European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2021.09.02.12

ICEST 2021
II International Conference on Economic and Social Trends for Sustainability of Modern Society

APPLYING MACHINE LEARNING TO STATISTICAL R&D COSTS ACCOUNTING

E. V. Dmitrishina (a)*, R. G. Smirnov (b), I. S. Fedorov (c)
*Corresponding author

(a) ANO Institute for Public Finance Reform, Moscow, Russia, evd@irof.ru, 0000-0002-3561-9650
(b) Lomonosov MSU Center for Storage and Analysis of Big Data, Moscow, Russia, r.smirnov.mailbox@gmail.com
(c) Lomonosov MSU Center for Storage and Analysis of Big Data, Moscow, Russia, ifedorov99@mail.ru

Abstract

The conducted research using machine learning shows that a number of research projects are not identified as such in the unified information system on procurement, but are defined as services, which leads to a statistical underestimation of internal costs for research and development, which underestimates performance indicators systems of scientific and technological development of the country. We propose how, using machine learning and natural language processing methods, a search is carried out for purchases placed in the Unified Information System in the field of procurement, which do not have a direct indication of the purchase of R&D or have a direct indication of the purchase of a service, while having a procurement subject corresponding to R&D. The interpretation of the results of applying the model is an increase in the recorded volume of purchases related to research and development carried out through the unified information system of public procurement by four times (up to 421 billion rubles), as well as an increase in the recorded volume of such purchases ordered by a business structure by 11 times. In order to ensure a simplified accounting of internal expenditures on research and development (IR&D) and regular monitoring of various sources to identify unaccounted IR&D, it is necessary to develop and implement a special algorithm for the search and classification of works.

Keywords: IR&D, R&D, public procurement, machine learning

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Research and development funding is one of the key conditions for long-term economic growth. In modern strategic planning documents, in particular the Strategy for the Scientific and Technological Development of the Russian Federation, as well as in the state program "Scientific and Technological Development of the Russian Federation" special attention is paid to indicators of IR&D, which characterize the direct funding of scientific research and development, as well as indicators related to commercialization the results of intellectual activity, the development of applied solutions focused on solving current business problems, for example, the indicator of internal costs for research and development at the expense of funds from extra-budgetary sources (at the expense of own funds of organizations in the commercial sector). In general, the IR&D at the expense of funds from extra-budgetary sources is an indicator of the effectiveness of the country's scientific and technological development system.

2. Problem Statement

IR&D funding is always associated with a set of advantages and limitations. An organization that finances R&D has the right to enjoy tax incentives. The use of tax incentives leads to increased attention to the activities of the organization on the part of the supervisory bodies. At the same time, the financing of IR&D at the expense of funds from budgetary sources causes the need for a scientific examination of the Russian Academy of Sciences, which acts as a bureaucratic restriction for funding of IR&D, on the one hand, and as a tool for quality control of R&D, on the other hand. Thus, incentives are formed for the classification of scientific works as works not related to scientific activity, and, accordingly, the implementation of their financing in a form not related to R&D, for example, the purchase of services, which negatively affects the indicators of IR&D.

3. Research Questions

The issues of statistical accounting of indicators of scientific and technological development are extremely relevant for Russia (Gokhberg et al., 2019). Funding remains one of the most important factors in the potential for such development, along with human resources and infrastructure. Researchers pay attention to the scale and dynamics of science funding, its sources and the budget component (Vlasova et al., 2018).

In the context of the development of the strategic planning system and the implementation of the program documents, the correct interpretation of indicators of the effectiveness of the development of science is extremely important (Dmitrishina et al., 2018). Although in a number of existing works on this topic it is proposed to consider the possibility of revising the methods for assessing the effectiveness of the implementation of the tasks of developing the country's scientific and technological complex and to modernize some indicators included in the existing methods, however, the issue of the correctness of the collection of statistical indicators is not fully disclosed (Mikhailets et al., 2019).

The scientific literature often raises the question of unaccounted intellectual property objects, for example, within the framework of work (Nurakhov & Telemtaev, 2011) it is mentioned that virtually any

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

educational institution is engaged in the commercialization of unaccounted intellectual property objects in the form of taught methods and courses. Similarly, the question arises that many organizations providing funding for various research projects can both intentionally and unintentionally refuse to account for the corresponding funding as research and development costs, thereby influencing the IR&D indicator.

Within the framework of this research the choice of a source of big data in the form of the Unified information system on procurement (hereinafter - EIS) is also not a unique practice. However, as a rule, data from the EIS are used to carry out research that is not related to the identification of unaccounted IR&D or the use of natural language machine processing methods. For example, within the framework of the study (Pererva et al., 2018), data from EIS are used to build models (including machine learning methods) necessary to determine the factors of the effectiveness of public procurement.

Since this research assumes work with text data obtained from EIS, there is a need to use natural language processing technologies. The basic element for implementing natural language processing is obtaining vector representations of words for further machine processing. The main approaches to the implementation of the vector representation of words were considered in (Joulin et al., 2017; Robertson, 2004).

Separately, at the preparatory stages of the study, the problem of forming a training sample was identified: if forming a sample of R&D titles does not pose a problem through keyword search or exporting data from R&D databases, then forming a sufficient sample of works not related to R&D is difficult. To solve this problem, it was decided to use the methods of iterative extraction of observations from the general population, in accordance with the methodology proposed in the research (Liu et al., 2003).

4. Purpose of the Study

The purpose of this study is to calculate the approximate volume of IR&D not accounted for in official statistics (including IR&D at the expense of funds from extrabudgetary sources), due to the classification of research and development expenditures in the form of the others.

A unified procurement information system was chosen as a source of information for the research. According to federal legislation, certain purchases, in particular, purchases of government agencies, state corporations and companies, natural monopolies, are published in EIS. Accordingly, within the framework of this study, it is shown how, using machine learning and natural language processing methods, a search is carried out for purchases placed in EIS that do not have a direct indication of the purchase of R&D or have a direct indication of the purchase of a service, while having a procurement subject corresponding to R&D.

5. Research Methods

5.1. Volumes of IR&D in various data sources

In various ways, we have determined the volumes of IR&D (including those at the expense of funds from extra-budgetary sources) for 2019: data from the Federal State Statistics Service, data from the unified state information system for recording research, development and technological work for civilian purposes

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

(where it is currently located information on only a part of the ongoing R&D), data from the materials of state programs of the Russian Federation, data from EIS (Table 1). At the same time, the given data from EIS were obtained as follows:

- IR&D, total - the cost of all purchases placed in EIS for 2019, the name of which contains the words "scientific" and "work", "scientific" and "research" or "R&D" and their derivatives;
- IR&D at the expense of funds from extra-budgetary sources - the total cost of purchases used to calculate “IR&D, total”, the name of the customer which contains the words “LLC”, “company”, “joint stock”, and their derivatives.

It should be noted that, in accordance with the modern practice of accounting for IR&D, organizations performing work placed on the UIS, based on the results of participation in competitive procedures, report on the use of funds received as funds from extrabudgetary sources.

Table 1. Internal expenditures on R&D in different data sources

	Internal expenditures on R&D in different data sources, mln.rbl.	Internal expenditures on R&D from extrabudgetary sources.
Federal State Statistics Service	1 134 786.7	403 984.1
Unified state information system for accounting of scientific research, experimental design and technological works for civil purposes (hereinafter - EGISU R&D)	270 255.6	14 752.6
Analysis of federal budget and extrabudgetary funds for scientific research, development and technological work for civil purposes provided for in the state programs of the Russian Federation	578 912. 3	138 426.1
Analysis of the Unified Procurement Information System (EIS)	96 915.8	333.4

Source: authors' calculations, (Federal State Statistics Service, 2020a, 2020b; Report on the progress of implementation and evaluation of the effectiveness of state programs of the Russian Federation, 2019; Unified State Information System, 2020).

The data of the Federal State Statistics Service are calculated based on the Consolidated Form of Federal Statistical Observation No. 2-Science "Information on the implementation of scientific research and development." Form No. 2-science is filled out by legal entities, except for small businesses, who carried out research and development in the reporting year and have a type of economic activity in accordance with the All-Russian Classifier of Economic Activities (OKVED2 OK 029-2014 KDES Ed. 2)) scientific research and development (code 72) (main or additional); higher education (code 85.22); training of highly qualified personnel (code 85.23); other types of economic activities that have received subsidies (grants) for the implementation of scientific research and development; as well as according to the list established by the Ministry of Science and Higher Education of the Russian Federation.

The analysis of federal budget allocations and extrabudgetary funds for research, development and technological work for civilian purposes, provided for in the state programs of the Russian Federation, is

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

based on an analysis of the costs of 32 state programs for which R&D expenditures were carried out in 2019 Internal expenditures on research and development at the expense of extra-budgetary sources are indicated only in two state programs: GP-16 "Development of industry and increasing its competitiveness" and GP-47 "Scientific and technological development of the Russian Federation." For other state programs, information on extrabudgetary funding for 2019 has not been published in the public domain.

5.2. Implementation of the valuation of unaccounted IR&D

The data collection was based on the data retrieved from the EIS portal for 2019. The following data were retrieved: names of purchases, names of customers, dates of purchase and the starting maximum price for each contract. At the same time, within the framework of the implementation of this study, for the sake of simplification, the hypothesis was adopted that the use of a limited description of each placed purchase, formulated in the name of the purchase, would be sufficient to solve the problem. That is, the texts of the terms of reference for each of the purchases or other documents posted on the EIS were not analyzed within the framework of the current study. The total amount of extracted data was about four million purchases; the volume of the extracted text data was about 5.5 GB.

Since data collection was carried out using automatic data extraction methods, observations were eliminated, in which the value of the contract was not determined. In addition, as part of the preparation of the data, from the data obtained, contracts were excluded, the cost of which is indicated in currencies other than the Russian ruble, which was also done to simplify the task and eliminate the need to carry out currency conversions at the exchange rate as of the date of each separate tender.

General descriptive statistics on the collected data and their cleaning are presented below:

- The total volume of purchases placed on the EIS, the cost of which is determined and the currency is the Russian ruble (99% of purchases) - 19 167 416 million rubles;
- The volume of excluded observations as a result of errors in data extraction algorithms - 171 observations out of 3 814 650 observations;
- The total volume of purchases placed on the EIS, the cost of which is determined and the currency is the Russian ruble, with an explicit indication of the performance of research work – 96 916 million rubles. (6276 purchases). Hereinafter, an explicit indication of the performance of research work means the use of the following words or phrases in the name of the purchase:
    - "R&D", taking into account variations in the location at the beginning, in the middle or at the end of the purchase name;
    - "scientific" and "research", taking into account the different forms of these words.
- The most common words in the names of purchases placed on the EIS, the cost of which is determined and the currency is the Russian ruble, with an explicit indication of the performance of research work: "state", "budgetary", "institution", "institute", "Russian "," education "," higher".

As a result of the initial testing of a set of models, analysis of the resulting sample, containing only 6276 applications with an explicit indication of the performance of research work in the name. It was

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

decided that it was necessary to enrich the data to increase the sample size characterizing the class of titles of work related to the performance of R&D (hereinafter - Class 0). To implement the enrichment, a method was developed for automated data extraction from EGISU R&D, where the names of a large number of implemented research and development projects, experimental design and technological works are given. The volume of data received from EGISU R&D amounted to 9569 unique R&D items. Thus, the total volume of Class 0 observations was about 15,000 units.

The preparation of the collected data was based on the use of standard methods for normalizing textual data, adopted in the field of natural language processing when solving comparable problems, namely:

- Change of uppercase to lowercase (all letters in words are lowercase);
- Removal of non-letter and non-whitespace characters (removal of punctuation, figures)
- Removal of extra spaces (double spaces, spaces at the beginning and end of lines) and white space characters (for example, tabs);
- Removal of stop words, in accordance with special existing dictionaries of stop words for individual languages (in the framework of this study, a stop word dictionary for the Russian language was used, implemented in the NLTK package for Python 3.x);
- Stemming, which means processing each individual word in such a way as to eliminate prefixes and endings and bring words into similar word forms (unlike lemmatization, which leads words to the first form, stemming is less resource-intensive; in this study, "Snowball Stemming" was used, implemented by in the NLTK package for Python 3.x);
- Splicing “not” with the next word (changing the grammatical structure to reflect the semantics, meaning).

Training a classification model requires training such a model in multiple classrooms. If the formation of Class 0, containing names related to the implementation of scientific research works, is described above, then for the formation of Class 1, containing the names of other works, it is necessary to use methods of iterative extraction of observations from the general population or to form the corresponding class in manual mode. Within the framework of this study, it was decided to use the methods of iterative extraction of observations from the general population, in accordance with the methodology proposed in the study (Liu et al., 2003).

The idea of iterative extraction is based on the formation of Class 1 from the total set of observations, data in such a way that the first step selects the observations that are most different from the observations in Class 0. In the subsequent stages, the observations are selected that are the least similar to the Class 0 observations and the most similar to observations, which have already been selected for Class 1. The implementation of the described method in the framework of this study relied on the application of logistic regression to the vector representations of purchase items obtained by the TF-IDF method, which is described in (Robertson, 2004).

The total volume of the received training samples:

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

- Class 0 - about 15,000 objects;
- Class 1 - 1,670,000 objects.

The total volume of the received test samples (test samples were formed in manual mode):

- Class 0 - 483 objects;
- Class 1 - 483 objects.

The used method of iterative extraction, based on logistic regression, allows you to determine the individual words that most determine the belonging of an object to a particular class. Table 2 are represented by five words that most strongly influence the belonging of an individual object to Class 0 and Class 1, respectively. The given list of words was obtained at the last iterations of the used iterative extraction method.

Table 2. The words most strongly influencing the classification of an object into Classes 0 and 1

Words with data preparation procedures (in non-vector notation)	
Words that most strongly influence the classification of an object in Class 0	Words that most strongly influence the classification of an object in Class 1
Scientific	supply
R&D	service
Research	made
Research	repair
Research	need

Source: compiled by the authors.

Also, among the most influencing words for assigning an object to Class 0 were such words as "analysis", "method", "technologist".

Within the framework of modeling, various approaches to obtaining vector representations of words were tested, including:

- TF-IDF;
- FastText is an approach that allows obtaining vector representations of words based on machine learning and dividing words into subwords of 3 characters. Based on the vector representations of words, the vector representation of the sentence is determined. The FastText method is described in (Joulin et al., 2017);
- FastText for words, vector representations of sentences, taking into account the weighted assessment of vector representations of words by weights obtained using TD-IDF - this particular approach to was recognized as the most effective within the framework of the task at hand.
- To solve the classification problem (determining whether an individual object belongs to Class 0 or Class 1), various models were also tested, including:
- Logistic regression;
- Random forest;

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

- Gradient boosts: XGBoost, LGBM, CatBoost.

The most effective approach is based on applying gradient boosting based on XGBoost to the resulting vector representations of words. The final XGBoost model at 1000 iterations achieved the following accuracy rates:

- ROC-AUC: 0.9945;
- Recall: 0.9151;
- Precision: 0.9977;
- Accuracy: 0.9565;
- f1: 0.9546.

Thus, the resulting model makes it possible to assess the likelihood that a separate object, characterized by the name of the purchase, is associated with the performance of research work. This model makes it possible to estimate the volume of IR&D, which are not included in the statistical indicator of IR&D due to the classification of the relevant work at the initiation stage as work not related to science (for example, their funding is carried out in the form of procurement of services).

To obtain results and further formulate conclusions using the resulting model, all the collected data were marked up.

A qualitative indicator of the model's accuracy, in addition to the previously given quantitative values, may be the following. In the preparation of Class 0, procurement titles containing R&D words in the text were not included. At the same time, almost all relevant purchases were marked by the model as related to the performance of research and development work, assigned to Class 0, the description of purchases and the likelihood of their assignment to Class 0 are presented in the table below (Table 3):

Table 3. R&D procurement model classification examples

№	Purchase Name Taking Into Account Data Preparation Procedures (In Non-Vector Representation)	Probability Of Being Classified As Class 0
1	Performance Of Experimental Design Work "Development Of A Technology For The Manufacture Of Germ Transistors On An Algan / Gan Heterostructure On A Silicon Substrate, To Create Power Amplifiers With A Breakdown Voltage Of More Than 150 V And A Technology For Creating Heterobipolar Transistors With A Hetero Emitter Based On Ingap	0.99
2	Performance Of Experimental Design Work	0.96
3	Performance Of An Integral Part Of The Experimental Design Work, Code "Center 2020-Gamma"	0.99
4	Performance Of A Component Part Of The Experimental Design Work: "Development Of High Specific Strength Spheroplastics And Technology To Ensure The Creation Of A Composite Lightweight Filler Based On Spheroplastics And Ceramic Macrospheres For The Buoyancy Systems Of Deep-Sea Equipment" Code "Sphere-N -"	1.00
5	Open Request For Proposals In Electronic Form For The Right To Conclude An Agreement To Perform Experimental Design Work To Reconfigure The Layout Of The Passenger Cabin In The Layout Scheme 25 Passengers	0.01

Source: compiled by the authors.

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Since the developed model determines the likelihood that a separate purchase, characterized by the name, is related to the performance of research work, the following are results for which the estimate of the probability of class 0 classification is higher than 0.99, that is, higher than 99% (excluding purchases with an explicit indication in the name of the performance of research work - such purchases were not subject to re-classification).

Thus, with a probability higher than 99%, the total cost of additionally defined as R&D purchases amounted to 324,430 million rubles. (12199 purchases). The cost of purchases defined as R&D and at the same time indicating the implementation of the purchase in the form of purchasing a service (that is, the name of the purchase contains the word "service") amounted to 11,006 million rubles. (2257 purchases). The most frequently used words in the names of purchases: institution, state, federal, company, budgetary, joint-stock.

Examples of identified activities classified as Class 0 are:

- provision of services for conducting an expert and analytical study on the topic: “monitoring the enforcement of legislative acts at the federal level and the level of the entities of the Russian Federation”;
- provision of services for the performance of bacteriological examination of biological material for streptococcus without studying the biochemical properties;
- provision of services for the implementation of an integral part of the development and design work on the creation of a medium-haul automobile and passenger environmentally friendly ferry on an electric ship, the code “eco-ferry”;
- services for the development of resource and technological models according to types of construction, taking into account the changes made to the federal estimate - normative framework;
- provision of services in the field of development work on the development of laboratory technological regulations for obtaining a culture-cell vaccine against sheep anaplasmosis;
- providing services for the creation and maintenance of a "process factory", which is a platform that provides practical training in the principles and tools of lean manufacturing by simulating real production and auxiliary processes.

Separately, based on the analysis of the names of customers, it is possible to determine purchases that are likely to be financed from extra-budgetary sources. In such cases, the customer must be a commercial organization. Thus, the assessment of the amount of funds from extra-budgetary sources within the framework of the identified purchases classified as Class 0 is based on the presence in the name of the customer of the corresponding purchases of the words "joint-stock", "company", "LLC", "company" and their derivatives. The total cost of the respective purchases reaches RUR 3,488 million.

6. Findings

To ensure the completeness of accounting of IR&D, it is necessary to provide institutional mechanisms that simultaneously contribute to: raising the awareness of organizations (both the private and

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

the public sector) about the possibilities and the need to take into account the expenditures associated with research and development in the form of IR&D, and not in the form of expenditures on purchasing services or other purposes; reduction of incentives for intentional concealment of IR&D. To awareness raising of organizations, it may be advisable to implement a program for increasing the literacy of organizations in terms of accounting for work related to research and development, while simultaneously forming a classifier of works that relate to research and development, ensuring compliance with international practice. To reduce the incentives for deliberately concealing the IR&D, mechanisms may be envisaged for attributing certain expenditures on the IR&D without additional burdens with expertise and provision of benefits under certain conditions.

In order to ensure a simplified accounting of IR&D and regular monitoring of various sources to identify unaccounted for IR&D, it is necessary to develop and implement a special algorithm for the search and classification of such works. At the same time, the aspect related to the distribution of expenditures in the framework of performing work between related research and development and other components of the work becomes critical. To ensure such a distribution, it is necessary to separately develop methods of labour rationing when performing modern research and development. The development of appropriate methods of labour rationing is possible only within the framework of processing a large amount of data on recorded and detailed R&D, in connection with which the development of a unified information system for recording research and development work, which in Russian practice is the EGISU R&D system, becomes critical.

Thus, in order to ensure the completeness of accounting for IR&D, it is necessary to implement a set of works related to institutional reforms, the development of methodologies, algorithms, monitoring systems and a unified accounting system for scientific research, experimental design and technological work for civil purposes, which is advisable to implement on the basis of EGISU R&D.

The presented approach has some limitations, which provide opportunities for improving the model and significantly increasing the accuracy of the estimate. To increase the accuracy of the model, in particular, the following measures can be implemented:

- Development of a model based on additional information about the procurement, including the text of the terms of reference, the draft contract in addition to the name of the procurement;
- Implementation of sample markup in manual mode, rather than using an iterative approach that increases the likelihood of distortions at the data markup stage.

Separately, for a detailed analysis of the sources of financing for various purchases, it is possible to conduct a detailed analysis of customers, their belonging to the private or public sector within the framework of individual purchases.

7. Conclusion

In 2019, about 3.8 million purchases were placed on the EIS for a total value of about 19 trillion. rubles, of which about 6 thousand purchases for a total cost of about 97 billion rubles. have an explicit indication in the name of the purchase of the performance of research work.

https://doi.org/ 10.15405/epsbs.2021.09.02.12
Corresponding Author: E. V. Dmitrishina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Using the 99% probability level of the model used, it can be argued that the total volume of unaccounted purchases identified as related to research and development is 12,119 purchases for a total cost of about 324 billion rubles. Of these, 2,257 were purchases for a total cost of about 11 billion rubles have an explicit indication of the purchase in the form of a service in the name of the purchase; and purchases for a total cost of about 3.5 billion rubles. have an indication of the use of funding from extrabudgetary sources.

The interpretation of the results of applying the model is an increase in the recorded volume of purchases related to R&D carried out through the EIS by four times (up to 421 billion rubles), as well as an increase in the recorded volume of purchases, the customer of which is a commercial structure, by 11 times. At the same time, according to the existing practice of accounting for funds for research and development, any "competitive" funds are for the organization funds from extra-budgetary sources. Then it can be argued that the use of the model makes it possible to identify and statistically record as R&D about 300 billion rubles. extra-budgetary funds for research and development, implemented through the EIS and in a form other than the procurement of research and development.

Thus, it can be stated that, according to the applied approach and the resulting model, a significant amount of R&D is not taken into account in the collected statistics of the IR&D, and the real indicators of the socio-economic development of Russia are significantly underestimated.

Acknowledgments

The publication is prepared based on of the results of the scientific and methodological support of measures to consolidate financial resources for the implementation of scientific and technical policy (registry number – 730000Ф.99.1.БВ16АА02001, topic ‘Scientific and methodological support of measures to support the implementation of the state program "Scientific and technological development of the Russian Federation"), conducted with financial support from the Ministry of Science and Higher Education of the Russian Federation.

References

Dmitrishina, E. V., Uskov, D. A., Yagovkina, V. A., & Mikhaylova, A. A. (2018). Peculiarities of Financial Provision for the Implementation of State Programs with Scientific and Technical Components. *European Research Studies Journal, 21*(1), 614-623.

Gokhberg, L. M., Ditkovsky, K. A., Dyachenko, E. L., Kotsemir, M. N., Kuznetsova, I. A., Lukinova, E. I., Martynova, S. I., Nefedova, A. I., Ratai, T. V., Rosovetskaya, L. A., Sagieva, G. S., Streltsova, E. A., Suslov, A. B., Tarasenko, E. I., Fridlyanova, S. Yu., & Fursov, K. S. (2019). *Indicators of science 2019: statistical collection.* Higher School of Economics. NRU HSE.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2*, 427-431.

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. *Third IEEE International Conference on Data Mining* (Melbourne, FL, USA), 179-186. https://doi.org/10.1109 / ICDM.2003.1250918

Mikhailets, E. B., Radin, I. V., & Shurtakov, K. V. (2019). Interim assessment of the degree of achievement of the planned targets of the Federal Target Program "Research and Development in Priority Areas

of Development of the Scientific and Technological Complex of Russia for 2014-2020". *Economics of Science, 5*(4), 234-247.

Nurakhov, N. N., & Telemtaev, M. M. (2011). Concept and structure of the Intellectual Property Cadastre. *Management of economic systems: electronic scientific journal, 8,* 25.

Pererva, O. L., Stepanov, S. E., & Nezimova, S. S. (2018). Using Big Data Analysis to Determine the Factors of the Effectiveness of the Public Procurement Process. *Bulletin of Eurasian Science, 10*(3), 32.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Vlasova, V. V., Gokhberg, L. M., Dyachenko, E. L., Kuznetsova, I. A., Kuznetsova, T. E., Martynova, S. I., Nefedova, A. I., Ratai, T. V., Rud, V. A., Sagieva, G. S., Streltsova, E. A., Suslov, A. B., & Fursov, K. S. (2018). *Russian science in figures.* Higher School of Economics. NRU HSE.