Adv Syst Sci Appl 2021; **03**; 101-112
Published online at https://ijassa.ipu.ru.

Neural Network Technologies and Topological Analysis of Social Media Data

Nailia Gabdrakhmanova[1*], Maria Pilgun[2]

[1)] *Peoples Friendship University of Russia (RUDN University), Moscow, Russia*

E-mail: gabdrakhmanova_nt@rudn.ru

[2)] *Institute of Linguistics Russian Academy of Sciences, Moscow, Russia*

E-mail: mpilgun@iling-ran.ru

Abstract: Currently, due to the successful use of neural network technologies for analyzing data of various formats, the range of problems that can be solved using mathematical modeling methods has significantly expanded. The paper deals with the topical task of analyzing speech perception by social media data and assessing the level of conflictogenity, social approval / stress of the residents regarding urban planning projects. The purpose of the paper is to develop and show on a practical example the effectiveness of the integration of neural network and mathematical models for solving such tasks. Mathematical models are built using the methods of mathematical statistics and topological data analysis.

Keywords: neural network, persistence diagrams, time series, speech perception, social media data

1. INTRODUCTION

Analysis of speech perception by social media data is a topical task, since it makes it possible to reveal the attitude of users to certain events, conduct predictive analytics and identify conflict situations in the early stages of crisis situations, which is a demanded task in various fields. There are studies of possible adaptive sampling mechanisms for haptic data compression aimed at applications like tele-operations and tele-surgery [3]. Machine understanding is being actively studied [6]. Notably, fuzzy and natural language based approaches and solutions for a more human consistent dealing with decision support, time series analysis, forecasting, clustering, etc. are discussed [2]. Much attention is paid to the study of the features of perception of robots [4, 7].

The purpose of the paper is to develop and show on a practical example the effectiveness of the integration of neural network and mathematical models for the analysis of user content and assessment of the level of conflictogenity, social approval/stress of the residents regarding urban planning projects or district conflicts based on the analysis of perception concerning the construction of the Fiztekh subway station (Moscow, Russia).

In our work, we show the potential of automatization of the conflictogenity level assessment according to social media data. This work is a continuation of the previous works of the authors [7]. To build mathematical models, methods of mathematical statistics and differential equations are used in the works of the authors [7]. In this paper, we supplement the developed methods and algorithms with methods of topological data analysis. Such an extension of the algorithm was required due to the specifics of the data in the time series of the problem being solved.

[*] Corresponding author: gabdrakhmanova_nt@rudn.ru

Topological data analysis (TDA) is a fairly new direction in the field of data analysis. TDA makes it possible to find the structure in time series data. Persistence diagrams introduced by H. Edelsbunner [8] are the most important tool in computational topology that allow, for example, to obtain qualitative information about the "topological dynamics" of a time series. An important feature of persistence diagrams is their robustness to "noise". Methods of computational topology were further developed in the works of G. Carlsson, A. Zomorodian [9,10]. Topological data analysis can be combined with methods in machine learning (including deep learnin) as well as statistical methods [11].

1.1. Data

The material for the study was the data of social media, microblogs, blogs, instant messengers, videos, forums and reviews concerning the implementation of the project of the terminal station for the Lyublinsko-Dmitrovskaya line (construction of the Fiztech subway station) in Moscow (Russia). The data was collected between January 1, 2019 and December 22, 2019 (see Table 1.1).

Table 1.1 Data characteristics

Number of messages:	7063
Max number of messages per day:	614
Number of authors:	933
Activity (posts per author):	7,57
Number of sources:	213

1.2. Method

The study involved several parallel models. With the help of the semantic neural network model, a text analysis was performed; the topic structure and the semantic network were analyzed, as well as the sentiment and the level of aggression; digital conflictogenic zones were identified; and the indices of social stress and well-being were calculated. In parallel with the neural network model, a statistical analysis of experimental data was performed and dynamic models of processes were built. The following models were used for data analysis: a model based on the neural network paradigm of using neural-like elements with time summation of signals [5], methods of mathematical statistics and topological data analysis [9]. Analysis of solutions for all models showed the consistency of solutions.

1.2.1. Topological Data Analysis

Persistence diagrams are an important tool in TDA. In this section, we will give basic information about persistence diagrams, following [9].

Let X be a triangulated topological space, f a continuous function on X. For $a \in R$, we denote the set of points of the sub-level: $X_a = f^{-1}(-\infty, a]$.

Figure 1 shows a graph of the construction of sub-levels for two levels $a=0.05$ (green); $a=0.1$ (purple).

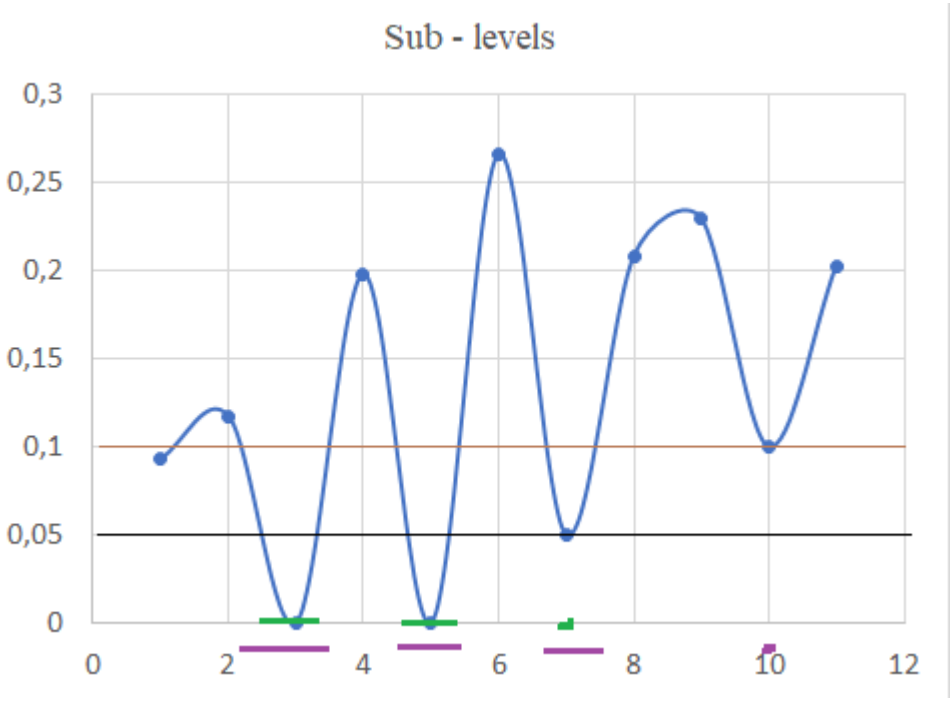

Fig. 1.1 Sub-levels for levels a=0.05 (green) ; a=0.1 (purple).

The number $a \in R$ is called a homological critical value if for some *k* the homomorphism induced by the embedding of $f_*: H_k(X_{a-\varepsilon}) \to H_k(X_{a+\varepsilon})$ is not an isomorphism for any sufficiently small $\varepsilon > 0$ (homology groups are considered with coefficients in Z_2). A continuous function *f* is called tame if it has only a finite number of homological critical values, the homology groups $H_k(X_a)$ are finitely generated for all k and all $a \in R$, and if the segment *[x, y]* does not contain homological critical values of the tame function, then the embedding-induced homomorphism of $f_x^y: H_k(X_x) \to H_k(X_y)$ is an isomorphism.

Suppose $f: X \to R$ is a tame function. Let $a_1 < a_2 < \cdots < a_n$ be its homological critical values. A persistence diagram $D(f) \subset R^2$ of function *f* is the set of points (a_i, a_j) , $i, j = \overline{1,n}$ *i, j = 1, . . . , n* combined with the set of points of the diagonal $\Delta = \{(x,x) | x \in R\}$. A remarkable feature of the persistence diagram *D(f)* is its stability with respect to perturbations of the function *f*.

Persistence diagrams can be used to calculate the lengths of the barcodes of connectivity components. Here the term barcode stands for the component lifetime. Let us denote the summarized lengths of barcodes of two homology groups H_0 and H_1 as L_0 and L_1 correspondingly. Then the mean of the Euler characteristic can be determined [9] as

$$\chi = L_0 - L_1 \qquad (1.1)$$

To identify conflictogenity by social media data, it is proposed to use the estimates of Euler's characteristics. It is proposed to choose a certain step on the time axis and find estimates of the Euler characteristics at each interval. Based on the dynamics of the estimates of Euler's characteristics, it is proposed to develop an algorithm for classifying the situation. In Section 2.3.3, we show the possibility of classifying the situation according to the dynamics of estimates of the Euler's characteristics of time series.

1.2.2. Time series analysis

To study the dynamics of processes according to observational data, methods of time series analysis are used [12]. A fairly general mathematical model for the time series *x(t)* is a model of the form:

$$x(t) = u(t) + v(t) ,$$

where *u(t)* is a deterministic sequence or systematic component, *v(t)* is a random component. The purpose of mathematical modeling of time series is to study the dynamics and predict the values of *x(t)* several steps ahead. In this work, regression models were used to estimate *u(t)* and autoregressive models to estimate *v(t)*. The results of building models are presented in paragraph 2.3.2.

2. RESULTS AND DISCUSSION

2.1. General description of the content

Key topics of the content related to the construction of the Fiztech subway station that had been generated by users, were found mostly in the social network VKontakte. It should also be noted that users' interest in the topic increased in 2020 compared to 2019 (Figures 2.1)

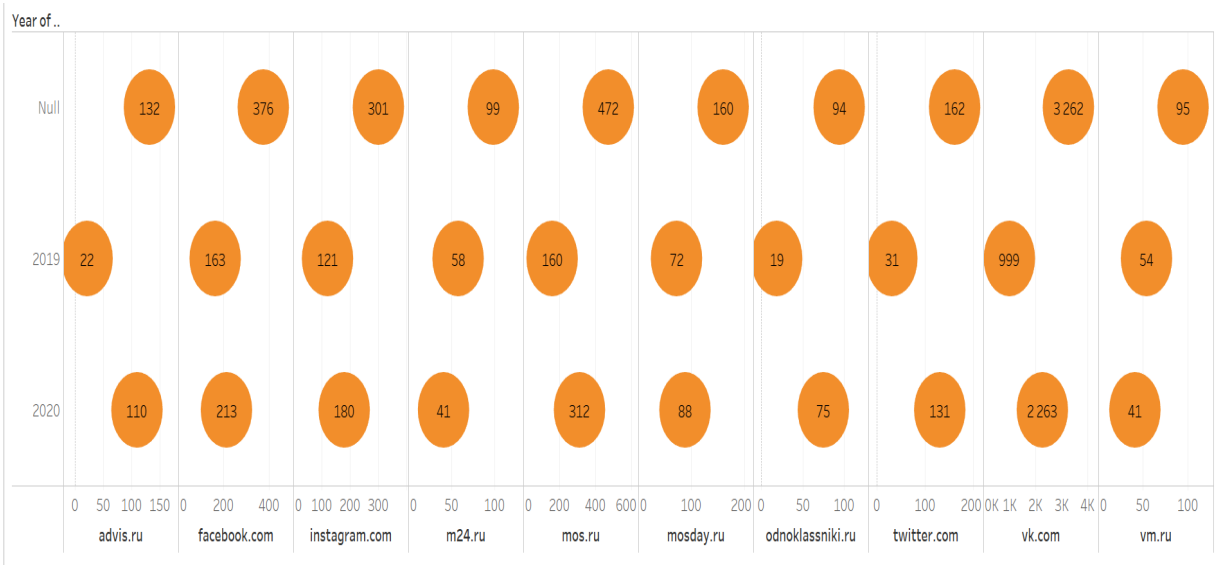

Fig. 2.1. Digital platforms where actors generated content

The peak of growth in the authors' activity dynamics (value of 107), the total number of messages (614) and unique messages (436) falls on May 25, 2020 (Figures 2.2; 2.3; 2.4) and is associated with information that the entire Lublinsko-Dmitrovskaya line of the Moscow subway will become underground. Analysis of the views dynamics shows that, in addition to this news, the greatest attention of users was also attracted by messages from August 20, 2020 (value of 2 524 187) regarding the mayor's approval of the names of seven stations under construction on the Big Beltline of the subway and from November 30, 2020 (value of 2 860 008) about how the new subway stations will look like (Figures 2.5).

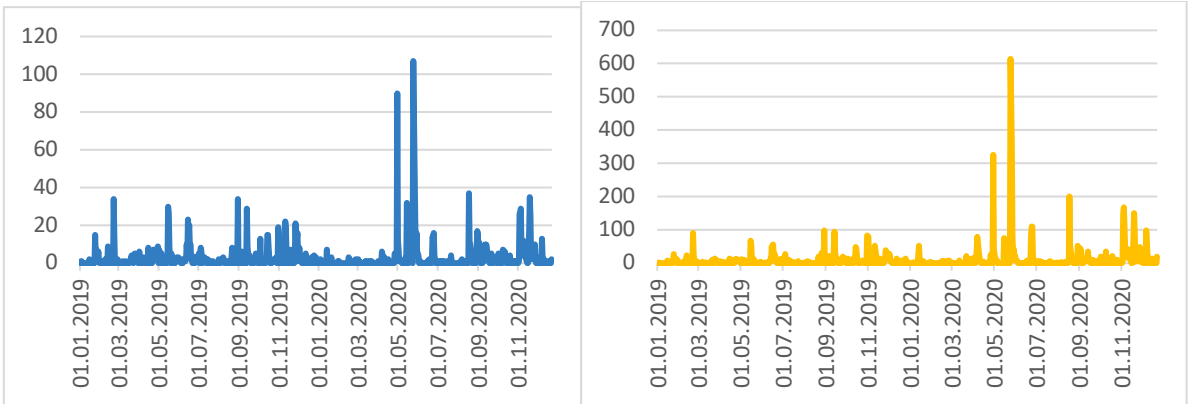

Fig. 2.2. Dynamics of users' activity (SEC)

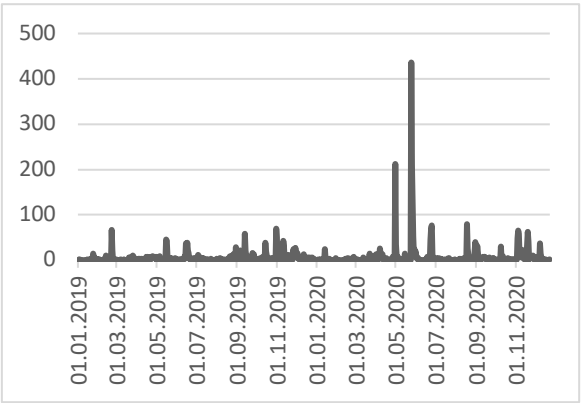

Fig. 2.3. Dynamics of the total number of messages

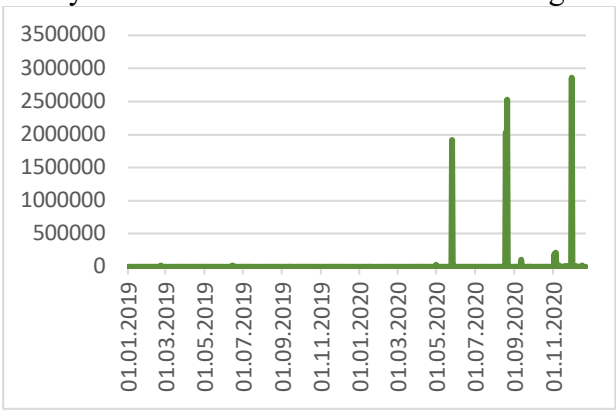

Fig. 2.4. Dynamics of unique messages

Fig. 2.5. Views dynamics

2.2. Semantic neural network model

2.2.1. Key topics

There is a small list of key topics related to complaints and dissatisfaction of the residents, which indicates a calm situation around this urban construction project:

Negative cluster:

- Surface construction of the subway from the Lianozovo subway station to the Severny settlement.
- Incorrect holding of public hearings on the project to extend the construction of the Lyublinsko-Dmitrovskaya line.
- Risk of destruction of buildings located in the vicinity of the construction sites.
- Environmental deterioration due to the project implementation.
- Deterioration of the traffic situation in the district during construction.
- Violation of the architectural appearance of the district and crudity of the design.
- The residents are afraid that after the appearance of new subway stations, their yards will be filled with unauthorized cars.
- The statements of the mayor's office representatives about construction works in the capital during the pandemic were negatively perceived by users.

Positive cluster:

- The strongest positive reaction was caused by the news of the abandonment of the elevated subway line.
- Messages about high rates of the construction.
- Improving the transport accessibility of districts after the commissioning of a new subway station.

2.2.2. Sentiment analysis

The overwhelming majority of the content with various types of sentiment was posted on the social network VKontakte. Messages with negative sentiment were also generated on Facebook and Instagram, and messages with positive sentiment - on mos.ru and m24.ru; while those with neutral sentiment were on mos.ru and Facebook (Figures 2.6; 2.7; 2.8.)

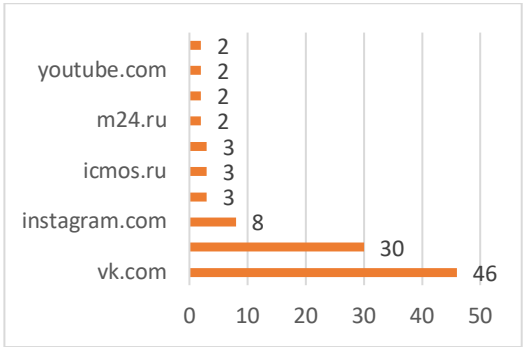

Fig. 2.6. Top-10 sources of negative sentiment

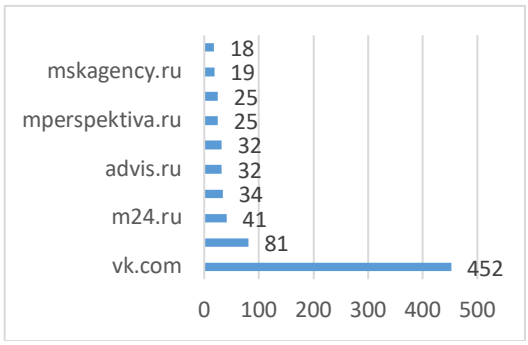

Fig. 2.7. Top-10 sources of positive sentiment

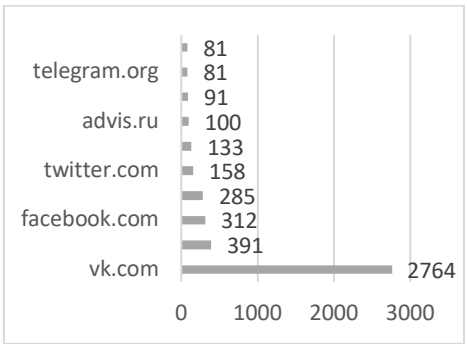

Fig. 2.8. Top-10 sources of neutral sentiment

Most digital footprints and audience reach are characterized by neutral sentiment, which indicates a calm situation around these objects and users' calm perception (Figures 2.9; 2.10.).

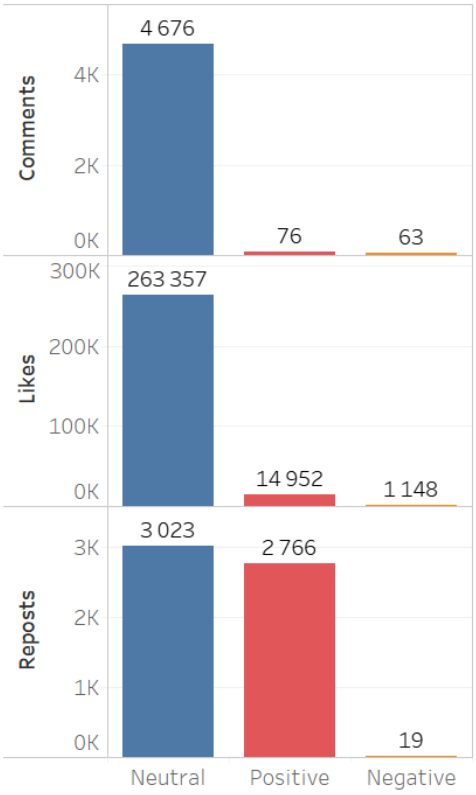

Fig. 2.9. Sentiment of the digital footprints

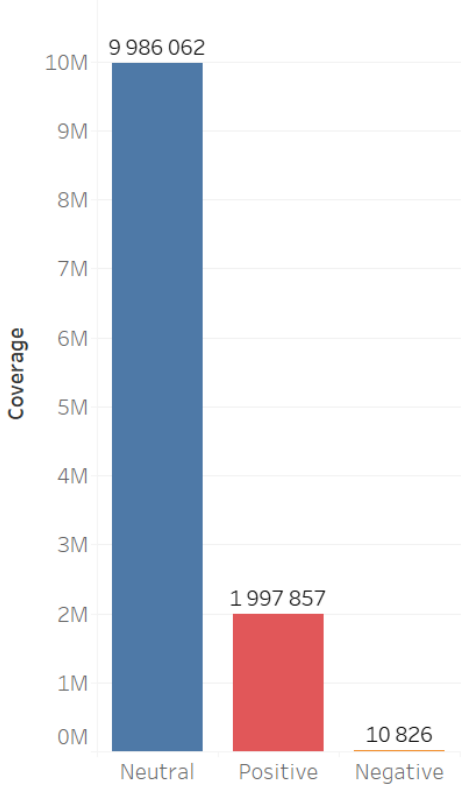

Fig. 2.10. Sentiment of the audience reach

2.2.3. Formation of the social tension rating regarding the construction of the Fiztech subway station

The dynamics of references by sentiment indicates a calm situation around the object, since there is a significant predominance of neutral reactions. Negative messages make up a negligible portion of the content (Figures 2.11.).
It is natural that the analysis of social stress regarding the construction of the Mosinzhproekt facility, the Fiztech subway station, according to the database (collected between January 1, 2019 - December 22, 2019), showed the presence of social stress with an extremely low index of 1.95 and a high index of social well-being of 18.92 (Figures 2.12.).

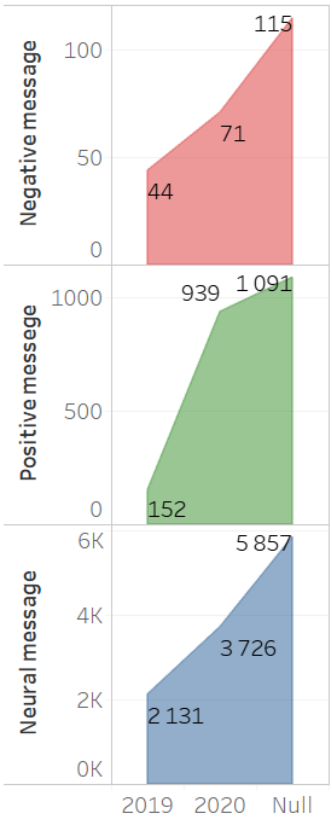

Fig. 2.11. Dynamics of references by sentiment

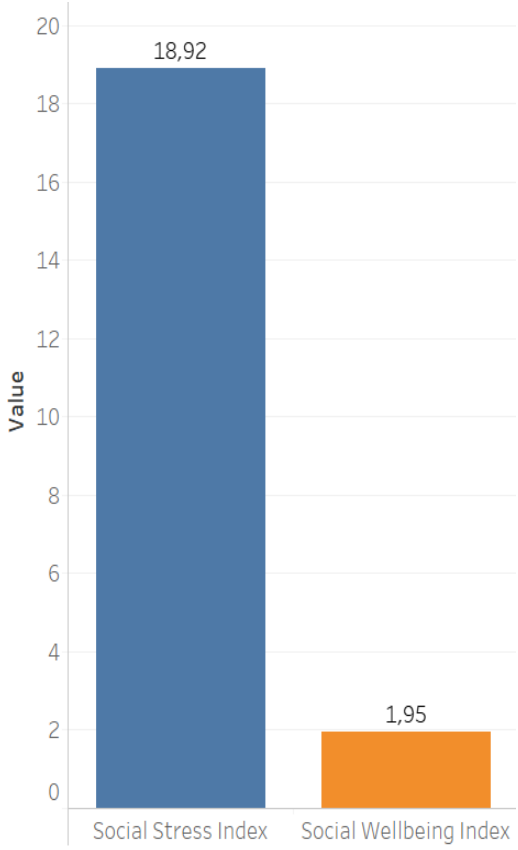

Fig. 2.12. Social tension and well-being indices

2.3. Mathematical models

2.3.1. Statement of the problem

As a result of the implementation of the steps described in 2.1 and 2.2, sequences were obtained that represent the number of negative *NEG(t)*, positive *P(t)* and neutral *NTR(t)* messages at each time moment *t*. For the problem being solved, it is important to know the proportion of negative, positive and neutral messages in each moment of time and to study the dynamics of these data. For this purpose, at the preprocessing stage, the entire data array is converted into several proportions. Let *S(t)* be the number of posts at time t with a reference to the topic of the subway construction.
We introduce the following notations:

p(t) is the proportion of messages at time t with a positive attitude towards the subway construction;

neg(t) is the proportion of messages at time t with a negative attitude towards the subway construction;

ntr(t) is the proportion of messages at time t with a neutral attitude to the subway construction.

The data conversion formulas are as follows:

p(t)= P(t)/S(t); neg(t)=NEG(t)/S(t); ntr(t)=NTR(t)/S(t).

The constructed time series are notes as: *{p(t)},{neg(t)},{ntr(t)}, t=1,..,n.*

It is necessary, according to the data of time series *{p(t), {neg(t)}, {ntr(t)}* for the period *t=1,..,n*

1) estimate the trend of each time series; 2) classify the situation (the actors' attitude towards the construction): a) there is a conflict or b) there is no conflict.

2.3.2 Time series trend models

Regression analysis models were used to estimate the parameters of the systematic components.

For a time series with a positive attitude towards the construction, the following model was built:

$$p_m(t) = -0.006+0.01t \quad (2.1)$$

where $p_m(t)$ is a model value for the proportion of positive messages at time t. A positive coefficient in (2.1) before t indicates an increase in the proportion of positive messages.

For a time series with a negative attitude towards the construction, the following model was built:

$$neg_m(t) = 0.028 - 0.001t \quad (2.2)$$

where $neg_m(t)$ is a model value for the proportion of negative messages at time t. A negative sign before the coefficient t in (2.2) indicates a decrease in the proportion of negative messages over time.

For a time series with a neutral attitude to the construction, the following model was built:

$$ntr_m(t) = 0.98 - 0.01t \quad (2.3)$$

where $ntr_m(t)$ is a model value for the proportion of neutral messages at time t. A negative sign before the coefficient t in (2.3) indicates a decrease in the proportion of neutral messages over time.

The adequacy of models (1), (2), (3) was assessed using standard methods. In particular, for model (3), the multiple coefficient of determination R= 0.7, the hypothesis of the model inadequacy according to Fisher was rejected at a significance level of p=0.02. These calculations indicate a fairly good quality of model (3). The results of checking the adequacy of models (2) and (3) showed the satisfactory quality of the models.

Based on the prediction using the constructed models, one can claim that there is no conflict.

2.3.3 Topological analysis of time series data

For the calculations, the TDA RStudio package was used, which has some tools for topological data analysis.

Before calculations, the entire time interval was divided into several intervals with the same step. The effectiveness of this approach and algorithm was proposed in the work of the author of the paper [13]. At the next step, for each time section and for each time series, persistence diagrams and barcodes were built, and estimates of Euler's characteristics were calculated. Below are the results for one time interval.

On Fig. 2.13 we presented the main results of the topological data analysis for the time series *{p(t)}*. Fig. 2.13 a) shows the dynamics of the time series *{p(t)}*. In this plot, the abscissa is time, and the ordinate is the value of *p(t)*. Fig. 2.13 b) shows the persistence diagram of the time series *{p(t)}*. The component time of birth is plotted along the abscissa axis, and

the component time of death is plotted along the ordinate axis. On the diagram, points correspond to zero-dimensional simplexes, triangles to one-dimensional simplexes. The point on the diagram has the values of the component birth and death as coordinates. Fig. 2.13 **c)** shows the barcodes for each simplex. The black line corresponds to zero-dimensional simplexes, the red line corresponds to one-dimensional simplexes. The length of the barcode corresponds to the lifespan of the simplex. Based on the calculated values, the estimate of the Euler's characteristics for the series *{p(t)}* according to formula (1) is as follows:

$\chi = -0.56$.

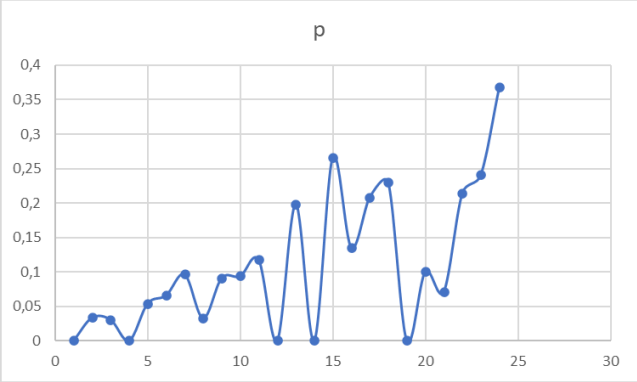

Fig. 2.13 a) Plot of the time series *{p(t)}*;

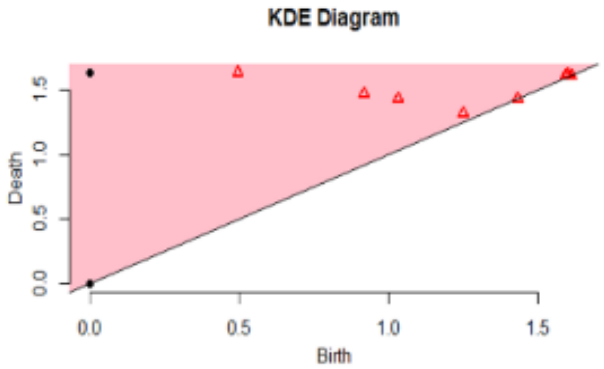

Fig. 2.13 b) Persistence diagram of *{neg(t)}*;

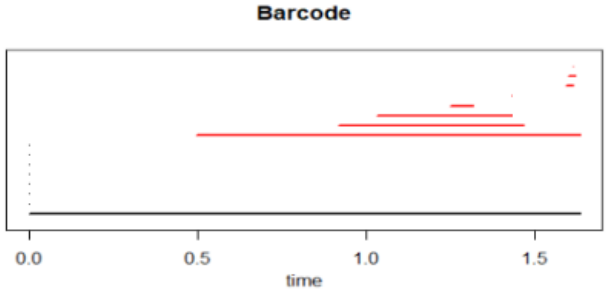

Fig. 2.13 c) Barcode diagram of p(t)

On Fig. 2.14, we presented the main results of the topological data analysis for the time series *{neg(t)}*. Based on the calculated values, the estimate of the Euler's characteristics for the series *{neg(t)}* according to formula (1) is as follows: $\chi = 1.77$.

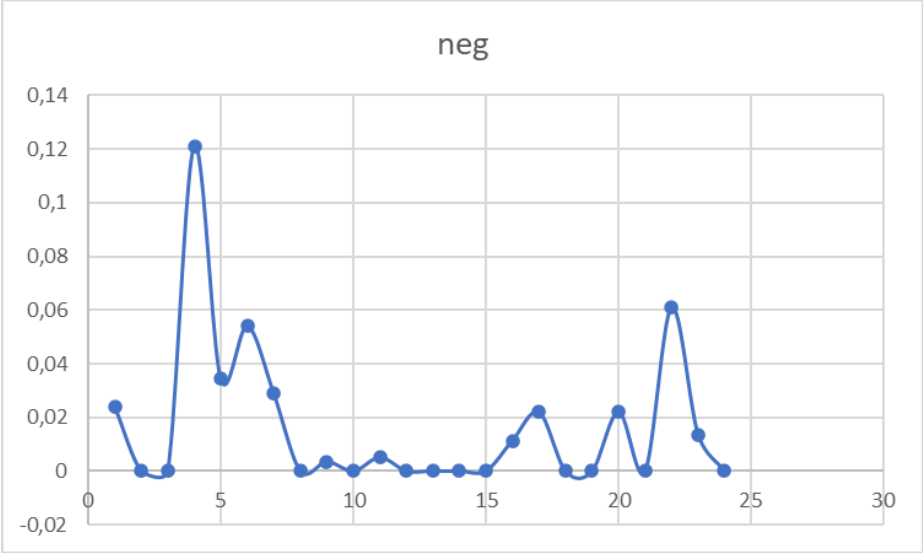

Fig. 2.14 a) Plot of the time series *{neg(t)}*;

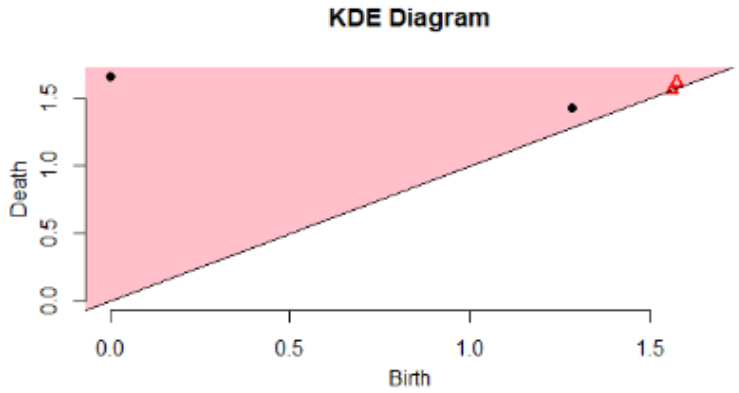

Fig. 2.14 b) Persistence diagram of *{neg(t)}*;

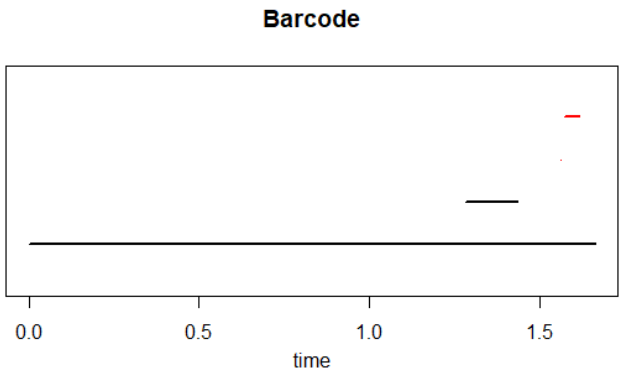

Fig. 2.14 c) Barcode diagram of *{neg(t)}*.

On Fig. 2.15, we presented the main results of the topological data analysis for the time series *{ntr(t)}*. Based on the calculated values, the estimate of the Euler's characteristics for the series *{ntr(t)}* according to formula (1) is as follows: $\chi = -0.73$.

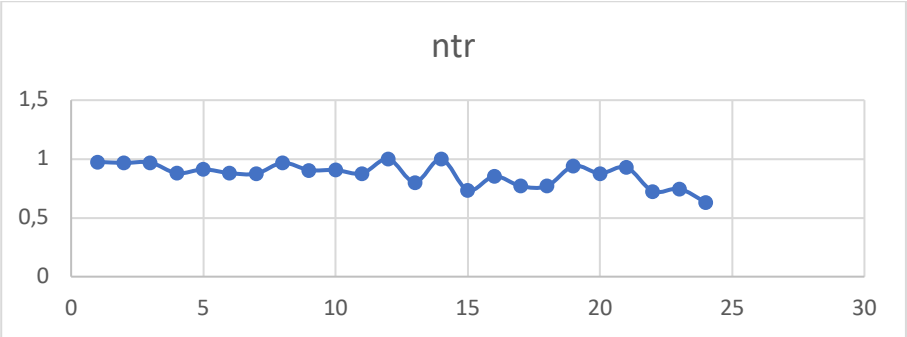

Fig. 2.15 a) Plot of the time series *{ntr(t)}*;

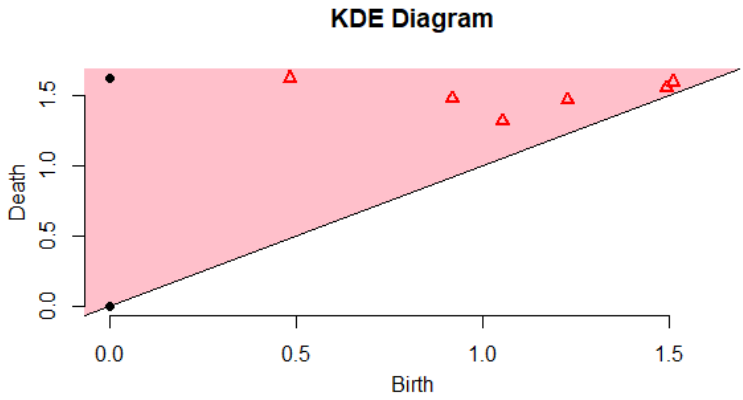

Fig. 2.15 b) Persistence diagram of *{ntr(t)}*;

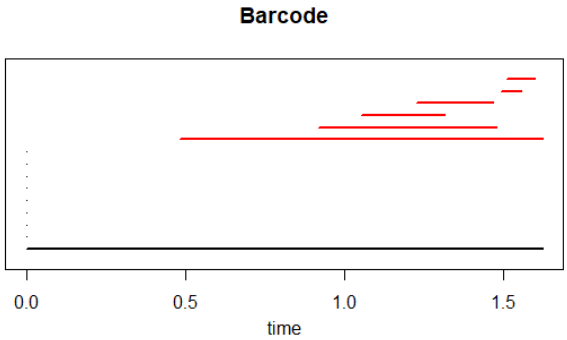

Fig. 2.15 c) Barcode diagram of *{ntr(t)}*.

From the above studies, we see that the highest value (1.77) of the Euler's characteristics corresponds to a number of message rates with a negative attitude towards the project. From the dynamics of this series, we observe rare outliers. The time series has a negative trend. These studies suggest the absence of a negative attitude towards the project. For a number of message rates with a positive attitude towards the project, the value of the Euler's characteristics is less (-0.56), the time series has a positive trend. Consequently, the positive attitude towards the project grows. For the series with a neutral attitude towards the project, the value of the Euler's characteristics is even less (-0.73), and the time series has a negative trend. Based on these studies, we believe that the Euler's characteristics in combination with regression and autoregression models make it possible to classify time series to solve the problem set.

3. CONCLUSION

Analysis of the data showed the absence of aggression, negative attitude towards the construction of the Fiztech subway station. Most of the actors express a neutral and positive attitude towards the project implementation. After solving the main problem that concerned the residents – that is, changing the subway type from elevated to underground, – the project is positively perceived by the residents as a planned improvement of the transport situation. Also, the residents were promised the construction of intercept parking lots near the new stations, since the residents fear that after the appearance of new subway stations, their yards will be filled with unauthorized cars. Taking into account the fact that the residents of the district have been waiting for the construction of the subway station for decades, the project justifiably receives social approval.

The calculation results obtained using the constructed mathematical models confirmed the absence of a conflict. The algorithms developed using TDA methods show the potential of using these methods to solve similar problems.

REFERENCES

1. Batyrshin, I., Sheremetov, L., Zadeh, L.A. (Eds.) (2007). *Perception-based Data Mining and Decision Making in Economics and Finance.* Berlin, Germany: Springer-Verlag Berlin Heidelberg.

2. Chaudhuri, S., Bhardwaj, A. (2018). *Kinesthetic Perception. A Machine Learning Approach*. Singapore: Springer Singapore.

3. Ferreira, J.F., Dias, J.M. (2014). *Probabilistic Approaches to Robotic Perception.* Cham, Switzerland: Springer International Publishing.

4. Kharlamov A.A., Pilgun M. (Eds.) *Neuroinformatics and Semantic Representations. Theory and Applications.* Newcastle upon Tyne: Cambridge Scholars Publishing.

5. Les, Z., Les, M. (2020). *Machine Understanding. Machine Perception and Machine Perception MU*. Cham, Switzerland: Springer International Publishing

6. Liu, H., Sun, F. (2018). *Robotic Tactile Perception and Understanding. A Sparse Coding Method*. Singapore: Springer Singapore.

7. Gabdrakhmanova, N., Pilgun M. (2020) Development of unified approaches to building neural network and mathematical models based on digital data. *Advances in Systems Science and Applications*, 20(4), 113-124. https://doi.org/10.25728/assa.2020.20.4.

8. H. Edelsbunner, D. Letscher, A. Zomorodian (2002) Topological persistence and simplification, *Discrete Comput. Geom.,* 28, 511–533. MR 1949898

9. Carlsson, G., Zomorodian, A. (2004) Computing persistent homology, *Proc. 20th Ann. Sympos.Comput. Geom*, 347–356.

10. D. Cohen-Steiner, H. Edelsbrunner, J. Harer (2007) Stability of Persistence Diagrams, *Discrete &Computational Geometry*, 37:1, 103–120. MR 2279866

11. Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1), 77–102.

12. Borovikov, V.P., Ivchenko, G.I. (2006) Forecasting in the Statistica system in the Windows environment. *Fundamentals of theory and intensive practice on a computer. Finance and Statistics*. Moscow.

13. Gabdrakhmanova N. (2018) Forecasting time series using topological data analysis, *ITISE 2018 International Conference on Time Series and Forecasting Proceedings of Papers* 19-21 September 2018 Granada (Spain), 1367-1374.

