Application of Automated Means of Content Analysis of Data from Geoinformation Networks to Study the Accessibility of Landscaping Facilities

Boris A. Nizomutdinov, Vladimir A. Kazak, Petr A. Begen

ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

Abstract

The article presents a method developed by the authors to assess the accessibility of urban improvement facilities for low-mobility groups of the population based on the analysis of text data from social networks and the socio-psychological well-being of city residents. The object of the study was the profiles of landscaping objects in Google Maps located in the Petrogradsky district of St. Petersburg, 25 urban landscaping objects (parks, gardens, squares) were selected. During the study, the total number of comments for each improvement object was determined, reviews were analyzed, reviews were identified that describe the impressions and experiences of elderly and low-mobility groups, and the tonality of these messages was evaluated. The analysis of the reviews showed that the comments contain information from low-mobility groups of the population describing the problems of improvement facilities, in particular accessibility. The conclusion is made about the great potential that can be extracted from the use of automated data collection from geoinformation social systems. Findings show that overlapping data from Google Maps enriches the analysis that would previously have relied on a single source.

Keywords

Geoinformation social networks, reviews, accessibility for people with limited mobility, parsing, data analysis

1. Introduction

In most regions, various programs and methods have been developed to assess the accessibility of the urban environment. Their main goal is to ensure unhindered access to priority facilities and services for the disabled and other low-mobility groups of the population. However, the process of identifying problem areas has a considerable number of nuances.

Today, there are various ways to assess the accessibility of the environment, for example, surveys, conducting observations, studying project documentation, Internet surveys. But all these methods are labor-intensive, that is, there is a need to use significant human resources for a long time and, depending on the area under study, their volume may be different. Often, in resource-saving mode, these methods are not used by the city authorities.

In this study, an original approach to assessing the accessibility of urban improvement facilities for low-mobility groups of the population based on data from their social networks is proposed.

Improving the quality of life of the population as one of the main tasks of the socio-economic development of the city is a consequence of the successful interaction of social institutions and residents of the city in maintaining public relations to solve urgent problems of the city. For monitoring, comprehensive ratings have been developed to assess the quality of life of cities, which are aimed primarily at assessing the human potential of active working residents of the city and practically does not affect the interests of vulnerable groups of the population (elderly people, women, parents with young children, adolescents, youth, disabled people, etc.). In megacities, due to large flows of information, "communicative gaps" arise between residents and social institutions, which leads to an increase in social distance and a decrease

SSI-2021: XXIII All-Russian Conference on Scientific Services & Internet, September 20–23, 2021, Moscow (on-line), Russia

EMAIL: boris@itmo.ru (B.A. Nizomutdinov); kazakvauniversity@gmail.com (V.A. Kazak); petyabegen@mail.ru (P.N. Begen)
ORCID: 0000-0002-4090-9564 (B.A. Nizomutdinov); 0000-0002-2158-5031 (V.A. Kazak); 0000-0002-0613-3133 (P.N. Begen)

in understanding of aspects of urban existence relevant to these groups of citizens. Now the introduction of "hybrid management systems" is relevant, assuming a subject-subject system of relations between the city and its residents, based on an experimental study of the ideas of city residents about their well-being and the introduction of automated monitoring systems based on them. Among the studies in which various sources of information are presented, it is possible to distinguish between those whose data are obtained from "voluntarily provided" information and those obtained from "not voluntarily provided" information. Voluntarily provided information is generated by users, for example, social media data in reviews and comments, while non-voluntarily provided information comes from sources that collect data on user activity, for example, data from mobile operators. The proposed method uses data that users generate independently, in the public domain.

In their work [1], the authors eliminate the knowledge gap by proposing a method for determining urban opportunities for urban regeneration, which includes pre-processing, analysis and interpretation of separate and overlapping LBSN data. A twofold point of view is accepted – based on people and on the spot. Data from four LBSNS – Foursquare, Twitter, Google Places and Airbnb – reflect a people-based approach as it provides insights into individual preferences, usage and activities.

A group of scientists [2] has developed a method for predicting the urban area based on the geospatial activity of users in a social network. One of the most popular social networks Instagram was taken as a source of spatial data. Two large cities with different features of online activity were selected as target cities – New York, USA, and St. Petersburg, Russia, a convolutional neural network based on three-dimensional convolution layers is used for processing.

The study [3] shows that low-mobility groups of the population actively participate in interaction in social processes using ICT, are included in online social discussions and are included in democratic processes in electronic forms.

In this study, one of the most popular geoinformation social networks Google Maps was taken as a source of spatial data, and St. Petersburg was chosen as the target city. The introduction of active progress and the widespread dissemination of the SmartCity concept leads to the need to develop systems capable of accurately predicting the future state of the urban environment and landscaping facilities. Forecasting the state of an urban area requires the use of various data sources, new data sources are emerging in the new world, and social networks are one of such sources. Their social media data has become a valuable addition to the input data of a modern decision support system. Having data on the problems of accessibility of improvement facilities for low-mobility groups of the population in urban areas, researchers could extract information about the current situation and detect potential problems and develop recommendations for their elimination.

This study is based on existing methodologies for analyzing and interpreting data from geoinformation social networks to identify potential accessibility issues. Reviews in Google Maps about landscaping objects are considered as layers of information for analyzing the surroundings inside the city.

This document is structured as follows. Firstly, the theoretical basis for this work is based on previous studies in which a parser was developed to collect information. Secondly, the sources and the general method of preparing data for analysis are described. Third, data analysis is described. Finally, the results are presented, followed by a discussion, the main conclusions of the study and a discussion of the limitations of the study.

2. Research methodology

In this article, we focused on studying the accessibility of facilities for vulnerable social groups – disabled people, families with wheelchairs and pensioners. We want to find out if they write reviews about parks that have accessibility information. For the study, we chose the 1st district of St. Petersburg – Petrogradsky district. We have compiled a list of parks and squares for this area, found a card in Google Maps for each improvement object, excluded all objects that do not have text reviews. The final database contained 21 objects of urban improvement in the Petrogradsky district.

The development of a new methodology will make it possible to improve public space and make it modern and accessible to low-mobility groups of the population. The proposed methodology implies:

- determining the method of downloading data from Internet resources;
- determining the audience of users of urban objects;
- compilation of a dictionary that includes words that can characterize the accessibility of an object;

- analysis and processing of the received data.

The general scheme of the analysis is shown in Figure 1.

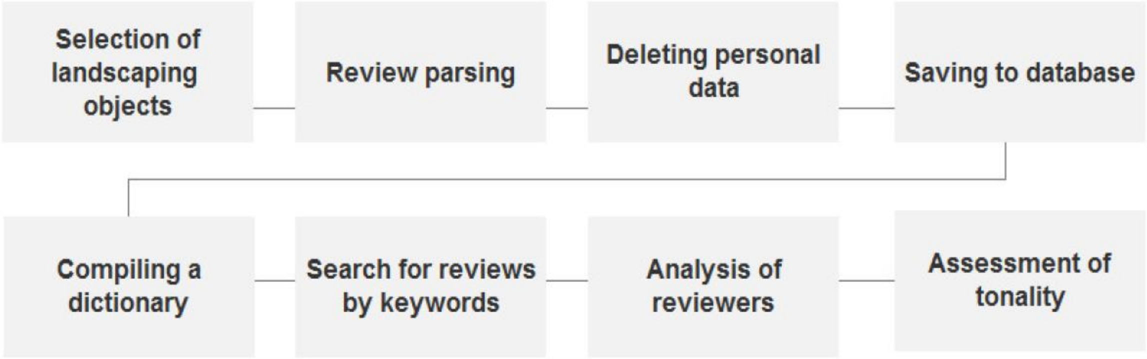


Figure 1: Search for landscaping objects on the map

At the step of collecting information, it was planned to use the Google API, however, during a detailed study and preparation of the parser, it turned out that the Google API has a limitation and is able to give only the last 5 reviews, which is not suitable for this study. To solve the problem of automated feedback collection, a ready-made Outscraper parser was used. With this parser, you can upload reviews by object ID to Google Maps, it does not have such critical limitations as the standard Google API.

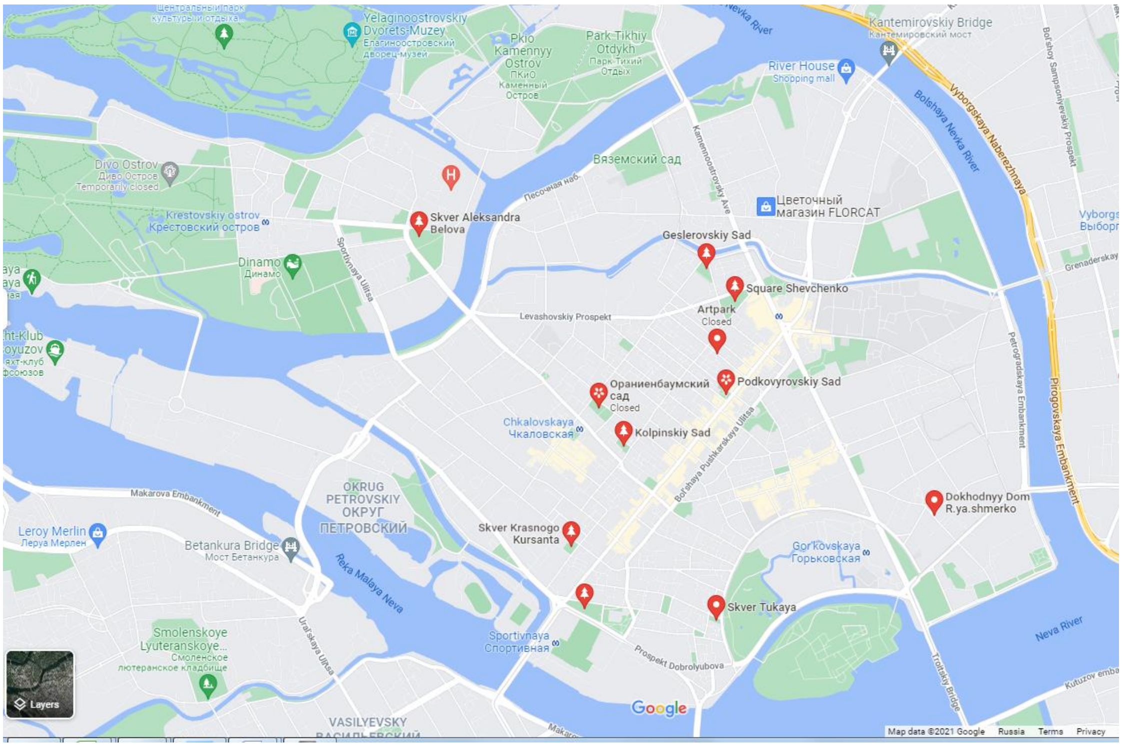


Figure 2: Search for landscaping objects on the map

An important aspect that was taken into account when collecting information is personal data. From the point of view of legislation, projects providing for automated data collection from social networks affect both special legislation on personal data (in terms of the object of research) and related intellectual rights of the creators of the social network (in terms of the data source for research). The difficulty of interpreting legal relations from the point of view of the law arises from the fact that changes in legislation lag far behind the development of technology.

It follows from the legislation on personal data and information that the information posted by the users themselves on social networks is publicly available from a legal point of view, but the courts, in some cases, come to a different conclusion.

The best option for researchers is not to collect the user's full name, and to depersonalize all the collected data using and storing only statistical information. When the parser was running, data about the user's full name was not saved, each record was assigned an ID, which excluded the personalization of data, in this form the information can be considered impersonal. This approach avoids the collection, storage and processing of users' personal data from social networks.

During the collection of information, 21 objects were processed, 4900 reviews were collected, all reviews were depersonalized and stored in a single database.

2.1. Search for thematic reviews about accessibility issues

At the next stage, we compiled a dictionary that includes 60 words that can characterize the accessibility of the object, for example, the words “ramp”, “barrier”, “wheelchair”, etc. were included in the dictionary. Then, using a script, we performed a search on the collected database using these words and various word forms. We have selected all the reviews that contain terms from our dictionary. During the search, 450 reviews describing accessibility were selected. This is very valuable information that can help improve these facilities.

Additionally, for the development of this direction, we tested the use of machine learning methods to determine the subject of comments. To solve the problem of automatic determination of the subject of the review, an algorithm is being developed to solve the problem of text clustering. Clustering is the splitting of a set of similar documents into clusters – subsets, the parameters of which are unknown in advance. The number of clusters can be arbitrary or fixed (set by the user at the initial stage). The clustering task refers to the well-known approach of unsupervised learning, unsupervised learning (learning on data not marked up by experts).

Using the implementation of machine learning methods in the algorithm, the result is achieved in the form of the formation of the nth number of groups (clusters) into which the source text array can potentially be divided. The resulting n-clusters should be further analyzed on the basis of the news corpuses that have fallen into them or by a list of keywords specific to each of the clusters. To implement the solution of the clustering problem, the KMeans method (k-means method) was used.

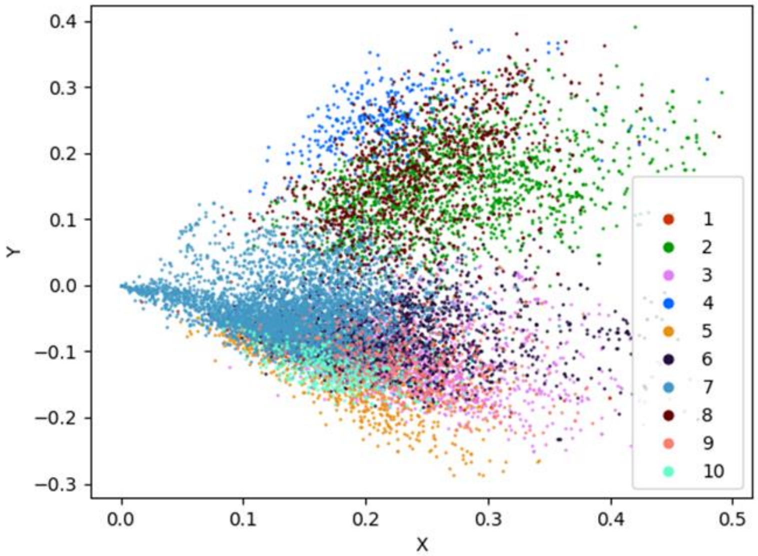


Figure 3: Partitioning into clusters

The operation of the algorithm is to minimize the total quadratic deviation of cluster points from the centers of these clusters themselves. To implement the algorithm, the Kmeans class from the sklearn.cluster library was used. Further, in order to train the algorithm on the collected data, it is necessary to pre-process them (remove punctuation marks, remove noise, etc.) and present them in vector (numeric) form. To do this, the basic methods and approaches for natural language processing (Natural Language Process) are used.

The pandas library was used to extract the collected news array into the program. The "text" field from the original news array was selected as the training text data. The received data were pre-processed as follows: punctuation marks, invisible symbols were removed using regular expressions, Latin letters, single letters were removed, extra spaces were removed.

With the help of the pymorphy2 library, all words were reduced to normal form (for example, the adjective word "electronic" is reduced to the form "electronic"). This allows you to reduce the dimension of the data array without losing significant features in the text. A list of stop words was also generated using the TfidfVectorizer class from the sklearn library to remove unnecessary noise in the source data (the words are presented in the file stopwords.txt).

The conversion of text into a vector (numeric) form was also carried out using the TfidfVectorizer class. This class converts text into a vector form by compiling a matrix of weights for each word based on the tf-idf approach. Then the processed data was transferred to the KMeans class algorithm to solve the clustering problem. The trained finished model was saved to a file for further use in other tasks.

During the analysis, we collected reviews which described the accessibility of facilities for people with reduced mobility, all of them were grouped by the main topics related to accessibility, for example: strollers, wheelchairs, disabled people, accessibility, family members and the elderly, restrictions. We had to combine the terms "wheelchair for children" and "wheelchair for the disabled", since they have the same name in Russian.

The "Restrictions" group includes general conditions related to access, functions and restrictions. Separately, we can single out a group associated with reviews in which they wrote about the problems faced by older people, there were 37 such reviews.

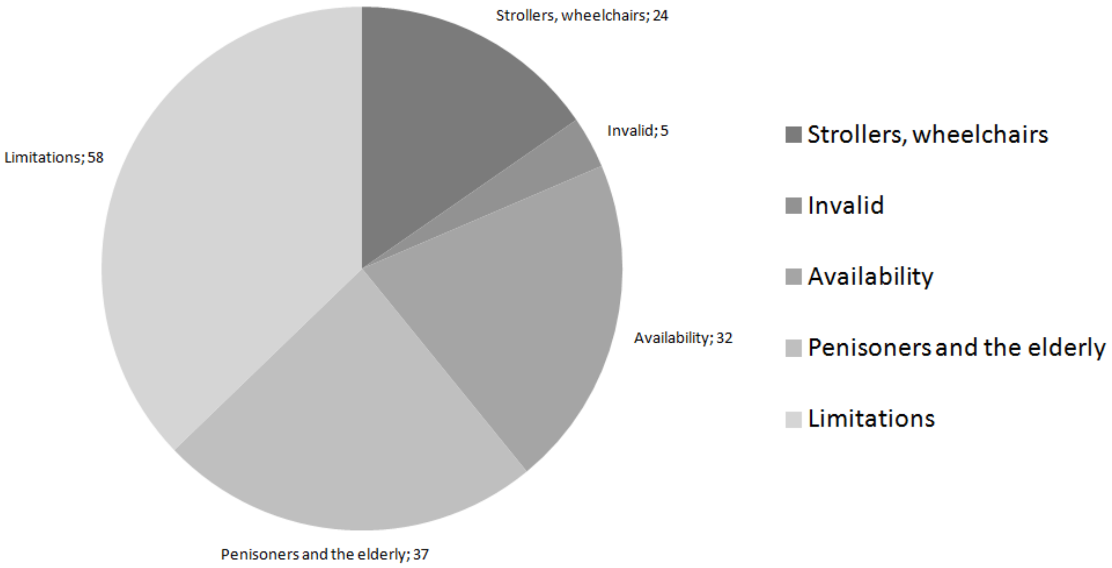
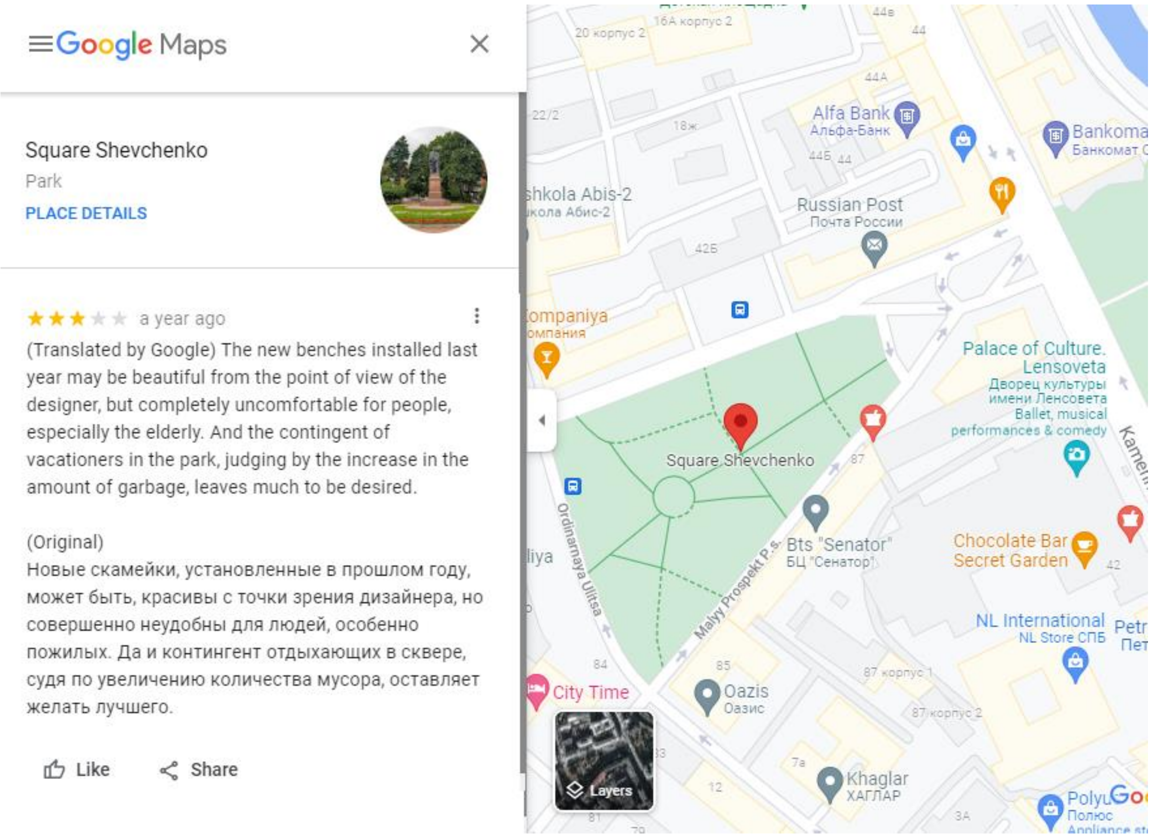


Figure 4: Partitioning into clusters

Reviews from people with limited mobility contain both a description of the advantages and an indication of the disadvantages of parks and squares in the area. Among the problems found, difficulties with moving along park paths, inconvenient entrances, uninformative information boards were mentioned, as well as mentions of uncomfortable benches for low-mobility groups of the population and other problems requiring attention.



★★★☆☆ a year ago

(Translated by Google) The new benches installed last year may be beautiful from the point of view of the designer, but completely uncomfortable for people, especially the elderly. And the contingent of vacationers in the park, judging by the increase in the amount of garbage, leaves much to be desired.

(Original)
Новые скамейки, установленные в прошлом году, может быть, красивы с точки зрения дизайнера, но совершенно неудобны для людей, особенно пожилых. Да и контингент отдыхающих в сквере, судя по увеличению количества мусора, оставляет желать лучшего.

Figure 5: Example of a review about a park

2.2. Definition of tonality

It is difficult to solve the problem of determining the tone of a comment by clustering the text into groups, because the text contains many signs by which they can be classified, and the algorithm does not yet know which comment is "positive" and which is "negative".

To solve such a problem, the supervised learning approach is most often used (training on an array of data marked up by experts), which boils down to solving the problem of text classification (the distribution of objects into previously known groups, categories). However, the data we have received does not have a preliminary markup for a "positive" or "negative" tone, so you should use another available and prepared array of texts. There are very few such corpora for the Russian language, because marking up large text bodies requires considerable time and human resources from researchers.

As a Russian-language array with positive and negative texts marked up, we used an array collected by Yulia Rubtsova from the Twitter site, which contains user reviews and comments on a variety of topics (politics, economics, IT, sports, medicine, etc.). For training, a training corps consisting of 114,911 positive and 111,923 negative entries was used. To solve the problem of determining the tonality of text messages, a ready-made implementation from the company's researchers was used Mail.ru, freely available for research (https://github.com/sismetanin/sentiment-analysis-of-tweets-in-russian). In this algorithm, a convolutional neural network (CNN (convolutional neural network)) was implemented, which showed an average accuracy of 78.1% in determining the tonality of the text, which is good enough for solving such problems.

In the received database, more than 35% of the reviews had a negative tone. It is planned to increase the accuracy of determining the tonality.

3. Conclusion

In the course of the work done, a methodology for assessing the accessibility of urban improvement facilities was presented and described. As an illustrative example, the objects of the urban environment (parks, squares, gardens) located on the territory of the Petrogradsky district were taken. Based on the results of the work done, an analysis of the feedback received was carried out and based on them it was revealed that, in general, the state of urban improvement facilities is not in quite proper condition, because 23% of negative reviews indicate that there are hard-to-reach territories for low-mobility groups in the area, to which the administration of this area should pay attention and promptly correct the situation.

This study combined the methodological foundations of traditional socio-psychological research and the possibilities of modern information technologies to substantiate value-oriented management of urban infrastructure development. The selected source of information in Google Maps showed that users generate a large number of reviews, and some of them about the problems of accessibility of urban facilities for low-mobility groups of the population. This project is a practical and promising solution in poorly formalized fields of knowledge. In addition, the use of such a solution at the national level can be an example of the introduction of digital technologies and platform solutions in the areas of public administration, business and society. During the processing of information, it was possible to identify informative reviews that describe the problems of accessibility of individual parks for low-mobility groups of the population.

The proposed methods of data extraction and processing have shown good results, at the next stage it is planned to collect feedback on all districts of St. Petersburg, with the construction of a heat map.

The study has a number of limitations that are planned to be worked on in the future, in particular, the study does not consider the reliability of reviews and so-called "fake reviews" that can be left by unscrupulous citizens. However, we believe that when discussing urban improvement facilities, the proportion of fake reviews is lower than in the commercial sector. In the commercial sector, there is a whole industry for writing positive and negative reviews (SERM), but for city parks there is simply no need to write paid and fake reviews, therefore, they can be considered reliable. In addition, certain limitations of the study are associated with the unavailability of some information about the socio-demographic characteristics of users. This difficulty is directly related to the personal data processing policy, we have not collected or processed personal data.

The results of the study can be used to develop recommendations for the management and development of the city's infrastructure. In this regard, the project has scientific, educational and educational value and contributes to the implementation of one of the priority directions of the city development related to improving the quality of the urban environment and ensuring the effectiveness of management and development of the urban environment.

Additionally, the data obtained can be used in Smart City projects for targeted operational monitoring of the actual needs of the population.

4. References

[1] P. Martí, C. García-Mayor, L. Serrano-Estrada. Identifying opportunity places for urban regeneration through LBSNs. Cities 90 (2019) 191–206. https://doi.org/10.1016/j.cities.2019.02.001

[2] K. D. Mukhina, A. A. Visheratin, Gali-Ketema Mbogo, D. Nasonov. Forecasting of the Urban Area State Using Convolutional Neural Networks. 2018 23rd Conference of Open Innovations Association (FRUCT). https://doi.org/10.23919/FRUCT.2018.8588075

[3] I. Grigoryeva, L. Vidiasova, D. Zhukб Seniors' Inclusion into e-Governance: Social Media, e-Services, e-Petitions Usage. ICEGOV '15-16: Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, March 2016, pp. 173–176. https://doi.org/10.1145/2910019.2910022

[4] N. Hochman, L. Manovich, Zooming into an Instagram City: Reading the local through social media. First Monday 18 (7) (2013). https://doi.org/10.5210/fm.v18i7.4711.

[5] D. Arribas-Bel, K. Kourtit, P. Nijkamp, J. Steenbruggen, Cyber Cities: Social Media as a Tool for Understanding Cities. Applied Spatial Analysis and Policy 8(3) (2015) 231–247.

[6] L. Mitchell, The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PloS one 5 (8) (2013) 64–71.

[7] A. E. Nenko, A. M. Semenova, A. A. Galaktionova, Evaluation of the quality of public spaces according to reviews in Google Maps. Scientific service on the Internet: proceedings of the XXII All-

Russian Scientific Conference (September 21–25, 2020, online). Moscow: IPM named after M.V. Keldysh, 2020, pp. 473–485.

[8] S. Van Canneyt, S. Schockaert, O. Van Laere, B. Dhoedt, Detecting places of interest using social media. Proceedings 2012 IEEE/WIC/ACM International Conference on Web Intelligence. 2012, pp. 447–451.

[9] F. Mairesse, M. Walker, M. Mehl, R. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30 (2007) 457–500. https://doi.org/10.1613/jair.2349

[10] A. Korneeva, Yu. Zeremskaya, O. Loyko, Virtual space as a sphere of the personal identity's formation. Journal of Economics and Social Sciences 8 (8) (2016) 31–35.

[11] Y. Kim, J. H. Kim, Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study. Information Processing & Management 6 (54) (2018) 1101–1114.

[12] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, A. Gribov, RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. Proceedings of COLING 2018, pp. 755–763.