European Proceedings of
Social and Behavioural Sciences
EpSBS

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

DOI: 10.15405/epsbs.2022.08.13

**ICEST 2022**
**III International Conference on Economic and Social Trends for Sustainability of Modern Society**

# PREDICTIVE MODELING FOR IMPROVING THE QUALITY OF EDUCATIONAL AND METHODOLOGICAL DOCUMENTATION

Vladimir Okrepilov (a,b), Anna Stepashkina (a)*
*Corresponding author

(a) State University of Aerospace Instrumentation, 67, Bolshaya Morskaya str., Saint Petersburg, 190000, Russia, okrepilov@test-spb.ru
(b) Institute for Regional Economics Studies Russian Academy of Science, Serpukhovskaya str., 38, Saint Petersburg, 190013, Russia, stepashkina.anna@yandex.ru

## Abstract

The preparation of educational and methodological documentation in educational institutions is an important element of the educational process. The purpose of the scientific article is to develop quality criteria for the educational and methodological complex of the discipline. We pay attention on checking the compliance of the document content. The paper presents the criteria for the effectiveness of the developing the educational-methodical complex of disciplines. We propose the method for improving the quality of educational and methodical documentation based on the work of a neural network. The method allows automating the process of checking educational and methodical documentation for the content. Natural language processing (Bert tokenizer) and machine learning methods (regression methods) are used as a tool. The problem of the quality of educational and methodological documentation can be solved by introducing specialized automated systems and their modules, which allow not only to generate appropriate templates for educational and methodological documentation but also to check the internal content of documents.



*Keywords:* Educational and methodological complex of disciplines, higher education, quality assessment

https://doi.org/10.15405/epsbs.2022.08.13
Corresponding Author: Anna Stepashkina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

## 1. Introduction

Every year higher education organizations prepare a large number of new work programs of disciplines or update existing ones. The methodological departments are engaged in the quality control of the educational and methodical complex of the discipline (EMCD) which includes the working programs (plans) of the disciplines (WP) and the funds of evaluation means (FOM). Methodologists, as a rule, check for compliance with the EMCD to a number of criteria including compliance with the federal state educational standard (FSES) of the training area, curriculum, discipline competencies, relevance of the list of references and software licenses for the implementation of the course. It is difficult to check the content of the annotation and the WP and FOM itself. The compilers of the EMCD are specialists in their narrow field. So, this task may seem easier for them. But we are also talking about large volumes. Natural language processing and machine learning techniques can simplify the work of checking the quality of content.

## 2. Problem Statement

For a long time now the development of the WP has not been completely done by teachers and methodologists manually. For this, for example, internal automated information systems AIS or LMS-Learning Management System are used (Kosmachyova et al., 2016; Rogov et al., 2017).

LMS is a software application designed for creating training courses. The system allows to structure the training material by section, lay out assignments, conduct online seminars, track the progress of students, etc. (Belozyorova & Chyiko, 2019; Cole & Foster, 2007; Dias et al., 2013; Kats, 2010; Menshikov et al., 2006; Vong & Prokhorova, 2015).

For these purposes the National Research University Higher School of Economics (NRU HSE) successfully implements a special module-constructor "Programs of Academic Disciplines" available to teachers in the personal accounts of the LMS. The module is based on the approved rules for drawing up curricula for academic disciplines. The module helps the teacher to design the program according to the approved format, publish the abbreviated version on the HSE portal for external users and the full version within the corporate system for use by the participants of the educational process: students, teachers, administration. The base of programs of academic disciplines is formed and stored in electronic form (Shalakmov & Starichkova, 2015).

SUAI has a specially developed information system coupled with LMS Moodle named AIS SUAI. AIS SUAI allows to form WP templates. The templates already have a certain partially completed form. Information is indicated on the number of hours allocated for the discipline, semester, in accordance with the curriculum, competencies, type of control. The teacher fills in the content part exclusively by hands. Then the programs are uploaded manually in the AIS SUAI. At these stages technical errors are not excluded:

i. the teacher mixed up the file, entered incorrect text or accidentally copied the material;
ii. a person who is responsible for uploading files to AIS SUAI can make mistakes in cells, upload mathematics to physics, as an example.

https://doi.org/10.15405/epsbs.2022.08.13
Corresponding Author: Anna Stepashkina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

It is possible to solve the issue of improving the quality of the content of the EMCD by introducing specialized intelligent digital systems: machine learning and natural language processing methods.

## 3. Research Questions

Machine learning methods are widely used to solve various technical and business tasks such as the creation of recommender systems, image analysis and processing, big data analysis, etc. The paper proposes a method that allows you to check the correctness of textual information based on natural language processing and supervised machine learning. The principle of operation is based on the use of supervised machine learning tools, i.e. initially; a software product is prepared, trained in advance for the topic.

The method can be divided into the following steps:

i. Preparation of a training sample by topic;
ii. Preparation of the model, program code, training of the model:
iii. Converting textual information to tensor for machine learning model,
iv. Application of the machine learning method.
v. Assessment of the quality of the result;
vi. Testing on real data.

## 4. Purpose of the Study

Let us consider the method of automated verification of EMCD using the example of the discipline "Metrology". The preparation of the training dataset takes a significant amount of time: a table of two columns is compiled from authoritative publications (Bavikin, 2019; Dehtyarev, 2021; Gribanova, 2019; Okrepilov et al., 2021). The columns named as;

**Text**, which has a text information;
**y / n**, shows how the text information corresponds to the topic of the dataset (in this case it is metrology).

An example of a training dataset is shown in Table 1 for five rows. Each row of the column **Text** contains no more than two or three sentences related to the discipline Metrology in accordance with authoritative sources. In addition to information corresponding to the discipline "Metrology", lines with similar terminology from related disciplines such as mathematics, physics, chemistry, etc. have been collected. Also data from other areas not directly related to metrology have been used: cultural studies, foreign language, history, etc.

The column **y / n** contains binary information: 1 and 0. One (1) corresponds to information related to metrology, 0 means that the text belongs to another scientific field.

When creating a training dataset, attention should be paid to the number of characters in a line. For the code quick work the value of characters should not exceed 300 items. Lines that are too long are discarded or split into shorter

https://doi.org/10.15405/epsbs.2022.08.13
Corresponding Author: Anna Stepashkina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

**Table 1.** Training data set for the discipline "Metrology"

| Text | y/n |
|---|---|
| establishment of units of physical quantities, state standards and exemplary measuring instruments | 1 |
| Metrology is the science of measurements, methods and means of ensuring their unity and ways to achieve the required accuracy. | 1 |
| Theoretical metrology is a branch of metrology the subject of which is the development of the fundamental foundations of metrology. | 1 |
| development of theory, methods and means of measurement and control; | 1 |
| ensuring the uniformity of measurements | 1 |

When examining a manually prepared training dataset consisting of 5000 lines the most common words associated with the discipline of metrology are measurement, means, result, method, magnitude, etc.

## 5. Research Methods

Initially our dataset is the text that is not understandable to the machine. To do this we will translate the text into a machine language. After this procedure we can use machine learning models. Thus we are talking about sequential work with two models and the dataset from one model to another is transmitted in the form of vectors (matrix) of a certain dimension.

Modern tasks of natural language processing, transforming the text into a vector, are quite well solved using the PyTorch-Transformers library (the library of modern pre-trained models). Models allow to prepare unlabeled text dataset for further building machine learning models (Liu et al., 2019). A mechanism that studies the contextual relationship between words (or subwords) in a text. Transformer includes two separate mechanisms - an encoder (which reads the input text) and a decoder (which makes a prediction for the task).

The BERT model is currently one of the most advanced and is based on the Google search engine. Soon after the release of the document describing the model the team also opened the source code of the model and made available for download versions of the model that had already been pretrained on massive datasets.

BERT is a model that has broken several records for solving natural language processing problems. BERT is a pretrained TransformerEncoder stack. There are two main BERT models in different sizes:

BERT BASE
BERT HUGE

Both sizes of the BERT model have a large number of encoder layers.

BERT accepts a sequence of words (a sentence) as input and outputs the sentence as a numeric array. This matrix is used as input for the selected machine learning model: regression, classification, etc.

Classic BERT requires a lot of computing resources, data volumes, computer performance. Due to which it can be difficult to run on any device. The classic BERT serves as a good starting point for modernizing your code. Its derivatives RoBERTa which can improve performance and DistilBERT which

https://doi.org/10.15405/epsbs.2022.08.13
Corresponding Author: Anna Stepashkina
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

increases the output speed, have shown good results in different areas (Sanh et al., 2020; Vaswani et al., 2017). Let's consider each variation in more detail.

DistilBERT is an open source minified version of BERT. It is a lighter and faster version of BERT and is roughly in line with its specifications. The algorithm is based on the approximation of a large neural network to a smaller one. The optimization function is based on the Kullback-Leiber divergence (Vaswani et al., 2017). DistilBERT reduces the size of the BERT model by 40% while still delivering 97% performance. DistilBERT is able to understand natural language 60% faster. Thus DistilBERT allows increasing the speed with some loss in quality, insignificant in some tasks.

Another model (RoBERTa) was given sufficient attention by the developers to forecasting metrics, due to which the model can provide high accuracy. The RoBERTa approach is a reworking of BERT with improved training methodology (Sanh et al., 2020). RoBERTa allows for more data processing and has more processing power. The model differs from the classic BERT by dynamic masking.

To select the optimal method for transforming a natural language into a machine language 400 rows were randomly selected from the training dataset. Then three modifications of the BERT algorithm were applied (RoBERTa, DistilBERT, BERTBase). The resulting matrices were processed by the Logistic Regression algorithm and the accuracy of the decision was estimated by the Score metric as well as by the assessment of the cross-validation sample. The results are shown in Table 2.

**Table 2.** Compression of natural language processing algorithms

| Algorithm | Model trainingtime | Cross validationscore (n = 5) |
|---|---|---|
| BERT Base | 4 min 46 sec | array([0.822, 0.733, 0.711, 0.711, 0.8]) |
| DistilBERT | 2 min 6 sec | array([0.689, 0.733, 0.844, 0.733, 0.8]) |
| RoBERTa | 6 min 36 sec | array([0.844, 0.822, 0.778, 0.822, 0.8]) |

The DistilBERT algorithm is indeed significantly faster than the classic BERT Base but it loses in accuracy, only slightly. RoBERTa has the expected high accuracy but the speed of working and the resources required are too big. Let's stop on the DistilBERT algorithm.

Next the training dataset was used to train the stack of natural language processing models DistilBERT and machine learning logistic regression.

## 6. Findings

Let's consider how a trained model works on real data. For this dataset of WP, EMCD or WP annotations are loaded. The program processes the data from a table, the test dataset. After that the test dataset is converted into a numeric array using the DistiBERT algorithm as was done for the training dataset. The result is predicted for each line, averaged and rounded to an integer value, the answer will be 0 or 1. If the content of the text matches its subject the screen displays “Content matches the discipline”, otherwise the message “Error, check file” appears. It takes no more than 22 seconds to check one EMCD.

## 7. Conclusion

A method for automated verification of quality indicators related to the substantive part of the EMCD is proposed: compliance of the content of the document with the discipline. The method is based on natural language processing algorithms DistilBERT and machine learning algorithm LogisticRegression. The method allows to significantly reduce the time for checking the EMCD, thereby improving the quality and increasing the efficiency of the process of developing the EMCD.

## Acknowledgments

The paper was prepared with the financial support of the Russian Foundation for Basic Research in the framework of research under the project of the RFBR No. 19-010-00968

## References

Bavikin, O. B., Fedorova, O. V., Dmitrievich, G. D., Zajcev, S. A., Parfenova, I. Ye., & Tolstoc, A. N. (2019). *Metrology. Forum.* https://doi.org/10.12737/textbook_5be96d68d333e2.71218396

Belozyorova, S. I., & Chyiko, O. I. (2019). Experience of using LMS Moodle for the creation and maintenance of training courses. *Sovremennie problem nauki I obrazovaniya-Modern problems of science and education*, *1*. https://www.science-education.ru/ru/article/view?id=28448

Cole, J., & Foster, H. (2007). *Using Moodle*, *2nd*. UK, Edition – O'Reilly Media, Inc. https://www.oreilly.com/library/view/using-moodle-2nd/9780596529185/

Dehtyarev, G. M. (2021). *Metrology, standartizaton and sertification.* KURS. https://znanium.com/catalog/product/1584617

Dias, S. B., Diniz, J. A., & Hadjileontiadis, L. J. (2013). *Towards an Intelligent Learning Management System Under Blended Learning: Trends, Profiles and Modeling Perspectives*. International Publishing. https://doi.org/10.1007/978-3-319-02078-5

Gribanova, D. D. (2019). *Fundamentals of Metrology, Certification and Standardization: A Study Guide.* Nauka.

Kats, Y. (2010). Learning Management System Technologies and Software Solutions for Online Teaching: Tools and Applications. *Information Science Reference.* https://doi.org/10.4018/978-1-61520-853-1

Kosmachyova, I. M., Kvyatkovskaya, I. Yu., & Sibikina, I. V. (2016). Automated system for the formation of work programs of academic disciplines. *Vestnik Astrakhanskogo gosudarctvennogo technicheskogo universiteta - Bulletin of the Astrakhan State Technical University*, *1*, 90-97. https://cyberleninka.ru/article/n/avtomatizirovannaya-sistema-formirovaniya-rabochih-programm-uchebnyh-distsiplin/viewer

Liu, Y., Ott, M., Goyal, N., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer L., & Veselin, S. (2019). RoBERTa: A RobustlyOptimized BERT Pretraining Approach. *Computation and Language*. https://arxiv.org/pdf/1907.11692.pdf

Menshikov, V. A., Lysyi, S. R., & Menshikova, L. V. (2006). System of distance learning, training and retraining of specialists based on modern computer and telecommunication technologies. *Sovremennie informatsionnie tehnologii I IT-obrazovanie - Modern problems of science and education, 4,* 63-65. https://science-education.ru/ru/article/view?id=453

Okrepilov, V. V., Antokhina, Yu. A., Ovodenko, A. A. & Sulaberidze, V. Sh. (2021). Fundamentals of metrology. SUAI.

Rogov, I. E., Adonyev, A. A. & Starichkova Yu. V. (2017). Development experience, trends and implementation of information systems to support the main educational process. *Sovremennie informatsionnie tehnologii I IT-obrazovanie - Modern problems of science and education, 4*(13), 82-90. https://doi.org/10.25559/SITITO.2017.4.628

Sanh, V., Debut, L., Chaumond, J. Th., & Wolf, V. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Computation and Language. Machine Learning*. https://doi.org/10.48550/arXiv.1910.01108

Shalakmov, S. A., & Starichkova, Yu. V. (2015). Experience in the development and implementation of a module for automating the process of creating and approving programs of academic disciplines within the information educational environment for supporting the main educational process. *Vestnik Rossiiskogo universiteta druzhby narodov, Bulletin of the RUDN University, 4,* 67-75. https://arxiv.gaugn.ru/s2312-86310000619-6-1-ru-507/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Kaiser, L. (2017). Attention Is All You Need. *Computation and Language. Machine Learning*. https://arxiv.org/pdf/1706.03762.pdf

Vong, V. V., & Prokhorova, O. A. (2015). The usage of LMS Moodle in teaching a foreign language in graduate school in the framework of mixed and distance education. *Vestnik Kemerovskogo gosudarstvennogo universitets, Bulletin of the Kemerovo State University, 2*(62), *3,* 27-30. https://cyberleninka.ru/article/n/ispolzovanie-lms-moodle-pri-obuchenii-inostrannomu-yazyku-v-aspiranture-v-ramkah-smeshannogo-i-distantsionnogo-obrazovaniya/viewer