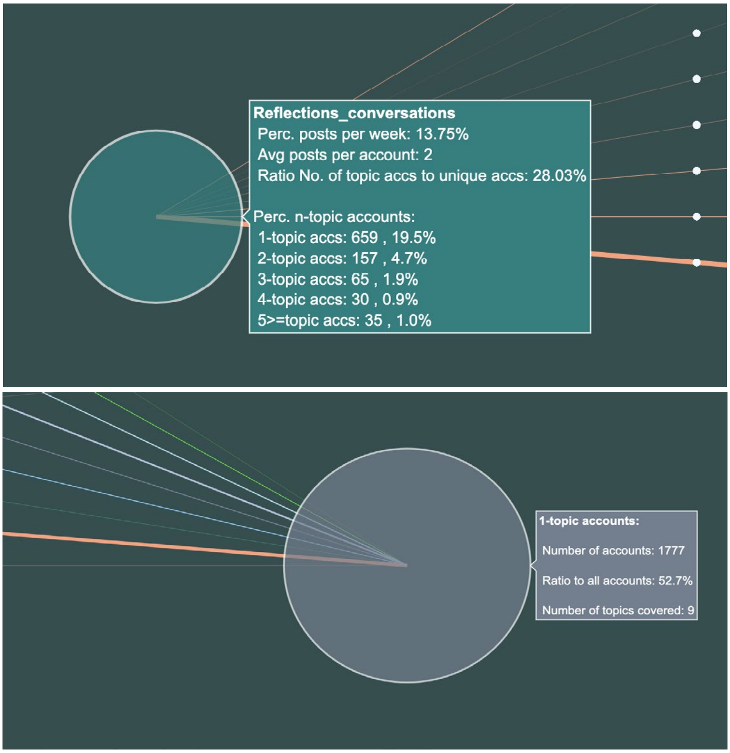
[21] Network Analysis in Python. *NetworkX*. URL: https://networkx.org/ (accessed: 15.05.2021).

[22] For more details, see URL: https://plotly.com/ (accessed: 15.05.2021).

The size of topic bubbles is proportional to the ratio of accounts in a given topic in a given cluster to the total number of unique accounts in a given week. The hover information for each node shows the percentage of posts per week, the average number of posts per account, as well as the number and percentage of each n-topic group. The size of the nodes on the other side (n-topic groups) is proportional to the ratio of the number of accounts in a given group to the total number of unique accounts per week. The hover information shows, for a given group, the number of accounts, the ratio to all accounts, and the number of topics covered. The hovering information on edges (false nodes) shows the contributing group and the ratio of texts in this contribution to all week's texts. The width of edges is proportional to the ratio of contributed texts for a given n-topic users group.

The dashboard application can be run directly from the Google Colab[23]. The Google Colab page also provides links for all the obtained plots, graphs, and tables. The datasets and topic dynamics tables are stored on our GitHub project page.

Fig. 8. Hover information on topic nodes (left) and user dispersion nodes (right)



Reflections_conversations
Perc. posts per week: 13.75%
Avg posts per account: 2
Ratio No. of topic accs to unique accs: 28.03%

Perc. n-topic accounts:
1-topic accs: 659 , 19.5%
2-topic accs: 157 , 4.7%
3-topic accs: 65 , 1.9%
4-topic accs: 30 , 0.9%
5>=topic accs: 35 , 1.0%

1-topic accounts:

Number of accounts: 1777

Ratio to all accounts: 52.7%

Number of topics covered: 9

[23] DASHBOARD APP: Weekly Changes of User Attention to COVID-Related Topics (March 22 — June 1, 2020). *Colab. Research.Google.* URL: https://bit.ly/3tcMmuX (accessed: 25.07.2021).

The produced interactive graphs combine all of the data representation features for further analysis, however, the use of their captions in the paper is not appropriate due to their size and interactive elements. Therefore, we plot the main weekly statistics for the analysis section.

Data Analysis

We perform the analysis of user activity dynamics with respect to the inferred topics for the period of ten weeks. The correspondence of week numbers to dates is given in the "Textual Data Preprocessing" part. Since the topics are directly associated with real events, we rely on a timeline[24] of these events that took place in Russia during the considered period and are found to be mirrored in SM (see table 2).

Table 2. ***Timeline of the main events that took place within the considered period***

№	Date (2020)	Event
1	March 25	The first Address of the President to the Nation (introduction of the self-isolation regime, non-working week is announced, 2020 Russian Constitutional Referendum is postponed)
2	March 27	International flights were grounded after the government ordered the civil aviation authority to suspend all regular and charter flights to and from the country
3	March 29	Mayor of Moscow Sergey Sobyanin issued a stay-at-home order starting the next day
4	March 30	Similar orders or recommendations were announced in numerous other federal jurisdictions, with many more announcing such restrictions over the next few days. The same day, the border was shut, with all border crossings closed
5	April 02	The second Address of the President to the Nation (establishment of penalties for the violation of the self-isolation regime, the announcement of the non-working month)
6	April 11	Moscow's mayor, Sobyanin, signed a decree introducing a digital pass system to enforce the coronavirus lockdown
7	April 28	President's announcement of the prolongation of non-working days until May 11
8	May 09	Air show instead of the 2020 Moscow Victory Day Parade
9	May 11	President Putin announced the end of the national non-working period
10	May 12	Announcement of additional support measures
11	May 27	Sobyanin announced that some restrictions in Moscow would be eased on June 1

Tweets are more "spontaneous": they are short (1—2 sentences) and not elaborate and polished as LJ posts. They appear more similar to spontaneous speech, reflecting more clearly the inner state of a person, his/her reactions to events and triggers, anxiety, loss of interest, joy. In the first place, this study uses the obtained data (based on the Twitter corpus) visualized on graphs to monitor the process of adaptation to new life conditions, and the corresponding dynamics of the anxiety levels. The anxiety levels

[24] COVID-19 Pandemic. Chronology of Events. *Interfax*. URL: https://www.interfax.ru/chronicle/novyj-koronavirus-v-kitae.html (accessed: 25.07.2021).

are associated with the intensity of user contribution to the information space. For the purposes of this case study, we chose to focus on the Twitter data due to its nature and a large number of users as compared to LJ. In LJ (after the consolidation of LDA topics for both networks), the main activity is observed in the same topics, however, their distributions differ. The dynamics of cross-week user activity in topics in LJ and a detailed comparison of both networks are subject to future research.

The intensity of user activity is measured based on the following data:

(1) the number of active users and contributions (posts) per week,

(2) cross-week dynamics of topic diversity, changes in user activity, and shifts in per-topic distributions of relevant words in each topic,

(3) user dispersion.

Let us justify the above statements:

(1) Anxiety generates the need for communication, self-expression, and discussion [Folkman, Moskowitz, 2004]. The latter enables an increase in the number of posts/messages.

(2) Two criteria are taken into account:

(2.1) the demand for a certain topic diversity and the number of covered topics: while a user is engaged in a topic, he generates posts and actively participates in the discussions related to this topic (or its subtopics). With the decline in interest towards COVID-19, the variety of topics and the corresponding activity decreases; the topics that have the greatest impact on life are the most talked about,

(2.2) dynamics in the topic space contents, particularly, the transition from the topics related to one's own well-being ("self comes first") to broader contexts ("external" topics).

(3) The narrowing of the thematic diversity of user-generated texts, as well as the transition from topics related directly to personal experiences to less emotionally colored and more rational topics.

Based on the analysis of the obtained data we suggest the following. The examined period is divided into two: weeks 1—6 (the end of March and April) and weeks 7—10 (May).

Week 1. The growing concern amid voluntary measures (voluntary self-isolation), transition to distance learning and work, COVID-19 spread, and the first Address of the President.

Week 2. The first policy-induced restrictive measures significantly affected everyday life and the household economy. It hit people like a bolt from the blue and caused a strong response from the SM audience — an additional burst of activity.

Weeks 3—6. Adaptation and acceptance of the introduced measures, private household issues followed by a switch to a new mode of life characterized by the transition to a broader context (online shopping, world events, aid to health professionals, etc.). A gradual drop of COVID-19-related interest rates that we associate with the decrease in anxiety levels. The turn of attention towards external topics and broader context, as well as the fading of the topic related to introspection, the need for self-expression ("Reflections_conversations") also characterize the decrease of stress.

Week 7. Festive days (the Victory Day and Easter) and the fading of attention towards COVID-19.

Week 8. The termination of non-working days and recommencement of work.

Weeks 9—10. The COVID-19-related topic diversity finally collapses to one major discussion on statistical reports that we associate with the complete decay of interest towards COVID-19.

Let us consider in more detail the changes in topic-related user activity that took place along the considered ten weeks using the above-listed criteria.

The number of active users and contributions per week[25]

The highest number of active users and posts is reached at weeks 1 and 2, with a peak at week 2. Over half of the total number of week 1 users are in the cluster of temporal activity (peak at week 1), meaning that their activity either drops or stops completely after this week. Further, starting from week 2, the number of active users drops each subsequent week until the end of April (week 6). In the following weeks, the drop either slows down significantly or stops. We assume that by May 2020 the distribution of active users across the topics approximates a standard (pre-COVID) pattern after the burst of activity in early April, so the change in the number of users is no longer as noticeable as in April. The number of active users drops for all topics inferred by the model and consolidated by domain experts. Additionally, user activity gets redistributed simultaneously with the decrease in the number of active users, which will be discussed below. Starting from week 2, the percentage of active users in the topic “Statistics” begins to grow (at the expense of a decrease in the percentage of active users in other topics). Starting from week 7, the number of active users in this topic is actively growing, and by weeks 9 and 10 it becomes the only major discussion (about 80 % of active users publish tweets/posts on this topic).

Thus, we observe a transition from the distribution of attention and interest between various topics to a monotopic focus. At the same time, the total number of active users is significantly decreasing, that is, over time, the number of users tweeting/posting on COVID-19 (tagged with “coronavirus”) gets smaller.

Cross-week dynamics of topic diversity
Changes in user activity

The plots illustrating the analysis are placed in Appendix B due to their size. Weekly sums of account ratios displayed in the corresponding plots can exceed 100 % since the same account often contributes to multiple topics and the total percentage of posts per week is always equal to 100 %. Therefore, the per-topic number of posts (blue bars) is often lower than the per-topic percentage of users (orange bars).

In week 1 (Appendix B., fig. 1), the highest productivity (over 23—24 % of posts) is observed for the topics “Reflections_conversations” (over 40 % of users), “Virus_control_measures (public level)” (over 35 % of users) and “Statistics” (over 23 % of users). Topic descriptions are given in Appendix A. It implies strong concern about COVID-19 and the need for discussing it. On the other hand, public protective measures are actively talked about. People are concerned with the shortage of protective means, uncontrollable prices, and the agiotage. Among other topics, household issues are the most actively discussed (“Solutions_to_household_problems”, over 13 % users and

[25] See figure 4 — left and middle.

6 % of posts) due to the ongoing changes in everyday life and voluntary self-isolation. As mentioned above, over 50 % of the week 1 users are in the cluster of temporal activity. These are mostly one-topic dispersion users (52.7 % of week 1 accounts) who contributed to all nine topics giving much higher priority to "Reflection_conversations" (28 % of accounts and 14 % of week 1 posts) and "Virus_control_measures (public level)" (17 % of accounts and 10 % of week 1 posts). This spark in activity is supposed to be triggered by the first Address to the Nation that took place on March 25. This first-of-its-kind event served as an indicator of the importance and severity of the coronavirus threat, and the unprecedented measures to be introduced. The topic diversity of this activity can speak of the strength of emotional tension and general anxiety that began to spread in society this week.

In week 2 (Appendix B., fig. 2), the introduction of the first policy-induced restrictive measures sparks the crowds. Indeed, this is the largest topic discussed ("Virus_control_measures (public level)"): over 39 % of the posts are assigned to it and over half of the accounts wrote about it. The week was shocking due to the introduction of a significant number of strict obligatory quarantine measures (the so-called self-isolation regime) with justification for their necessity. Such a sharp change caused an urgent need to discuss it, which is characterized by a steep increase in activity on the topic. The latter speaks of the high significance and impact of the introduced measures on people and their lives and perception of things.

The discussions on statistics reports remain almost at the same level of activity (around 20 % of users and posts), which indirectly indicates the stability of this topic. During this period, the absolute number of accounts writing on the topic of statistics practically did not change, while the number of texts decreased insignificantly. In the rest of the topics, the activity is weak with "Reflections_conversations" being the most discussed (about 15 % of users), which points to the existing concern about social interaction on COVID-19 issues.

In week 3 (Appendix B., fig. 3), there remains a slightly increased participation of users in the discussion of public protective measures, although the topic's popularity declined, engaging only 15 % of active users. We assume that this is a "residual effect" of the second week's maximum. It should be mentioned that the acceptance of measures as inevitable circumstances occurred already in week 1. The topic "Statistics" gains major popularity this week in terms of both the number of accounts and posts/tweets since the government's decisions are based on the growing number of infected people. Also, the number of texts grew faster than the number of accounts, which indicates greater per-user activity. This week, the discussion on improving everyday life ("Solutions_to_household_problems") returns and is prevalent among the rest of the topics (about 21 % of users published about 9 % of posts on it). As we observed, this topic appears at the beginning of the period, in week 1, in the context of voluntary measures, and in week 3 in the context of the introduced mandatory measures. The gap in week 2 is explained by the attention shift to new uncertain events about to occur related to the introduction of quarantine and restrictions. With the acceptance of quarantine as an inevitable condition, there is a return to vital issues, and, in the first place, to the establishment of everyday life in new conditions. Also, there is another important topic related to personal well-being in this week — work and education, including distance education ("Work").

In week 4 (Appendix B., fig. 4), after a decline of active users in week 3, the topic "Virus_control_measures (public level)" gained popularity again (approximately 35 % of users and 22 % of posts), although significantly less than in week 2. In the subsequent weeks (5—7) it remains approximately at the same level. The activity in the "Solutions_to_household_problems" drops this week. Thus, the discussion of general (public level) restrictive measures shifts the focus of attention from solving specific practical problems of everyday life to a broader context. Also, the growth of activity in this topic appears to be due to the introduction of the system of digital permits in Moscow (the decree as of 10.04 put into effect as of 13.04). Except for these two most prominent topics (statistics and public measures), the topic "Reflections_conversations" comes back to the fore (19 % of users and 11 % of the posts). There appears a local (occasional) topic "Holidays" due to the Easter celebration.

This week does not feature the topics that were active during weeks 2 and/or 3 — "Solutions_to_household_problems", "Economic_issues", "Work", "Entertainment_and_leisure", which may imply the normalization of people's understanding of the new conditions — distance learning and work, online shopping, the shift to local and more general topics.

In all of the following weeks, the dynamics of the "Statistics" topic remains the same (the highest activity rates). Therefore, the analysis describes other prominent features of the dynamics.

In week 5 (Appendix B., fig. 5), the highest activity remains in "Reflections_conversations" and "Virus_control_measures (public level)", the number of users in "Reflections_conversations" being noticeably higher (approximately 37 % versus 27 %). The topic "Economic_issues" reappears (approximately 8 % of posts and 13 % of users).

In week 6 (Appendix B., fig. 6), the topic "Reflections_conversations" drops sharply (about 8 % of active posts and 18 % of users). The topic of "Virus_control_measures (public level)" is still the leading one (approximately 27 % of posts and 37 % of users). This week the topic related to distance learning and work ("Work") comes up again, which may be due to the extension of the non-working days until the end of the May holidays (announced at the end of April).

In week 7 (Appendix B., fig. 7), "Virus_control_measures (public level)" remains the leading topic. The local topic "Holidays" returns (related to the Victory Day on May 9), and a new topic appears — the support of medical workers ("Healthcare_professionals (incl.payments)"), most likely, as a reaction to the news reports about the introduction of incentive payments for healthcare workers as of 6.05, and internal policy. This week the activity in the "Reflections_conversations" topic stops and shows up in the following weeks with the activity of about 5 % of users.

By week 8 (Appendix B., fig. 8), the attention to public protective measures gradually fades covering about 17 % of posts, and 26 % of users and a clear increase of the topic "Statistics" starts (64 % of posts and 58 % of users). The activity in other topics also declines. Among the remaining topics, peace and information are leading, which indicates a shift of interest further and further from personal problems to the general information and foreign policy level.

Week 9 (Appendix B., fig. 9) basically features two topics: the growing "Statistics" (85 % of users and posts) and "Virus_control_measures (public level)" (10 % of posts and 20 % of users) whose activity keeps declining.

In week 10 (Appendix B., fig. 10), the trends of week 9 persist. Another topic that returns to the fore this week is the assistance to medical workers, "Healthcare_professionals (incl.payments)", with 22 % of posts and 10 % of accounts. This may be due to the shortage of protective suits and personal protective equipment observed in late May and early June in various regions of the Russian Federation.

The observed changes allow defining the main trends in the dynamics of user activity and topic contents in the examined period. In the first place, we notice the overall gradual fading of interest towards COVID-19 and the introduced measures that is manifested in:

(1) the decrease of active accounts and posts/tweets across the weeks and

(2) the shift of the contents of the topic space towards less emotionally loaded and more rational.

Secondly, news content gets quickly reflected in discussions (no longer than within one week, which may be due to the selected analysis unit) and rapidly disappears if the effects of the reported events are either short-term or weak (e. g., the discussions on the Addresses to the Nation and holidays keep being active for at most two weeks).

Thirdly, the power of experience and the relevance of the COVID-19 theme to the authors is represented by the high topic diversity at the beginning of the period. Even in week 1 before the introduction of the first restrictive measures, there are nine almost uniformly active topics (except for "Statistics" and "Virus_control_measures (public level)" that dominate the activity across all weeks). By contrast, at the end of the period, with a similar number of topics, only the leading topics remain actively discussed while the coverage of the rest of the topics is on average below 10 %.

Shifts in per-topic distributions of relevant words in each topic

Let us consider the weekly changes within the main topics:

"COVID (investigation, tests, treatment)". In the first 2—3 weeks of April, the discussions within this topic are mostly about private well-being (immune system boosting, COVID symptoms compared with flu and pneumonia, what different sources say about the virus). From week 4, the topic becomes more "detached", the attention shifts from studying the interaction with the virus at a household level to topics related to mass health, laboratory research, vaccines, and the vaccine market. In the second period (May), the topic arises in weeks 8 and 9, which may be associated with a reaction to local events: first, in connection with mass antibody tests in Moscow and vaccine testing that began to be reported in the media. To summarize, during the first 2—3 weeks (until May) the topic is perceived in a more personal and emotional way and after this period it becomes more abstract and susceptible to news feeds.

"Economic_issues". In weeks 2—3, the discussion concerns global problems relevant to personal safety and interests: economic crisis, the closure of borders. The local context covers the governmental support for families and businesses in a difficult economic situation (weeks 1 and 3). The topic is active in the initial period (weeks 1—3), then it drops off the discussions yielding to the topics "Statistics" and "Virus_con-

trol_measures (public level)". We associate it, firstly, with the general tendency of a decrease in user anxiety and loss of interest in the former variety of topics, as well as a shift in the focus of attention to a more thematically focused context dedicated to the disease itself (its statistics and restrictive measures).

"Hospitalization". In weeks 1—4, the topic content is quite predictable — hospitalization, patients, communal services, infection, artificial lung ventilation, etc. In weeks 5—6, the topic shifts the focus of attention: in week 5, Moscow appears in the topic and, in week 6, the topic covers the stories of hospital staff. Then the topic drops off and returns in week 8 in the context of the increase in hospital admissions in St. Petersburg, which caused an excessive burden on hospitals. The topic also reflects local events — hospitalizations of famous people. Thus, here we can also talk about a transition from more personal aspects that affect emotional stability, since there is a prospect of being admitted to a hospital, to some more detached view of what happens in hospitals.

"Information_sources". In week 1, a greater emphasis is on the topic of distance education (schoolchildren, students) and the Olympiad (Tokyo, 2020). In week 2, the topic shifts towards news sources of relevant statistics and tracking the situation in the world. In the fourth week, a discussion on TV series joins in. In week 6, the emphasis is shifted even more strongly to information on leisure (films, shops, subscriptions), and the topic of distance learning comes up again. The rest of the period it mostly contains reactions to news events (hospitalization of the Kremlin press secretary Dmitri Peskov, reaction to the Russian filmmaker Nikita Mikhalkov's program "Besogon TV" (a YouTube and TV show whose title means "driving out demons" that defends conspiracy theories around COVID-19), online outlet, news portals).

This topic illustrates well the turn of attention to COVID-19 in week 2 (sources of information and statistics become the focus of attention) when the virus spread starts to directly affect people's lives, pushing aside the problems of distance education. We carefully suggest that it was in week 2 that the pandemic began to be perceived as something real and impacting. Also, the emergence of interest in information on TV shows in week 4 indirectly indicates a decline in anxiety about the virus and a shift in attention to organizing leisure during the pandemic. In week 6, the leisure-related discussion intensifies, and the topic of distance learning arises again, which may be related to the hope for changes after the May holidays.

"Reflections_conversations". In week 1, the topic includes 2 parts: on one side, general discussions on the virus — its origin, current events, scientific advancements, vaccines, and, on the other side, how to spend time on quarantine with children and family, family leisure. In week 2, attention shifts to the economic and political aspects of the pandemic. In the third week, the topic of self-isolation with children returns, and the topic of the economic and political aspects of the pandemic remains. In addition, week 3 discusses the virus spread and what is happening in the country and in the world (with an emphasis on China and Italy). Week 3 is the most diverse in terms of subtopics. In week 4, various foreign news feeds are discussed. In week 5, the topic related to distance education for children reappears. In weeks 5—6, the subtopic on staying home in quarantine remains. In weeks 8—10, the discussions touch on lockdown life and concerns: family time (children), job, tests, payments, medicines.

The subtopic on family life on lockdown and children is the most stable. Its importance can be attributed to a sharply increased burden on parents and difficulties in organizing children's education and leisure at home. This is associated with an increase in anxiety. This subtopic persists in weeks 1—6 (except week 2) and its tokens are among thirty with the highest-ranking scores (relevance). Despite the fact that the main restrictions were announced in the second week, this discussion is strongly present only starting from week 3, which is probably due to a need to organize primary household tasks first and then there arise new problems related to the long stay in self-isolation with children and other family members in a limited space. After week 6, specific subtopics are not clearly expressed, which is assumed to indicate successful coping with the ongoing changes.

"Statistics". In week 1, the discussions focus on the spread of the virus in the world, particularly, infection and death statistics recovered for different countries including Russia. From the second week, the global context disappears, the discussion on domestic statistics remains, Ukraine is also present. In week 2, there are statistics for Moscow and, in week 3, for the regions. In weeks 4—5, the general trend continues, and world statistics reappears. Week 5 features statistics in Ukraine, Moscow, St. Petersburg, Moscow oblast, Belarus, and foreign cities. In week 7, the global context returns to the fore, the infographics and virus spread in April are discussed, and the subtopic on testing statistics appears. In week 8, there are 2 major subtopics: domestic and world statistics. In week 9, the upcoming second wave in the fall is talked about. Here, similarly to other topics, the scope narrows from global to domestic and more "personal" level through the transition to Russia, then Moscow and different regions (political subdivisions) with their specific locations. Then, from the fifth week, there is a reverse process of expanding the context with a gradual return to global statistics while domestic statistics persists.

"Virus_control_measures (personal level)". In week 1, the discussions touch on individual protection measures (wearing a mask, staying at home, washing hands). In weeks 2 and 3, the emphasis changes to purchasing personal protective equipment ("purchase", "production", "delivery"). In week 4, a sub-discourse on the protection of contact people (workers, doctors, salespeople, etc.) arises. It persists in week 5 while the sub-topic of acquiring personal protective equipment is replaced by the sub-topic of their use. In week 6, the topic is a mixture of discourses related to purchasing and obligatory wearing of personal protective equipment in public places. In weeks 8—10, the focus is on the mandatory use of personal protective equipment in public places.

As in the previous topics, there is a transition from personal everyday practical tasks (find, buy) in the initial period to the external (protection of other people and social obligations).

"Solutions_to_household_problems". The topic of improving everyday life is present in the first weeks of the period — up to five weeks. In general, the development of the topic comes from everyday issues of buying vital supplies (shops/pharmacies, buckwheat, toilet paper in week 1) through the organization of home leisure with children (three weeks) and then goes to online leisure and shopping (five weeks). The latter may indicate adaptation to a new lifestyle by the fifth or sixth weeks and even readiness for long-term quarantine when season tickets become relevant.

"Virus_control_measures (public level)". In the first week, the sub-discourses are related to the President's Address, the introduction of a non-working week with salary retention, measures in Moscow, border closure, and restrictions on movement. In the second week, the topic as a whole reaches its peak. It discusses the second Address of the President, the extension of non-working days, the introduction of the self-isolation regime, the closure of most stores, restrictions on visits to grocery stores, pharmacies, self-isolation fines, support measures, and spending time at home on lockdown. The self-isolation discourse (staying home, family relations) remains stable in weeks 3—7. In week 3, discussions concern the cancellation and delay of holidays (Easter, Victory Day), the introduction of the digital permit system in Moscow, and fines. In week 4, we observe a shift towards social support and business support, other week 3 topics remain. In week 5, the discussions touch on social protests and rallies (in Vladikavkaz and online), the extension of non-working days until the end of the May holidays, and public restrictions in May. In week 6, people discuss the empty capital, the self-isolation regime extension, and governmental support. Week 7 concerns the situation in Moscow.

In week 8, the attention is drawn to the easing of the quarantine regime, work commencement, governmental measures, lifting restrictions, and antibody testing. In week 9, there is a mixture of topics: self-isolation, work, staying home, testing, Moscow region, digital permits, testing for antibodies. Week 10 discusses a possible lifting of restrictions in June and the upcoming celebration of the postponed Victory Day.

In general, the topic remains quite stable throughout the entire period, and clearly reflects the reaction to external changes and introduced measures. In the middle of the period, the discourse on the imposed restrictive measures in Moscow becomes more prominent, which may be due to the fact that it was in Moscow that some of the most stringent measures were introduced and subsequently imposed at the regional level. In the discussions, the sequence — Moscow, Moscow oblast, regions — is observed.

"Work". The topic appears locally in weeks 2—3 and then reappears in week 6. In the first period, it is related to the start of non-working days and in week 3 it is fed by a discussion on remote work and learning. In week 6, the topic appears in a broader context — as a discussion of the prospects for starting work during or after the May holidays and salaries.

In general, the analysis of the topic content transformation confirms and expands the conclusions made during the examination of the dynamics of user activity within topics. The tendency of transition from a general broad context of week 1 to a personal perspective in weeks 2—4 is most clearly traced, followed by a return to a wide context with a gradual disappearance of personal discourses (even those that are characteristic of the second part of the period).

Also, the analysis of subdiscourses within topics shows that social networks respond quickly and clearly to all the main news feeds, and the reaction normally lasts one week. It is important to note that the topics that call for the personal involvement of users keep being present in discussions for a longer time, even if they are generated by a news feed. Thus, the duration of the presence of the topic can serve as a signal of its importance for society and a fairly strong emotional response to it. However, it is important to remember that such an observation is not an unambiguous indicator and

DOI: 10.14515/monitoring.2021.6.2044

cannot be perceived as a sufficient condition. It requires further meaningful analysis and correlation with the news context of the relevant period.

Topic dispersion groups in activity clusters

The observed user behavior in clusters of activity confirms our statement about the spontaneous nature of this social network. Throughout the entire period, the number of accounts that wrote on one topic significantly exceeds the number of accounts in other groups: their ratio is above 70 % in each week amid the overall sharp decrease in the number of accounts and contributions. Moreover, a large part of these accounts is usually in the temporal activity cluster, which means that their activity slows down or stops completely after the corresponding week. During the considered period, the percentage of this group in the temporal cluster drops from 53 % (week 1) to 32 % (week 10) due to the overall activity decline while in the cluster of users with uniform activity it increases from 23 % to 49 %. Moreover, the number of different dispersion groups drops from ten to four meaning that the accounts from other groups move to the monothematic group, which confirms our conclusion on the collapse of topic diversity by the end of the considered period.

Discussion and Conclusions

The proposed tool for the analysis of SM data complements the traditional approach since the achieved results not only agree with the conclusions made in survey-based studies but also provide additional data. Together with a visualization of adaptation mechanisms, it expands the understanding of the process of adaptation to new conditions since the processed texts are the product of a collective search for psychological meanings in the new reality. It is important to note that the selection of documents for the given corpus was keyword-based. Therefore, the trend of a gradual decrease in the overall discussion intensity and number of topics can be explained by the shift of the audience focus either to offline communication or to other topics (probably tagged with other keywords) that are not covered by the corpus. However, this trend can also represent the gradual social adaptation to the lifestyle changes imposed by COVID-19: the only concern left after the adaptation period is the pallid statistics, which is similar to checking the thermometer behind the window while life is going on. Indeed, during the study period (ten weeks from March 22 to June 1) we observe the gradual narrowing of the focus from a set of multiple diverse topics related to the COVID-19 spread to one dominating topic — infection and death statistics. Along the study period the multi-topic space with several major topics alternating their dominance (discussions on the introduced national-level measures, introspection) and other significant topics (including economics, foreign affairs, healthcare, COVID-19 vaccine, and tests) gets transformed into a monotopic one (at weeks 9 and 10) where over a half of the accounts publish posts related to only statistical reports. The transition process goes, in the first place, through the personal experience and adaptation to the introduced changes at a "micro level" (finding solutions to household and job-related problems). Then it turns to a wider range of topics related to domestic politics (business, economics, aid to healthcare professionals) and foreign affairs until it eventually collapses to one major discussion on statistics: over a half of users produce over 80 % of weekly posts on statistics and new cases reports. In other words, we deal with a topic space transformation from multitopic to monotopic through personal adaptation followed by the perception and acceptance of a wider context (the new reality at the macro level). It indicates the displacement of stress caused by the imposed restrictions and high anxiety levels characterized by the absence of a clearly defined subject of concern [Zevnik, 2017], the overall chaos and fragmentation, which is reflected in the increased user activity and the prevalence of the topic "Reflections (introspection) and conversations".

Similar conclusions were made based on the psychological measurements of anxiety and depression levels made during the first two weeks after the introduction of quarantine measures in Argentina. After two weeks the anxiety levels decreased significantly while the depression levels persisted almost unchanged [Canet-Juric et al., 2020]. It may reflect the ongoing adaptation process though accompanied by a persisting negative emotional state. Further, overcoming anxiety proceeds (from the thematic perspective) by solving, in the first place, more understandable problems of establishing personal life, organizing work, and learning in a distance format, which corresponds to the traditional way of overcoming anxiety — establishing a daily routine [Hiremath et al., 2020]. The next major change in the topic structure is the expansion of the topic space to cover country- and world-level context, which may indicate a gradual decrease in anxiety at the personal level and the search for support in awareness ("I am not alone") and understanding of the broader context of the problem, which constitutes the next stage in overcoming stress [Kar, Kar, Kar, 2021].

The measurements of user activity in the first weeks after the announcement of the virus control measures (in the first Address to the Nation on March 25, 2020) performed in our study show that the user activity dynamics coincides with the dynamics of stress levels published in psychological research by Rusch et al[26]. So, in the second week after the introduction of the most significant restrictive measures, there is a sharp increase in the number of the corresponding posts and the overall activity followed by a gradual decline during weeks 4—6. It may indicate the decrease in the interest rate and relevance of these problems for users. Thus, the proposed approach can be used as a complementary tool and, possibly, an alternative to traditional psychological and sociological methods.

References

Abd-Alrazaq A., Alhuwail D., Househ M., Hamdi M., Shah Z. (2020) Top Concerns of Tweeters during the COVID-19 Pandemic: Infoveillance study. *Journal of Medical Internet Research*. Vol. 22. No. 4. P. e19016. https://doi.org/10.2196/19016.

Al-Dmour H., Masadeh R., Salman A., Abuhashesh M., Al-Dmour R. (2020) Influence of Social Media Platforms on Public Health Protection against the COVID-19 Pandemic via the Mediating Effects of Public Health Awareness and Behavioral Changes: Integrated Model. *Journal of Medical Internet Research*. Vol. 22. No. 8. P. e19996. https://dx.doi.org/10.2196%2F19996.

[26] Rusch T., Han Y., Liang D., Hopkins A., Lawrence C., Maoz U., Paul L. K., Stanley, D. A. (2021) COVID-Dynamic: A Large-Scale Multifaceted Longitudinal Study of Socioemotional and Behavioral Change across the Pandemic. PsyArXiv Preprints. URL: https://psyarxiv.com/75eyx/ (accessed: 21.12.2021).

Blei D. M., Ng A. Y., Jordan M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*. No. 3. P. 993—1022.

Boon-Itt S., Skunkan Y. (2020) Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modelling Study. *JMIR Public Health and Surveillance*. Vol. 6. No. 4. P. e21978. https://doi.org/10.2196/21978.

Canet-Juric L., Andrés M. L., Del Valle M., López-Morales H., Poó F., Galli J. I., Yerro M., Urquijo S. (2020) A Longitudinal Study on the Emotional Impact Cause by the COVID-19 Pandemic Quarantine on General Population. *Frontiers in Psychology*. No. 11. P. 2431. https://doi.org/10.3389/fpsyg.2020.565688.

Chakraborty K., Bhatia S., Bhattacharyya S., Platos J., Bag R., Hassanien A. E. (2020) Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers — A Study to Show how Popularity is Affecting Accuracy in Social Media. *Applied Soft Computing*. Vol. 97. https://doi.org/10.1016/j.asoc.2020.106754.

Cho S., Park C.-U., Song, M. (2020) The Evolution of Social Health Research Topics: A Data-Driven Analysis. *Social Science & Medicine*. Vol. 265. https://doi.org/10.1016/j.socscimed.2020.113299.

Chu I. Y., Alam P., Larson H. J., Lin L. (2020) Social Consequences of Mass Quarantine during Epidemics: A Systematic Review with Implications for the COVID-19 Response. *Journal of Travel Medicine*. Vol. 27. No. 7. https://doi.org/10.1093/jtm/taaa192.

Cinelli M., Quattrociocchi W., Galeazzi A., Valensise C. M., Brugnoli E., Schmidt A. L., Zola P., Zollo F., Scala A. (2020) The COVID-19 Social Media Infodemic. *Scientific Reports*. Vol. 10. http://doi.org/10.1038/s41598-020-73510-5.

Clemente-Suárez V. J., Dalamitros A. A., Beltran-Velasco A. I., Mielgo-Ayuso J., Tornero-Aguilera J. F. (2020) Social and Psychophysiological Consequences of the COVID-19 Pandemic: An Extensive Literature Review. *Frontiers in Psychology*. Vol. 11. https://doi.org/10.3389/fpsyg.2020.580225.

Daly M., Robinson E. (2021) Psychological Distress and Adaptation to the COVID-19 Crisis in the United States. *Journal of Psychiatric Research*. Vol. 136. P. 603—609. https://doi.org/10.1016/j.jpsychires.2020.10.035.

Das S., Dutta A. (2020) Characterizing Public Emotions and Sentiments in COVID-19 Environment: A Case Study of India. *Journal of Human Behavior in the Social Environment*. Vol. 31. No. 1—4. P. 154—167. https://doi.org/10.1080/10911359.2020.1781015.

Fang, J., Partovi, F. (2021) Criteria Determination of Analytic Hierarchy Process Using a Topic Model. *Expert Systems with Applications*. Vol. 169. https://doi.org/10.1016/j.eswa.2020.114306.

Folkman S., Moskowitz J. T. (2004) Coping: Pitfalls and Promise. *Annual Review of Psychology*. Vol. 55. P. 745—774. https://doi.org/10.1146/annurev.psych.55.090902.141456.

Gozzi N., Tizzani M., Starnini M., Ciulla F., Paolotti D., Panisson A., Perra N. (2020) Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis. *Journal of Medical Internet Research*. Vol. 22. No. 10. P. e21597. http://doi.org/10.2196/21597.

Hagen L. (2018) Content Analysis of E-Petitions with Topic Modeling: How to Train and Evaluate LDA Models? *Information Processing and Management*. Vol. 54. No. P. 1292—1307. http://dx.doi.org/10.1016/j.ipm.2018.05.006.

Hasan M., Rahman A., Karim M. R., Khan M. S. I., Islam M. J. (2021) Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA). In: Kaiser M. S., Bandyopadhyay A., Mahmud M., Ray K. (eds) *Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, Vol. 1309*. Singapore: Springer. P. 341—354. https://doi.org/10.1007/978-981-33-4673-4_27.

Hiremath P., Suhas Kowshik C. S., Manjunath M., Shettar M. (2020) COVID 19: Impact of Lock-Down on Mental Health and Tips to Overcome. *Asian Journal of Psychiatry.* Vol. 51. https://doi.org/10.1016/j.ajp.2020.102088.

Hou K., Hou T., Cai L. (2021) Public Attention about COVID-19 on Social Media: An Investigation Based on Data Mining and Text Analysis. *Personality and Individual Differences*. Vol. 175. http://dx.doi.org/10.1016/j.paid.2021.110701.

Jelodar H., Wang Y., Orji R., Huang S. (2020) Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE Journal of Biomedical and Health Informatics*. Vol. 24. No. 10. P. 2733—2742. https://doi.org/10.1109/JBHI.2020.3001216.

Jordan S. E., Hovet S. E., Fung I. C., Liang H., Fu K., Tse Z. T. H. (2019) Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data.* Vol. 4. No. 1. https://doi.org/10.3390/data4010006.

Kastrati Z., Kurti A., Imran A. S. (2020) WET: Word Embedding-Topic Distribution Vectors for MOOC Video Lectures Dataset. *Data in Brief.* Vol. 28. https://doi.org/10.1016/j.dib.2019.105090.

Kar N., Kar B., Kar S. (2021) Stress and Coping during COVID-19 Pandemic: Result of an Online Survey. *Psychiatry Research.* Vol. 295. https://doi.org/10.1016/j.psychres.2020.113598.

Kurten S., Beullens K. (2021) #Coronavirus: Monitoring the Belgian Twitter Discourse on the Severe Acute Respiratory Syndrome Coronavirus 2 Pandemic. *Cyberpsychology, Behavior, and Social Networking*. Vol. 24. No. 2. P. 117—122. https://doi.org/10.1089/cyber.2020.0341.

Marutho D., Hendra Handaka S., Wijaya E., Muljono (2018) The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. In: *Proceedings of the 2018 International Seminar on Application for Technology of*

Information and Communication (ISemantic). Semarang: Universitas Dian Nuswantoro. P. 533—538. https://doi.org/10.1109/ISEMANTIC.2018.8549751.

Medford R. J., Saleh S. N., Sumarsono A., Perl T. M., Lehmann C. U. (2020) An "Infodemic": Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak. *Open Forum Infection Diseases*. Vol. 7. No. 7. https://doi.org/10.1093/ofid/ofaa258.

Mimno D., Wallach H., Talley E., Leenders M., McCallum A. (2011) Optimizing Semantic Coherence in Topic Models. In: Barzilay R., Johnson M. (eds) *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh: Association for Computational Linguistics. P. 262—272.

Porter K. (2018) Analyzing the DarkNet Markets Subreddit for Evolutions of Tools and Trends Using LDA Topic Modelling. *Digital Investigation*. Vol. 26. P. S87—S97. https://doi.org/10.1016/j.diin.2018.04.023.

Riesener M., Dölle C., Schuh G., Tönnes C. (2019) Framework for Defining Information Quality Based on Data Attributes within the Digital Shadow using LDA. *Procedia CIRP*. Vol. 83. P. 304—310. http://dx.doi.org/10.1016/j.procir.2019.03.131.

Röder M., Both A., Hinneburg A. (2015) Exploring the Space of Topic Coherence Measures. In: *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. New York, NY: Association for Computing Machinery. P. 399—408. https://doi.org/10.1145/2684822.2685324.

Ruggieri S., Ingoglia S., Bonfanti R. C., Lo Coco G. (2021) The Role of Online Social Comparison as a Protective Factor for Psychological Wellbeing: A Longitudinal Study during the COVID-19 Quarantine. *Personality and Individual Differences*. Vol. 171. https://doi.org/10.1016/j.paid.2020.110486.

Sievert C., Shirley K. (2014) LDAvis: A Method for Visualizing and Interpreting Topics. In: Chuang J., Green S., Hearts M., Heer J., Koehn P. (eds) *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, MD: Association for Computational Linguistics. P. 63—70. http://dx.doi.org/10.3115/v1/W14-3110.

Tagliabue F., Galassi L., Mariani P. (2020) The "Pandemic" of Disinformation in COVID-19. *SN Comprehensive Clinical Medicine.* No. 2. P. 1287—1289. https://doi.org/10.1007/s42399-020-00439-1.

Tsao S.-F., Chen H., Tisseverasinghe T., Yang Y., Li L., Butt Z. A. (2021) What Social Media Told Us in the Time of COVID-19: A Scoping Review. *The Lancet Digital Health*. Vol. 3. No. 3. P. E175-E194. http://dx.doi.org/10.1016/S2589-7500(20)30315-0.

Wallach H., Mimno D., McCallum A. (2009) Rethinking LDA. Why Priors Matter. In: Bengio Y., Schuurmans D., Lafferty J. D., Williams C. K. I., Culotta A. (eds) *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Association Inc. P. 1973—1981.

Wang P., Tian D. (2021) Bibliometric Analysis of Global Scientific Research on COVID-19. *Journal of Biosafety and Biosecurity.* Vol. 3. No. 1. P. 4—9. https://doi.org/10.1016/j.jobb.2020.12.002.

Wicke P., Bolognesi M. M. (2020) Framing COVID-19: How We Conceptualize and Discuss the Pandemic on Twitter. *PLoS ONE*. Vol. 15. No. 9. P. e0240010. https://doi.org/10.1371/journal.pone.0240010.

Xiong J., Lipsitz O., Nasri F., Lui L., Gill H., Phan L., Chen-Li D., Iacobucci M., Ho R., Majeed A., McIntyre R. S. (2020) Impact of COVID-19 Pandemic on Mental Health in the General Population: A systematic Review. *Journal of Affective Disorders.* Vol. 277. P. 55—64. https://doi.org/10.1016/j.jad.2020.08.001.

Xue J., Chen J., Chen C., Hu R., Zhu T. (2020) The Hidden Pandemic of Family Violence during COVID-19: Unsupervised Learning of Tweets. *Journal of Medical Internet Research*. Vol. 22. No. 11. P. e24361. http://dx.doi.org/10.2196/24361.

Zevnik A. (2017) From Fear to Anxiety: An Exploration into a New Socio-Political Temporality. *Law Critique.* Vol. 28. P. 235—246. https://doi.org/10.1007/s10978-017-9211-x.

Zhou R., Awasthi A., Cardinal J. (2020) The Main Trends for Multi-Tier Supply Chain in Industry 4.0 Based on Natural Language Processing. *Computers in Industry*. Vol. 125. http://dx.doi.org/10.1016/j.compind.2020.103369.

Appendix A. Descriptions of the manually consolidated topics and the main keywords that helped to distinguish between these topics

Topic	Assignment logic	Keywords
Domestic politics	Everything related to the President, the Government and the Addresses to the Nation	Putin, president, Vladimir, Mishustin
Healthcare professionals (incl. payments)	Discussions on the work of healthcare professionals during the COVID-19 and assistance to them (financial aid in the first place)	Physician, hospital, doctor, aid, receive, payment
Foreign affairs	Everything related to foreign countries. This theme gathers all the interactions with the outside world (outside Russia), both political and economic, as well as the news about the situation in foreign countries. These are mostly discussions of foreign news stories that do not concern Russian citizens; therefore, they are combined into one theme	Names of foreign countries
Information sources	Discussions of the information sources: news channels, LJ and Twitter accounts, news retelling	Internet, online, read, channel, news, information, journalist, newspaper, account (several of these keywords should co-occur in a given post)
COVID (investigation, tests, treatment)	Biomedical research and tests of COVID-19 vaccines, symptoms and their comparison to flu and pneumonia	Vaccine, antibody, test(ing), sars, virus, science, laboratory, analysis
Hospitalization	Hospitalization and treatment of COVID-19 in Russian health facilities	Physician, hospital, patient, treatment, medicine, artificial lung ventilation
Folk_medicine, mysticism, conspiracy	Fighting the virus with folk remedies and association of COVID-19 origin with mysticism and conspiracy theories. This group gathers esoterically oriented people prone to believe in the above things who both read and produce the corresponding content	Garlic, ginger, "Besogon" (a program about conspiracy theories), gates, bill, Hodos (Ukranian personality)
Virus_control_ measures (personal level)	Individual prevention and control measures. This theme is separated from the general (public level) measures and paid special attention, because, according to the results of psychological studies on this subject, the use of individual protective means is associated with the decrease in anxiety levels	Mask, glove, protective, means, sanitizer
Virus_control_ measures (public level)	General virus control measures introduced at the state level. In the first place, they include the introduction of the self-isolation regime in certain regions (Moscow and Moscow oblast), its variations and compliance with it in other regions, the rules for leaving home, the cancellation of festive events, and gradual restriction lifting	Quarantine, permit, pass, go out, home, self-isolation, confinement, restrict

Topic	Assignment logic	Keywords
Solutions_to_ household_problems	A small theme that gathers posts on the first response to the introduction of prevention measures and describes the related household problems and solutions to them	Shop, go out, home, deliver, food, goods, buckwheat, toilet paper (the last two words occurred in a significant number of humoristic posts on buckwheat and toilet paper shortage in supermarkets)
Holidays	Festive days (the Victory Day, Easter, Qurban Bayram)	Bayram, day, victory, easter
Work	Due to the introduction of non-working days, remote working and other work management measures, we decided to give prominence to the work topic (at the personal level)	Work, money, employee, work, receive payments, get paid, salary, business
Entertainment_and_ leisure	Everything related to the leisure options during the lockdown (online activity, sports news, and a subtopic about church attendance)	Free, online game, channel, book, video, access (here, each leisure type has its specific keywords, therefore, the assignment is made based on the co-occurrence of several terms)
Reflections_ conversations	Non-specific subtopics containing random reflections and discussions. They can be considered as an alternative to real communication that was limited during the lockdown	Know, talk, understand, criticize, want, panic, fear, sense (here, keywords are difficult to define in terms of topic focus, however, the most representative ones are related to states (emotional state in particular) and actions)
Regional_problems	Several topics clearly associated with Russian regions (political subdivisions)	Nizhniy Novgorod, Tatarstan, Ingushetiya, Bashkortostan, Rostov (topological names)
Statistics (infection, death)	New cases, infection spread and death dynamics	Spread, infection, new, case, number, infected, names of months
Economic_issues	Topics concerned with both domestic and international economies. What these topics have in common is the “macro-level” (global) type of discussion that concerns country and world economic problems and does not touch on one’s personal situation. This theme also includes the topic about small and medium business support in Russia.	Economics, crisis, market, business, oil, price, finance,

DOI: 10.14515/monitoring.2021.6.2044

Appendix B. Twitter: the per-topic percentage of users and posts in week 1 (1) and week 2 (2), week 3 (3) where accratio stands for the percentage of users in a topic in a given week and textratio is the percentage of texts in a topic in a given week

Fig. 1. Week 1

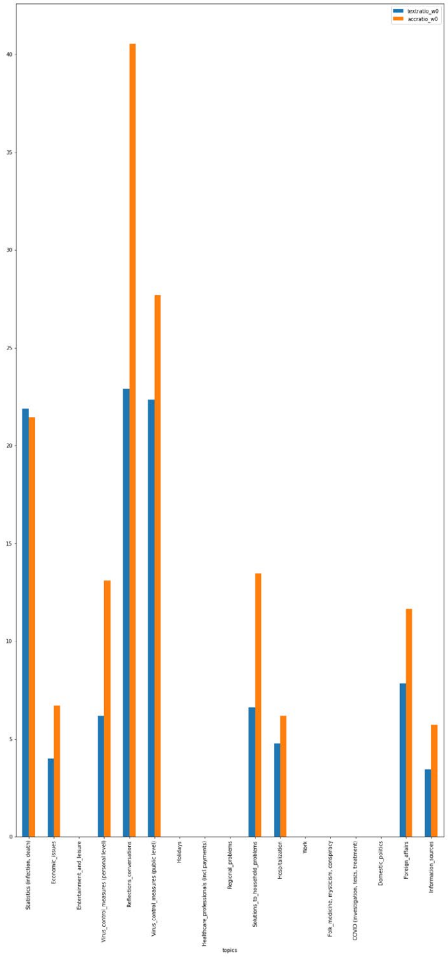


Fig. 2. Week 2

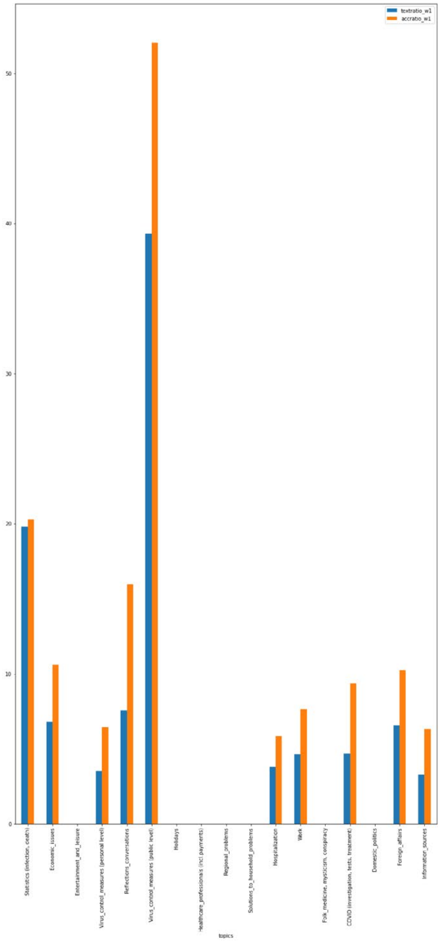


Fig. 3. Week 3

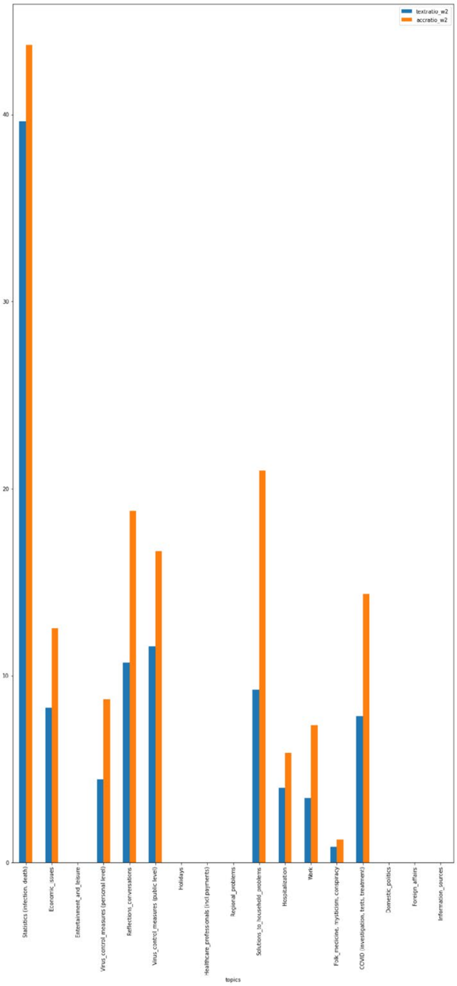


Fig. 4. Week 4

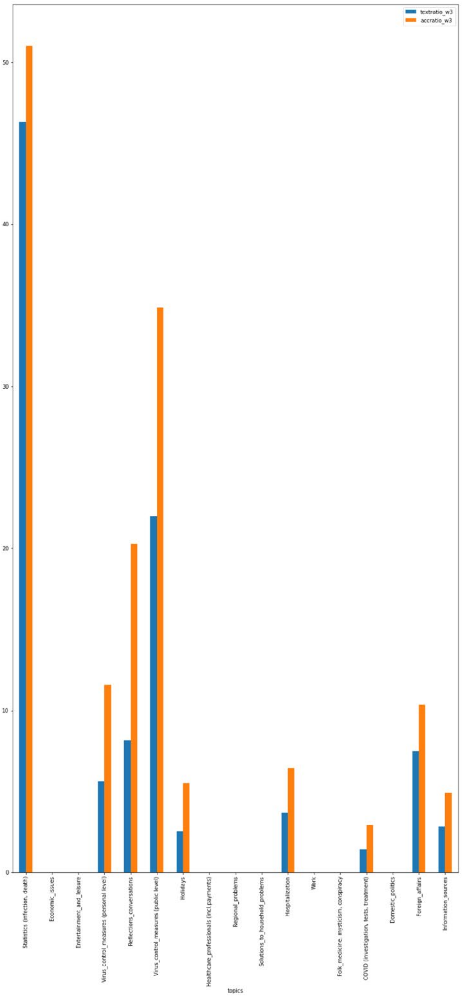


Fig. 5. Week 5

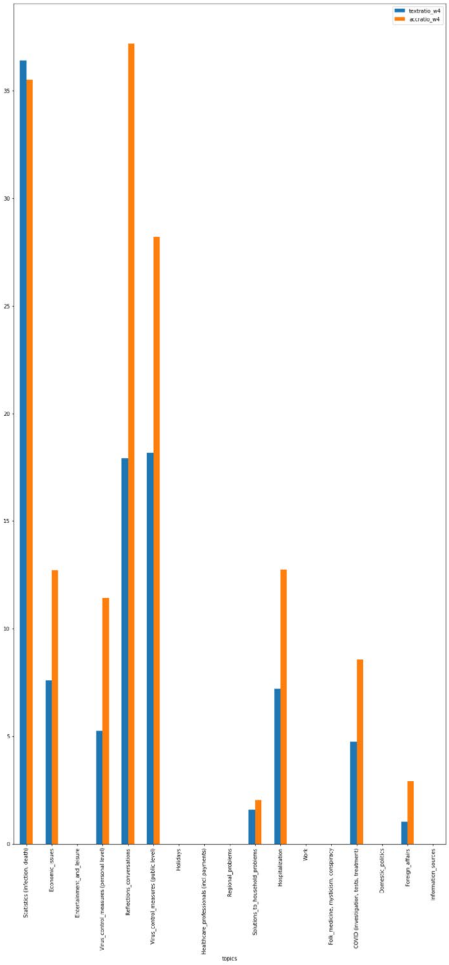


Fig. 6. Week 6

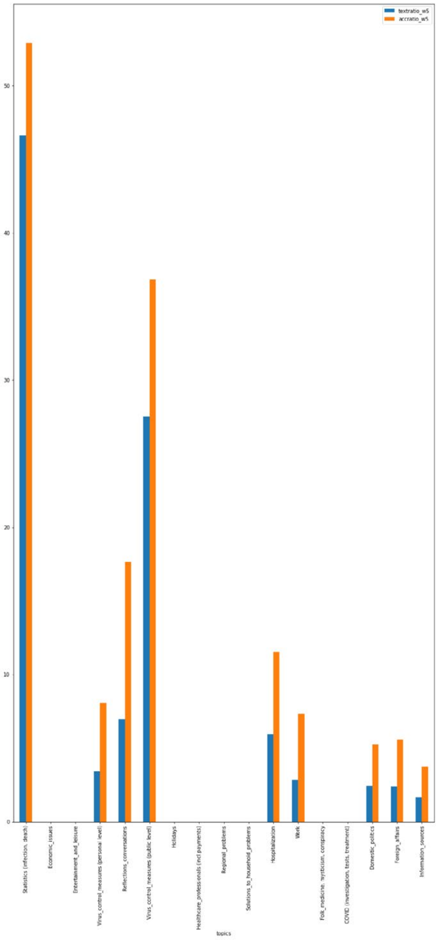


Fig. 7. Week 7

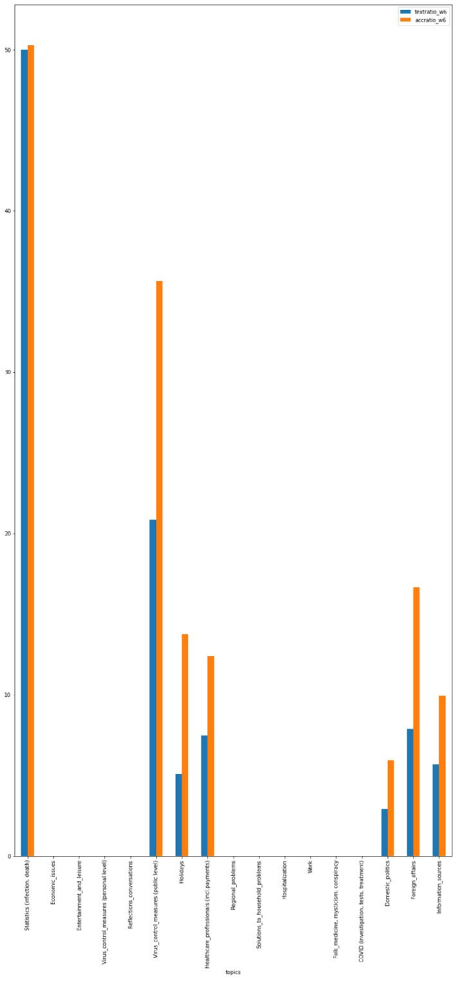
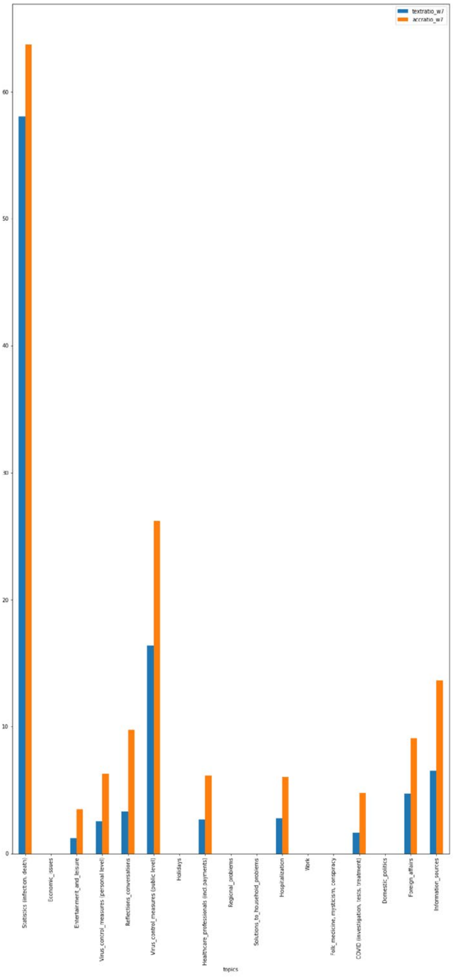


Fig. 8. Week 8



DOI: 10.14515/monitoring.2021.6.2044

Fig. 9. Week 9

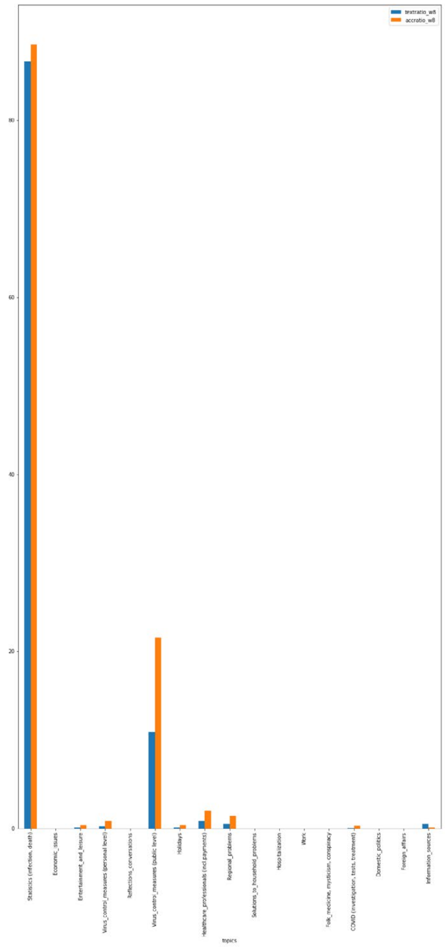
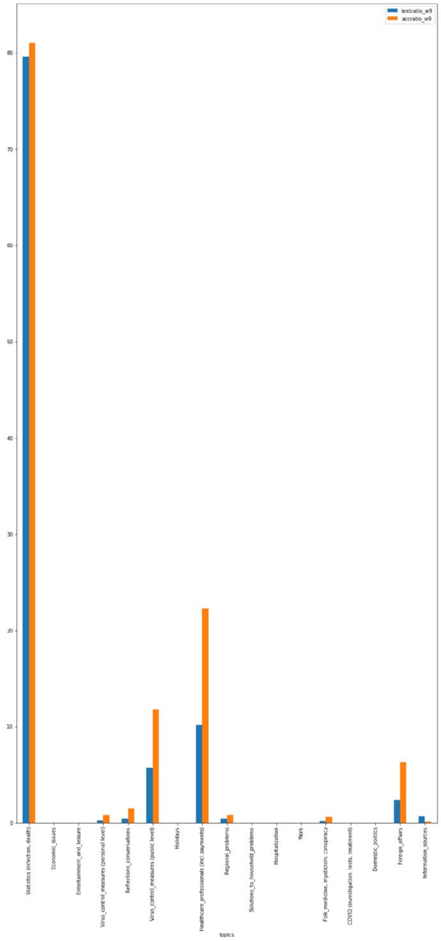


Fig. 10. Week 10


textratio_w9
accratio_w9
Statistics (infection, death)
Economic_issues
Entertainment_and_leisure
Virus_control_measures (personal level)
Reflections_conversations
Virus_control_measures (public level)
Holidays
Healthcare_professionals (incl.payments)
Regional_problems
Solutions_to_household_problems
Hospitalization
Work
Folk_medicine, mysticism, conspiracy
COVID (investigation, tests, treatment)
Domestic_politics
Foreign_affairs
Information_sources
topics