ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

ҮЛГІНІ ТАНУ ЖӘНЕ МАШИНАЛЫҚ ОҚЫТУ
РАСПОЗНАВАНИЕ ОБРАЗОВ И МАШИННОЕ ОБУЧЕНИЕ
PATTERN RECOGNITION AND MACHINE LEARNING

DOI 10.51885/1561-4212_2023_4_92
IRSTI 20.19.27

A.V. Belov[1], E.A. Egorova[2]
National Research University «Higher School of Economics», Moscow, Russia
[1]E-mail: avbelov@hse.ru
*[2]E-mail: eaegorova_8@edu.hse.ru**

MACHINE LEARNING APPROACH FOR SCIENTIFIC AND TECHNICAL EXPERTISE

ҒЫЛЫМИ-ТЕХНИКАЛЫҚ САРАПТАМА ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ НАУЧНО-ТЕХНИЧЕСКОЙ ЭКСПЕРТИЗЫ

Abstract. *When conducting scientific and technical expertise, it is necessary to analyze the texts of reports on scientific research work. The analysis is carried out in order to determine whether the research being conducted belongs to the class of scientific research and development work in the field of IT. The main purpose of this study is to improve algorithms for analyzing text documents during scientific and technical expertise. To achieve this goal, the tasks of binary classification of documents provided by companies using machine learning technologies are considered. As a result of the study, a comparative analysis was carried out and the most effective machine learning algorithms were identified. The proposed algorithms will be used in a system that automates the process of checking documents submitted to taxpayers by the tax office.*

Keywords: *R&D projects; decision support system; scientific report; machine learning; text classification.*

Аңдатпа. *Ғылыми-техникалық сараптама жүргізу кезінде ғылыми-зерттеу жұмысы туралы есептердің мәтіндерін талдау қажет. Талдау жүргізіліп жатқан зерттеу ақпараттық технологиялар саласындағы ғылыми-зерттеу және тәжірибелік-конструкторлық жұмыстар класына жататынын анықтау үшін жүргізіледі. Бұл зерттеудің негізгі мақсаты ғылыми-техникалық сараптама барысында мәтіндік құжаттарды талдау алгоритмдерін жетілдіру болып табылады. Осы мақсатқа жету үшін машиналық оқыту технологияларын қолдана отырып, компаниялар ұсынатын құжаттарды екілік жіктеу міндеттері қарастырылады. Зерттеу нәтижесінде салыстырмалы талдау жүргізіліп, машиналық оқытудың ең тиімді алгоритмдері анықталды. Ұсынылған Алгоритмдер салық инспекциясы салық төлеушілерге ұсынатын құжаттарды тексеру процесін автоматтандыратын жүйеде қолданылады.*

Түйін сөздер: *ғылыми-зерттеу жобалары; шешімдерді қолдау жүйесі; ғылыми есеп; Машиналық оқыту; мәтіндерді жіктеу.*

Аннотация. *При проведении научно-технической экспертизы необходимо анализировать тексты отчетов Анализ проводится для того, чтобы определить, относится ли проводимое исследование к классу научно-исследовательских и опытно-конструкторских работ в области информационных технологий. Главной целью данного исследования является усовершенствование алгоритмов анализа текстовых документов при проведении научно-технической экспертизы. Для достижения данной цели рассматриваются задачи бинарной классификации документов, предоставляемых компаниями с использованием технологий машинного обучения. В результате исследования произведен сравнительный анализ и выявлены наиболее эффективные алгоритмы машинного обучения. Предложенные алгоритмы будут использованы*

в системе, автоматизирующей процесс проверки документов, представляемых налогоплательщикам налоговой инспекцией.

Ключевые слова: *Научно-исследовательские проекты; система поддержки принятия решений; научный отчет; машинное обучение; классификация текстов.*

Introduction. The widespread development and implementation of innovative technologies in various fields of science and technology is encouraged by the State for economic development. This support can be provided by direct financing or indirect incentives. In Russian tax legislation, indirect incentives are implemented in the form of tax benefits. For example, paragraph 7 of Article 262 of the Tax Code of the Russian Federation establishes that "A taxpayer who carries out expenses for scientific research or development according to the list of scientific researches established by the Government of the Russian Federation has the right to include these expenses as a part of other expenses of the reporting taxable period in which such research or development was completed (separate stages of work), or the initial cost of amortized intangible assets in the amount of actual costs using a coefficient of 1,5." [1].

This tax benefit provides savings on the income tax of organizations performing work that has scientific novelty for the development of national science and technology as a whole. However, there are many cases when a taxpayer has unlawfully applied this tax benefit. The identification of the illegality of the application of the tax benefit provided for in Article 262 of the Tax Code of the Russian Federation is implemented by tax audit of the R&D reports provided by taxpayers. In addition to the analysis of financial documentation, scientific and technical expertise is carried out in order to determine the correctness of the tax calculation. This check is realized in order to determine the signs of scientific novelty of R&D and compliance of the work performed with scientific and technical directions established by the government.

Conducting a scientific and technical examination of the results of research activities is a difficult task that requires a lot of routine operations: checking the compliance of documents with the requirements of state standards, analysis of the sources used, verification of novelty of results, etc. These works are carried out by both tax inspectors and experts. In this regard, the task of supporting the decision-making of scientific and technical examinations is an urgent problem. Despite the fact that decision-making support systems are actively used in areas such as medicine, economics, education, human capital management, etc., such systems did not find widespread in the analysis of scientific and technical reports and work. Most researchers consider the methods of information support for experts, as well as the applicability of models covering key issues of organizational and methodological support at various stages of scientific and technical examination [2, 3, 4].

The paper [5] discusses the construction of a system for supporting decision-making to assess the effectiveness of scientific and technical projects. The algorithm is given for the decision-making system that helps to evaluate the effectiveness of scientific and technical decisions. The system includes data collection, their processing, decision -making and interface. Data collection is carried out using manual filling out information about the project. For data processing, a fuzzy logical conclusion according to the Takagi-Sugeno algorithm is used. A further step is the decision that remains with the user who has the result of the algorithm in his interface. However, the mathematical model for assessing the effectiveness of scientific and technical decisions is not suitable for making a decision on issuing tax benefits, since it does not use information about the topic of work and does not analyze the degree of its novelty.

The paper [6] describes the process of information support for expertise of scientific and technical projects in the framework of the federal target program. The approach proposed in the work is designed primarily for information and analytical support of the examination at the

stage of selection of projects applying for state support. The advantage of this approach is the presence in the proposed system of built -in analysis modules on incorrect borrowing and integrating sources of scientific and technical information. The analysis module for incorrect borrowing using several databases helps to check the work on the repeatability of the subject, as well as counteract direct and indirect borrowings in reporting materials. Using tools for visualizing the results of analysis in the form of graphs, it becomes easier to make decisions on the novelty and originality of the work under study. It is noted that most experts noted the convenience of using such a system that helps to conduct an examination of scientific and technical work. A significant drawback of the system is the lack of verification of the structure of the documentation provided by the requirements of GOST and industry standards.

The work [7] provides an overview of the technologies of information support for the work of the expert in Russian and foreign programs. In the software product “Expertise of Scientific and Technical Programs and Projects” [8], functions are implemented that allow you to create expert profile templates, select experts, form conclusions and summarize the examination. However, this development does not include data analysis, but only helps accompanied by the examination process.

The references analysis shows that the task of supporting decision-making of scientific and technical examinations is an urgent problem. There are many different decision-making systems, however, specialized support systems for conducting examinations of scientific and technical work are not enough. In addition, not one of the considered systems includes automatic verification of documents for compliance with the structure of reporting documents according to R&D with the requirements of local standards, simultaneously checking the elements of novelty and the importance of the results obtained. Also, in the systems under consideration, there are no functions of intellectual analysis of texts.

To improve the quality of scientific and technical expertise, the Decision Support System was created that automates the process of examination of accounting documentation provided by the taxpayer to the tax authorities to confirm the legality of the application of tax benefits [9]. The architecture of the system consists of a client and a server part. The client part is written in the JavaScript programming language using a framework Vue.js. Due to the Bootstrap library, the adaptability of the interface for mobile devices is ensured. The server application is written in Python programming language using Django framework and Django REST framework. The server application uses GUnicorn as a WSGI server and NGINX as a proxy server. Report analysis algorithms are performed asynchronously using Celery. Asynchronous execution allows you to perform analysis in the background without blocking other requests from users. PostgreSQL is used as the main database. Redis is used as a message broker for Celery. Figure 1 demonstrates the SW architecture of the System.

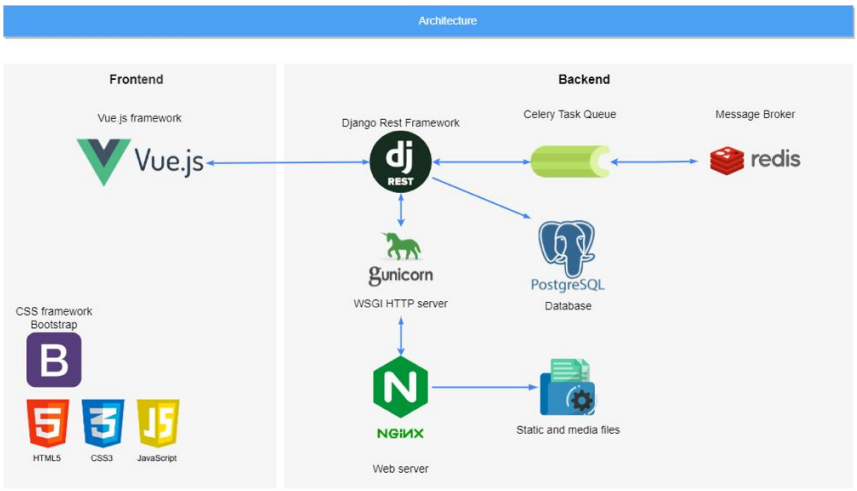

Figure 1. The System architecture

In this system, the taxpayer is given the opportunity to fill out a checklist in a WEB browser, which will allow an initial check of the submitted information. This online service will be integrated with other services of the digital platform of the Federal Tax Service. This will give the opportunity to the taxpayer to interact with the inspection of the Federal Tax Service, using secure Internet channels to transmit large amounts of textual information concentrated in scientific and technical reports. The developed service [9] should provide an opportunity not only to analyze the information filled in by the taxpayer in the format of a "Checklist", but also scientific and technical reports on completed research/R&D. The solution of this problem is impossible without the use of artificial intelligence methods for a meaningful analysis of the texts of reports to identify mandatory signs of R&D. It will significantly increase the efficiency of scientific and technical expertise to make a decision on the legality of the tax benefits applied.

The main purpose of this study is to improve the algorithms of text analysis for the support system of scientific and technical expertise. To achieve this goal, the problem of binary classification of documents, that are provided by the taxpayer, using machine learning technologies is considered. As the final result of the study, the most effective, in terms of technical and economic indicators, machine learning algorithms will be determined.

Literature Review. Currently, quite a large number of approaches are already being used to solve the problem of text classification. For example, in the study [3], the authors conduct a comparative analysis of such methods as the support vector machine, Bayes method and k-nearest neighbors to solve the problem of classifying Chinese news texts. For the vector representation of the text, the authors used the TF-IDF (term frequency-inverse document frequency) method. In the course of the study, the authors found out that the support vector machine provides the best classification quality for Chinese text.

The examples of similar work are the study [10, 11]. In [10], to solve the multiclass classification problem the authors compared the following models: the Bayes method, the support vector machine method, the decision tree and the k-nearest neighbors' method. The authors tested the classification result on Turkish news texts. In [11], the authors considered the problem of classifying BBC news texts in English using the random forest method, logistic regression and k-nearest neighbors.

In the study [12], the authors consider the problem of classifying Amazon customer reviews. Researchers analyze the impact of using N-grams, when vectorizing data, using Bag-of-Words (BOW). As a result of the study, it was revealed that the simultaneous application of unigrams, bigrams and three-grams provide a slight improvement accuracy in comparison with using only

unigrams.

However, the study [13] presents a different approach, using neural networks for the problem of classifying the tonality of short Russian-language texts. To build models of neural networks, the following architectures were studied: multilayer perceptron, networks with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Experiments demonstrated that the best quality was achieved using GRU and vector representation of the Word2Vec model trained on the WikiRuscorpora corpus.

The works [14, 15] are devoted to the study of the BERT model for solving the problem of classification messages in Russian language [16]. The main disadvantage of BERT for Russian-language texts is the limit of 512 tokens. The solution to this problem was proposed by the authors in [17]. The most important information in the text frequently is contained at the beginning and at the end of the document. Therefore, the methods of text truncation were analyzed. As a result, it was revealed that the best classification quality is demonstrated by the truncation method using the last 512 tokens. It is worth noting that neural networks should be used only with a sufficiently large and complete dataset.

The above studies were mainly conducted on texts with a non-academic style. However, it is worth mentioning the work [17], where the authors investigate the problem of classification of scientific articles and abstracts. This study was carried out for the purpose of automatic rubrication of research papers. For the extraction of features from the corpus, the authors used the word2vec model based on neural networks. During the experiments, the support vector machine method presented the best result in all quality indicators.

Thus, the following methods are suitable for solving the binary text classification problem: decision tree, random forest, logistic regression, k-nearest neighbors, support vector machine, Bayes method and multilayer perceptron. The vector representation of textual information is mainly modeled using the methods based on word counting. The main disadvantage of these vectorization methods is the problem of the large dimension of the feature space, so these methods are often used with dimensionality reduction methods.

Problem statement. The paper examines a set of documents that were provided by the taxpayer for the conduction of the expertise. The submitted documents can be divided into two classes based on the affiliation of the works described in the accounting documents to R&D. We denote the first class of analyzed reports as positive, i.e., having characteristic of R&D, and the second as negative, which includes reports that do not have signs of R&D. In the gathered dataset, 26 positive reports and 23 negative reports were presented. Thus, the problem of binary classification of documents is set.

At the first stage of the study, all files with analyzed reports were read and manual marking of document categories was carried out. In unprocessed texts, there are often words that do not have a semantic load. Since the extraction of features from the texts generates a large dimension of the feature space, it is necessary to get rid of stop words. The next step is basic text preprocessing that includes tokenization, lowercasing, lemmatization and removal of all digits and punctuation marks.

The final stage is the selection of classification metrics. Since we are considering a classification problem with two classes, it is advisable to use the F_1 – macro as a metric. To do this, the F_1 measure for each class will be calculated, and then the average for all classes will be calculated.

Methods. After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use

the scroll down window on the left of the MS Word Formatting toolbar.

Machine learning algorithms cannot work directly with unprocessed text: it needs to be converted to a numeric representation. To achieve this goal, several approaches to presenting texts in features have been implemented.

Bag-of-Words (BOW) is a text representation that describes the presence of words in a document. The model creates a dictionary of all the unique words in the corpus and calculates whether a certain word occurs in the text. However, with such a separation of features from the text, information about the order of words or the structure is lost.

A similar approach is the TF-IDF (Term frequency-inverse document) data vectorization method. Unlike the first approach, the counter in the description of features is replaced by the value TF-IDF:

$$tfidf_{wd} = tf_{wd} * \log \frac{|D|}{df_{wd}},$$

where tf_{wd} is the ratio of the number of occurrences of a word in a document to the total number of words in this document; df_{wd} is the number of documents containing this word; $|D|$ is the number of documents in the collection.

One of the popular methods of extracting features from the text is embedding words. The main essence of this method is that each word can be matched to a vector of real numbers; words appearing in a similar context will be closer in vector space. To extract a feature of one text, there are two approaches: the vector of each text can be equal to the average of the vectors of the words included in this text, or it can be the weighted average of the word vectors. In this paper, we used the pre-trained word2vec model from the RusVectores website. This model is trained on Wikipedia and National Corpus of the Russian Language for November 2021. As weights for words, we used the value of the IDF from the extraction of features TF-IDF.

As a result of literature research, the following methods were chosen to solve the problem of text classification:

1) Logistic regression;
2) Decision Tree;
3) Random forest;
4) Support vector machine;
5) K-nearest neighbors;
6) Multilayer perceptron.

One of the classical and linear classifiers is logistic regression. This algorithm uses a logistic function to estimate the probability that an object belongs to a certain class, and makes a prediction based on this estimate.

The Decision Tree (DT) method is one of the most popular machine learning algorithms for classification. This method allows to build a tree where each node represents a feature, each branch is a possible value of the feature, and the leaves are the predicted class. The essence of the method is to choose the optimal feature for splitting the data at each step of the tree construction to reduce the uncertainty in the classification.

Random Forest (RF) is a machine learning algorithm that is based on building an ensemble of decision trees. When classifying a new object, each tree in the forest provides its own answer, and the final decision is made based on the voting of the trees.

Support Vector Machines (SVM) method is a linear machine learning method for solving classification and regression problems. It is based on finding the optimal separating hyperplane

between classes, which maximizes the margin between them. The support vectors are examples of the training sample, which are located on the boundary of the class separation and determine the position of the separating hyperplane.

The k-Nearest Neighbors is a machine learning method that is used to classify objects. The essence of the method is that each object is classified based on the classes of k objects closest to it in the feature space. The number of the nearest neighbors k is selected in advance and it is a hyperparameter of the method. The object belongs to the class that is most common among the k nearest neighbors.

Results. In order to justify the choice of machine learning methods for the problem of classifying texts submitted for expertise, it is necessary to conduct computer experiments. That is, based on the processed texts, to train the model and determine the accuracy of classification on the test set. To implement all classifiers, we use the scikit-learn library, which is written in the Python programming language. To estimate the predictive ability of algorithms, we consider the results of classification on k-fold cross-validation with k =5.

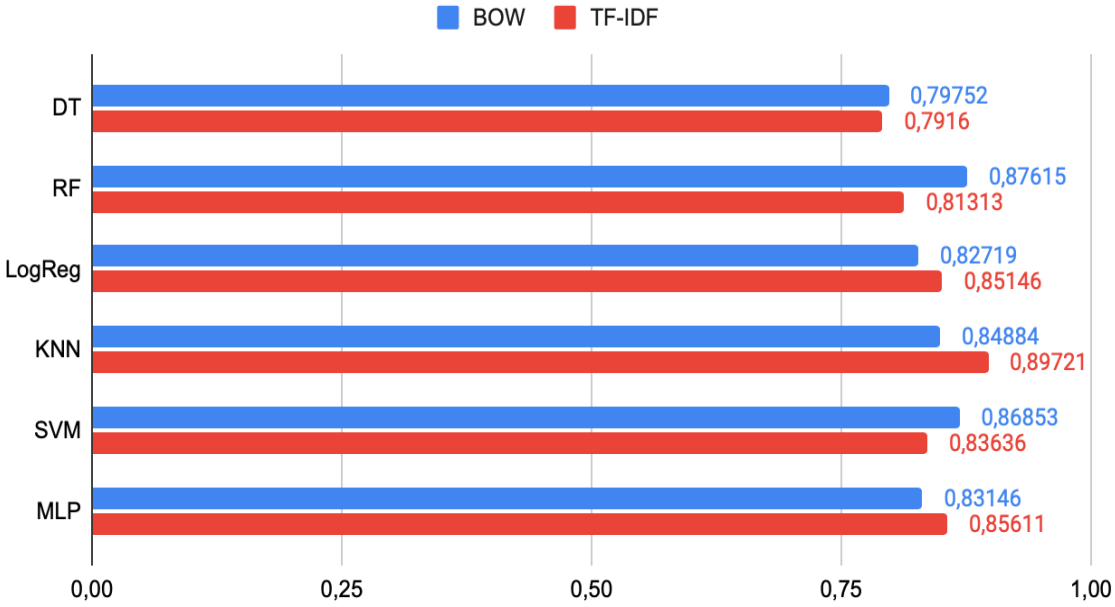

Figure 2. Classification results with two approaches BOW and TF-IDF

Figure 2 demonstrates that the TF-IDF technique improved the quality only for the classifiers of logistic regression, k-nearest neighbors and multilayer perceptron. The best classification quality was demonstrated by the k–nearest neighbors method with TF-IDF method. The results of classification using a different approach based on the vector representation of words are shown in Figure 3.

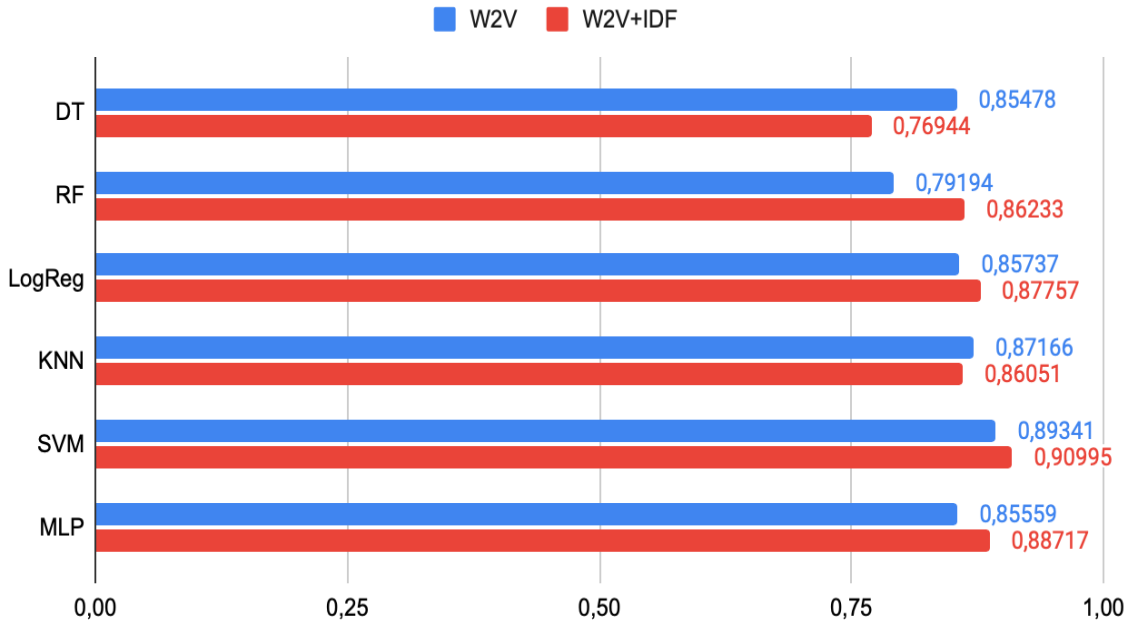

Figure 3. Classification results using the word2vec model

Overall, the vector representation using the word2vec model improved the quality for methods such as decision tree, logistic regression, support vector machine and multilayer perceptron. The best result with a value of F_1 – measure equal to 0,90995 was shown by the method of support vectors machine with weighted average vectors.

It is worth noting that practically all methods with different approaches have shown great results. When testing the word2vec model, the classification quality improved significantly compared to the methods based on word counting. Moreover, the approach with weighted average vectors improved the quality for the majority of algorithms. Based on the results of the study, the following classifiers can be distinguished: support vectors machine, k – nearest neighbors and multilayer perceptron. They consistently showed a good classification result and it is necessary to use them for further work.

To justify the choice of a machine learning algorithm for the system, it is also necessary to conduct research on computational complexity. That is, it is necessary to determine the time of preprocessing of the text collection, the time of vectorization of data with different approaches, the time of training and prediction of each classifier. Table 1 shows the average processing time for the entire collection of texts.

Table 1. Average processing time for the entire collection of texts

Methods	Time, sec
BOW/TF-IDF	186,9106
W2V	255,1518

For the BOW and TF-IDF techniques, preprocessing the entire set took approximately 186 seconds. Since another step of preprocessing is needed for the vector representation of words

using the word2vec model, the time has increased to 255 seconds. Table 2 shows the average time required to vectorize the data.

Table 2. Average time required for data vectorization

Methods	Time, sec
BOW	0,23758
TF-IDF	0,24343
W2V	1,44211
W2V+IDF	3,91884

The methods BOW and TF-IDF transform data in about the same time. It takes more time to vectorize the entire collection of texts using the word2vec model. The approach with weighted average vectors takes practically twice as long as the approach with the construction of vectors through the average. This is due to the fact that in order to determine the weight of the vector, it is necessary to determine the value of the IDF for each word. Table 3 shows the average time required to vectorize the data and Table 4 presents the average prediction time.

Table 3. Average training time in ms

Методы	DT	RF	LogReg	KNN	SVM	MLP
BOW	0,00219	0,06493	0,01892	0,00057	0,01572	0,25308
TF-IDF	0,00221	0,07184	0,004	0,0005	0,02028	1,04966
W2V	0,00279	0,07067	0,00265	0,00123	0,0018	0,11759
W2V+IDF	0,00113	0,07116	0,0029	0,00066	0,00114	0,12838

Table 4. Average prediction time in ms

Методы	DT	RF	LogReg	KNN	SVM	MLP
BOW	0,00032	0,00754	0,00024	0,00707	0,00185	0,00054
TF-IDF	0,00043	0,00797	0,00021	0,00266	0,00358	0,00068
W2V	0,00061	0,00628	0,00032	0,00139	0,00034	0,00051
W2V+IDF	0,00013	0,00589	0,00019	0,00124	0,00021	0,00024

For each classifier with a different technique of extraction features, the training and prediction time was different. From Table 3, it can be seen that the multilayer perceptron classifier requires much more times for training in comparison with other machine learning algorithms. However, it needed less time to predict than the random forest method. It is also worth noting the results of methods such as k-nearest neighbors and logistic regression. The k-nearest neighbors method demonstrated the shortest learning time for all data vectorization approaches. Logistic regression demonstrated the best prediction time for all vector representation methods, requiring approximately 0,24 milliseconds.

Thus, when testing the word2vec model, the time for training and predicting all classifiers was reduced in comparison with methods based on word counting. On the one hand, it took much more time to prepare and vectorize data for the word2vec model, however, this approach demonstrated the best classification accuracy and minimum time for training. All computer experiments were conducted on a computer with macOS operating system installed on a 4-core Intel Core i5 processor.

Conclusion. In the course of the study, a range of methods for preprocessing, vectorization and classification of data were applied in practice. Classification methods such as decision tree, random forest, logistic regression, k–nearest neighbors, support vector machine and multilayer perceptron were considered. Based on the results, the best classifier, in terms of technical and economic indicators, was found. It is SVM with vectorization of data using weighted average vectors. It should be noted that classical methods of vector representation of text also showed good results. The proposed approach to the classification of scientific and technical reports provided by the taxpayer to justify the application of tax benefits to the R&D performed will be used for the implementation of a software support system for scientific and technical expertise. This research will be continued in terms of the design and implementation of the service provided to the taxpayer to verify the reporting documentation on R&D.

References

1. Nalogoviy Kodeks RF (2022), Moscow, Prospect: 1168.
2. Tuzova S.YU., Mironova YA.S. Nauchno-tekhnicheskaya ekspertiza kak instrument realizacii gosudarstvennoj podderzhki nauchno-tekhnicheskoj deyatel'nosti, Vlast', t.26, №4, 2018, pp. 33-39
3. Divueva N.A., Gusev Y.U. MODELING OF ORGANIZATIONAL AND METHODOLOGICAL SUPPORT OF EXPERT AND ANALYTICAL SUPPORT OF MANAGEMENT OF SELECTION OF INNOVATIVE PROJECTS, Finansovaya ekonomika, №4, 2020, S. 360 – 366
4. Touzova S.Y., Musatov A.A. INFORMATION AND ANALYTICAL SUPPORT OF SCIENTIFIC AND TECHNICAL EXPERTISE. Aktual'nye problemy social'no-ekonomicheskogo razvitiya Rossii, №4, 2017, S. 87-90
5. Khalyasmaa A.I., Zinovieva E.L. Intelligent decision support system for technical solutions efficiency assessment // Proceedings of 2017 IEEE 2nd International Conference on Control in Technical Systems, CTS 2017. Institute of Electrical and Electronics Engineers Inc., 2017. - P. 247-250
6. MUSATOV A., MIRONOVA YA. INFORMACIONNAYA PODDERZHKA EKSPERTIZY NAUCHNO - TEKHNICHESKIH PROEKTOV.v sbornike statej Mezhdunarodnoj nauchno-prakticheskoj konferencii «Razvitie nauki i tekhniki: mekhanizm vybora i realizacii prioritetov», chast' 3, Izd. Aeterna, 2017
7. Lekh D.YU., Derkanosov M.A., ZHolobov P.A., Bandurov A.V. Special'noe programmnoe obespechenie dlya avtomatizirovannogo provedeniya nauchno-tekhnicheskoj ekspertizy «Ekspert-ERA». V sbornike statej II Vserossijskoj nauchno-tekhnicheskoj konferencii. Tom 2. Voennyj innovacionnyj tekhnopolis "ERA". Anapa, 2020, S. 164-171
8. Sistema ekspertiz nauchno-tekhnicheskih programm i proektov. Svidetel'stvo o gosudarstvennoj registracii programmy dlya EVM № RU 2021616367, 2021, Availableat: https://elibrary.ru/download/elibrary_45822528_93420236.PDF.
9. Belov A. V, Bikbaev B. I., Gevondyan M. S., Levitan D. A., Panina I. Yu. (2023) “Decision Support System for scientific and technical expertise,” in Proceedings of the 2023 Conference of Russian Young Researches in Electrical and Electronic Engineering (EIConRus). IEEE: 188-193
10. F. Miao, P. Zhang, L. Jin, and H. Wu (2018) “Chinese News Text Classification Based on Machine Learning Algorithm.” 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)
11. F. Gurcan (2018) “Multi-Class Classification of Turkish Texts with Machine Learning Algorithms.” 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)
12. K. Shah, H. Patel, D. Sanghvi, and M. Shah (2020) “A Comparative Analysis of Logistic

Regression, Random Forest and KNN Models for the Text Classification.” Augmented Human Research.
13. T. Pranckevičius and V. Marcinkevičius (2017) “Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification.” Baltic Journal of Modern Computing.
14. O.V. Oglezneva (2020) “Analysis of the tonality of short text messages,” in Information Technologies. Problems and solutions:113-118. (In Russian)
15. S.S. Maslenikova, V.V. Korotkov (2022) “The application of BERT for classifying messages to the SAP support service,” in Informatics problems, methods, technologies: 1164-1170. (In Russian)
16. N.D. Kropanev, A.V. Kotelnikova (2021) “BERT for analyzing the tonality of long texts on the example of Kaggle Russian News Dataset,” in Society. The science. Innovation: 256-259. (In Russian)
17. A. Romanov, K. Lomotin, and E. Kozlova (2019) “Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts.” Data Science Journal.