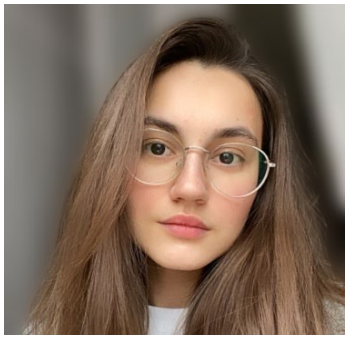
УДК 311.2

MEASURING SOCIAL WELL-BEING OF CITY DWELLERS BY DIGITAL MARKERS: TESTING THE METHODOLOGY



Polina Dymova

Tula State University (Tula, Russia); Dymova.pi@gmail.com

Anna Dombrovskaia

DSn (Soc. Sci.), Professor of the Department of Political Science, Director of the Center for Political Studies of the Financial University under the Government of the Russian Federation (Moscow, Russia); an-doc@yandex.ru

Abstract. *The article presents a method for measuring subjective well-being using digital markers based on data from online city communities from the VKontakte social network. Publications from online groups in Tula and the Tula region for the year 2023 were used in the study. The collected data was processed using the Python 3 programming language and additional libraries (pandas, spaCy, dostoevsky). As a result, information was obtained about the categorization of textual documents into subjective well-being categories, as well as their sentiment. In conclusion, the limitations of the proposed method are discussed, along with possible ways to overcome them to improve the technique for measuring subjective well-being using cybermetric analysis methods.*

Keywords: *social well-being; city community; online community; cybermetrics; digital markers.*

For citation: Dymova P.I., Dombrovskaia A.Yu. Measuring social well-being of city dwellers by digital markers: testing the methodology // Social Novelties and Social Sciences. – 2025. – N 2. – P. 36–46.

URL: https://snsen-journal.ru/archives DOI: 10.31249/snsneng/2025.02.03

Introduction

Social well-being is one of the many indicators of society's overall health, which is regularly monitored by both large public opinion centers and individual research groups. In the scientific literature, there is a large number of works devoted to study social well-being in a country or its integral part, in socio-demographic or professional groups.

An analysis of the available studies of social well-being allows to track some of the problems associated with evaluating it. First, in sociological science, there is no single definition and set of indicators for measuring social well-being. In a general sense, social well-being means an individual's subjective assessment of their own condition and position in society: how it is now and what they hope for in the future. A different set of objective and subjective indicators is used to measure well-being. Examples of objective indicators include: the level of material well-being and the state of health of an individual [Simonovich, 2003, p. 5–6], the real social status in various spheres of life [Mikhailova, 2010, p. 47] and the level of power [Dushatskii, 2004, p. 67–69]. Examples of subjective indicators are: an individual's confidence in the future [Grachev, Rusalinova, 2007, p. 9–10], satisfaction with the chosen life strategy [Petrova, 2000, p. 53], perception, assessment and attitude towards one's social status [Mikhailova, 2010, p. 47] and others.

The above indicators, both objective and subjective, certainly help to assess the social well-being of an individual in particular and a specific group as a whole, but can hardly explain the reasons for a particular assessment. A healthy and well-paid respondent may feel uncertain about the future and be dissatisfied with the way his life is going. Another, who has health problems and fewer material benefits, may assess life as quite happy and not worry about the future. It is also worth noting that most of the measured indicators, especially subjective ones, need additional operationalization and can be interpreted by each participant of the study in their own way [Rogozin, 2007, p. 111].

The second problem associated with the study of social well-being is manifested in a large number of sociological concepts that are synonymous in both form and content: social well-being and disadvantage, quality and standard of living, social mood, etc. On the one hand, this allows the researcher to discover interesting methods and approaches that can be translated into the field of studying social well-being, and on the other hand, it returns to the first, previously mentioned difficulty.

And, finally, the third problem is the research method used to assess social well-being. The most commonly used tool is a survey conducted in person or by remote interviews. The use of the survey methodology not only does not allow us to find out what stands behind the answer given by the respondent, but also does not exclude the influence of the interviewer on the answers.

The resolution of the first two indicated difficulties is not possible at the moment. However, the analysis of unprovoked statements in social networks by digital markers seems interesting and promising [Development of methodology and methods ..., 2017, p. 83–84]. The advantage of cybermetric methods for analyzing social and political processes lies not only in the abundance of data left by users of social networks, but also in the availability of ready-made tools for monitoring social media.

The study of user activity in online communities is widely praciced in public administration. In 2020, a Regional Management Center (hereinafter referred to as the RMC) was established in each constituent entity of the Russian Federation to promptly respond to urban planning problems. According to A. Kurmanov, the administrator of the Development Feedback Platform, the main function of an RMC is to anticipate the needs of citizens. Thanks to the analysis of the appeals of the online community audience, it becomes possible to conduct deep analytics by regions, create «heat maps», identify typical problems, and unify and optimize ways to solve them. Such analytical data can fully serve as a basis for decision-making by regional authorities and for the creation of regional development programs, taking into account the most important areas of urban development for citizens [Bolshakova, Klimova, 2022, p. 394].

Satisfaction with infrastructure is a significant factor in measuring the social well-being of the urban population [Tsvetkova, 2017, p. 115]. The degradation of urban infrastructure, according to the author, leads to a gradual decrease in social well-being. Long-term dissatisfaction with life and a generally negative atmosphere within a particular city or region can lead to an outflow of residents in an attempt to find a more suitable and comfortable place to live, as well as to a decline in psycho-emotional and physical health, an increase in crime, and the spread of deviant behavior.

As noted earlier, the sociological literature contains a fairly large number of concepts synonymous with social well-being. E.V. Shchekotin, M.G. Myagkov, and others used online activity data from VKontakte users to calculate an index of subjective well-being/ disadvantage in 43 constituent entities of the Russian Federation [Subjective Assessment of ..., 2020, p. 92–93].

Description of the methodology and method

In the context of the conducted study, the approach developed by the team of authors E.I. Golovakha, N.V. Panina, and A.P. Gorbachik was chosen to operationalize the concept of social well-being. This approach consists of measuring the integral indicator of social well-being based on judgments about the sufficiency of social benefits in 11 spheres of an individual's life [Golovakha, 1998, p. 49]. In the original study, a sufficient scale was used to assess the sufficiency of each component of social well-being, including the assessments «not enough», «difficult to say whether enough or not», «enough», and «not interested».

It's also worth noting that to measure the integrated index of social well-being researchers use questionnaires, either in an expanded or abbreviated format. Since the study did not involve a survey method,

it was decided to judge the adequacy of each component of social well-being based on the sentiment of the analyzed document. A positive sentiment indicates adequacy, a negative sentiment indicates a lack of social benefits, and a neutral sentiment is equivalent to the responses «difficult to say» and «not interested», which were also combined and assigned the same score in the original methodology [Golovakha, Panina, Gorbachik, 1998, p. 50–51].

The next stage in the study involved selecting online communities for document collection within the social network VKontakte. A manual selection process conducted in March 2024 resulted in a list of 28 communities in Tula and the Tula region. Two groups were then excluded because they did not fit the «urban community» criteria. Communities were identified based on the following criteria:

1. A sufficiently large number of subscribers relative to the municipality's population as of 2023.

2. Activity and the presence of recently published posts at the time of selection.

When selecting communities, we primarily focused on finding online groups of the «urban community» type. Unlike newsgroups, urban online communities contain more emotionally charged posts. Below is a table showing the distribution of communities by administrative-territorial divisions of the Tula region.

Table 1

The number of communities belonging to the administrative-territorial units within the Tula region*

Name of the administrative-territorial unit	Number of communities
Tula	10
Novomoskovsk	2
Yasnogorsk	2
Aleksin	1
Kimovsk	1
Efremov	1
Kireevsk	1
Belev	1
Uzlovaya	1
Shchekino	1
Not defined	5

* Source: compiled by the authors.

The next step was to develop software code in Python 3 to automate extraction of information about communities and their publications, as well as to normalize texts for further analysis.

Data on the selected VKontakte communities was obtained using VK API methods accessed through the «requests» module. The groups.getById method was used to retrieve characteristics such as the number of community members, ID, city, and others. The wall.get method was used to generate a list of publications for each selected community, yielding the following: publication date and time, text and attachments (photos and videos), author ID, number of views, comments, likes, and reposts, and much

more. Full details on the returned fields can be found in the documentation on the social network's official website.

The total dataset downloaded from the selected communities consisted of 472,855 publications: the latest document was dated April 14, 2024, the earliest – March 3, 2009. To calculate the social well-being index, the total number of 52,810 records from 2023 were selected.

As part of the preliminary analysis, 1,197 documents published between 2024 and 2021 were selected from the entire dataset. The objectives of the preliminary analysis were as follows:

1. Associate the text of each publication with a category of social well-being within the chosen approach.

2. Identify two categories from which publications appear most frequently.

3. Compile dictionaries of social well-being markers for a specific category and supplement them with keywords if necessary.

As a result, among the documents analyzed: 581 publications were marked as «garbage» due to containing inappropriate statements, 268 publications were news, 25 were job postings, 14 were advertising texts, 13 were missing persons postings, and four were duplicates of previously encountered posts. The remaining documents were distributed among social well-being categories as follows (Fig. 1).

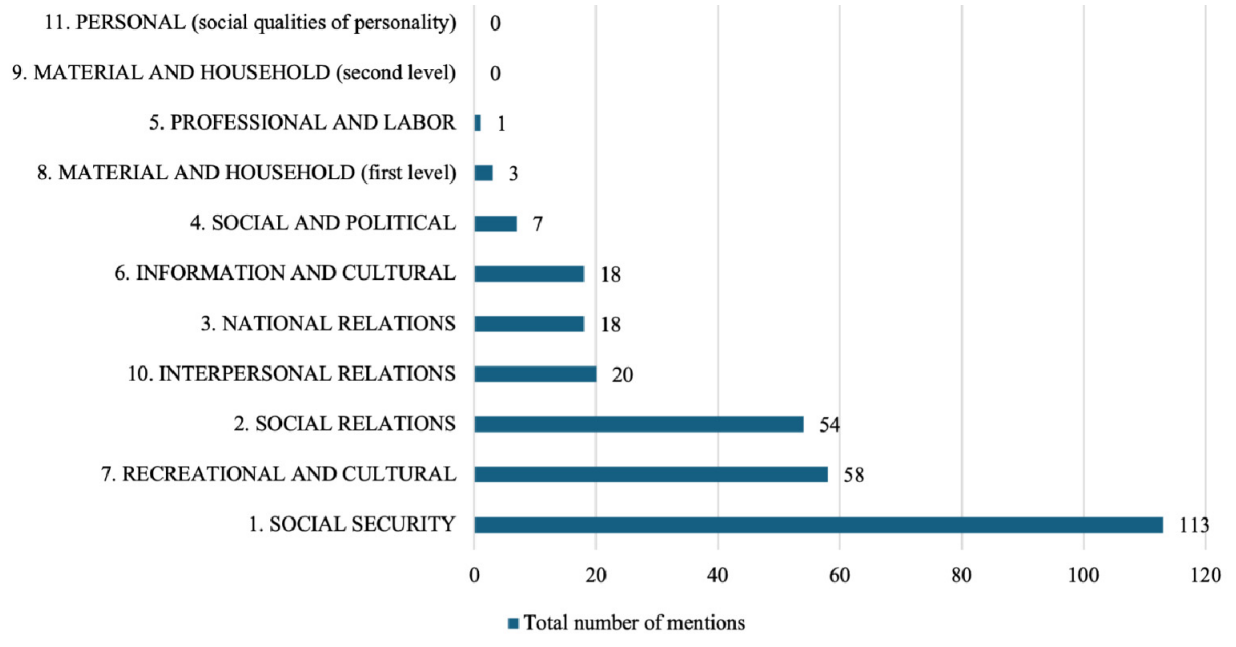


Fig. 1. Distribution of publications by categories of social well-being

* The integrated social well-being index has two material and household categories: levels 1 and 2. Level 1 includes: adequate housing, adequate clothing, ability to purchase essential food and furniture; level 2 includes: car, ability to eat according to one's tastes, fashionable and attractive clothing, and a garden plot.

Despite the fact that the recreational and cultural component of social well-being was in second place in terms of the number of publications, it was decided to conduct further analysis in the categories

of «social security» and «social relations», since the documents that fell into the category of «recreational and cultural» were neutral in nature and were more like announcements of cultural events.

To search for digital markers in the target array of documents, the text of publications was subjected to standard natural language processing procedures, which included:

1. Removing punctuation, hyphens, emojis, and any other additional symbols from the text.
2. Lowercase the text.
3. Tokenization and lemmatization.

The tokenization procedure is the division of the text into separate words and punctuation marks, if it has not been previously cleaned of them. The lemmatization procedure is the reduction of a word to its initial morphological form. Such processing simplifies the search for digital markers in text documents, since when compiling a dictionary, there is no need to use different forms of the same words, for example, declension forms of nouns, in order to take into account all possible variations of word forms.

The lemmatization procedure is often mentioned either together with or instead of the stemming procedure. Unlike lemmatization, in stemming, the basic form of a word is isolated by removing suffixes and endings. Below (Table 2) is an example illustrating the differences in stemming and lemmatization procedures on one of the selected publications.

Table 2

Comparison of the results of stemming and lemmatization procedures*

Cleaned text	Result of Stemming	Result of Lemmatization
масочный режим в учреждениях здравоохранения тульской области введен с сегодняшнего дня для посетителей и медицинского персонала	масочн реж в учрежден здравоохранен тульск област введ с сегодняшн дня для посетител и медицинск персона	масочный режим в учреждение здравоохранение тульский область ввести с сегодняшний день для посетитель и медицинский персонал

* Source: compiled by the authors.

In this work, the text of publications obtained after lemmatization appears to be the most convenient for searching for markers, and it is this text that is subsequently used to classify a document into a particular category of social well-being.

A search for markers from the dictionaries compiled for the categories «social security» and «social relations» yielded the following distribution of data (Figure 2): 20% of publications for 2023 fell into the «social security» category, 3% into the «social relations» category, 4% fell into both categories, and 73% were unclassified. Publications that did not fall into either social well-being category could be either «junk data» or those that could theoretically fall into the nine remaining categories.

The overlapping of categories within a single document may be due to the publication containing markers related to both categories of social well-being. This dataset will not be excluded from further analysis.

To determine the sentiment of messages, the dostoevsky library, designed for sentiment analysis of Russian-language texts, was used. The model used in this library was trained on the RuSentiment dataset, consisting of public posts from the social network VKontakte [RuSentiment…, 2018]. After processing the text, the model returns the probability with which the text belongs to a particular sentiment. The model's developers distinguish five sentiment categories: positive, negative, neutral, speech, and skip. The model skips posts in which the sentiment is difficult to determine, as well as jokes, meaningless statements, and texts not in Russian.

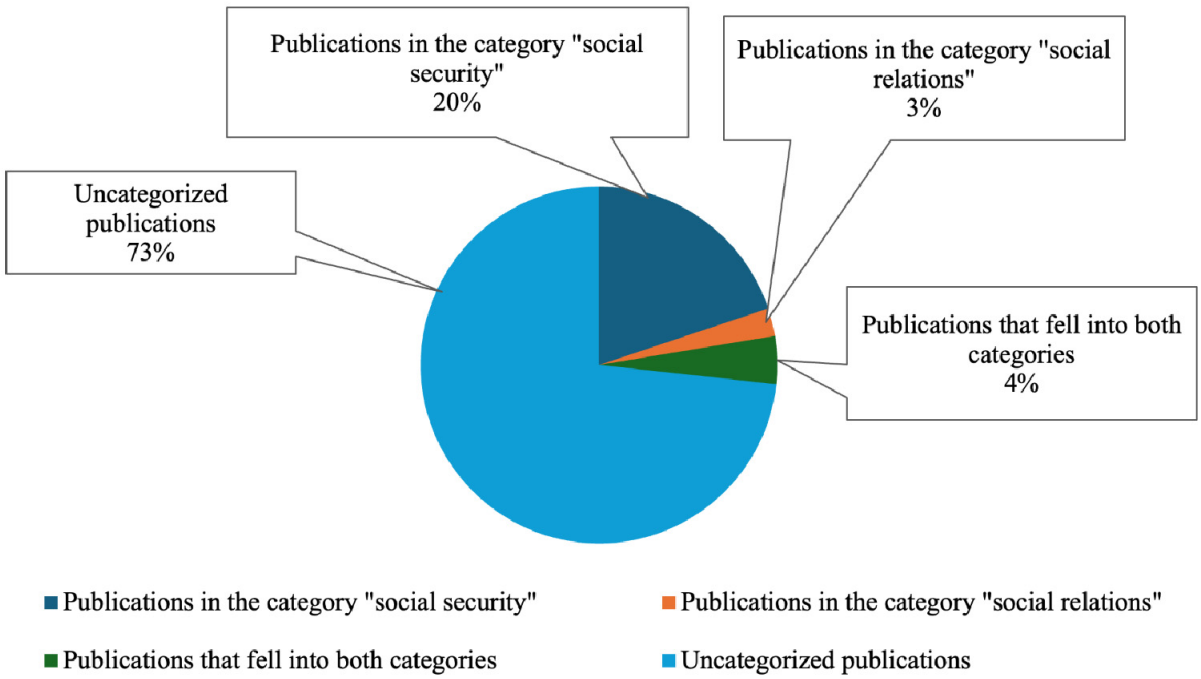


Fig. 2. Distribution of publications by categories of social well-being.

When processing text with the model, it is necessary to specify the parameter k, which takes values from 1 to 5 and shows the «confidence» of the model in the predicted sentiment. In this study, k = 1 was used to obtain one, the most probable, definition of sentiment. An increase in k in the calculations does not lead to a recalculation of the sentiment probability. When k = 2, the second most probable sentiment variant is given as a result of text processing, similarly, when k increases, the third, fourth, and fifth variants appear.

As a result of data processing, the sentiment of the overwhelming majority of publications was determined as neutral. In second place, but by a wide margin, are publications with a negative sentiment, with the exception of text documents in the «social relations» category. There, following the neutral sentiment, is a verbal statement (Fig. 3).

Next, the distributions of the number of publications and their sentiment by month in 2023 were compiled. Outbursts of activity are clearly well defined in publications with a neutral sentiment. Thus, most of the text documents in the 'social security' category fall on April and October 2023. At about the

same time, there is an increase in the number of publications in the 'social relations' category: May and November 2023. For publications in both categories, the peak of activity falls on March 2023.

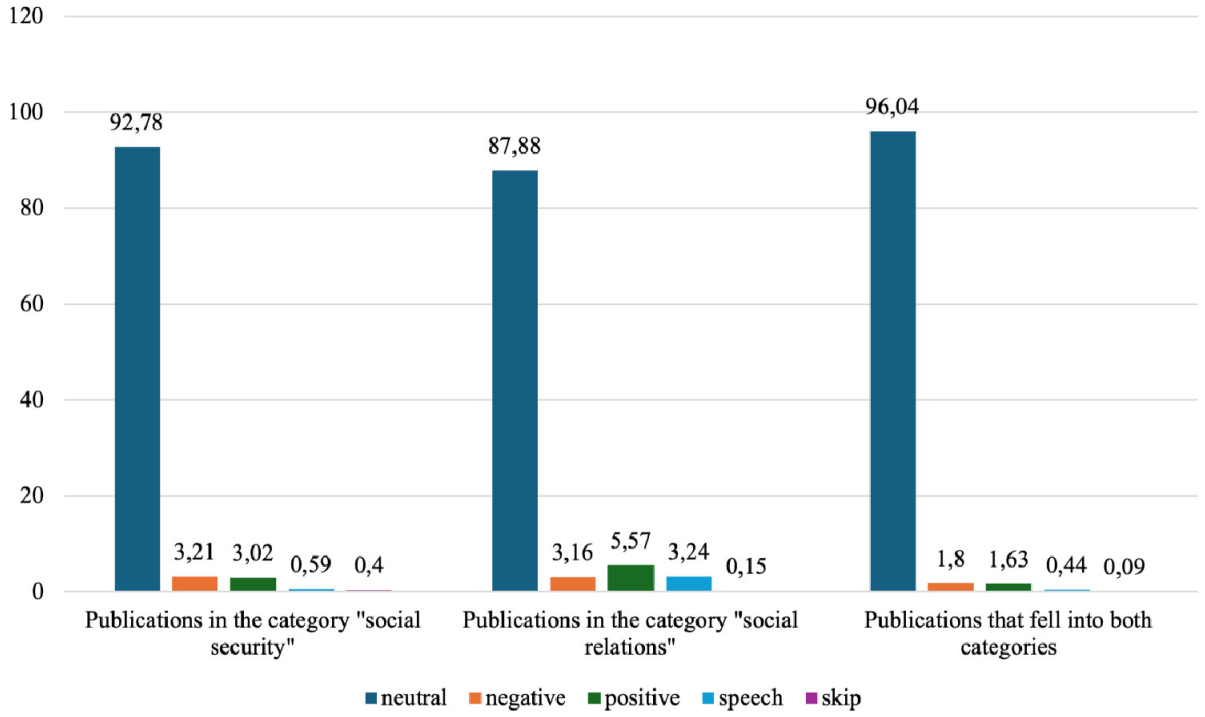


Fig. 3. Distribution of publications by sentiment within the category (relative indicators).

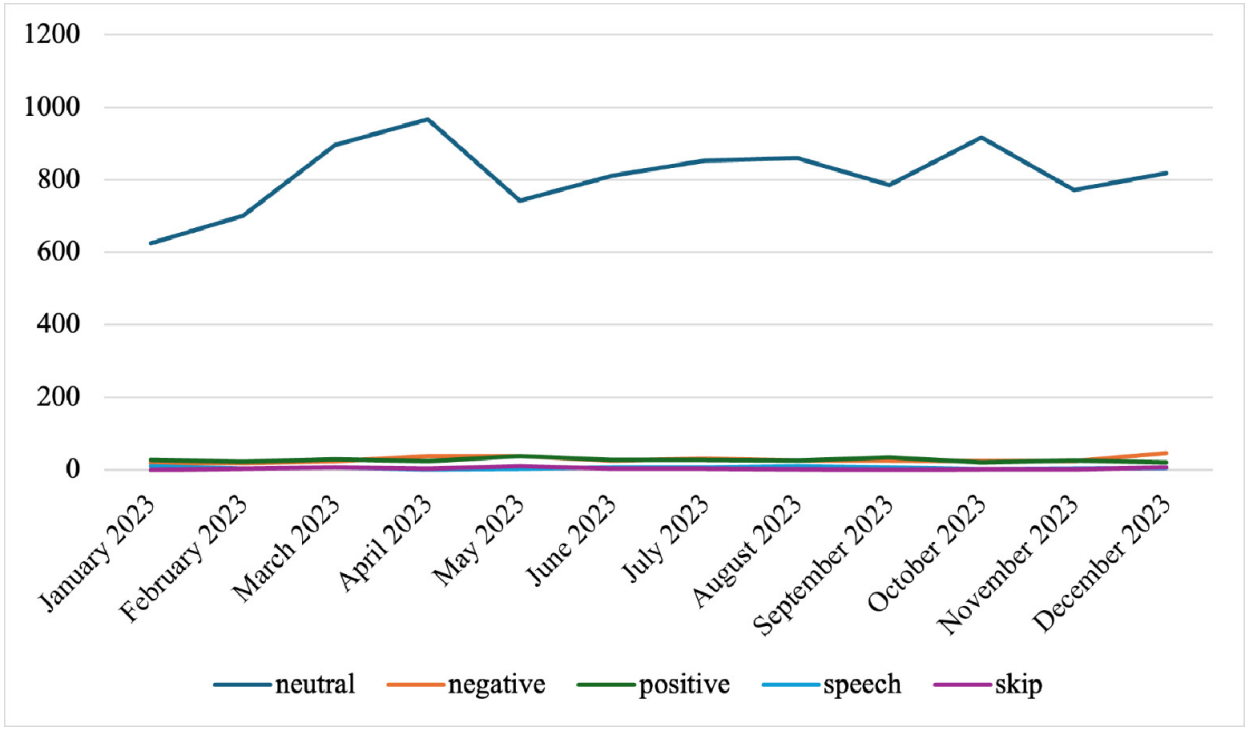


Fig. 4. Distribution of sentiment by month for publications in the «social security» category.

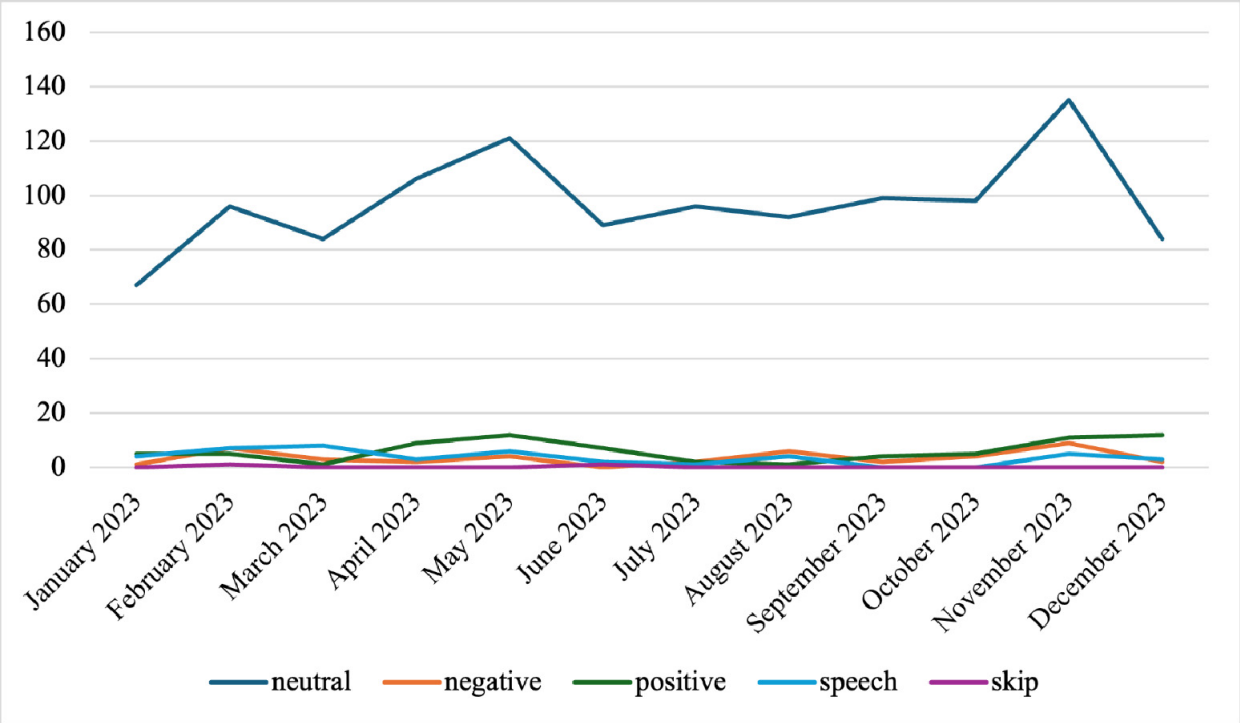


Fig. 5. Distribution of sentiment by month for publications in the «social relations» category.

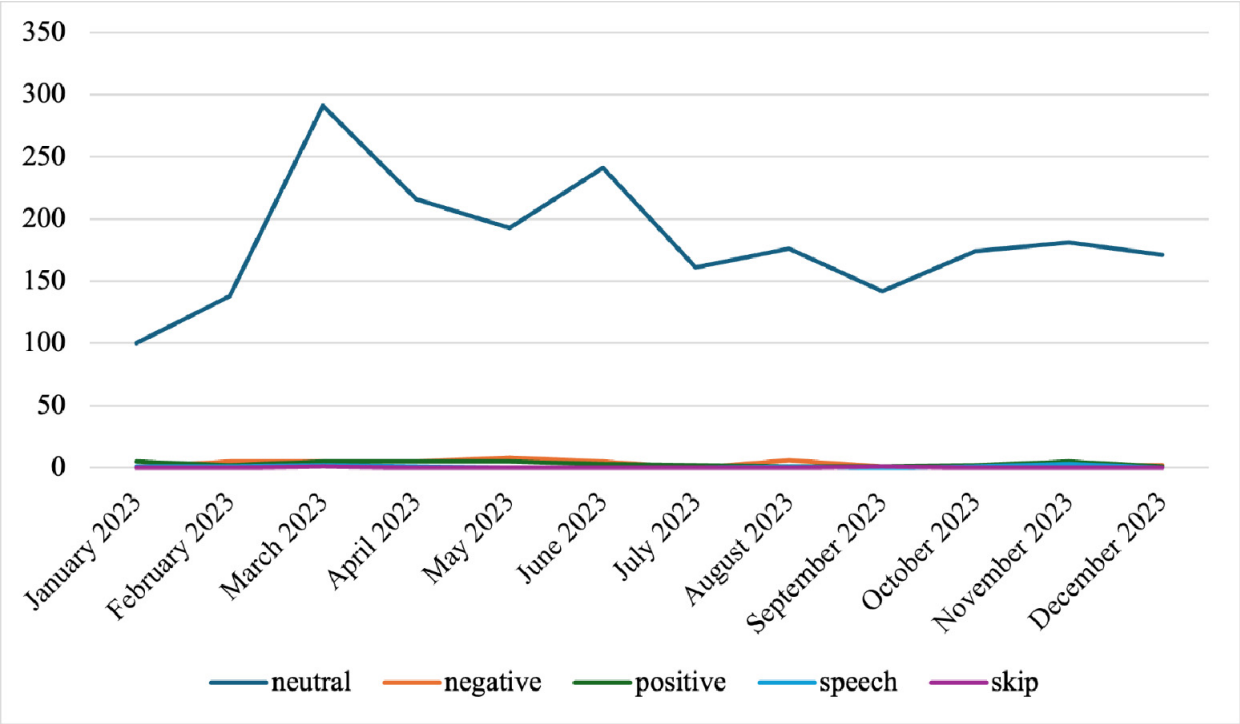


Fig. 6. Distribution of sentiment by month for publications that fell into both categories.

Results and discussion

Testing the methodology revealed several issues that require solutions to further improve cybermetric procedures in analyzing social well-being using digital markers. The first issue concerns the ambiguous classification of a text document into a particular social well-being category. Even with non-overlapping marker dictionaries, a text document may contain keywords from several dictionaries. Perhaps the assumption that every text document can be unambiguously assigned to a specific social well-

being category is inherently flawed, and a more appropriate approach is to determine the probability of a publication belonging to each group from the existing set.

The second problem relates to sentiment determination. The overwhelming majority of documents were characterized as neutral, which makes it difficult to definitively assess social well-being within the categories studied. It was previously noted that, within the approach used, the adequacy of social benefits was determined by scale values: «sufficient», «insufficient», «difficult to say whether sufficient or not», and «not interested». It was assumed that the ratings «sufficient» and «insufficient» correspond to positive and negative sentiments, while neutrality corresponds to the remaining two. This creates a contradiction: on the one hand, the fact that a large number of publications fell into the categories of «social security» and «social relations» during the preliminary analysis suggests the importance of these topics in the daily lives of city residents; on the other hand, it is difficult to determine the explicit emotional connotations of statements contained in them. This may be a shortcoming of the model, which, despite considerable accuracy, fails to capture subtle emotional nuances in the text. One way to address this problem is by developing a custom model trained on data used to analyze social well-being and related indicators. A less radical, but still interesting, option is to consider another approach. The sentiment model used in this study, as well as models used in other studies, excludes emojis from the analysis, which are frequently used both in private messages with loved ones and in the public space of online communities. Python libraries exist that convert emojis to the word they represent. Since emojis represent various emotional reactions, explicitly referring to laughter, joy, anger, or sadness in a text can influence the sentiment of a text. Another extension of this method is the use of comments under a post, if any. Currently, they are excluded from the analysis as units of text and are considered more as an indicator of community member activity [Subjective Assessment of..., 2020, p. 92–93].

The third issue is the need for further research into the temporal distribution of publications. Even documents with a neutral sentiment exhibit peaks and troughs in activity, and, as a result, rises and falls in the number of posts in online communities. The current study did not identify the events that influenced fluctuations in publication activity. Special attention must be given to domestic and foreign policy events that impact online communities.

Another issue worth highlighting is a technical limitation related to the structure of the online community publication data obtained from the VKontakte social network. This limitation stems from the way the author of a publication is identified within the group. Most posts are published on behalf of the community, and actual authors of proposed entries may only be identified if they choose to forego anonymity. Of all the data collected for 2023, only 9,628 (18,23%) contained author information; the remaining 43,182 (81,77%) were published anonymously. This complicates the compilation of a demographic profile of authors. However, it remains possible to analyze the age and gender characteristics of community members as a whole.

Conclusion

In today's research landscape, there is a significant diversification of strategies and methods for collecting and analyzing empirical data. Practicing empiricists are increasingly using automated database accumulation and processing. Sociological research is one of the first areas of social and humanities knowledge to increasingly utilize automation, programming, and machine learning algorithms to study digital traces of social sentiment, attitudes, values, opinions, and, ultimately, social well-being. This article presents a specific empirical case that demonstrates the capabilities and limitations of software for measuring the social well-being of Tula residents based on their online network markers. Key advantages of this and other automated formats for collecting and processing data include the ability to conduct comprehensive analysis within the required timeframe and social media platform, time savings, high data accuracy, and the absence of negative human factors (fatigue, inaccuracy, and errors during data collection). Among the limitations of automated formats for analyzing social processes we find the inevitable simplification while verifying the studied characteristics, and sometimes the simplification of indicators of the measured parameters due to the need to develop an algorithm for extracting relevant information. The prospects for solving these problems lie in improving data collection and processing methods using machine learning (for example, supervised learning).

References

1. *Bolshakova K.Yu., Klimova A.V.* Regional management centers as a new form of management activity // Vestnik Rossiiskogo universiteta druzhby narodov. Seriia Gosudarstvennoe i munitsipal'noe upravlenie. – 2022. – Vol. 9, N 4. – P. 394. (in Russian).
2. *Development of methodology and methods of intellectual search for digital markers of political processes in social media* / Brodovskaia E.V., Dombrovskaia A.Yu., Karzubov D.N. [et al.] // Monitoring obshchestvennogo mnieniia : Ekonomicheskie i sotsial'nye peremeny. – 2017. – N 5. – P. 79–104. (in Russian).
3. *Dushatskii L.E.* Material and power resources of Russians in self-esteem and social well-being // Sotsiologicheskiie issledovaniia. – 2004. – N 4. – P. 64–70. (in Russian).
4. *Golovakha E.I., Panina N.V., Gorbachik A.P.* Measuring social well-being: the IISS test // Sotsiologiia : metodologiia, metody, matematicheskoe modelirovanie (Sotsiologiia: 4M). – 1998. – N 10. – P. 45–71. (in Russian).
5. *Grachev A.A., Rusalinova A.A.* Social well-being of a person in an organization // Izvestiia RGPU im. A.I. Gertsena. – 2007. – N 30. – URL: https://cyberleninka.ru/article/n/sotsialnoe-samochuvstvie-cheloveka-v-organizatsii (accessed 01.03.2024). (in Russian).
6. *Mikhailova L.I.* Russians' social well-being and perception of the future // Sotsiologicheskiie issledovaniia. – 2010. – N 3. – P. 45–50. (in Russian).
7. *Petrova L.E.* Social well-being of young people // Sotsiologicheskiie issledovaniia. – 2000. – N 12. – P. 50–55. (in Russian).
8. *Rogozin D.* Testing questions about social well-being // Sotsial'naia real'nost'. – 2007. – N 2. – P. 97– 113. (in Russian).
9. *RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian* / Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. // Proceedings of the 27th International Conference on Computational Linguistics. – Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. – P. 755–763.
10. *Simonovich N.E.* Social well-being of people and its research technologies in modern Russia: abstract of the dissertation of Doctor of Psychological Sciences. – Moscow, 2003. – 37 p. (in Russian).
11. *Subjective assessment of the (non-) well-being of the population of the regions of the Russian Federation based on data from social networks* / Shchekotin E.V., Myagkov M.G., Goiko V.L., Kashpur V.V., Kovarzh G.Yu. // Monitoring obshchestvennogo mnieniia : Ekonomicheskie i sotsial'nye peremeny. – 2020. – N 1. – P. 78–116. (in Russian).
12. *Tsvetkova I.V.* Factors of social well-being of city dwellers // KNZh. – 2017. – N 1 (18). – URL: https://cyberleninka.ru/article/n/faktory-sotsialnogo-samochuvstviya-gorozhan (accessed 20.04.2024). (in Russian).