

Algorithm for Digitalising Technological Schedules of Train Processing Operations at Railway Stations



Zhivko YANEV



Nikita V. LUGOVSKY



Yuri O. PAZOYSKY



Sergey V. KALININ

Zhivko Yanev¹, Nikita V. Lugovsky², Yuri O. Pazoysky³, Sergey V. Kalinin⁴

^{1, 3}Russian University of Transport, Moscow, Russia.

^{1, 2, 4}Design & Research Institute for Information Technology, Signaling and Telecommunication on Railway Transport (JSC NIAS), Moscow, Russia.

¹ ORCID: <https://orcid.org/0000-0002-0856-341X>; Russian Science Citation Index SPIN-code: 8538-7511; Russian Science Citation Index AuthorID: 1092590.

³ Russian Science Citation Index SPIN-code: 5355-5945; Russian Science Citation Index AuthorID: 403168.

⁴ Russian Science Citation Index SPIN-code: 4957-2295; Russian Science Citation Index AuthorID: 403168.

✉ ¹ zivkoacter@yahoo.com.

ABSTRACT

Changing volumes of cargo transportation necessitate accelerated movement of cargo flows, including by modifying the technology of trains' transit through the railway infrastructure. In such conditions, it is necessary to create digital models, the purpose of which is to reproduce the work of the original. The results obtained following the operation of the digital model will serve as a rationale for further development of options for operation of the simulated object, aimed at fulfilling the up-and-coming indicators of operation of railway transport.

In this regard, as part of the study, an algorithm for digitalising technological schedules of train processing

operations. The purpose of the developed algorithm is to build software conform to its structure that ensures operational automated accounting of the technology of operations of original railway facilities in the conditions of transforming analogue information into its digital type. An algorithm based on a machine learning model was created using program, structural and system methods. The accuracy of determining the input technological operation is assessed by the purity of the information node, and more than 120 technological schedules of train processing operations at various railway stations were digitalised during the experiment.

Keywords: digital model of a railway station, digitalised technological schedule of train processing, stop word, lemmatisation, multiclass classification (random forest), text vectorisation, TF-IDF method, Gini index.

For citation: Yanev, Zh., Lugovsky, N. V., Pazoysky, Yu. O., Kalinin, S. V. Algorithm for Digitalising Technological Schedules of Train Processing Operations at Railway Stations. World of Transport and Transportation, 2024, Vol. 22, Iss. 4 (113), pp. 141–148. DOI: <https://doi.org/10.30932/1992-3252-2024-22-4-2>.

The original text of the article in Russian is published in the first part of the issue.

Текст статьи на русском языке публикуется в первой части данного выпуска.

BACKGROUND

The need for prompt train traffic processing, particularly when reorientating of cargo traffic volumes, requires adjustments to the technology of railway stations operations [1]. In this regard, it is necessary to promptly prove the correctness of the proposed measures to change the infrastructure configuration or technology of railway station operations. It is possible through creating digital models of railway stations consisting of a block (unit) of a digital infrastructure model of a railway station and a block describing the technology of processing trains and resource elements included in it. A detailed structure of a digital model of a railway station is provided in the work [2]. The proposed blocks should ensure a high level of identity of the behaviour of the original in the virtual world, which, by creating digital infrastructure models and their digitalised technological schedules of train processing operations, will allow for accelerated or on-line planning and management of railway stations' operations, which is a very urgent task facing the railway transport, namely in the Russian Federation.

The works [3; 4] provide a definition of a digital model (digital twin¹) of a railway station, according to which digital models are generally intended to simulate the behaviour of the original under conditions of changing initial values and operational technology.

Digital models of a railway station can also be used to develop scenarios for the behaviour of objects within the framework of on-line management of railway transport. The authors of [5–7] concluded that the information base of technological parameters should include technological standards for operation of railway stations for subsequent operational forecasting.

The authors of [8] provide a list of the departments of JSC Russian Railways that are primarily subject to implementation and use of digital and quantum tools to improve customer service by advancing on-line decision-making. The study notes the importance of using digital modules in the process of planning the operation of the station, which confirms the relevance of developing a module for digitalising the technological

schedule for train processing operations.

To simulate the behaviour of an existing railway station (the original) under conditions of constant changes in initial values and operational technology, it is necessary to promptly adjust the technological schedules for train processing operations.

Hence, the *objective* of the study is development of a procedure for digitalising technological schedules of train processing operations automatically and with minimal time losses.

The proposed algorithm should be universal for subsequent use at the stages of preparing large-scale concepts and projects for digital transformation of companies [9].

The architecture of the algorithm for digitalisation of technological schedules of train processing operations was outlined using the *structural approach*, as well as its substantive part, consisting of five main blocks: input of initial information, byte transformation, lemmatisation, vectorisation, and issuing of a ready-made digitalised technological schedule of train processing operations.

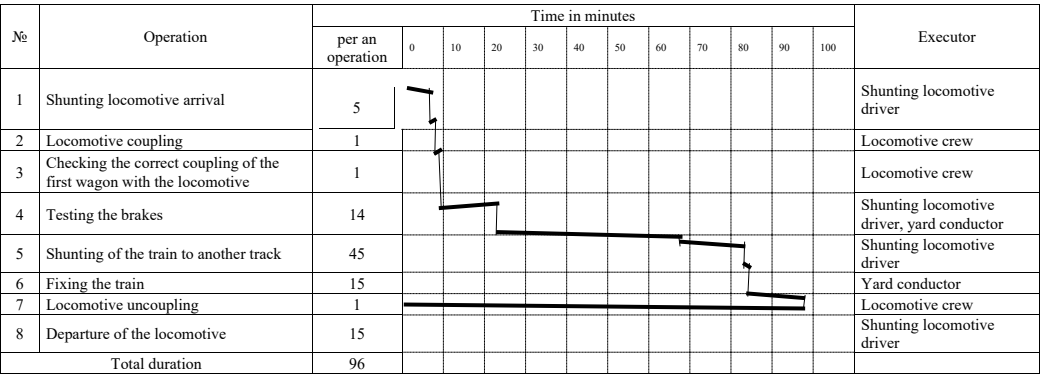
The system approach allowed us to formulate a list of requirements for the developed algorithm.

Using *the programming method*, the blocks were provided with the necessary functionality; several experiments were conducted with the developed software to digitalise the corresponding technological schedules of train processing operations. Based on the results of the use of the software, the purity of the information node is assessed, digital process chains are formed, and the latter are loaded into the multilevel management system software to further simulate the railway station operation.

When adjusting, updating or describing the operation technology of existing or newly created digital models of railway stations, modeller engineers² need from several hours to several days to recreate the technological behaviour of a real railway station. In this regard, this article proposes an algorithm for constructing and transforming technological schedules of train processing operations into

¹ Digital twins of RZD. How Russian Railways Use Digital Twins. [In Russian]. [Electronic resource]: <https://twins.rzddigital.ru/twins>.

² A modeller engineer is an employee with competencies and skills in the field of railway operations; experience in creating digital models of railway facilities and having a total work experience of at least two years. – *authors' note*.



Pic. 1. Example of a technological schedule of train processing operations [performed by the authors].

their digital type. The algorithm will result in a digitalised technological schedule of train processing operations³.

RESULTS

The data entry is input source file containing the technological schedules of train processing operations (Pic. 1).

Technological schedules for train processing operations are provided for in the regulatory documents of each railway station (as a technological process, technological map). They can be presented in the form of a drawing, graph, chart or table indicating the duration and sequence of execution of relevant operations.

The software developed according to the proposed algorithm (Pic. 2) should have the following functionality:

- To recognise the received information of two types (text and image).
- To convert the received type of information into byte format.
- To process text sentences and images for training the machine learning model.
- To vectorise the text using Term Frequency-Inverse Document Frequency method (the TF-IDF method; the choice of the TF-IDF method is explained by the fact that, for instance, the Word2vec method is more suitable for vectorising text in neural networks, and by the reasons described below).
- To create a digitalised technological schedule of train processing operations.

³ A digitalised technological schedule of train processing operations is a digital twin of an analogue technological schedule of train processing operations obtained through a specialised algorithm that considers the standard duration and sequence of operations with a train or other facilities at a railway station. – *authors' note.*

After entering the initial information in the form of text or image, in accordance with the algorithm proposed in Pic. 2, the required information is selected by removing stop words from the sentences. After cleaning the processed text from stop words, the proposed algorithm launches the lemmatisation mechanism [11].

The algorithm for digitalising the technological schedule of train processing operations is built on the software using the Python programming language.

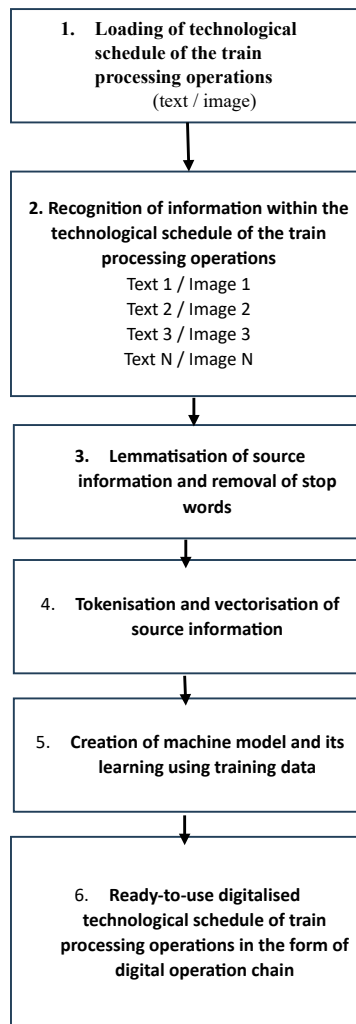
To recognise information obtained from the technological schedule of train processing operations, the algorithm must include two functions:

- The function of reading data from text: if the technological schedule is presented in the form of a table or a document with text, then the function takes the form shown in Pic. 3.
- The function for reading data from an image: if the technological schedule is presented in the form of an image or raster graphics, then the function takes the form shown in Pic. 4.

After the software has received information from two types of sources (text or image), the subsystem for converting the corresponding information into byte form is launched, as shown in Pic. 5.

The successful and error-free operation of the proposed text processing algorithm is influenced by tokenisation and vectorisation of the text [12]. Tokenisation is performed to separate the name of the technological operation into individual words while preserving their order. Text vectorisation is necessary to ensure a high degree of perception of the information provided by the machine learning model. The TF-IDF method [13; 14] was chosen for several reasons since it:





Pic. 2. Algorithm for digitalising the technological schedule of train processing operations [performed by the authors].

```

def read_tables_from_docx(docx_file):
    # Load the Word document
    doc = Document(docx_file)

    all_tables = []

    # Iterate through all tables in the document
    for table in doc.tables:
        table_data = []

        # Iterate through each row in the table
        for row in table.rows:
            row_data = []

            # Iterate through each cell in the row
            for cell in row.cells:
                # Get the text in each cell and append to row_data
                row_data.append(cell.text)

            # Append the row data to table_data
            table_data.append(row_data)

        # Append the table data to all_tables
        all_tables.append(table_data)

    return all_tables
  
```

Pic. 3. A fragment of the function for reading data from the text of the technological schedule initially composed in analogue format [performed by the authors].

```
def extract_images_from_docx(docx_file):
    doc = Document(docx_file)
    images = []

    for s in doc.inline_shapes:
        blib = s._inline.graphic.graphicData.pic.blipFill.blip
        rId = blib.embed
        document_part = doc.part
        image_part = document_part.related_parts[rId]
        image_bytes = image_part.blob
        images.append(image_bytes)

    return images
```

Pic. 4. A fragment of the function for reading data from the drawing of the technological schedule [performed by the authors].

No	Name of an operation	Duration of an operation	Digital form
1	Showing locomotive arrival	5	...
2	Showing locomotive coupling	1	...
3	Checking the correct coupling of the first wagon with the shunting locomotive	1	...
4	Testing the brakes	14	...
...

Pic. 5. An example (fragment) of transformation of the corresponding technological operation into byte form [performed by the authors].

- considers the importance of the word in the context of the document;
- identifies keywords;
- eliminates frequently occurring words;
- scales well (can be applied to large text corpora).

Vectorisation was performed using the TF-IDF method according to formula (1):

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right), \tag{1}$$

where $w_{x,y}$ – the value of words;
 $tf_{x,y}$ – the proportion of repetitions of a word x in one sentence in relation to the total number of such words x in all sentences;

N – number of sentences in the sample;
 df_x – the number of sentences from the total sample containing the word x .

Text vectorisation is the transformation of tokens into numbers. This action is necessary for further training, additional training and adjustment of the algorithm during operation of the machine learning model.

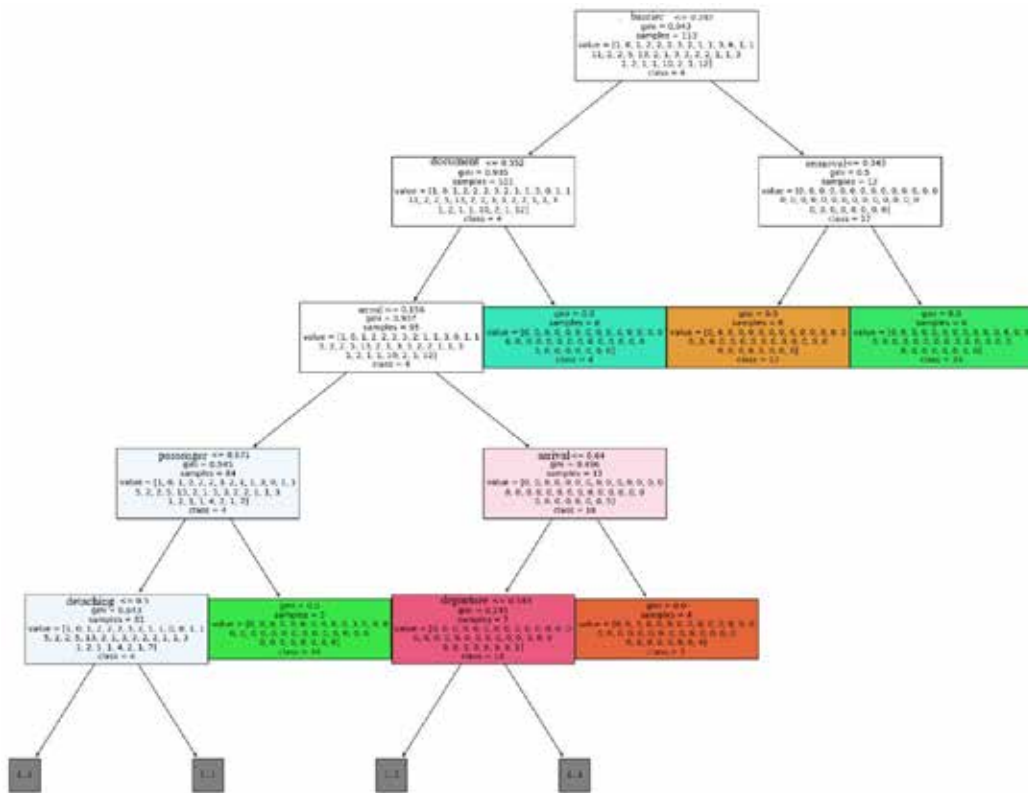
After successful tokenisation and vectorisation of the text, the technological schedule for train processing operations previously shown in Pic. 5 takes the form shown in Pic. 6.

Correct result of the machine learning model can be obtained through processing a larger



	shoe	blank	crew	car	departure	delivery	landing	stretching	hump	readiness	—	tgel	technical	brake	brake
0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
1	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
2	0.0	0.0	0.0	0.357503	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
3	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.707107	0.0
4	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
...
108	0.0	0.0	0.0	0.000000	0.0	0.000000	0.725470	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
109	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
110	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
111	0.0	0.0	0.0	0.336182	0.0	0.000000	0.000000	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0
112	0.0	0.0	0.0	0.000000	0.0	0.449994	0.535292	0.0	0.0	0.0	—	0.0	0.0	0.000000	0.0

Pic. 6. Results of tokenisation and vectorisation of the text of the technological schedule of train processing operations [performed by the authors].



Pic. 7. Structure of a machine learning model using multiclass classification (random forest) [performed by the authors].

amount of training data. Therefore, it is necessary to select an appropriate machine learning model that will help to form a set of skills, competencies and knowledge aimed at producing a digitalised technological schedule for train processing operations.

The machine learning model was trained in this study using multiclass classification (random forest), which ensures good processing of large data, stability of retraining and treating diverse data. The structure of the machine learning model is shown in Pic. 7.

The proposed multiclass classification machine learning model structure used Gini

impurity measurement [15; 16]. It measures the probability that a randomly selected element from the set will be incorrectly classified if it was randomly labelled according to the distribution of labels in the node. For a node with many elements belonging to different classes, Gini impurity is calculated as follows: If all the elements in the node belong to the same class, Gini impurity will be 0, which indicates the best possible purity of the node. However, if the elements are uniformly distributed across different classes, Gini impurity will be close to 0,5, which indicates the maximum uncertainty or heterogeneity of the node purity [15; 16].

```

parameters={'max_depth': [2,5,7,10], 'n_estimators': [10,50,100], 'min_samples_split': [10,20,30], 'min_samples_leaf': [10,20,30]}

scoring = {'accuracy': 'accuracy',
           'precision': 'precision_macro',
           'recall': 'recall_macro',
           'f1': 'f1_macro'}

clf = RandomForestClassifier()

grid_search_cv_clf=GridSearchCV(clf, parameters, cv=5, scoring = scoring, refit='accuracy')#добавление еще процесса-выбора лучшей с помощью cv
grid_search_cv_clf.fit(vectorize_text, numeric_labels)
best_params = grid_search_cv_clf.best_params_

```

Pic. 8. List of parameters showing the accuracy of a machine learning model [performed by the authors].

The accuracy of the machine learning model for digitalising the technological schedule of train processing operations can be formed based on the parameters (tree depth, number of trees, minimum number of samples, and minimum number of split sheets) shown in Pic. 8.

CONCLUSIONS

The machine learning model was applied to 127 experiments on digitalisation of technological schedules of train processing operations. Pic. 9 shows the final parameters of the machine learning model quality assessment regarding digitalisation of the technological schedule of train processing operations.

The accuracy assessment of the machine learning model for digitalising the technological schedule of train processing operations using training data is shown in Pic. 10 and amounted to 0,938 conventional units, which corresponds to a high level of accuracy in determining, perceiving, and converting the information received.

The implementation of the proposed algorithm resulted in a digitalised technological schedule of train processing operations in the form of a digital technological chain used in the process of modelling the operation of a railway station (Pic. 11).

The proposed algorithm for digitalising the technological schedule of train processing operations can be used for:

- Processing and transmitting orders, instructions and messages on board a locomotive, as well as to dispatchers (the proposed algorithm for processing analogue information is universal, therefore it allows for the digitalisation of documents that provide advance warning to the train crew about upcoming emergency restraints affecting the running speed of the train).

- Developing technological schedules for train processing operations at non-public railway stations. (Technological schedules for train processing operations at non-public railway

best_params

```

{'max_depth': 5,
 'min_samples_leaf': 10,
 'min_samples_split': 10,
 'n_estimators': 100}

```

Pic. 9. Numerical values of the model parameters providing the best values of accuracy metrics during the experiments conducted [performed by the authors].

```

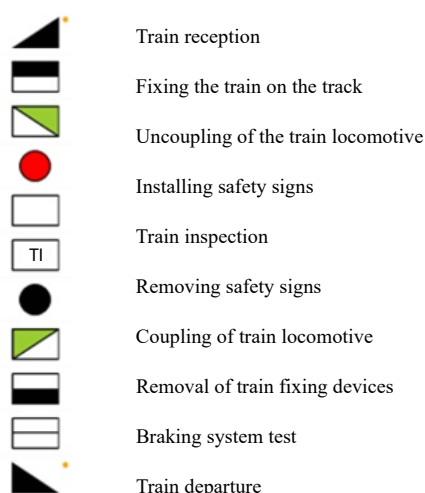
accuracy_score(clf.predict(vectorize_text), numeric_labels)

0.9380530973451328

```

Pic. 10. Assessment of accuracy of the artificial intelligence machine learning model for digitalising the technological schedule of train processing operations [performed by the authors].

stations are stored in analogue form. At the stage of describing the technology of operation of such facilities, it is necessary to create digital technological chains manually. This process takes from one week to two months, depending on the scale of the modelled object. Reduction of time costs for describing the technology of operation in this case will be achieved using the



Pic. 11. An example of a digital technological chain [performed by the authors].



proposed algorithm and software module for digitalising technological schedules of train processing operations).

– Shaping technological chains for train processing within the framework of projects of the Digital Railway Station (the multivariant mode of operation of the facility in the process of modelling requires prompt changes in the time standards for the implementation of technological schedules. The change in the technological schedules for train processing operations, as well as in the entered time standards, can be performed using the proposed software module and algorithm).

REFERENCES

1. Chigrin, N. S., Borisenkov, E. S., Grefenstein, A. P. Characteristics of the problems of cargo transportation by rail on Eastern direction [In Russian]. *Technique and technology of transport*, 2023, Iss. 4 (31). EDN: ZKCJKE.
2. Yanev, Zh. About the structure of a digital model of a railway station [IN Russian]. *Transport business of Russia*, 2024, Iss.3, pp. 186–190. EDN: OJZHAY.
3. Kostenko, V. V., Bogdanovich, D. E. Practical application of a digital model for a railway station reconstruction feasibility study. *Bulletin of scientific research results*, 2021, Iss. 1, pp. 61–73. DOI: 10.20295/2223-9987-2021-1-61-73.
4. Golovnich, A. K. Determining the concept of «digital twin» in 3D models of railway stations [*Determinatsiya ponyatiya «tsifrovoy dvoynik» v SD-modelyakh zhheleznodorozhnykh stantsii*]. *Transport technician: education and practice*, 2023, Iss. 4 (2), pp. 184–192. DOI: <https://doi.org/10.46684/2687-1033.2023.2.184-192>.
5. Erofeev, A. A. Normative and reference information in the train formation system [*Normativno-spravochnaya informatsiya v sisteme poezdooobrazovaniya*]. *Bulletin of Belarusian State University of Transport: Science and Transport*, 2007, Iss. 1–2 (14–15), pp. 54–59. EDN: YSDPLF.
6. Lysikov, M. G., Olshansky, E. N., Rozenberg, E. N., Rozenberg, I. N. Patent No. 2723051 C1 Russian Federation, IPC B61B 1/00. Operational control system for transit train traffic: No. 2019131595: application 08.10.2019: published 08.06.2020. Applicant: Joint-Stock Company «Research and Design Institute of Informatization, Automation and Communications in Railway Transport». EDN: UTJFTN.
7. Kostenko, O. A., Parkhomenko, N. V. Principles of constructing a software digital model of a marshalling yard and its study on a digital computer [*Printsipy postroeniya programmoi tsifrovoy modeli sortirovochnoi stantsii i ee*

issledovanie na ECVI] [In Russian]. Moscow, Nauka publ., 1967, pp. 118–133.

8. Khomonenko, A., Khalil, M. M. Quantum computing in controlling railroads. In: E3S Web of Conferences: International Scientific Conference Transport Technologies in the 21st Century (TT21C-2023) «Actual problems of Decarbonization of Transport and Power Engineering: Ways of Their Innovative Solution», Rostov-on-Don, Russia, 05–07 April 2023, Vol. 383. A. Bieliatynskiy and A. N. Guda (Eds.). Rostov-on-Don, Russia: EDP Sciences, 2023, art. 01010. DOI: <https://doi.org/10.1051/e3sconf/202338301010>.

9. Lakemond, N., Holmberg, G., Pettersson A. Digital Transformation in Complex Systems. *IEEE Transactions on Engineering Management*, 2024, Vol. 71, pp. 192–204. DOI: 10.1109/tem.2021.3118203.

10. Potyupkin, A. Yu., Chechkin, A. V. Artificial intelligence based on information and system redundancy [*Iskusstvennyi intellekt na baze informatsionno-sistemnoi izbytochnosti*]. Moscow, Scientific and Publishing Center INFRA-M LLC, 2019, 384 p. ISBN 978-5-907064-44-7.

11. Zherdeva, M. V., Artyushenko, V. M. Stemming and lemmatization in lucene.net. *Bulletin of the Moscow State Forest University – Forest Bulletin*, 2016, Vol. 20, Iss. 3, pp. 131–134. EDN: WKNMTN.

12. Chelyshev, E. A., Otsokov, Sh. A., Raskatova, M. V., Shchegolev, P. Comparing classification methods for news texts in Russian using machine learning algorithms. *Proceedings in Cybernetics*, 2022, Iss. 1 (45), pp. 63–71. DOI: 10.34822/1999-7604-2022-1-63-71.

13. Oskina, K. A. Optimisation of TF-IDF text classification method by introducing additional weighting coefficients. *Bulletin of Moscow State Linguistic University. Humanities*, 2016, Iss. 15 (754), pp. 175–187. EDN: OPWIAU.

14. Polyakova, A. S., Lipinsky, L. V., Poplauhina, M. A. [et al.] Certificate of state registration of a computer program No. 2021681722 Russian Federation. An intelligent system for solving natural language analysis problems based on the text vectorization procedure using machine learning models, a genetic algorithm and the TF-IDF method: No. 2021681149: application 17.12.2021: published 24.12.2021; applicant: Federal State Budgetary Educational Institution of Higher Education «Siberian State University of Science and Technology named after Academician M. F. Reshetnev». EDN: GVNFKK.

15. Disha, R. A., Waheed, S. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. *Cybersecurity*, 2022, Vol. 5, Iss. 1, pp. 1–22. DOI: 10.1186/s42400-021-00103-8.

16. Demetriou, D., Michailides, C., Onoufriou, T., Papanastasiou, G. Coastal zone significant wave height prediction by supervised machine learning classification algorithms. *Ocean Engineering*, 2021, Vol. 221, art. 108592. DOI: 10.1016/j.oceaneng.2021.108592. ●

Information about the authors:

Yanev Zhivko, Senior Lecturer at the Department of Railway Stations and Transport Hubs of Russian University of Transport; Chief Simulation Modelling Specialist of Scientific and Technical Complex for Digital Modelling (STC DM) named after V. I. Umansky of Research and Design Institute of Informatization, Automation and Communications in Railway Transport (JSC NIIS), Moscow, Russia, zivkoacter@yahoo.com.

Lugovsky, Nikita V., Analyst of Scientific and Technical Complex for Digital Modelling (STC DM) named after V. I. Umansky of Research and Design Institute of Informatization, Automation and Communications in Railway Transport (JSC NIIS), Moscow, Russia, lnvlnikit@gmail.com.

Pazoysky, Yuri O., D.Sc. (Eng), Professor, Head of the Department of Railway Stations and Transport Hubs of Russian University of Transport, Moscow, Russia, pazoyskiy@mail.ru.

Kalinin, Sergey V., Deputy Director of Scientific and Technical Complex for Digital Modelling (STC DM) named after V. I. Umansky of Research and Design Institute of Informatization, Automation and Communications in Railway Transport (JSC NIIS), Moscow, Russia, hart82@gmail.com.

Article received 29.03.2024, approved 10.09.2024, accepted 08.10.2024.