Application of Machine Learning Methods to Compare Disciplines Content Using Text Data

Roman Kupriyanov, Dmitry Zvonarev, Ruslan Suleymanov
Moscow City University
Moscow, Russia
kupriyanovrb@mgpu.ru, zvonarevd@mgpu.ru, sulejmanovrs@mgpu.ru

***Abstract*—The paper investigates one of the approaches based on machine learning methods aimed at finding and identifying similar disciplines. In the research we used two most popular methods of machine learning to process text data – BERT and Doc2Vec. Machine learning was conducted using the datasets of various disciplines with the total of 2,5 million entries. To assess the quality of the developed models, 30 experts from different scientific fields were engaged in the study to evaluate the level of similarity between the disciplines defined by the trained models. Based on the results of the research, both methods trained using identical datasets generated similar outputs. Another algorithm Doc2Vec, trained on a relatively small data sample with 15 000 entries of the target discipline database that included disciplines descriptions and curriculums, showed better results which justifies the need for developing specific solutions for particular tasks. Further development of machine learning methods and models design to solve specific tasks in the educational field will promote digitalization of education within the area of university operations management.**

I. Introduction

The problem of mutual recognition of education exists both at the national and international levels within two dimensions: recognition of professional and academic learning outcomes [1].

Recognition of academic learning outcomes has become a pressing issue in the last 10–15 years due to a number of reasons. For example, universities tend to exercise more autonomy in training programmes design not so much to ensure their correspondence to a programme track or an academic field, but to satisfy the needs of the end-consumers and society which is additionally confirmed by the fact that a training programme has turned into an actual commercial product [2].

Within the Russian academic community, universities expand the application of benchmarking technologies to increase their capacity of timely and rapid response to the labour market's needs and improve their competitive positions at the Russian and international markets of educational services [3].

Students are also a driver of this process by being proactive in designing their own personalized learning tracks and expanding academic mobility opportunities. It is noteworthy that the academic mobility rates have not declined even in the conditions of the pandemic, which stimulates the new ways of being involved in the academic mobility, such as distance learning technologies, developing and utilizing online courses [4].

Taking into account the processes described above, various educational stakeholders, such as students and academics, might find useful an intellectual digital system that supports decision-making by being able to identify training programmes or components of training programmes with similar learning outcomes.

Applying machine learning algorithms for comparison of learning content with anticipated learning outcomes can be difficult due to the implicit context meanings within the descriptions of training programmes. The use of identical terminology in similar-titled disciplines or disciplines with similar content does not guarantee their full correspondence with each other when viewed from the perspective of specific scientific or professional fields. Therefore, the task of developing machine learning models to analyse disciplines text descriptions and learning outcomes boils down to not so much obtaining quantitative parameters of text similarity but developing algorithms for their semantic analysis that would include contextual components.

Within the course of this study, we have conducted a comparative analysis of three machine learning algorithms developed to search for similar disciplines studied by Moscow City University's (MCU) students.

II. Related Work

Finding and identifying similar disciplines is an important task for any educational institution. The study [5] provides an account of a research aimed at finding similar disciplines and defining semantic similarity of text descriptions. The authors of the study plan to use the obtained values of similarity as a guideline system for university students to facilitate their choice of elective disciplines. The research was based on the data of 94 disciplines, 572 discipline units / PDF textbooks from the Faculty of Mathematics and Computer Science. To generate word embeddings, Paragraph Vectors was used (one of Doc2Vec versions).

A similar task in the educational field is finding similar exercises within online education systems. In the research [6], a novel Multimodal Attention-based Neural Network (MANN) framework was developed for finding similar exercises in large-scale online education systems with the use of text embedding and pictures. This solution is based on using neural

networks, such as long short-term memory network and convolutional neural network.

Efficient search of similar text documents is in demand in many fields. One of the most common fields is processing customers' feedback. The paper [7] analyses text similarity of users' feedback to improve the recommendation system. The study used different methods and techniques (TF-IDF, embedding-based matching, LSTM, SIF, etc.) with the best result achieved by LSTM.

In the work [8] the authors conduct a comparative analysis of neural embedding approaches, such as Google Sentence Encoder, ELMo, and GloVe, which apply traditional similarity scales. The results of the analysis showed that Google Sentence Encoder and ELMo were most efficient among others. Besides comparing algorithms and their combinations, there is research [9] that applies approaches based on combining arrays of similarities into one, which also increases the final accuracy of similar text documents search.

To work with the text data, the following steps were taken at the stage of data pre-processing: removing punctuation marks, tokenizing words, removing stop-words, normalizing words (lemmatizing), lowercasing all words. The next step after the text data pre-processing is feature engineering when textual representation is transformed into digital representation. The methods often used to do this are bag-of-words (BOW) and word embedding. To conduct search of similar courses there are various methods, for example, fuzzy clustering algorithms based on Latent Dirichlet allocation (LDA) that uses the topic model approach [10]. In the research [11] dedicated to analysing semantic similarity of text data, two categories of approaches are defined: non-deep learning and deep learning.

Within the non-deep learning approaches, the Doc2Vec algorithm has proved efficient since it can represent full text documents in the form of a vector. Particularly, a team of scientists in the research [12] used the Doc2Vec algorithm to search for similar issues in Jira when processing customer requests. It is worth pointing out that despite the fact that Doc2Vec is a recent development, it is one of the most cited methods in scientific papers according to the survey [11]. Within the deep learning approaches, one of the most recent methods available for application is BERT (Bidirectional Encoder Representations from Transformers) [13], which shows high performance in solving text data processing tasks and is able to train context-dependent models.

III. METHODS

To identify the most efficient algorithm, the results generated by the algorithms were analysed and compared by a team of experts. The analysed sample included 450 disciplines from 8 scientific fields:

- History and Social Studies
- Physical Education and Sports
- Linguistics
- Culture and Arts
- Law and Management
- General Education Science and Psychology
- Special and Correctional Pedagogy
- Mathematics and Information Science

For each discipline, the algorithms selected 5 matching disciplines identified as highly similar in terms of content and learning outcomes. In total, the study was based on the sample of 7200 disciplines from the university database of 50 000 disciplines taught university-wide. The selection of the disciplines did not account for the professional fields that encompassed various disciplines.

To conduct the expert evaluation of the results generated by the algorithms, 30 experts from different scientific fields were engaged in the study (minimum of 3 experts for one scientific field). All experts are lecturers at the university and are highly qualified specialists. Every expert was asked to conduct a pair-wise comparison of the disciplines from the original dataset with the sets of matching disciplines generated by the algorithms, using the following criteria: concordance of the disciplines' structure and content; concordance of the prospective learning outcomes (excluding the professional field as one of the objectives of the training programme that comprises the discipline within it); concordance of the disciplines' scope that ensures obtaining the described learning outcomes.

The task for each expert was to compare at least 10 disciplines from the original dataset with the sets of matching disciplines generated by the algorithms. It is important to mention that 3 disciplines from each scientific field were to be analysed by all experts representing the same scientific field so that to evaluate the coherence of the expert opinions [14].

The similarity between the disciplines was evaluated across all the criteria in an integrated manner using the 5-grade scale with grades from 1 ('very different') to 5 ('this is exactly the same'). The 5-grade scale was chosen because, on the one hand, it is sufficient to differentiate between the expert opinions, on the other, it is a common tool used within the academic community to evaluate students' knowledge. The numerical values of the scale were converted into words so that the experts had no doubt in interpreting their meanings.

To train the training models based on the BERT and Doc2Vec algorithms, two datasets were used:

- The first dataset (MES DB) included the list of lessons uploaded in the Moscow Electronic School (the MES). The dataset included text descriptions and lesson scripts. The total sample size was over 2.5 million entries.
- The second dataset (MCU DB) included discipline descriptions uploaded in the MCU database. The dataset included titles of disciplines and descriptions of study units within disciplines (study schedules). The total sample size used to train the algorithm model included 15 000 entries.

To conduct the study, we applied the BERT and Doc2Vec algorithms to develop three models aimed at search of similar disciplines – two models are basic and one as additional:

- BERT-DB MES – a model based on the BERT algorithm that was additionally trained using the MES database.
- Doc2Vec-DB MES – a model based on the Doc2Vec algorithm that was additionally trained using the MES database.

- Doc2Vec-DB MCU – a model based on the Doc2Vec algorithm, that was additionally trained using only the MCU database.

To train the algorithms BERT and Doc2Vec, we used Python and corresponding machine learning libraries – tensorflow/PyTorch and gensim. The models were trained using CPU. The training time for BERT-DB MES was approximately 10 hours, for Doc2Vec-DB MES – one hour.

To work with BERT, we used a pre-trained model rubert_cased_L-12_H-768_A-12_pt [15], which was trained on the Russian Wikipedia texts and news articles. The following parameters were used to work with this model: Batch size equals 32; Learning rate equals 2e-5; Epochs equals 4. As an optimizer we used AdamW – optimizer that implements the Adam algorithm with weight decay [16].

The training of the model based on Doc2Vec-DB was performed with the following algorithm parameters: ignores all words with total frequency lower than 1; dimensionality of the feature vectors equals 100; the maximum distance between the current and predicted word within a sentence equals 3; number of iterations (epochs) over the corpus equals 10.

IV. Results

To identify the disciplines similar to the disciplines in the initial dataset, we applied cosine similarity which is one of the most widely used similarity measurements and great for creating a baseline for further improvement. In cosine similarity each document is represented by a vector comprised of components, each corresponding to a word from a dictionary. The component equals 1 if the corresponding word is included in the text, otherwise it equals 0. The value of the cosine between the two vectors depends on the number of similar words that are present in both documents. If there are two vectors of attributes, A and B, the cosine similarity $cos(\theta)$ can be described by means of scalar product and norm [17]:

$$sim = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

The following factor was used as the metrics to assess the quality of the results generated by the algorithms:

$$S\text{–}meanscore_i = \frac{\sum_e \sum_a C_{iae}}{E_i * A_i * 5},$$

where C_{iae} describes similarity between discipline i and discipline a assessed by expert e using the grades from 1 to 5. E_i is the number of experts assessing disciplines i, A_i is the number of disciplines assessed in terms of similarity to discipline i.

Thuswise, we calculated the factor with the values from 0.2 to 1 that describes the overall quality of the sets of disciplines identified by the algorithms as similar to the initial dataset. The higher the value of the factor the higher the similarity between the tested dataset and the initial dataset, according to the expert assessments.

The factor values are standardized and do not depend on the number of experts or alternatives. The results of the experiment are shown on Fig. 1.

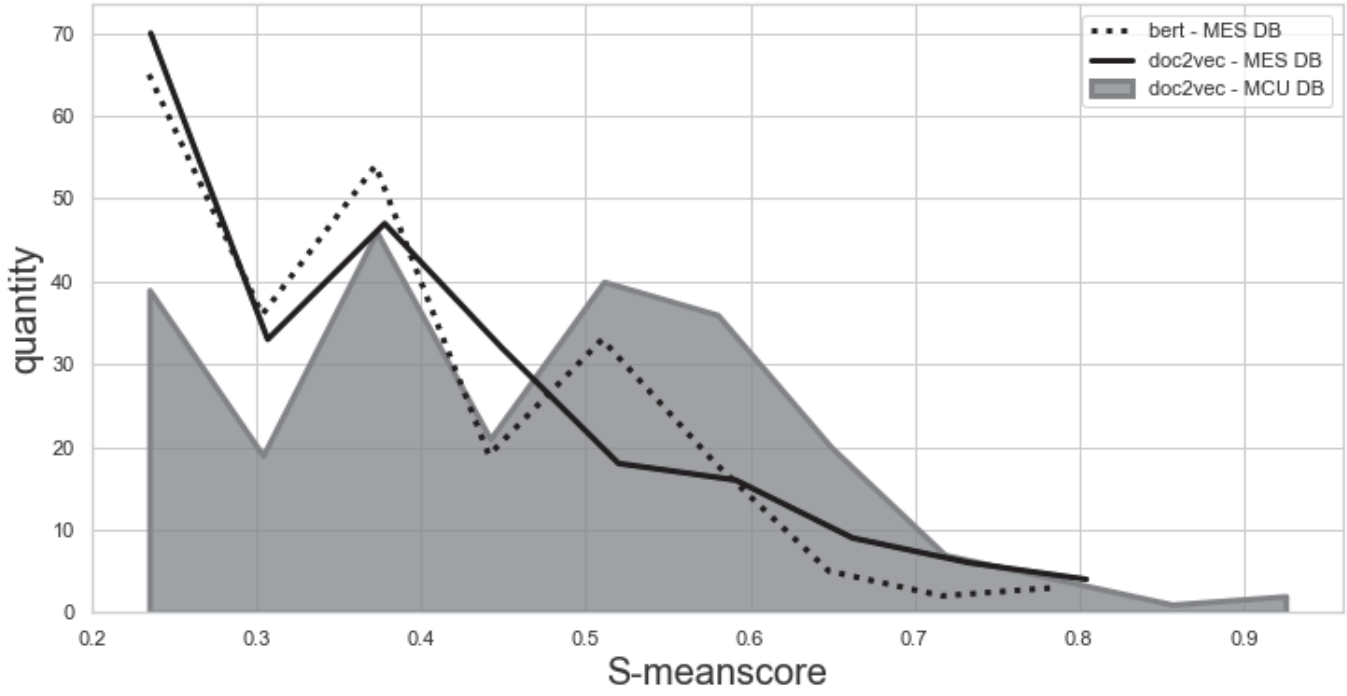

Fig. 1. Results of the comparative analysis of 3 algorithms

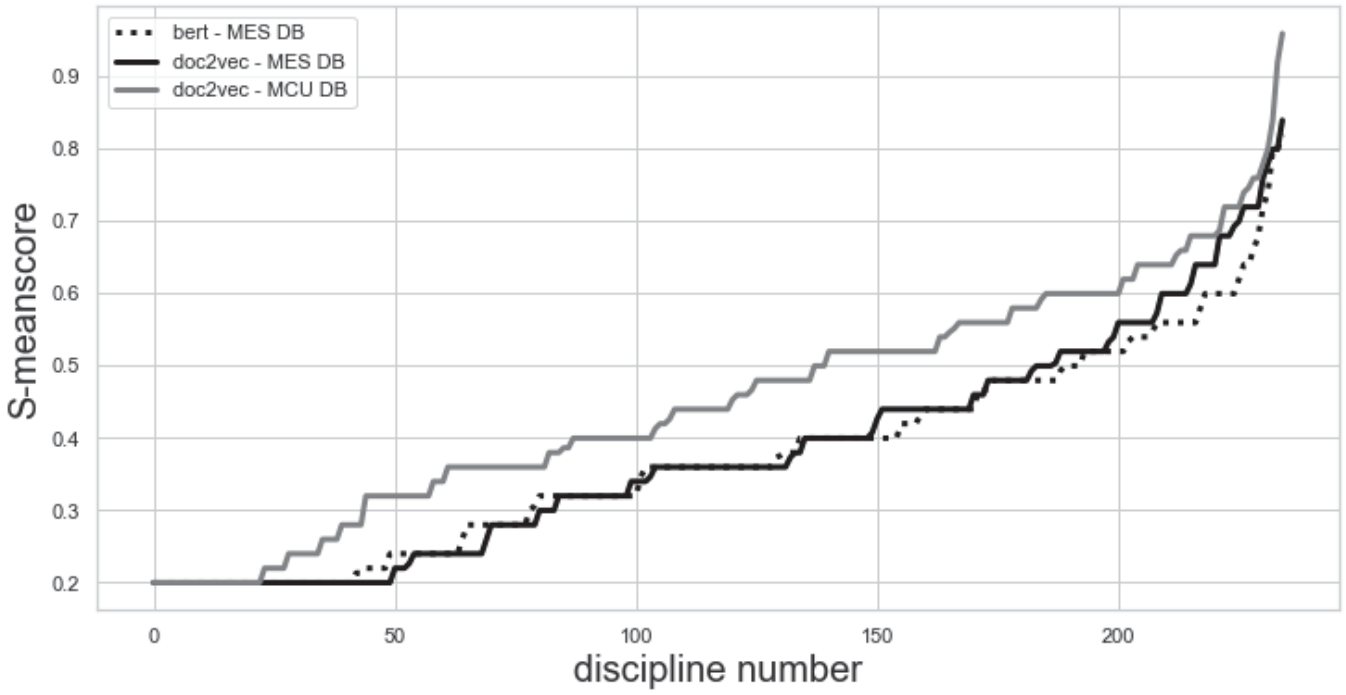

Fig. 2. S-meanscore values in ascending order

Using the metrics applied, we calculated the values for every algorithm within each of the selected scientific disciplines in Table I.

TABLE I. STATISTICS OF SIMILARITY WITHIN THE DISCIPLINES CALCULATED BY THE ALGORITHMS

Algorithms / Discipline	BERT-MES DB	Doc2Vec-MES DB	Doc2vec-MCU DB
History and Social Studies	0.26	0.24	0.36
Special and Correctional Pedagogy	0.40	0.39	0.48
Physical Education and Sports	0.36	0.46	0.44
Linguistics	0.40	0.36	0.44
Culture and Arts	0.30	0.28	0.40
Law and Management	0.34	0.34	0.40
General Education Science and Psychology	0.35	0.39	0.52
Mathematics and Information Science	0.43	0.47	0.56

According to Table I, the highest performance was shown by the algorithm Doc2Vec-MCU DB which yielded best results for 7 out of 8 scientific disciplines.

The following graph shows the values of S-meanscore factor arranged in ascending order for 3 tested algorithms (Fig. 2). Based on the visual presentation, we can make a conclusion that the algorithm Doc2Vec-MCU DB is more efficient in every data point of the graph according to expert opinions, since the curve of the Doc2Vec-MCU DB algorithm is located higher than the curves of other algorithms.

The correlation of the *S-meanscore* values of different algorithms is shown in Table II, while the comparison of every pair of algorithms by the *S-meanscore* value is shown on Fig. 3, Fig. 4, Fig. 5.

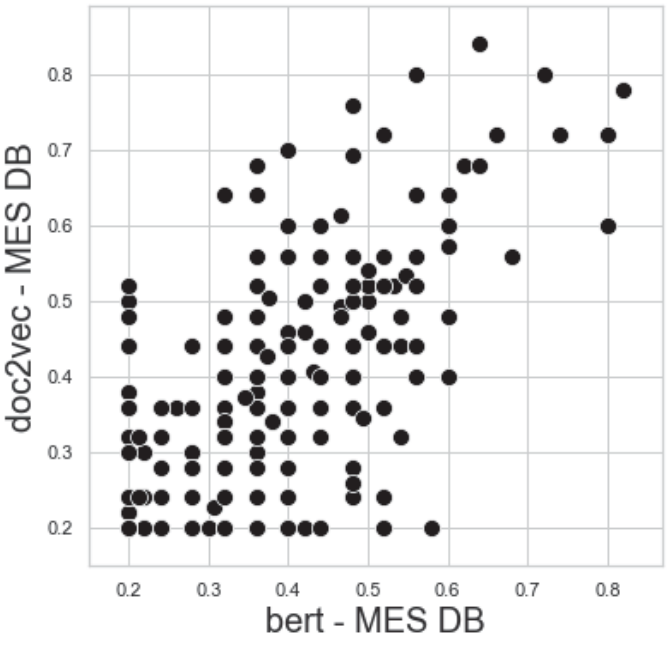

Fig. 3. Pair-wise comparison of S-meanscore values for “bert - MES DB” and “doc2vec - MES DB”

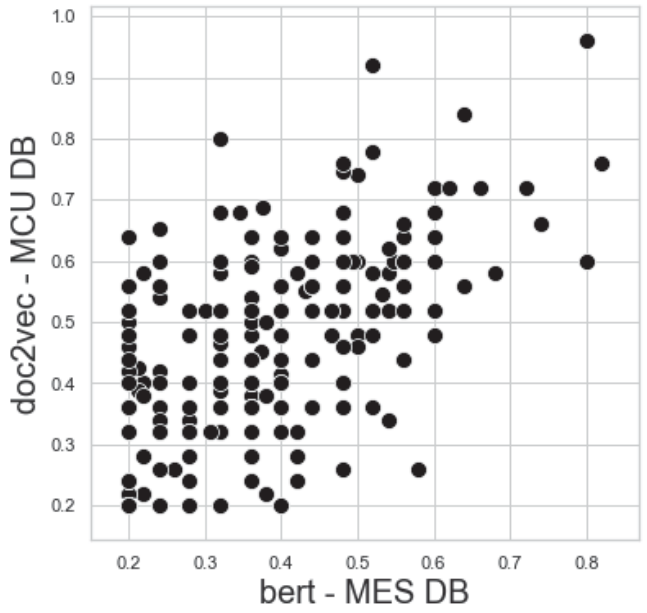

Fig. 4. Pair-wise comparison of S-meanscore values for “bert - MES DB” and “doc2vec - MCU DB”

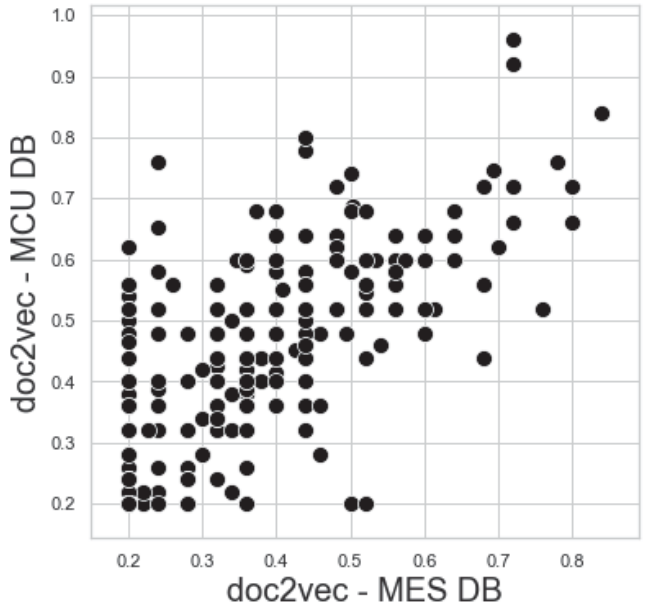

Fig. 5. Pair-wise comparison of S-meanscore values for “doc2vec - MES DB” and “doc2vec - MCU DB”

TABLE II. THE CORRELATION OF EXPERT ASSESSMENTS ACROSS ALGORITHMS

	bert-MES DB	**doc2vec-MES DB**	**doc2vec-MCU DB**
bert-MES DB	1.00	0.68	0.60
doc2vec-MES DB	0.68	1.00	0.66
doc2vec-MCU DB	0.60	0.66	1.00

To test the significancy of the differences between the results of the expert assessment, we conducted the Mann–Whitney U-test for all pairs of algorithms. The results are shown in Table III.

TABLE III. PAIR-WISE TESTS OF THE RESULTS OF ALGORITHMS EXPERT ASSESSMENT

Pairs compared	**Ratio value**	**p-value**
Doc2Vec-MCU DB / Doc2Vec-MES DB	35090	< 0,001
Doc2Vec-MCU DB / BERT-MES DB	35636	< 0,001
Doc2Vec-MES DB / BERT-MES DB	27664	0. 972

It can be assumed form Table III that the difference between the results of Doc2Vec-MCU DB and other algorithms is significant, while for the algorithms Doc2Vec-MES DB and BERT-MES DB the distributions are equal under the null hypothesis.

V. DISCUSSION

The research results can be used when designing personalized learning tracks, developing guidelines for students to choose elective courses of interest. Such systems can prove useful for faculty committees responsible for recognising credits obtained by students during exchange programmes at other educational institutions (i.a. in other countries) or upon completing MOOCs. Another application is for students to support them in choosing exchange programmes or MOOCs that would be most useful for their further studies [18].

VI. CONCLUSION

The final output of the study is development of a personalized toolkit for text data processing based on the machine learning algorithms enabling to perform the described tasks for all educational stakeholders.

The development of the solutions suggested in this paper is aimed at digitalization of designing personalized learning tracks that provide optimal learning outcomes and take into account individual learning needs of students, as well as ensure transparency of the university’s internal educational environment and interactions between the university students and external stakeholders of the market of educational services.

ACKNOWLEDGMENT

This research was supported by the Moscow Department of Education.

REFERENCES

[1] A.O. Tchetverikov, “Pravovoe regulirovanie vzaimnogo priznaniia obrazovaniia i kvalifikatsii v evropeiskom Soiuze: opyt sistemnogo analiza [Legal Regulation of the Mutual Recognition of Education and Qualifications in the European Union: a Systemic Analysis]”, *Courier of Kutafin Moscow State Law University (MSAL)*, 2018, vol. 5, pp. 133-145. (In Russ.).

[2] N.S. Mushketova, “Kontseptsiia marketinga vuza: soderzhanie, printsipy, funktsii v sovremennykh usloviiakh [The Marketing Concept of University: Contents, Principles, Functions in Modern Terms]”, *Uchenye zapiski Orlovskogo gosudarstvennogo universiteta. Seriia: Gumanitarnye i sotsial'nye nauki [Proceedings of Orel State University. Series Humanitarian and Social Sciences]*, 2018, vol. 5, pp. 123-149. (In Russ.).

[3] T.S. Kovylnikova and L.S. Pavlova, “Proektirovanie osnovnoi obrazovatel'noi programmy vysshego obrazovaniia pri realizatsii FGOS VO (FGOS 3+) s uchetom trebovanii professional'nykh standartov [Designing Principal Educational Program in Higher Education to Implement 3+ FSES Taking into Account Professional Standard Requirements]”, *Vestnik Tverskogo gosudarstvennogo universiteta. Seriia: Pedagogika i psikhologiia [Bulletin of Tver State University. Series Education Science and Psychology]*, 2017, vol. 1, pp. 42-64. (In Russ.).

[4] Yu. Cherepanova, “Mezhdu pervoi i vtoroi: onlain-obrazovanie na volne pandemii [Between the first and the second: online education on the wave of the pandemic]”, *Forbes Education*, Web: https://education.forbes.ru/authors/online-education-vs-covid/ (In Russ.).

[5] N. Seidel, C.M. Rieger and T. Walle, “Semantic Textual Similarity of Course Materials at a Distance-Learning University”, *Proceedings of 4th Educational Data Mining in Computer Science Education (CSEDM) Workshop co-located with the 13th Educational Data Mining Conference (EDM 2020)*, Virtual Event, July 10, 2020.

[6] Q. Liu, Zai Huang, Zhenya Huang, C. Liu, E. Chen, Y. Su and G. Hu, “Finding Similar Exercises in Online Education Systems”, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*. Association for Computing Machinery, New York, NY, USA, 2018, pp. 1821-1830. https://doi.org/10.1145/3219819.3219960

[7] N. Ghasemi and S. Momtazi, “Neural text similarity of user reviews for improving collaborative filtering recommender systems”, *Electronic Commerce Research and Applications*, vol. 45, 101019, 2021. https://doi.org/10.1016/j.elerap.2020.101019

[8] M. Hendre, P. Mukherjee and M. Godse, “Utility of Neural Embeddings in Semantic Similarity of Text Data”, *Evolution in Computational Intelligence. Advances in Intelligent Systems and Computing*, vol. 1176, 2021. Springer, Singapore. https://doi.org/10.1007/978-981-15-5788-0_21

[9] M. Farouk, “Measuring text similarity based on structure and word embedding”, *Cognitive Systems Research*, vol. 63, 2020, pp. 1–10. https://doi.org/10.1016/j.cogsys.2020.04.002

[10] I.A. Kuznetsov and A.I. Guseva, “A method for obtaining a type of scientific result from the text of an article abstract to improve the quality of recommender systems”, *Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019*, 2019. pp. 1888-1891 doi: 10.1109/EIConRus.2019.8656806

[11] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun and C. Gao, “A survey on the techniques, applications, and performance of short text semantic similarity”, *Concurrency Computat Pract Exper*, 2020, vol. 33-5, e5971. https://doi.org/10.1002/cpe.5971

[12] A. Kovalev, N. Voinov and I. Nikiforov, “Using the Doc2Vec Algorithm to Detect Semantically Similar Jira Issues in the Process of Resolving Customer Requests”, *Intelligent Distributed Computing XIII. IDC 2019. Studies in Computational Intelligence*, vol. 868, 2020, Springer, Cham. https://doi.org/10.1007/978-3-030-32258-8_11

[13] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv* 2018, arXiv:1810.04805

[14] T.Ya. Danelyan, “Formal"nye metody ekspertnykh otsenok [Formal Methods of Expert Estimations]”, *Applied Informatics. Economics, Statistics and Informatics*, vol. 1, 2015, pp. 183-187.

[15] DeepPavlov: an open source conversational AI framework, Web: https://deeppavlov.ai/.

[16] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization”, *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6-9, 2019.

[17] A.E. Silaeva, G.A. Gabrielyan, I.A. Isaeva and E.V. Nikulchev, Intellektual'nyi analiz tekstovykh otvetov v massovykh oprosakh [Intelligent analysis of text responses in large-scale surveys]. vol. 6, 2019, pp. 779-788. (In Russ.).

[18] S. Vachkova, R. Kupriyanov, R. Suleymanov and E. Petryaeva, “The Application of Text Mining Algorithms to Discover One Topic Objects in Digital Learning Repositories”, *2021 28th Conference of Open Innovations Association (FRUCT)*, 2021, pp. 502-509, doi: 10.23919/FRUCT50888.2021.9347611.