DOI: 10.17323/2587-814X.2024.2.35.47

Thematic modeling and linguistic analysis of text messages from a social network for information and analytical support of logistics business

Ludmila A. Borisova [a]
E-mail: la.borisova@hse.ru

Yury I. Kostyukevich [b]
E-mail: y.kostyukevich@skoltech.ru

[a] Graduate School of Business, HSE University, Moscow, Russia
[b] Skolkovo Institute of Science and Technology, Moscow, Russia

Abstract

In the modern economy, the success of a business is largely determined by the company's ability to analyze consumer preferences, consumer attitudes towards the company's products, as well as the ability to quickly respond to changing preferences or negative trends. Social listening is a technology for analyzing conversations, text messages and any kind of mention of a company, its products or brand. Currently, it is most effective to carry out social listening by monitoring social networks (VKontakte, etc.), which are the largest sources of text messages from millions of users. The purpose of this work is to analyze the practices of using social listening technology, as well as common approaches to the use of social networks by domestic and foreign companies. Based on the specialized software developed by the authors, an analysis of more than 50 000 news reports published in 2021–2024 was carried out on companies of different levels and

specialization. Using linguistic analysis of the corpus of text messages for various companies and sectors of the economy, the most common words were identified, thematic modeling was carried out, and the dynamics of news reports and their relationship with external factors were studied.

Keywords: logistics, social listening, linguistic analysis, business, brand, monitoring

Citation: Borisova L.A., Kostyukevich Y.I. (2024) Thematic modeling and linguistic analysis of text messages from a social network for information and analytical support of logistics business. *Business Informatics*, vol. 18, no. 2, pp. 35–47. DOI: 10.17323/2587-814X.2024.2.35.47

Introduction

The fourth technological revolution or Industry 4.0 [1–3] requires companies to significantly revise their business processes, create new products, flexibly interact with consumers, analyze, and predict consumer demand and preferences, as well as consumer attitudes towards their brand or company development strategy in order to remain competitive. Recently, the concept of social listening has been defined as an active process of paying attention, observing, interpreting and responding to various stimuli through indirect, electronic and social channels, and many companies actively use the method of social listening to assess the perception of a company, its products and brand by a wide consumer base [4–7]. Thus, Pomputius [8] investigated the use of social listening by healthcare organizations, universities and government organizations.

In the work of Spitale et al. [9] the authors studied the correspondence of users in Telegram messenger chats to analyze the reasons for the rejection by certain groups of citizens of the introduction of a special access regime in Italy during the COVID-19 pandemic. As a result, common beliefs peculiar to such citizens were revealed. Picon et al. [10] used a social listening approach based on the analysis of the Reddit social network to study the opinions of users with COVID-19 symptoms regarding the course of the disease, the therapy used, etc. Thus, the task was to develop new methods for obtaining additional necessary information that is not available via traditional methods. Habib et al. [11], analyzing messages on the social network Twitter, investigated how the COVID-19 pandemic changed people's preferences in choosing a vehicle.

Jang et al. [12] used the concept of social listening to determine the momentary involvement of consumers when watching movies. The authors analyzed text comments left by viewers while watching a film to analyze consumer preferences. As a result, the general concept of such a study was formulated – moment-to-moment synchronicity (MTMS).

In the work of Petrova and Trunin [13], a study was conducted of the entire body of press releases of the Bank of Russia in order to determine the tone of statements on monetary policy, identify signals about easing or tightening monetary policy, as well as the impact of press releases on key indicators of the money market. Similar work was also carried out to analyze the statements of the US Federal Reserve on economic indicators, statements by the Bank of England and others [14].

Large public B2B companies such as Coca-Cola, Nike, etc. are actively using social listening to analyze the perceived image of the company, as well as to study the attitude of consumers to the company's activities

and initiatives [15–19]. Tollinen et al. [20] investigated the use of social media monitoring to assess consumer attitudes towards companies operating on the B2B principle.

Social listening plays a special role in the logistics sector, where it allows firms to increase the efficiency of building supply chains and their design, accumulate consumer assessments of the quality of the service provided, reduce the uncertainty of planned demand and solve the problem of optimizing the search for intermediaries and suppliers (Galaskevich [21]). Gal-Tzur et al. [22] described a methodology for the automatic collection and analysis of news reports published on social networks related to the transport sector to study consumer sentiment and preferences. In the work of Guner et al. [23], more than 2 million messages published by users on social networks between 2011 and 2021 were analyzed to assess the quality of service on Turkish railways from the consumer's perspective. The authors paid special attention to the changing trends in consumer preferences. Jing et al. [24] analyzed more than 40 000 posts on China's main social networks (Sinaweibo and TikTok) to study differences in public perception of vehicles of different levels of autonomy before and after accidents involving them.

Battacharzha et al. [25] analyzed the interaction of online trading companies with consumers on the social network company Twitter. More than 200 000 communications related to logistics issues were analyzed. The result was recommendations to companies on the basic principles of building interaction with consumers on the social network, as well as prompt satisfaction of their priority requests. The potential role of the social network Twitter in the practical activities of a logistics company was also explored by Tea [26]. The analysis of more than 22 000 messages with the #supplychain label allowed us to identify key topics associated with the activities of the logistics company. Ahmadi et al. [27] analyzed more than 37 million messages on Twitter and developed approaches that allow companies to accumulate consumer feedback about their product, analyze sentiment (tonality) and use this information to develop the most effective returns logistics strategy. The study was conducted using the example of Apple phones. Later, the same authors [28] expanded their research to study the returns logistics of personal computers.

In the previous work of the authors [29, 30], a study was conducted on the use of the VKontakte social network in Russian logistics practice. An analysis of more than 30 000 news reports published in 2014–2019 by various logistics market participants in terms of level and specialization allowed us to draw conclusions about preferences in fundamental approaches to the use of social networks by various companies, as well as to study the adaptation of the Russian transport industry to the COVID-19 pandemic.

In this paper, we continue the study of the VKontakte social network with a focus on news reports published in the period 2021–2024, and, unlike our previous research, computer methods of text analysis based on NLPT and Gensim libraries for the Python programming language, which have recently been developed, are used. This paper carried out automatic thematic modeling based on Latent Dirichlet allocation (LDA) in order to determine which topics are present in the sets of downloaded messages, and which words characterize each topic.

1. A software system for analyzing text messages on the VKontakte social network

1.1. Using the API of the VKontakte social network

The general architecture of the system developed is shown in *Fig. 1*. The VKontakte social network provides the user with special software tools for developing their own applications (application programming interface, API). First of all, in order to access the interface, the user must create an application inside the VKontakte

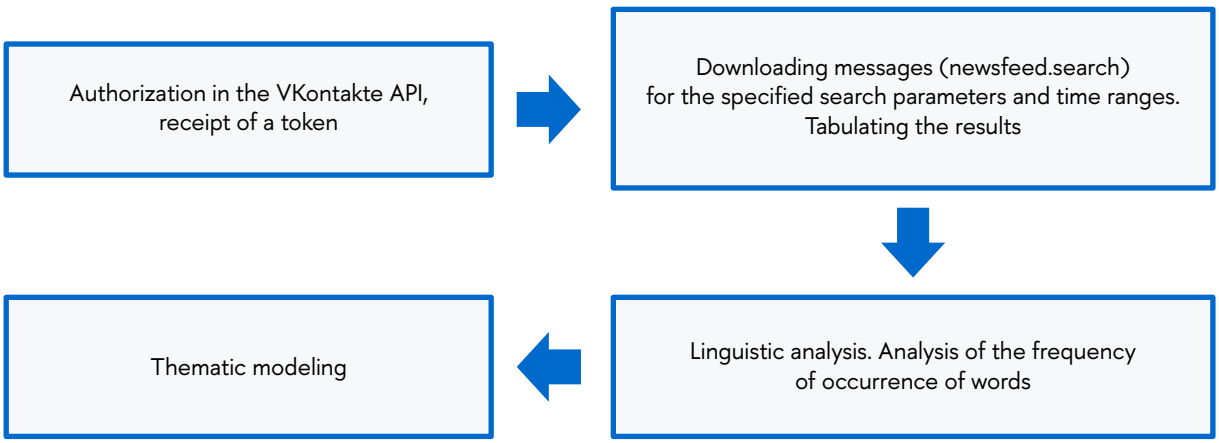

Fig. 1. The architecture of the developed system.

network and receive its identifier (MyAppID). Next, you need to make a request: https://oauth.vk.com/authorize?client_id=MyAppID&displaypage&redirect_uri=https://oauth.vk.com/blank.html&response_type=token&v=5.199

After that, the system will provide an access key (token) for using the interface. For automated downloading and analysis of news reports, we have developed software code in Python. The "newfeed.search" function provided by the VKontakte and API was used to search for news reports. The "newsfeed.search" function accepts the following arguments: a token, a search query in the form of a string, the time range for which to search. The function allows you to download up to 200 messages per search query.

1.2. Linguistic analysis of downloaded texts

For a given hashtag, we downloaded 200 news messages for each calendar week starting in 2021 (the VKontakte network does not allow downloading earlier messages). For each news item, they were automatically identified: the text of the message, the date of publication, the number of views, the ID of the community that published it, and all the hashtags. The results were compiled into a table.

The message text has been processed as follows: all characters have been converted to lowercase, punctuation, hyphenation, tabs, etc. have been removed. Then, using the NLTK library (Natural Language Toolkit), the text was converted into a sequence of individual words or other text elements that have a semantic meaning (so-called tokens). Then the tokens were lemmatized (put into dictionary form) by the *pymorphy2* package, after which standard, frequently occurring and non-meaningful words (so-called "stop words") were removed. We used a list of 561 stop words. The texts processed in this way were further analyzed, namely, the frequency of occurrence of individual words or hash tags was determined, as well as the number of publications made by each community.

1.3. Thematic modeling

For thematic modeling, the libraries "*spacy*" (preliminary text preparation and lemmatization), "*gen-*

ism" (calculations), and "*pyLDAvis*" (visualization) were used. The optimal number of topics was chosen experimentally.

2. Text analysis of VKontakte social network messages

According to MediaScope, as of 2023, the most popular social networks in Russia are VKontakte, Telegram, TikTok, Odnoklassniki, Instagram and Facebook (the latter two belong to META, recognized as extremist and banned in the Russian Federation). The summary data is shown in *Fig. 2*.

Based on the data presented in *Fig. 2*, it can be noted that in almost all age groups, with the exception of the 65+ group, the average Russian spends more than a few hours online, of which 18% of the time is spent on social networks. The most popular social network in Russia at the moment is VKontakte, with a daily reach of more than 43% of the population. Thus, the choice of the VKontakte social network as an object of research is justified.

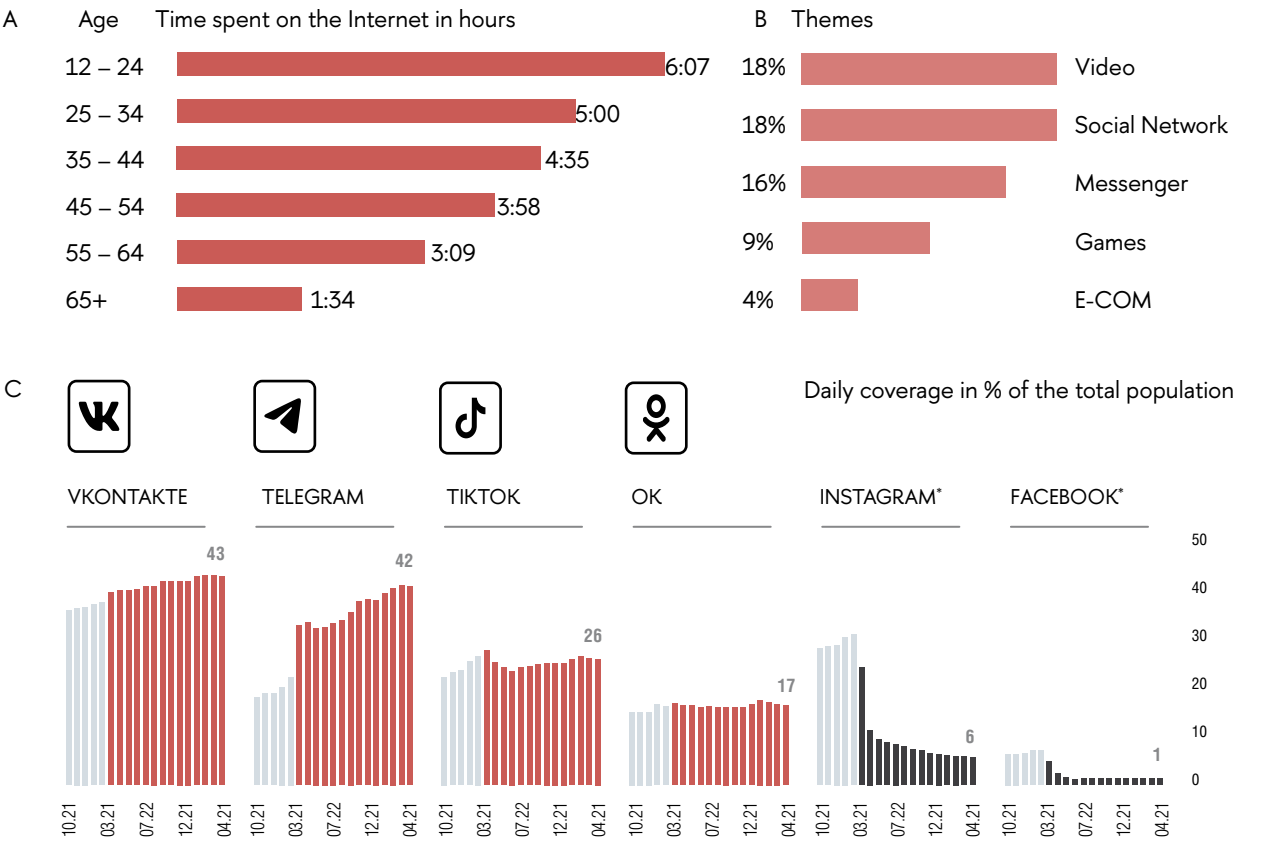

Fig. 2. (A) – the average time spent by the user on the Internet for different age groups;
(B) – the distribution of topics on which the user spends his time;
(C) – the dynamics of the average daily coverage (as a percentage of the population)
for the most popular social networks.

Table 1.

Official communities of large Russian companies on the Internet

Company	Annual turnover, billion rubles per year	Number of employees, thousands	The official VKontakte group and the number of subscribers, thousand	The official Telegram channel and the number of subscribers, thousands
Rosneft JSC	8760	330	rosneftru; 114	rosneftofficial; 18,5
Sberbank PJSC		288	sber; 3200	sberbank; 651
AvtoVAZ JSC	160	35	lada; 236	lada_rf; 9
Gazprom PJSC	8000	468	gazprom; 158	gazprom; 35
JSC Russian Railways	2500	711	rzd_official; 437	telerzd; 40
Yandex	522	25	yandex; 302	yandex; 151
PJSC Severstal	728	50	severstal; 54	severstal; 8

Numerous Russian companies, including large ones, have been actively developing their communities on the Internet for a long time, including on the VKontakte social network and the Telegram messenger. The number of subscribers in official communities can exceed several million (*Table 1*).

We have downloaded news reports published on the VKontakte social network for the period from 2021 to 2024. 200 messages were downloaded for each calendar week. The texts of these messages were processed and analyzed.

Figure 3 shows the results of a frequency analysis of the occurrence of words in downloaded texts for the queries #HSE, #RZD, #Vkusvill and #Fesco. One can see that the image of the HSE in the VKontakte social network is, as expected, associated with the words "business," "education," "project," "research." The image of the Russian Railways company is associated with the words "train," "station," etc. However, the frequent presence of the words "training," "salary," "vacancy" indicates that the company actively uses the VKontakte social network to search and hire new staff. For the request #Vkusvill the most common words are "45000," "month," "consultant," "seller," "vacancy," which also indicates the search for employees. However, it is worth paying attention to the group of words "promo code," "discount," "order," which indicates that these messages are intended to advertise and attract new consumers. At the request of #Fesco, the most commonly used terms are: "Vladivostok," "port," "service," "marine," "container," "China," "transportation," "logistics," which also corresponds to the image of this company as a large intermodal logistics operator.

For visual analysis, it is convenient to use the "word cloud" representation method. An example of

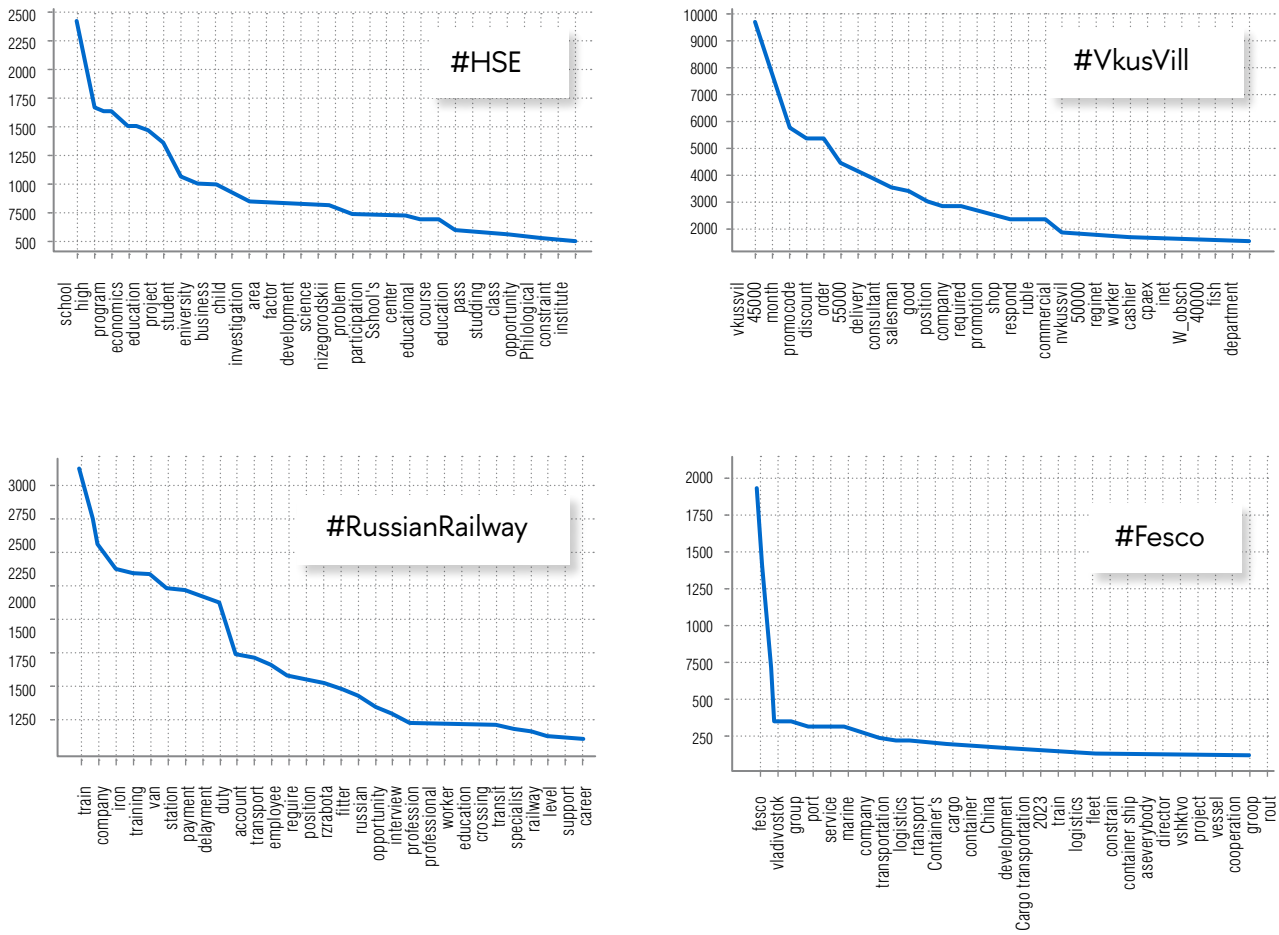

Fig. 3. Analysis of the frequency of occurrence of words
for some companies operating in various sectors of the economy.

such visualization for texts downloaded by the query #Logistics is shown in *Fig. 4*.

You can see that the most common words in texts associated with the topic of logistics are "vacancy," "order," "courier," "driver," "application," etc. It can be concluded that companies and individuals use social networks in the field of logistics to post vacancies, search and offer delivery services, mainly by road.

We also conducted thematic modeling based on Latent Dirichlet allocation (LDA). Thematic modeling allows you to automatically determine which topics are present in a set of texts, and which words characterize each topic. This is very important for the effective description and interpretation of large amounts of textual information. When conducting a thematic analysis, an important point is to choose the number of topics. Unfortunately, this parameter is determined experimentally by training the LDA model for a different number of topics and choosing the number at which we see the maximum value of the coherence of the model.

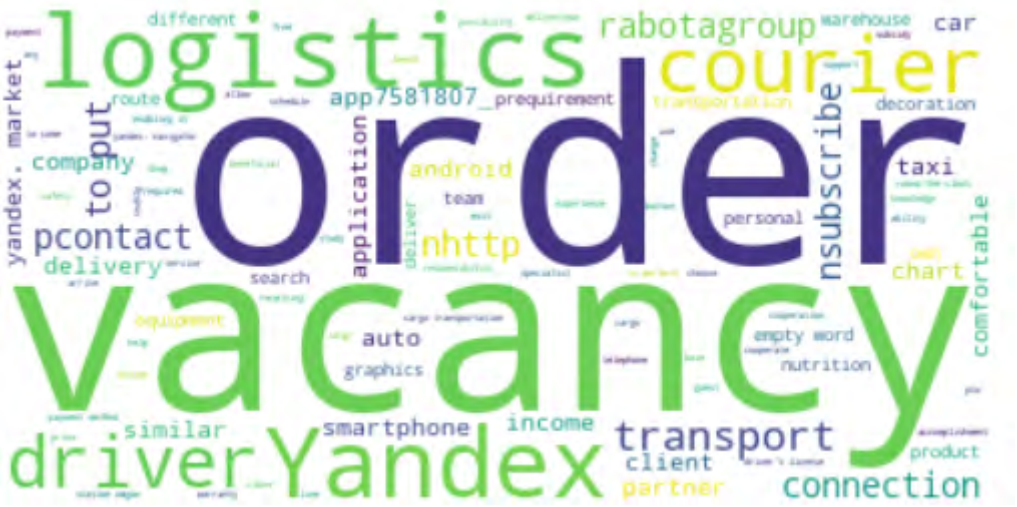

Fig. 4. Visualization of the frequency of occurrence of words using the "word cloud" method for texts downloaded on the request #Logistics.

For the corpus of texts downloaded on request #Transport, we have trained LDA models for topic values from 3 to 15. At the same time, the maximum coherence value corresponded to 11 topics. The visualization results are shown in *Fig. 5*. You can see that each topic is described by its own set of words. At the same time, topic 1 corresponds to the topic of urban infrastructure development (mainly in Moscow), topic 3 – courier job ads, topic 5 – delivery services, topic 6 – bus transportation of passengers, topic 8 – warehouse services, topic 10 – traffic accidents. As a result, you can see which transport-related topics are the most popular among VKontakte users.

Social listening also allows you to explore the response of companies, consumers, or even entire sectors of the economy to external challenges. Indeed, in the case of intervention by some significant external factor, the content, tone of news reports, as well as their number, will change markedly. This effect has been studied using the example of the logistics industry. *Figure 6* shows monthly statistics on the number of news reports that simultaneously contained the hashtags #Transport and #Logistics, as well as the most active communities that sent messages. The number of subscribers is indicated for the communities. The analysis showed that the overwhelming number of news reports with these hashtags were announcements of a courier job offer. Most of the messages were posted on the Rabota.Group community network.

On the graph of the number of news reports per month, you can see characteristic peaks corresponding to February 2022, October 2022, May 2023. These peaks can probably be interpreted as the industry's reaction to external macroeconomic and social phenomena.

Conclusion

Social listening based on automated monitoring of social networks has become an important tool that allows companies to obtain operational information to analyze consumer attitudes towards their brand, prod-

Fig. 5. Thematic analysis of the corpus of texts downloaded on request #Transport.
A map of the distances between the topics
and the 10 most characteristic words corresponding to the selected topics.

ucts, service level, etc., thereby facilitating the solution of one of the most urgent and difficult tasks in business – demand modeling. Indeed, in the modern world, almost every user has an account in one or more social networks and spends a significant part of their time reading news reports or creating their own content. Companies, government and non-profit organizations also have their own communities on social networks and regularly publish news reports aimed at informing users about new events, innovations, goods or services. Social networks are also used as platforms for recruiting, searching for suppliers, and promoting goods or services. Global practice of recent years indicates active use by large businesses, including logistics (for example, in distribution and retail, in the food sector), of this tool, usually provided by specialized high-budget digital agencies, which is difficult for small businesses to access. At the macroeconomic level, social listening allows us to study the reaction of society and economic sectors to external factors.

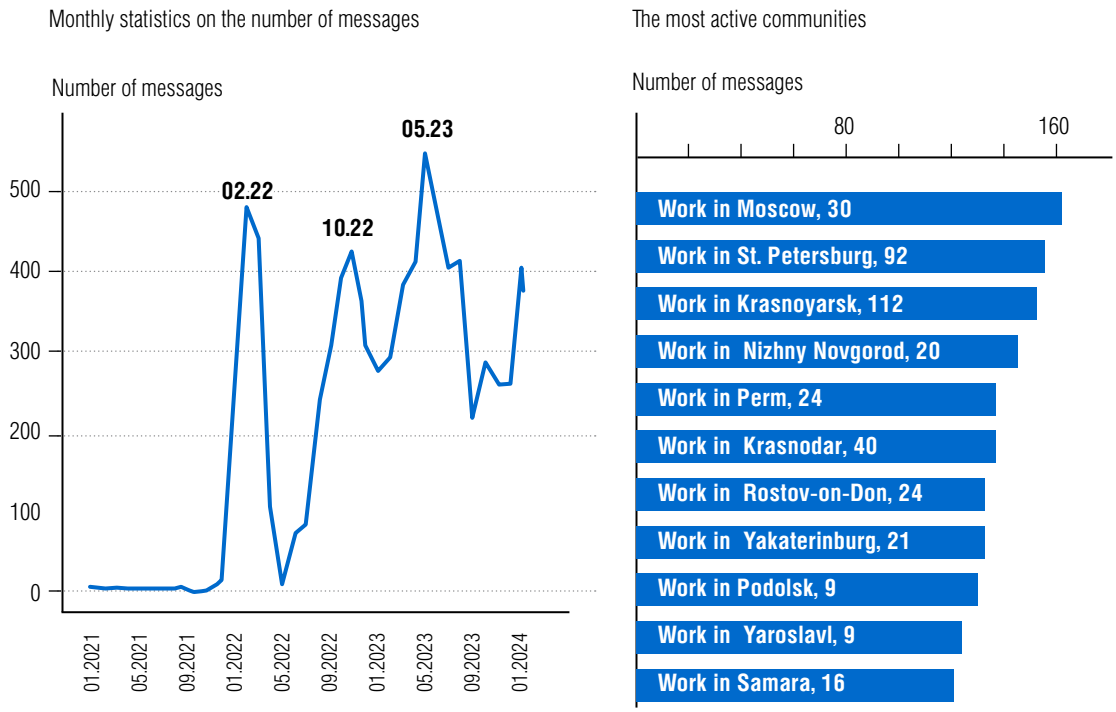

Fig. 6. Monthly statistics on the number of news reports that simultaneously contained the hashtags #Transport and #Logistics, as well as the most active communities that sent messages.

The number of subscribers is indicated for the communities. These communities are members of the Rabota. Group community network.

This paper presents the results of a study of the effectiveness of social listening based on automated search and linguistic analysis of all published news reports on the VKontakte social network. We show the possibility of quick and cost-effective identification of words and key topics discussed associated with a particular brand, company or industry. This tool is especially relevant for small businesses, as a preferred alternative to approaches that require additional investments to find solutions to enter certain markets based on more reliable demand forecasting. In addition, the growing need for such effective tools for studying and planning demand based on open information in social networks is intensifying in the logistics sector, while it is expected that the trends of recent years regarding the radical transformation of supply chains will only strengthen this trend.

It is worth noting, however, that the retrospective analysis of news reports is complicated by the fact that published news reports on social networks can be deleted or changed at the user's request. For example, numerous Western companies, when shutting down their business in Russia, also closed their official social media communities, which led to the deletion of all news reports published in them. As a result, these messages are not currently available for analysis.

Predicting the further development of this process, one can expect such an additional effect of the active introduction of social listening to be the realization of the possibility of consumers directly, by actively expressing their preferences and wishes, to stimulate companies to develop new products and services or make other adjustments to their activities. ■

References

1. Barreto L., Amaral A., Pereira T. (2017) Industry 4.0 implications in logistics: an overview. *Procedia Manufacturing*, vol. 13, pp. 1245–1252. https://doi.org/10.1016/j.promfg.2017.09.045
2. Speranza M.G. (2018) Trends in transportation and logistics. *European Journal of Operational Research*, vol. 264, no. 3, pp. 830–836. https://doi.org/10.1016/j.ejor.2016.08.032
3. Winkelhaus S., Grosse E.H. (2020) Logistics 4.0: A systematic review towards a new logistics system. *International Journal of Production Research*, vol. 58, no. 1, pp. 18–43. https://doi.org/10.1080/00207543.2019.1612964
4. Westermann A., Forthmann J. (2021) Social listening: a potential game changer in reputation management How big data analysis can contribute to understanding stakeholders' views on organisations. *Corporate Communications: An International Journal*, vol. 26, no. 1, pp. 2–22. https://doi.org/10.1108/CCIJ-01-2020-0028
5. Cano-Marin E., Mora-Cantallops M., Sánchez-Alonso S. (2023) Twitter as a predictive system: A systematic literature review. *Journal of Business Research*, vol. 157, article 113561. https://doi.org/10.1016/j.jbusres.2022.113561
6. Yang J., Xiu P., Sun L., Ying L., Muthu B. (2022) Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management*, vol. 59, no. 1, article 102751. https://doi.org/10.1016/j.ipm.2021.102751
7. Saura J.R., Ribeiro-Soriano D., Palacios-Marqués D. (2022) Evaluating security and privacy issues of social networks based information systems in Industry 4.0. *Enterprise Information Systems*, vol. 16, nos. 10–11, pp. 1694–1710. https://doi.org/10.1080/17517575.2021.1913765
8. Pomputius A. (2019) Can you hear me now? Social listening as a strategy for understanding user needs. *Medical Reference Services Quarterly*, vol. 38, no. 2, pp. 181–186. https://doi.org/10.1080/02763869.2019.1588042
9. Spitale G., Biller-Andorno N., Germani F. (2022) Concerns around opposition to the green pass in Italy: social listening analysis by using a mixed methods approach. *Journal of Medical Internet Research*, vol. 24, no. 2, article e34385. https://doi.org/10.2196/34385
10. Picone M., Inoue S., DeFelice C., Naujokas M. F., Sinrod J., Cruz V. A., Stapleton J., Sinrod E., Diebel S.E., Wassman E.R. (2020) Social listening as a rapid approach to collecting and analyzing COVID-19 symptoms and disease natural histories reported by large numbers of individuals. *Population Health Management*, vol. 23, no. 5, pp. 350–360. https://doi.org/10.1089/pop.2020.0189
11. Habib M.A., Anik M.A.H. (2023) Impacts of COVID-19 on transport modes and mobility behavior: Analysis of public discourse in twitter. *Transportation Research Record*, vol. 2677, no. 4, pp. 65–78. https://doi.org/10.1177/03611981211029926

12. Zhang Q., Wang W., Chen Y. (2020) Frontiers: In-consumption social listening with moment-to-moment unstructured data: The case of movie appreciation and live comments. *Marketing Science*, vol. 39, no. 2, pp. 285–295. https://doi.org/10.1287/mksc.2019.1215

13. Petrova D.A., Trunin P.V. (2021) Analysis of the impact of central bank communications on money market indicators. *Business Informatics*, vol. 15, no. 3, pp. 24–34. http://doi.org/10.17323/2587-814X.2021.3.24.34

14. Hansen S., McMahon M. (2016) Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, vol. 99, pp. 114–133. https://doi.org/10.1016/j.jinteco.2015.12.008

15. Liu Y., Lopez R.A. (2016) The impact of social media conversations on consumer brand choices. *Marketing Letters*, vol. 27, p. 1–13. https://doi.org/10.1007/s11002-014-9321-2

16. Austin L.L., Gaither B.M. (2016) Examining public response to corporate social initiative types: A quantitative content analysis of Coca-Cola's social media. *Social Marketing Quarterly*, vol. 22, no. 4, pp. 290–306. https://doi.org/10.1177/1524500416642441

17. Reid E., Duffy K. (2018) A netnographic sensibility: Developing the netnographic/social listening boundaries. *Journal of Marketing Management*, vol. 34, nos. 3–4, pp. 263–286. https://doi.org/10.1080/0267257X.2018.1450282

18. De Luca F., Iaia L., Mehmood A., Vrontis D. (2022) Can social media improve stakeholder engagement and communication of Sustainable Development Goals? A cross-country analysis. *Technological Forecasting and Social Change*, vol. 177, article 121525. https://doi.org/10.1016/j.techfore.2022.121525

19. Vitellaro F., Satta G., Parola F., Buratti N. (2022) Social Media and CSR Communication in European ports: the case of Twitter at the Port of Rotterdam. *Maritime Business Review*, vol. 7, no. 1, pp. 24–48. https://doi.org/10.1108/MABR-03-2021-0020

20. Töllinen A., Järvinen J., Karjaluoto H. (2012) Social media monitoring in the industrial business to business sector. *World Journal of Social Sciences*, vol. 2, no. 4, pp. 65–76.

21. Galaskiewicz J. (2011) Studying supply chains from a social network perspective. *Journal of Supply Chain Management*, vol. 47, no. 1, pp. 4–8. https://doi.org/10.1111/j.1745-493X.2010.03209.x

22. Gal-Tzur A., Grant-Muller S.M., Kuflik T., Minkov E., Nocera S., Shoor I. (2014) The potential of social media in delivering transport policy goals. *Transport Policy*, vol. 32, pp. 115–123. https://doi.org/10.1016/j.tranpol.2014.01.007

23. Güner S., Taşkın K., Cebeci H.İ., Aydemir E. (2022) Service quality in rail systems: listen to the voice of social media. *PREPRINT (Version 1) available at Research Square*. https://doi.org/10.21203/rs.3.rs-1980183/v1

24. Jing P., Cai Y., Wang B., Wang B., Huang J., Jiang C., Yang C. (2023) Listen to social media users: Mining Chinese public perception of automated vehicles after crashes. *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 93, pp. 248–265. https://doi.org/10.1016/j.trf.2023.01.018

25. Bhattacharjya J., Ellison A., Tripathi S. (2016) An exploration of logistics-related customer service provision on Twitter: The case of e-retailers. *International Journal of Physical Distribution & Logistics Management*, vol. 46, nos. 6/7, pp. 659–680. https://doi.org/10.1108/IJPDLM-01-2015-0007

26. Chae B.K. (2015) Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, vol. 165, pp. 247–259. https://doi.org/10.1016/j.ijpe.2014.12.037

27. Ahmadi S., Shokouhyar S., Shahidzadeh M.H., Papageorgiou E.I. (2022) The bright side of consumers' opinions of improving reverse logistics decisions: a social media analytic framework. *International Journal of Logistics Research and Applications*, vol. 25, no. 6, pp. 977–1010. https://doi.org/10.1080/13675567.2020.1846693

28. Ahmadi S., Shokouhyar S., Amerioun M., Tabrizi N.S. (2024) A social media analytics-based approach to customer-centric reverse logistics management of electronic devices: A case study on notebooks. *Journal of Retailing and Consumer Services*, vol. 76, article 103540. https://doi.org/10.1016/j.jretconser.2023.103540

29. Borisova L., Kostyukevich Y. (2020) Digitalization of logistics: what is the role of social networks? *Logistics and Supply Chain Management*, no. 3, pp. 44–50 (in Russian).

30. Borisova L. (2021) Logistics – the Eurasian Bridge. Materials of the *XVI International scientific and practical conference (April 28 – May 01, 2021, Krasnoyarsk, Yeniseisk)*. Krasnoyarsk: Krasnoyarsk State University, pp. 15–18 (in Russian).

About the authors

Ludmila A. Borisova

Cand. Sci. (Econ.);

Associate Professor, Department of Operations Management and Logistics, Graduate School of Business, HSE University, 26–28, Shabolovka St., Moscow 119049, Russia;

E-mail: la.borisova@hse.ru

ORCID: 0000-0003-0691-4519

Yury I. Kostyukevich

Dr. Sci. (Chem.);

Associate Professor, Skolkovo Institute of Science and Technology, the territory of the Skolkovo Innovation Center, Bolshoy Blv., 30, p. 1, Moscow 121205, Russia;

E-mail: y.kostyukevich@skoltech.ru

ORCID: 0000-0002-1955-9336