CHAPTER 19

# Social Media-based Research of Interpersonal and Group Communication in Russia

*Olessia Koltsova, Alexander Porshnev, and Yadviga Sinyavskaya*

## 19.1 Introduction

Social media and, in particular, social networking sites (SNS) have become an important source of research data both in Russia and worldwide which, correspondingly, has given rise to new research methods and approaches. Social media data serve as sources of two major types of data: first, the data about the "offline" reality, such as migration, electoral outcomes or mental disorders, and second, the data about human behavior within social media, which includes online self-presentation, networking, media consumption or purchasing behavior. Russian social media, unusual both in terms of their market configuration and data access opportunities, create a slim, but an interesting stream of research.

In this chapter, we critically review the most exemplary works in Russian social media studies. Our goal is to discuss strengths and weaknesses of different research designs and methods that are seldom reported in the papers focused on the research results. As of now, Russian social media studies can be classified along two major lines. The first line differentiates between studies using Russian SNSs as a source of data about human behavior in general, and those that aim at studying specifically Russian society or Russian-speaking community with the data from different social media, including the Russian SNSs.

The second line of differentiation is disciplinary, and we can single out three disciplines that have contributed most to Russian social media research:

O. Koltsova • A. Porshnev • Y. Sinyavskaya (✉)
Higher School of Economics (HSE University), Saint Petersburg, Russia



D. Gritsenko et al. (eds.), *The Palgrave Handbook of Digital Russia Studies*, https://doi.org/10.1007/978-3-030-42855-6_19

psychology and health studies, sociology and political science. The latter has focused specifically on Russia, and especially on the relations between social media and protests. Sociology has addressed a wide range of topics, such as virtual demography, education and ethnic relations whose results reflect specific Russian context, but can often be extrapolated beyond the Russian society. Psychological research has mostly contributed to fundamental psychology by studying the relation of social media to such universal psychological phenomena as depression or personality traits. Health studies may be situated between psychology and sociology.

This review chapter focuses on the three above-mentioned disciplines. We find that they use a wide range of data types and data-collection techniques: self-reported data (from surveys and experiments collected off-line or on-line) and SNS activity user data (incl. user texts and their metadata, such as timestamps and geolocation, the data from group accounts, data on links between accounts and various external statistics). Methods of data analysis vary from traditional discourse analysis and classical statistics to social network analysis (SNA), supervised and unsupervised machine learning, and various combinations of those.

We should also note that Russian social media are actively used in computational linguistics. Though this research community works with the Russian language, it does not focus either on the Russian society or a broader Russian-speaking community, tackling such problems as optimization of information retrieval, text clustering and summarization, named entity recognition or automatic translation. It thus forms a distinct stream of literature addressed in this book (for more, see Chaps. 26, 19, 29, 25, 23 and 24) that we leave out in this review. We also omit the large and important topic of "research ethics is social media studies" as it is a subject for a separate contribution.

The rest of the chapter is structured as follows. First, we briefly introduce the context of social media development in Russia and show how it has resulted in their unique position within the global SNS landscape. The next three sections are devoted to the disciplinary overviews of political science, sociology, health and psychology. We conclude with summarizing both research opportunities and limitations created by social media.

## 19.2 Social Media in Russia

Russian Internet landscape is unique in terms of its "home-grown" character. According to Forbes.ru, Russian companies compete with global players nearly in all spheres of Internet business, including search engines, mailing services and social media (Forbes.ru 2019). Unlike in China, where the closed Internet ecosystem owes most of its success to the policy of technical, political and economic isolation known as the Great Chinese Firewall, Russian information technologies (IT) industry has until recently developed without any protectionist barriers.

**Table 19.1** SNS use in Russia according to media research (October 2018, in mil)

| *SNS* | *Average daily reach (desktop) (Mediascope 2018)* | *Active users[a] per month (Brand Analytics 2018)* | *Monthly messages[b] (Brand Analytics 2018)* |
|---|---|---|---|
| Vkontakte | 16.82 | 36.4 | 1096 |
| Youtube | 13.84 | 1.9 | 15.9 |
| Odnoklassniki | 8.22 | 15.8 | 364 |
| Facebook | 3.23 | 2.3 | 122 |
| Instagram | 2.70 | 23.7 | 304 |
| LiveJournal | 1.24 | 46 | 4.6 |
| Twitter | NA | 0.8 | 59.6 |
| My World ("MojMir") | NA | 0.099 | 7.2 |

[a]Active user: user who wrote at least one public message per month
[b]Message: any publicly available post—status, wall post or comment, post or comment in online group etc. The analysis did not include private messages

As a result, social media "diets" of Russian and Russian-language users are substantially different from global trends (see Table 19.1). Social networking site VKontakte (VK), a Russian replica of Facebook, has by far higher reach and especially higher user activity than all other SNSs, followed by popular, but much less active Odnoklassniki. Facebook (FB) in Russia is a niche network that attracts higher-educated audiences oriented towards international integration, business and, to some extent, more oppositional political views (e.g. Enikolopov et al. 2018). VK, on the other hand, has become a universal tool for everyday communication, practical task-solving and small business. A typical VK user consumes news from a few large entertainment and/or political public pages run by media organizations and also belongs to a multitude of smaller groups that include everything from school classes and self-help communities to pages of local businesses that use VK to promote their services.

Consequently, political and market pressures experienced by VK are to some extent different from those faced by Facebook. Unlike Facebook, VK has not been accused of illegitimate influence on elections, since it is widely accepted that electoral outcomes in Russia depend on very different things (Gel'man 2014). Combined with relatively low importance of privacy among the Russian population (Kisilevich et al. 2012), this until recently has been creating incentives for VK to privilege data sharing over privacy protection. As a result, the amount and diversity of data available through VK application programming interface (API) is incomparably higher than in FB, and thousands of business and research actors use it on a daily basis. It is this unique data availability that has made possible such large-scale virtual demography projects as Webcensus (Zamyatina and Yashunsky 2018) (see further below). Surprisingly, such opportunities have attracted nearly no attention from international scholars, which is why most VK-based research has been done by Russian researchers.

Another important trend is the fragmentation of the Russian-speaking online environment. For a while, VK was an integrating medium for

Russian-speakers on the Post-Soviet space, but this capacity has been severely hindered by the ban of VK in Ukraine in 2017 and the overall deterioration of Russia's relations with the rest of the world. It is plausible to expect that in the near future comparative studies that include Russia might be increasingly dominated by research based on global SNSs which will decrease the value of VK as a data source.

## 19.3 Political Science

Political science research on social media is the largest among the three mentioned disciplines. It focuses on a number of subfields, such as the role of social media for political protest and civil activity (Enikolopov et al. 2018; Koltsova and Selivanova 2019), mapping political discourses and agendas formed both by lay and professional SNS users, including media professionals and politicians (Bodrunova et al. 2018; Goncharov and Nechay 2018; Bulovsky 2019; Kelly et al. 2012; Koltsova and Koltcov 2013) or topic-specific political discussions (Filer and Fredheim 2016), and the newly emerged topic of SNS-channeled propaganda (Barash and Kelly 2012; Kelly et al. 2012; Stukal et al. 2017; Badawy et al. 2018; Sanovich 2017).

Studies based on discourse, agenda and discussion mapping are by their nature, mostly descriptive; and when based on manual text analysis only, usually not scalable. However, manual approach can be enhanced with automated text analysis, albeit in such case often at cost of its depth. Goncharov and Nechay (2018) benefit from applying such a combination to a collection of about 45 thousand tweets related to the anti-corruption protests organized by a prominent Russian oppositionist Alexei Navalny in spring 2017. To evaluate the mobilization potential of Twitter, they apply keyword-determined time-constrained sampling and then use topic modeling to reveal content-based clusters and social network analysis (SNA) to find link-based communities. They convincingly demonstrate the dominance of an oppositional and a loyalist meta-cluster in both partitions. An important methodological note of the authors is that hashtags seldom occur in their data, which questions the frequently used hashtag-based sampling in Twitter research. At the same time, the authors do not specify the type of links used, neither they explain how the two partitions of their data are related. Most importantly, they do not find an answer for their main question about mobilization effect of Twitter, since cluster analysis is a method suitable for descriptive, not for inferential investigation.

This limitation is shared by most SNS research based on clustering techniques and some research based on manual coding. Thus, Koltsova and Koltcov (2013) illustrate the growth of political topics at the expense of other topics in the Russian-language LiveJournal (LJ) top blogs on the eve of the Russian parliamentary elections in 2011 which, nevertheless, does not lead to any hypothesis testing. Bodrunova et al. (2018) go further by formulating the hypotheses about prevalence of certain media roles among professional authors of tweets discussing politicized violent events in four different countries,

including Russia. Tweets are classified manually; in principle, automatically clustered tweets might have equally been tested for prevalence, but no statistical procedures for doing so have been introduced in this research. The authors, nevertheless, offer a multitude of interesting details that help understand the structure of the political discussion in the four countries in a comparative perspective, which is still quite rare in quantitative media studies. In particular, they echo with Goncharov and Nechay (2018) in observing that Russian media on Twitter, unlike those of other countries, cluster along the pro-government—anti-government axis.

A successful attempt to do inferential research based on blog texts is presented by Bulovsky (2019). He fits a regression model to find out the difference between Twitter communication used by authoritarian and democratic political leaders across 144 countries, including Russia. He finds that the former have a significantly lower number of posts per day and a significantly smaller proportion of replies to other users.

Another difficulty with SNS research is a possible lack of context and the reliability of non-SNS-based data. Enikolopov et al. (Enikolopov et al. 2018) perform a most rigorous statistical inference to evaluate the influence of VK on protests in Russia using the data on the large rallies against electoral fraud and on voting in the electoral cycle of 2011–2012. They find out that, paradoxically, VK penetration increases pro-government voting in the respective cities, but simultaneously has a positive effect on both the probability of protests and the number of protesters. There is not enough data to test possible alternative explanations of this effect, such as a polarizing influence of higher SNS penetration levels on the population. At the same time, the reliability of both voting data (given the unknown character of electoral fraud) and the protest data (given that they were taken from the media where the numbers reported by the protesters and by the police dramatically diverged) is highly questionable. Reuter and Szakonyi (2015), who use offline survey data only, obtain somewhat different results and find that the usage of international social networks Twitter and Facebook increased the awareness about electoral fraud, while the usage of domestic VK and Odnoklassniki did not.

Koltsova and Selivanova (2019) solve the problem of contextual enrichment of SNS data with the deep involvement of one of the researchers into the social movement they study. Similar to Goncharov and Nechay (2018), their goal is to evaluate the mobilization potential of VK communities. In particular, they study the effect of VK on the turnout of the movement activists at the poll stations in the role of independent observers on the voting day in 2014. As the movement created VK groups responsible for each of the 17 administrative districts in St. Petersburg, the researchers investigate them all and find that their size, activity and density are positively related to the overall turnout; however, offline observers are neither more active nor more connected members of online groups. The authors offer two alternative interpretations of this effect, based on their experience with the movement, still the problem of reliability of the offline turnout data is one of the unresolved issues.

Importantly, SNS data, though more "objectified" than self-reported or hand-coded data, are not always reliable either. Comparing Twitter discussions devoted to two resonant political murders in Russia and Argentina, Filer and Fredheim (2016) find out that a significant proportion of the Russian tweets, unlike Argentinean messages, are automatically generated (2016, 13). This has two implications: first, the value of Twitter as the data source in Russian political research is limited because of the network's limited penetration. Second, the study of Filer and Fredheim leads us to a whole range of related research topics that include online state propaganda, fake news, bot-generated content, influence of all those factors on electoral outcomes and the problem of the use of personal SNS data for political purposes.

Of this stream of research, Russia-related research has a number of special features. First, some research on political propaganda is performed as bot detection (Stukal et al. 2017). However, not all bots are political (some are commercial), and not all propaganda is robotized (some is manual). Second, the phenomena that researchers try to trace—for example, trolls—are hard to define conceptually and even harder to find empirically. For instance, Badawy et al. (2018) perform an interesting descriptive research of a collection of tweet accounts identified as Russian "trolls." Using one of the algorithms of automatic classification known as label propagation, they identify most trolls as conservative-leaning, while Botometer software (botometer.iuni.iu.edu/#!/) (another classification algorithm) allows them to determine that the majority of those who retweeted trolls were not bots. Those valuable findings are to a certain extent limited by the data used. While trolls are defined as "malicious accounts created for the purpose of manipulation" (Badawy et al. 2018, 258)—that is, intentionally deceptive—the authors use a list of Twitter accounts taken from the website of the US congress Committee for Intelligence via a publication at the www.recode.net website. The list contains only Twitter IDs and usernames, and no other information is searchable. It is thus unclear whether the list creators followed the authors' definition of trolls, and if so, how they learnt about users' purposes. More broadly, finding empirical referents for concepts whose definitions are based on intention (e.g. deception) is a challenge, while taking out intention from the definition of political trolls deprives them from their core meaning and makes them lumped together with authors of unconventional, still legitimate opinions.

Third, this type of research is sometimes not entirely free from politicization. For instance, Sanovich (Sanovich 2017) refers to Barash and Kelly (2012) and Kelly et al. (2012) as the research that for the first time identified a "large-scale deployment of pro-government bots and trolls in Russia" in favor of President Medvedev. However, this is not exactly what the sources suggest. First of all, while the word "bot" is mentioned only once, "troll" is never mentioned in either of the sources. Second, Barash and Kelly (2012) show unusual distributions of activity around tweets related to Medvedev's innovation policy program, while Kelly et al. (2012), using the same data, note that the whole "innovation cluster" of tweets disappears when they filter out "instrumental"

accounts—those that are likely to use search engine optimization (SEO) and automation (bots). The authors do not offer any interpretations, but this suggests that accounts that tweeted about innovation were likely to use commercial promotion methods, in particular automation. Since Kollanyi et al. (2016) show that automation was spread both among pro-Trump and, to a lesser extent, pro-Clinton Twitter accounts during the 2016 US electoral campaign, such accounts might be equally termed trolls. However, there is more sense in distinguishing between trolls and bots than placing them in one category. This does not mean that the Russian government has never used trolls but suggests a certain lack of accuracy when it comes to the Russian computational propaganda research.

To sum up, social media data may significantly enrich the repertoire of research in the field of political science in Russia by providing access to large volumes of data and thereby conducting large-scale research with the usage of automated methods of data analysis. In particular, the access to textual and network data makes it possible to grasp the substance of political discussions and track communication flows between different political parties. At the same time, the results of such research should be viewed through the lens of existing technical and methodological constraints. First, often accessible data allows producing only a limited range of conclusions, for example, descriptive ones. Different approaches to the conceptualization of key concepts which sometimes leads to inconsistent results are also of highly relevant issue. Finally, online data from social media are not free from specific limitations which may affect its reliability.

## 19.4 Sociology

### *19.4.1 Virtual Demography and Structure*

The relative openness of VK data has enabled a number of large-scale studies investigating VK population structure, composition and patterns of communication. By far the largest of them is the project "Virtual population of Russia" (Zamyatina and Yashunsky 2018). It is based on the analysis of approximately 200 million VK accounts and 3.5 billion friendship links, although only 88 million accounts have been found to claim their location in Russia. The most valuable outcome of this project is an interactive website webcensus.ru that contains various subsets of the initial sample at different levels of aggregation and visualizes the most important distributions in the form of charts and maps. The data include age, gender, education, friendship patterns, migration routes and others, along with their relations, for example, distribution of average friendship connectedness over the Russian regions. This data is an incredible resource for researchers seeking to assess the difference of their samples from the total VK population of Russia and thus to statistically test various hypotheses about specific features of the studied sub-populations. Russia is, to our knowledge, the only country for which such detailed virtual census exists. However, this data

has a serious limitation: it dates back to 2015 and, as such data collection is extremely expensive and is unlikely to be updated.

Some studies aim at restoring missing data in SNS accounts based on the available data from other accounts. Such data may include age, gender, geolocation and others. As this research mostly lies in the sphere of computer science, we omit it here, with an exception of two related papers based on the full VK data from the Russian city of Izhevsk. The first paper studies the impact of missing geolocation data on the features of the city friendship network (Kaveeva and Gurin 2018), while the second does the same in relation to fake accounts (Kaveeva et al. 2018). In the first paper, the authors train a classifier that restores users' city of residence based on the accounts in which this data is present. They find out that while the city friendship network grows substantially when the missing users are added, most of its important metrics, do not change, while modularity and the number of communities in the largest connected component experience a very modest growth. This suggests that using incomplete data based on the users who choose to report publicly is a valid research strategy in social network analysis. In the second paper the authors train a classifier to recognize fake accounts. After deleting them, friendship network experiences the reverse change, as compared to the first paper, that is modularity and the number of communities drop a little. One limitation of the second paper is the nature of its training set that is based on self-reported data from 32 users who were asked to assess their friends. The problem of fake account identification is close to troll and bot detection and has no easy solution as all these types of accounts mutate quickly in their attempts to imitate real users.

Other research devoted to SNS structural features attempts inferential design (Kisilevich et al. 2012; Rykov et al. 2018). Kisilevich et al. (2012) examine how age and gender are related to the amount of disclosed information, based on 16 million accounts from a Russian SNS My World (Moj Mir, www.my.mail.ru). This research allows not only to evaluate the completeness of self-reported user data but also to investigate their self-disclosure behavior. The authors report that self-disclosure dramatically drops with age, but find no substantial gender difference which, as they claim, differentiates the Russian SNS users from the Western users. However, the statistical procedure used to claim no gender difference is somewhat unclear, while the presented plots suggest that females may be less frequently sharing information about their political views but are more often sharing quite a number of other types of information.

### *19.4.2 Social Issues and Problems: Education, Ethnicity, Urbanity*

While virtual demography and related studies are interested in the online population per se, various research tackling more specific sociological topics uses SNS data to obtain results about offline reality, or about the role of SNS in the respective problem or issue, be it education, migration, ethnic relations, or urbanity.

For instance, Smirnov (2018) uses VK data on 4400 Russian students to predict their scores in Programme for International Student Assessment (PISA) test—an international test assessing learning outcomes in reading, mathematics and science among 15-year old students. He obtains the data on 73 thousand online communities which students belong to and reaches the correlation of about 0.5 between the predicted and the real PISA scores. The most interesting conclusion is that groups contributing most to high scores are related to arts and science, while those contributing to low scores are related to humor, sex and horoscopes. Although 0.5 is a high correlation in social science, it also means that VK group membership cannot be used as the only predictor of PISA scores. Since VK group membership cannot be used as a causal explanation of PISA performance either, the significance of such type of research should be treated with caution. In general, the predictive power of such models tends to drop the more the farther in time other studied datasets are from the original dataset.

Alexandrov et al. (2018) use VK self-reported geolocation data to study factors influencing outgoing educational migration. They examine a larger number of student migration destinations over a sample of 85 thousand VK users aggregated at the city level. They find that, quite predictably, Far East and Southern Siberia are gravitated to China, North-West—to the Nordic countries, and Muslim regions—to the Middle East. It is interesting that among significant predictors they find both offline factors, such as religious and geographic proximity, and online factors, such as the number of VK groups related to the country of destination in the donor city. However, of course, it is unknown how well self-reported data on users' secondary schools and universities reflect the overall educational migration flows. This poses a broader question of the extent to which online data represent offline reality.

Ethnicity has been another important topic in Russia as a multi-ethnic society, with studies asking how communication in social media either reflects ethnic tensions or influences ethnic relations. Bodrunova et al. (2017) study the posts of top LiveJournal bloggers to determine how different ethnic groups are treated. They first extract ethnicity-related topical clusters via topic modeling and then hand-code 30 most relevant texts in each of 33 ethnicity-related clusters. Among other things, they find that Central Asians are treated as relatively positive aliens, with North Caucasians being presented both as negatively assessed aliens and aggressors. But while North Caucasians are also sometimes victimized, Americans are always both negative and aggressive, which suggests that global political conflicts overshadow local inter-ethnic tensions. This research is one of the few addressing the problem of instability of topic modeling by running the algorithm several times and choosing only stable topics, which is virtually never done in empirical social research. However, it has problems with representativity both in terms of the size of the coded sample and in terms of the choice of LiveJournal popular bloggers as a source.

Urban studies is a subfield that is widely believed to benefit from social media data. However, it often results in the simple plotting of social media data

on geographical city maps that, unlike the mapping of political discussions or virtual demography, is often less useful. Human movement in urban spaces is much better detected with mobile phone data than with social media data, and the content or sentiment of SNS messages is not always related to the places where they have been created. Some studies still attempt to tie geomapping to practical purposes of urban planning. Thus Petrova et al. (2016) examine some hundred thousand posts and check-ins from different SNSs in the city of Samara in order to generate town planning recommendations.. They find that messages are concentrated in the city center and are differentiated by gender, type of place, and topic, while locations also differ by check-in intensity, predominant sentiment of messages and the prevailing type of visitors—either locals or tourists. The authors suggest to create more attractive places in the city periphery and also to unite the most visited and the most positively assessed places in the center by a single pedestrian pathway. This conclusion seems to be based on the visual examination of maps, as the paper describes neither analytic procedures nor any methods of posterior evaluation of the efficiency of the suggested town planning strategy (for more, see Chap. 32).

An interesting result about the nature of urban civic activity is presented in Voskresenskiy et al. (2016). The authors analyze 41 restricted-access and 132 open-access VK groups run by the neighbors sharing the same apartment blocks in St. Petersburg. Based on topic modeling, restricted-access groups are found to prefer such topics as mutual help, socialization and apartment repairs, while open access groups favor city-level initiatives, contentious initiatives, including court disputes with the city administration, and, paradoxically, the maintenance of their apartment blocks and yards. This research is a rare comparison of closed and open SNS groups in an urban context, although one should keep in mind that the majority of the former (91 of 132) had denied access to the researchers.

Compared to political research, sociological research of Russian SNSs is less numerous. While political science has one of its important objects of research, namely political discourse and discussion, readily available in the form of online content, sociological focus demands more links to offline reality, which makes the overall tasks more difficult. Additionally, sociological problems of Russia generally provoke less interest from the international research community than the country's political problems.

## 19.5 Psychology and Health Studies

### *19.5.1 Health Studies*

Health studies, as a field of research lying at the intersection of sociology, policy studies, psychology and medicine, is very young in Russia, and the works in E-Health are few and mostly exploratory. A series of papers has been devoted to the VK groups of acquired immune deficiency syndrome (AIDS) denialists, people who deny the existence of AIDS or its relation to human

immunodeficiency viruses (HIV), and other AIDS-related groups. The first work by Meylakhs et al. (2014) uses netnography (an online variant of ethnography) to examine the largest community of AIDS-denialists in VK. The authors obtain a policy-relevant result that the motives of newcomers are often far from irrational and may result from negative experience with doctors or atypical medical history. In addition, persuasive strategies of the "old" community members are described which makes the authors set a new research task—to find a method that would discern the core of the community from its periphery. The significance of this task is based on the assumption that, while core members cannot be convinced to change their views, the periphery could be re-oriented. This task is addressed in Rykov et al. (2017) with the help of SNA and regression analysis. The authors indeed find a correlation between some network and activity measures, on the one hand, and the core-periphery status of a user as determined by hand coding of his/her messages, on the other. However, as in (Smirnov 2018) the set of the examined predictors is not sufficient to classify the users correctly, which is why hand coding seems to be still needed.

Meylakhs' conclusions about the motivation of AIDS-denialism newcomers echo qualitative research on coping strategies of HIV-positive people (Dudina and Artamonova 2018). The authors exploit the anonymous character of the respective Russian-language forum obtaining confessions that, in the authors' opinion, would have never been possible in face-to-face interviews. On the whole, this research describes well-known stages of coping with chronic illness and the related problems, such as shock, denial, acceptance, status disclosure and stigmatization.

An attempt to describe interests of drug users based on social network data is made in Yakushev and Mityagin (2014). The authors apply a keyword-based search while crawling Russian-language LiveJournal accounts in order to find users who write about drugs. They also exploit an LJ feature allowing users to include tags representing their interests and perform a statistical test to find out which interests are typical among those users who write about drugs, as compared to those who do not. The main problem with this approach, as the authors themselves indicate, is the lack of equivalence between those who write about drugs and those who actually use them.

Thus, studying of various online communities in Russian social media is one of the research avenues in health studies, which open up great prospects for in-depth study of the structure, communication network, and leadership phenomenon in different, including hidden and hard-to-reach, populations.

### *19.5.2 Psychology*

As mentioned earlier, psychological research based on Russian-language social media is least focused on Russia, but more often attempts to establish online-offline connections by seeking to predict psychological traits or conditions with social media data. Thus, Semenov et al. (2015) try to predict depression

propensity with the data from VK accounts, including different network metrics, and reach area under ROC (receiver operating characteristic) curve (AUC) metric of 0.84 which is comparable to other research in the field. The main problem of this research, similar to Yakushev and Mityagin (2014), is that the training set of users with depression propensity is compiled of those who contributed to discussion threads on being suicidal in depression- and suicide-related VK communities. This problem may be resolved in two different ways: by either collecting ground truth on psychological conditions outside social media, as in Panicheva et al. (2016) or by using social media data not as an indicator of "true" psychological condition but as a source of users' self-representations, as in Bogolyubova et al. (2018).

In the latter work, the authors compare Instagram images used by Russian-speaking and anglophone users to express psychological distress. The truthfulness of those expressions is thus left out; this lets the authors concentrate on the observable data and make interesting findings about significantly more frequent use of images containing text by anglophone users. The authors connect this finding to the lack of culture of verbal psychological self-expression in Russia. It should be noted that in this research the agreement between coders who manually assessed the images was not very high, which is a common problem for this type of studies based on human labeling.

Panicheva et al. (2016) develop a Facebook application to collect the ground-truth data about users' psychological traits—in their case, the so-called dark triad. They manage to obtain both the completed questionnaires and the text data from almost 2000 Russian-speaking users which is a huge number for psychological research. Using text data to predict the dark triad, the researchers, however, refuse from constructing a single model: as they are interested in evaluating the effect of each linguistic feature, not in the accuracy of prediction; they acknowledge the problem of distortion of significance levels in the models with too many predictors and apply a special correction procedure. For some reason, this problem is seldom raised outside psychological community, although it is typical for other tasks using such high-dimensional data as texts, and attention to this problem is a special value of the research by Panicheva and colleagues. At the same time, in their work, as only a limited number of user messages are available, and the sample is not described, the biases that might be introduced by these factors may be in fact more significant than those that the authors are struggling against.

An interesting extension of this research is presented in Bogolyubova et al. (2018) where the authors relate users' linguistic behavior, their psychological traits and the propensity to engage in harmful online behavior. They find out that one of the dark triad components—psychopathy—is the best predictor of such behavior. They also use a different strategy to deal with linguistic features by first representing words as word vectors (lists of words most closely associated with each given word) and then clustering them into 182 clusters. They use these clusters as harmful behavior predictors with the same procedure of significance correction as in their first paper. It should be said that, just like

topic modeling (for more, see Chaps. 23, 25 and 24), both word embeddings and clustering algorithm used are unstable, and when combined can produce an indefinite number of different solutions with the same data.

A task similar to Panicheva (2016) is addressed in Rubtsova et al. (Rubtsova et al. 2018): the authors seek to find associations between user account features in VK and the types of teenager personality accentuations as defined by Lichko (1983). A major limitation of this research is that while Lichko's classification contains 11 types of personality, the authors manage to survey only 88 teenagers. This again raises the problem of big data collapse into small data due to constrained access to one type of data needed. This problem is also present in the research by Belinskaya and Bronin (2015) which is a reduced replica of the famous study by Youyou et al. (2015). Both teams use quasi-experimental design to measure the accuracy of perception of the most important personality traits—the so-called Big Five (Piedmont 2014)—by FB or VK users, respectively. For this, they ask one group of subjects (the assessed) to fill in the Big Five questionnaire, while the other group of subjects (assessors) are asked to fill in the same questionnaire on behalf of the assessed subjects. The differences between the two studies are, however, more significant than their similarities. First, Kosinski's team tests the accuracy of those who know the assessed people well, while Belinskaya and Bronin focus on those who have only met their friends online. Second, Kosinski and colleagues, by developing and promoting a FB application, manage to collect 86,220 observations, while the Russian authors collect only 30 offline assessments from 15 assessors. This turns the problem of big data collapse into a problem of digital divide in science: while collecting big data online seems cheap at the first glance, this is not the case in practice. On the contrary, substantial financial resources and time are needed to conduct large-scale research with social media data.

## 19.6 Conclusion

In this chapter we reviewed both the works on the Russian-language social media and the Russia-related topics that can be studied with social media data in general. We have shown that Russian SNSs give very broad opportunities for research—broader than most international SNSs do. However, this potential stays somewhat underused due to a number of factors, including the lack of resources for researchers within Russia and the lack of interest to the opportunities given by the Russian SNSs from the international scholars. The sphere that generates the largest interest from the international researchers is Russian politics, and this is reflected in the dominant position of this topic among Russian SNS-based studies. Sociological research is somewhat fragmented, and psychology studies are least of all related to Russia, with some strong studies done by the Russian scholars not using Russian data at all (Buraya et al. 2018).

In our review we focused both on the opportunities and problems of social media research. Our goal was to go beyond the strengths and limitations of

concrete works and to highlight the common trends, especially the limitations of the field because they are seldom spoken about in research papers, since they tend to report success rather than failures.

The opportunities include the ability to obtain large observational data collected in a non-intrusive manner and the ability to scale the research that otherwise would be bound to very small laboratory experiments or qualitative field work. Additionally, the fact that the data of the Russian-language SNSs come mostly from the Post-Soviet space gives an opportunity to study various political, social and psychological phenomena outside the Western context where most social research data come from. Finally, social media are an important key to a society where other types of data are often less available than in more transparent countries.

The limitations are, however, also large. First, online digital traces, in order to be meaningful, often have to be combined with other types of data that are not so easy to collect and that become a bottleneck on the way to large samples. This is where we observe the effect of big data collapse. Second, SNS data have various problems of representativity in terms of their ability to represent both offline and online phenomena. Sampling network data and especially textual data is generally a poorly developed methodological area, while these types of data are the core of digital traces left by humans on social media.

Finally, methods of SNS data analysis are lagging behind the techniques available for data collection. The existing approaches are very complex, and they hide many caveats that social scientists are often unaware of. Instability of the majority of text-clustering techniques, absence of statistical inference methods for non-independent (networked) data, lack of approaches to work with power-law distributions so common for SNS data compromise the validity of many of the existing studies without social scientists being fully able to grasp the scale of the problem. Nevertheless, an open discussion of these methodological difficulties can enrich our understanding of the field of social media research and enhance its development.

**Acknowledgments** This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

## References

Alexandrov, Daniel, Viktor Karepin, Ilya Musabirov, and Daria Chuprina. 2018. Educational Migration from Russia to the Nordic Countries, China and the Middle East. Social Media Data. In *Companion of the The Web Conference 2018—WWW'18*, 49–50. Lyon, France: ACM Press. https://doi.org/10.1145/3184558.3186923.

Badawy, Adam, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265. IEEE.

Barash, Vladimir, and John Kelly. 2012. Salience vs. Commitment: Dynamics of Political Hashtags in Russian Twitter. In *SSRN Scholarly Paper ID 2034506*. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2034506.

Belinskaya, E.P., and I.D. Bronin. 2015. Accuracy of Interpersonal Perception in Mediated Contacts in Social Media. *Social Psychology and Society* 6 (4): 91–108. https://doi.org/10.17759/sps.2015060407.

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

Bodrunova, Svetlana S., Anna A. Litvinenko, and Ivan S. Blekanov. 2018. Please Follow Us: Media Roles in Twitter Discussions in the United States, Germany, France, and Russia. *Journalism Practice* 12 (2): 177–203. https://doi.org/10.1080/17512786.2017.1394208.

Bogolyubova, Olga, Philipp Upravitelev, Anastasia Churilova, and Yanina Ledovaya. 2018. Expression of Psychological Distress on Instagram Using Hashtags in Russian and English: A Comparative Analysis. *Sage Open* 8 (4): 2158244018811409. https://doi.org/10.1177/2158244018811409.

Brand Analytics. 2018. SNS in Russia (Report). *Brand Analytics 2018*. https://br-analytics.ru/blog/wp-content/uploads/2018/12/Sotsseti-Rossiya-osen-2018.pdf.

Bulovsky, Andrew. 2019. Authoritarian Communication on Social Media: The Relationship between Democracy and Leaders' Digital Communicative Practices. *International Communication Gazette* 81 (1): 20–45. https://doi.org/10.1177/1748048518767798.

Buraya, Kseniya, Aleksandr Farseev, and Andrey Filchenkov. 2018. Multi-view Personality Profiling Based on Longitudinal Data. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ed. Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, vol. 11018, 15–27. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-98932-7_2.

Dudina, Victoria I., and Kristina N. Artamonova. 2018. Stigmatization of People Living with HIV/AIDS and the Problem of Disclosure: Analysis of the Online Forum Discussions. *Sociology* 11 (1): 66–78. https://doi.org/10.21638/11701/spbu12.2018.106.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2018. Social Media and Protest Participation: Evidence from Russia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2696236.

Filer, Tanya, and Rolf Fredheim. 2016. Sparking Debate? Political Deaths and Twitter Discourses in Argentina and Russia. *Information Communication & Society* 19 (11): 1539–1555. https://doi.org/10.1080/1369118X.2016.1140805.

Forbes.ru. 2019. Vremâ Edinorogov. Kak v Runete sozdaetsâ biznes stoimost'û $1 mlrd | Tehnologii. Forbes.Ru, February 21, 2019. https://www.forbes.ru/tehnologii/372571-vremya-edinorogov-kak-v-runete-sozdaetsya-biznes-stoimostyu-1-mlrd.

Gel'man, V. 2014. The Rise and Decline of Electoral Authoritarianism in Russia. *Demokratizatsiya* 22 (4): 503.

Goncharov, D.V., and V.V. Nechay. 2018. Anti-Corruption Protests 2017: Reflection in Twitter. *The Journal of Political Theory, Political Philosophy and Sociology of Politics Politeia* 88 (1): 65–81. https://doi.org/10.30570/2078-5089-2018-88-1-65-81.

Kaveeva, Adelia, and Konstantin Gurin. 2018. 'VKontakte' Local Friendship Networks: Identifying the Missed Residence of Users in Profile Data. *The Monitoring of Public Opinion: Economic&Social Changes* 3: 78–90. https://doi.org/10.14515/monitoring.2018.3.05.

Kaveeva, Adelia, Konstantin Gurin, and Valery Solovyev. 2018. How 'VKontakte' Fake Accounts Influence the Social Network of Users. *International Conference on Digital Transformation and Global Society*, 492–502. Springer.

Kelly, John, Vladimir Barash, Karina Alexanyan, Bruce Etling, Robert Faris, Urs Gasser, and John G. Palfrey. 2012. Mapping Russian Twitter. In *SSRN Scholarly Paper ID 2028158*. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2028158.

Kisilevich, Slava, Chee Siang Ang, and Mark Last. 2012. Large-Scale Analysis of Self-Disclosure Patterns among Online Social Networks Users: A Russian Context. *Knowledge and Information Systems* 32 (3): 609–628. https://doi.org/10.1007/s10115-011-0443-z.

Kollanyi, B., P.N. Howard, and S.C. Woolley. 2016. Bots and Automation Over Twitter during the First US Presidential Debate. *Comprop Data Memo* 1: 1–4.

Koltsova, Olessia, and Sergei Koltcov. 2013. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal. *Policy and Internet* 5 (2): 207–227. https://doi.org/10.1002/1944-2866.POI331.

Koltsova, Olessia, and Galina Selivanova. 2019. Explaining Offline Participation in a Social Movement with Online Data: The Case of Observers for Fair Elections. *Mobilization: An International Quarterly* 24 (1): 77–94. https://doi.org/10.17813/1086-671X-24-1-77.

Lichko, A.E. 1983. *Psihopatii i akcentuacii haraktera u podrostkov*. Medicina.

Mediascope. 2018. WEB-Index: The Audience of Internet Projects. The Results of the Measurement: Desktop, November 2018, Russia 0+. https://mediascope.net/data/?download=3179&date=2018%2011&set_filter=Y&FILTER_TYPE=internet.

Meylakhs, Peter, Yuri Rykov, Olessia Koltsova, and Sergey Koltsov. 2014. An AIDS-Denialist Online Community on a Russian Social Networking Service: Patterns of Interactions with Newcomers and Rhetorical Strategies of Persuasion. *Journal of Medical Internet Research* 16 (11): e261. https://doi.org/10.2196/jmir.3338.

Panicheva, Polina, Yanina Ledovaya, and Olga Bogolyubova. 2016. *Lexical, Morphological and Semantic Correlates of the Dark Triad Personality Traits in Russian Facebook Texts*. In 2016 IEEE artificial intelligence and natural language conference (AINL); 1-8. IEEE.

Petrova, Marina, Aleksandra Nenko, and Kirill Sukharev. 2016. Urban Acupuncture 2.0: Urban Management Tool Inspired by Social Media. In *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, 248–257. ACM.

Piedmont, R.L. 2014. Five Factor Model of Personality. In *Encyclopedia of Quality of Life and Well-Being Research*, ed. A.C. Michalos. Dordrecht: Springer.

Reuter, Ora John, and David Szakonyi. 2015. Online Social Media and Political Awareness in Authoritarian Regimes. *British Journal of Political Science* 45 (1): 29–51. https://doi.org/10.1017/S0007123413000203.

Rubtsova, O., A.S. Panfilova, and V.K. Smirnova. 2018. Research on Relationship between Personality Traits and Online Behaviour in Adolescents (With VKontakte Social Media as an Example). *Psihologičeskaâ Nauka I Obrazovanie [Psychological Science and Education]* 23 (3): 54–66. https://doi.org/10.17759/pse.2018230305.

Rykov, Yuri, Peter A. Meylakhs, and Yadviga E. Sinyavskaya. 2017. Network Structure of an AIDS-Denialist Online Community: Identifying Core Members and the Risk Group. *American Behavioral Scientist* 61 (7): 688–706. https://doi.org/10.1177/0002764217717565.

Rykov, Yuri, Yadviga Sinyavskaya, and Olessia Koltsova. 2018. Accumulating Social Capital in an Online Urban Network: The Effects of User Behaviors. *SSRN Electronic Journal*. Higher School of Economics Research Paper No. WP BRP 83, 1–27. https://doi.org/10.2139/ssrn.3301266.

Sanovich, Sergey. 2017. Computational Propaganda in Russia: The Origins of Digital Misinformation. *Working Paper, Computational Propaganda Research Project*. Oxford Internet Institute. http://www.philosophyofinformation.net/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf.

Semenov, Aleksandr, Alexey Natekin, Sergey Nikolenko, Philipp Upravitelev, Mikhail Trofimov, and Maxim Kharchenko. 2015. Discerning Depression Propensity Among Participants of Suicide and Depression-Related Groups of Vk.Com. In *Analysis of Images, Social Networks and Texts*, ed. Mikhail Yu, Natalia Konstantinova Khachay, Alexander Panchenko, Dmitry Ignatov, and Valeri G. Labunets, vol. 542, 24–35. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26123-2_3.

Smirnov, Ivan. 2018. Predicting PISA Scores from Students' Digital Traces. *Twelfth International AAAI Conference on Web and Social Media*.

Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. 2017. Detecting Bots on Russian Political Twitter. *Big Data* 5 (4): 310–324. https://doi.org/10.1089/big.2017.0038.

Voskresenskiy, Vadim, Ilya Musabirov, and Daniel Alexandrov. 2016. Private and Public Online Groups in Apartment Buildings of St. Petersburg. In *Proceedings of the 8th ACM Conference on Web Science—WebSci'16*, 301–306. Hannover, Germany: ACM Press. https://doi.org/10.1145/2908131.2908173.

Yakushev, Andrei, and Sergey Mityagin. 2014. Social Networks Mining for Analysis and Modeling Drugs Usage. *Procedia Computer Science* 29: 2462–2471. https://doi.org/10.1016/j.procs.2014.05.230.

Youyou, Wu, Michal Kosinski, and David Stillwell. 2015. Computer-Based Personality Judgments are More Accurate than Those Made by Humans. *Proceedings of the National Academy of Sciences* 112 (4): 1036–1040.

Zamyatina, Nadezhda, and Alexey Yashunsky. 2018. Virtual Geography of Virtual Population. *The Monitoring of Public Opinion: Economic&Social Changes* 1: 117–137. https://doi.org/10.14515/monitoring.2018.1.07.