Method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP

Iurii Krak[1,2], Olexander Mazurets[3], Maryna Molchanova[3], Olena Sobko[3], Daryna Hardysh[3] and Olexander Barmak[3]

[1]*Taras Shevchenko National University of Kyiv, 64/13 Volodymyrska Str., Kyiv, 01601, Ukraine*
[2]*Glushkov Institute of Cybernetics of NAS of Ukraine, 40 Glushkov Ave., Kyiv, 03187, Ukraine*
[3]*Khmelnytskyi National University, 11 Instytutska Str., Khmelnytskyi, 29016, Ukraine*

Abstract

The paper is devoted to creation and approbation of method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP was proposed. The use of various approaches to selecting key terms, such as recognition for named entities, context analysis, and use of lemmatization methods, allows for comprehensive consideration of various aspects of educational materials, which ensures high level of correspondence of tests to educational goals. This method not only improves the content correspondence of tests, but also significantly reduces the resources required to check test tasks, making this process more automated and convenient for use in educational systems. In addition, the method's adaptability to different disciplines and languages opens up opportunities for its use in international educational environments. This not only contributes to improving the testing quality, but also stimulates the improvement of structure and content of educational materials, which, in turn, increases the overall quality of education. The research contributes to achievement of Sustainable Development Goals No. 4 (Quality education), No. 9 (Industry, innovation and infrastructure), No. 10 (Reduced inequality) and No. 12 (Responsible consumption and production). The results of the conducted experimental research of developed method for automated evaluation of test tasks correspondence to semantic structure of educational materials of number of STEM-disciplines showed the correlation of the results of method and conducted cluster analysis with visual interpretation. As result of comparing the averaged estimates of experts with the estimates obtained automatically, it was found that the developed method allows evaluating the correspondence of test tasks to the semantic structure of educational materials in STEM-disciplines with average accuracy of 94.6%; within individual topics of STEM-disciplines the minimum accuracy was 71.8%, and the maximum accuracy was 97.4%.

Keywords

NLP, test tasks, STEM, educational materials, adaptive testing, named entities recognition, KeyBERT

1. Introduction

The rapid development of artificial intelligence technologies, in particular natural language processing (NLP), creates new opportunities for automating processes in the field of education. In the context of the constant growth of the volume of educational materials and the need to systematize them, automated analysis methods are becoming necessary to ensure the relevance of the content of tests and curricula, since the control of results plays a fundamental role in the educational process [1, 2].

In STEM disciplines, the accuracy and relevance of educational materials play an important role, because any inconsistencies or inaccuracies can lead to serious errors in the educational process and negatively affect the assimilation of the material.

STEM@ICon-MaSTEd 2025: 4th Yurii Ramskyi STE(A)M Workshop co-located with XVII International Conference on Mathematics, Science and Technology Education, May 14, 2025, Ternopil, Ukraine

yuri.krak@gmail.com (I. Krak); exe.chong@gmail.com (O. Mazurets); m.o.molchanova@gmail.com (M. Molchanova); olenasobko.ua@gmail.com (O. Sobko); darinka.gardisch@gmail.com (D. Hardysh); alexander.barmak@gmail.com (O. Barmak)

0000-0002-8043-0785 (I. Krak); 0000-0002-8900-0650 (O. Mazurets); 0000-0001-9810-936X (M. Molchanova); 0000-0001-5371-5788 (O. Sobko); 0009-0002-6831-2337 (D. Hardysh); 0000-0003-0739-9678 (O. Barmak)

Thanks to NLP [3] it is possible to conduct a deep analysis of text data, which allows you to identify lexical, semantic and contextual inconsistencies between test tasks and the main educational materials. This allows for improving the quality of educational tests, ensuring their greater accuracy and compliance with the goals of the educational process [4].

Using NLP to assess the suitability of test tasks saves teachers time and automates verification processes, increasing efficiency and reducing the likelihood of human errors. Given the integration of intelligent systems into the educational process, developing such methods becomes an important component for improving the quality of STEM education in a global context [5].

The study is closely related to the Sustainable Development Goals defined by the United Nations. In particular, it contributes to the achievement of SDG No. 4 "Quality education" – ensuring comprehensive, inclusive and equitable education, as well as opportunities for lifelong learning for all [6]. In the context of globalization and expanding access to educational resources, the automation of test assessment makes it possible to provide high-quality and adapted learning that meets the needs of different groups of students, regardless of their socio-economic status or geographical location [7]. In addition, the study is in line with SDG No. 9 "Industry, innovation and infrastructure", which aims to develop sustainable infrastructure, promote innovation and industrial modernization, SDG No. 10 "Reduced inequality", which contributes to reducing inequalities in education, and SDG No. 12 "Responsible consumption and production" to ensure sustainable consumption and production [8]. The use of modern NLP technologies in the field of education contributes to the creation of intelligent systems that can integrate the latest scientific achievements into the educational process while reducing the cost of time and resources [9]. This creates the basis for building a sustainable educational infrastructure that can quickly adapt to changing requirements and implement innovative teaching methods [10]. The use of such methods can help eliminate educational disparities between countries and regions, ensuring the same quality of knowledge testing and test tasks, regardless of infrastructure or resource constraints. By automating the processes of testing and analysis of tests, resource consumption is reduced, particularly in terms of time, paper, and human effort. This allows not only the reduction of the costs of testing tests but also the implementation of a more environmentally sustainable approach to creating and using educational materials [11].

The aim of the article is to improve the processes of monitoring and evaluating the quality of educational materials, which allows for more accurate and complete coverage of key concepts and topics of STEM disciplines being studied and to optimize educational resources.

The main contribution of article is proposed method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP, which ensures the achievement of goals.

2. Related works

Automated evaluation of test tasks correspondence to semantic structure of educational materials is an important problem in modern scientific research in the field of educational technologies, in particular in STEM disciplines [12]. Tests, as one of the main means of assessing students' knowledge, are of particular importance in crisis situations, such as pandemics, armed conflicts and other extraordinary circumstances. In such conditions, traditional assessment methods that require the personal presence of teachers and students become difficult or even impossible due to restrictions on mobility, security or access to educational institutions. Tests, in particular automated ones, allow the preservation of the assessment process, providing the possibility of distance learning and testing students' knowledge without needing physical presence [13]. At the same time, NLP allows for deep linguistic analysis, including morphological, syntactic and semantic analysis of texts [14], which allows not only checking grammatical correctness but also to assessing the content compliance of tasks with STEM educational materials [15].

The research [16] considers the use of learning progressions (LP), performance assessment and artificial intelligence to improve STEM education. Particular attention is paid to how LP can contribute

to the effective development of curricula and assessment systems that help students understand scientific concepts more deeply and apply them in practice. The problem is the high cost and complexity of assessing such tasks. Therefore, to improve this process, the use of AI is proposed, particularly machine learning methods, such as unsupervised learning and semi-supervised learning for initial validation of LP, as well as supervised machine learning for more accurate diagnostics at more mature stages. In addition, the use of generative artificial intelligence is proposed to develop tasks that will correspond to LP, as well as automatic feedback systems for personalized learning and teacher support.

In [17] the functional structure of an intelligent system for linguistic analysis of text responses is developed and tested using artificial intelligence models. An algorithm for fuzzy semantic comparison of text information – students' answers to questions in natural language with correct answer options is presented, which formalizes the description of the linguistic structure of educational content and answers. The algorithm automatically converts student answers from natural language into intersystem form, the formation of lexical units of the text, after which morphological, syntactic, semantic and pragmatic analysis are implemented. At semantic and pragmatic analysis stages, artificial intelligence models are used to compare text information. A semantic network is created as a result of semantic analysis – a structure for representing knowledge in the form of nodes connected by arcs. Pragmatic analysis determines whether the response belongs to a specific subject area. The proposed stages are implemented using neural networks, a universal tool that can adapt to comparing texts from different subject areas. Unlike the well-known methods of semantic and pragmatic analysis, algorithms based on artificial intelligence models provide more significant opportunities for checking automated text responses in the form of free text in natural language with greater confidence.

The research [18] presents the method for automated assessment of the complexity of test items for social science tests. In particular, the complexity of multiple-choice test items consisting of a question and alternative answer options is investigated. For this purpose, a method for creating a semantic space using word embedding technologies is used. The texts of the task elements are projected into this space to obtain the corresponding vectors. The semantic characteristics of the tasks are determined by calculating the cosine similarity between the vectors of the task elements. The obtained semantic features are transferred to the classifier for training and testing. Based on the classification results, a model for assessing the complexity of the task is created. The results show that the semantic similarity between the main part of the question and the answer options has the greatest impact on the complexity of the task. In addition, the proposed method demonstrates an advantage over the traditional method of pre-testing, which allows us to hope for its further application in the future as an alternative or addition to previous methods of assessing complexity.

In [19] discusses methods for automated prediction of question difficulty for pedagogical tests that measure different skills of students at different levels. Predicting question difficulty is an important aspect of creating test items, as it allows for a qualitative and objective assessment of students' knowledge. Traditional methods of difficulty assessment, such as expert assessments and pre-testing, are criticized for their high cost, time-consuming and subjectivity. Therefore, more and more attention is paid to automated approaches, particularly based on text analysis, to assess the difficulty of tasks. An overview of scientific research in this field is provided, the use of automatic text-based difficulty prediction models is described, and the role of linguistic characteristics, such as syntactic and semantic features, in determining the difficulty of tasks is analyzed. The authors also note the need for publicly available standardized datasets and further research into alternative difficulty prediction models.

The research [20] addresses the challenges of managing large repositories of multiple-choice questions (MCQs), which are often used in educational assessments and professional certification exams. One of the challenges is the presence of questions that duplicate concepts with different wording, making it difficult to detect such duplications using syntax checks while not adding value to the repository. The authors propose a workflow for detecting and managing potential duplicate questions in large repositories of MCQs. The process includes three main steps: preprocessing the questions, calculating similarities between the questions, and graph analysis of the resulting similarity values. Three strategies are used for preprocessing: removing answer options, adding the correct answer to each question, or adding all answer options. Similarities between questions are calculated using deep learning-based

natural language processing (NLP) techniques, particularly the Transformers architecture. Finally, a new approach to community-based graph analysis is proposed to explore similarities and relationships between questions. The article illustrates the approach using the example of the "Competenze Digitali" program, a large-scale assessment project initiated by the Italian government.

Thus, from the analysis of current scientific achievements, there is a lack of automated methods capable of performing deep linguistic analysis at semantic level, which creates a gap in existing approaches to assessment. Therefore, the development of method automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP is highly relevant. Such approach will not only increase the accuracy and efficiency of evaluation but also take into account the context and content semantic connections in educational materials, which are key for STEM-disciplines.

3. Method design

Method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP is designed to automatically check the extent to which test tasks cover key terms and concepts contained in educational materials, thereby ensuring their compliance with educational objectives and improving the quality of assessing student knowledge. This method works with test tasks that support the structure shown in figure 1.

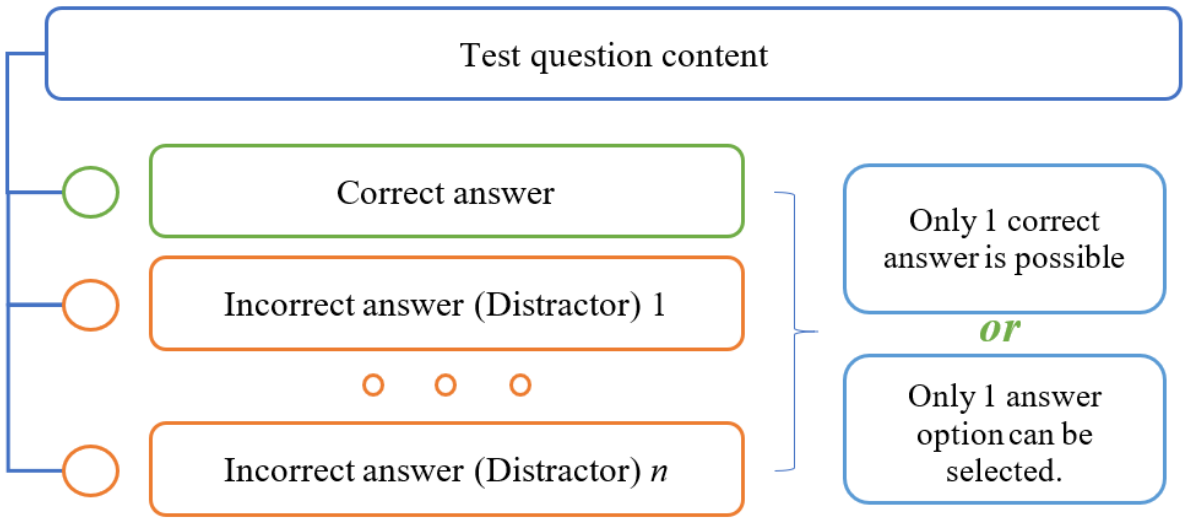

Figure 1: Typical test task scheme for analysis using the proposed approach.

Accordingly, each test task contains the question text, one correct answer and one or more incorrect answers (distractors) [21]. In this case, it is possible to choose only one answer option.

The method scheme for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP is shown in figure 2.

The input data of method are educational materials and test tasks for educational materials.

Stage 1 is responsible for pre-processing of text input data. Pre-processing includes removing punctuation marks and converting to lowercase [22]. Pre-processing is carried out for both test materials and educational materials.

Stage 2 recognition of key terms. In tests, all terms are considered key, and lists of words without repetitions are formed for each topic separately and in general.

For educational material, key terms recognition is carried out in several stages:

- recognition of named entities;
- recognition of general key terms for the document;

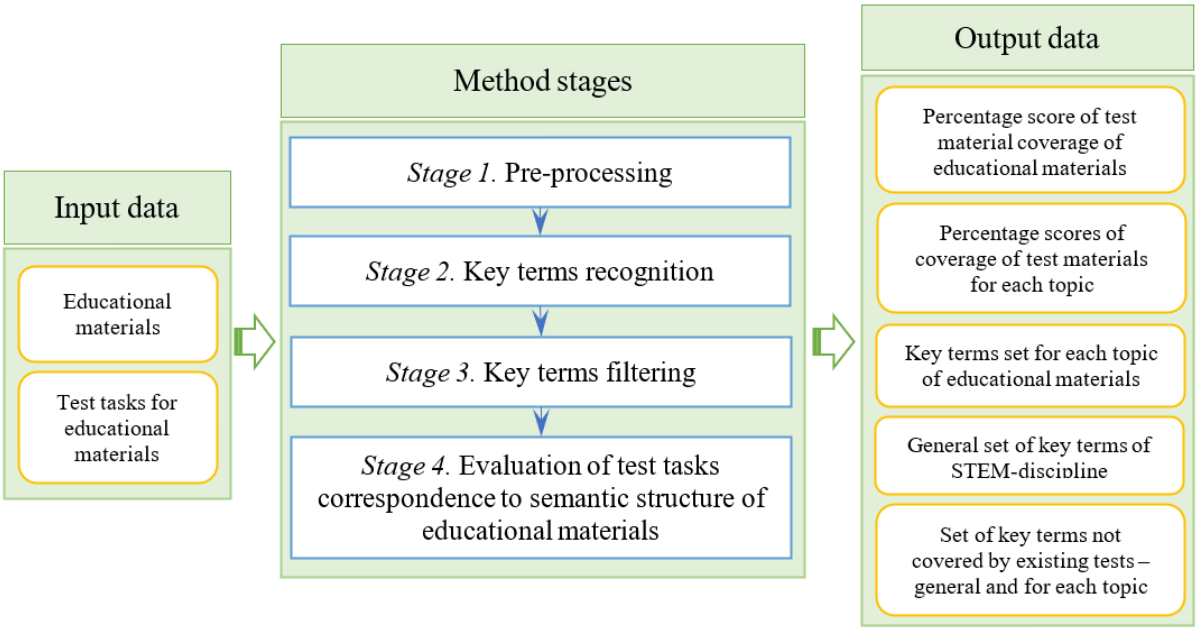

Figure 2: Scheme of method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials.

- recognition of key terms taking into account the context of the educational materials.

Within the framework of research, the neural network library "Stanza" [23] will be used to search for named entities, the TF-IDF [24] of the "Sklearn" [25] library will be used to search for general key terms in the document, and "KeyBERT" [26] will be used to search for key terms taking into account the context of educational materials from the library of the same name "KeyBERT".

Based on key terms of educational materials found separately for each topic and the general one, a list without repetitions is formed for each topic and the general one. The scheme of stage 3 is shown in figure 3.

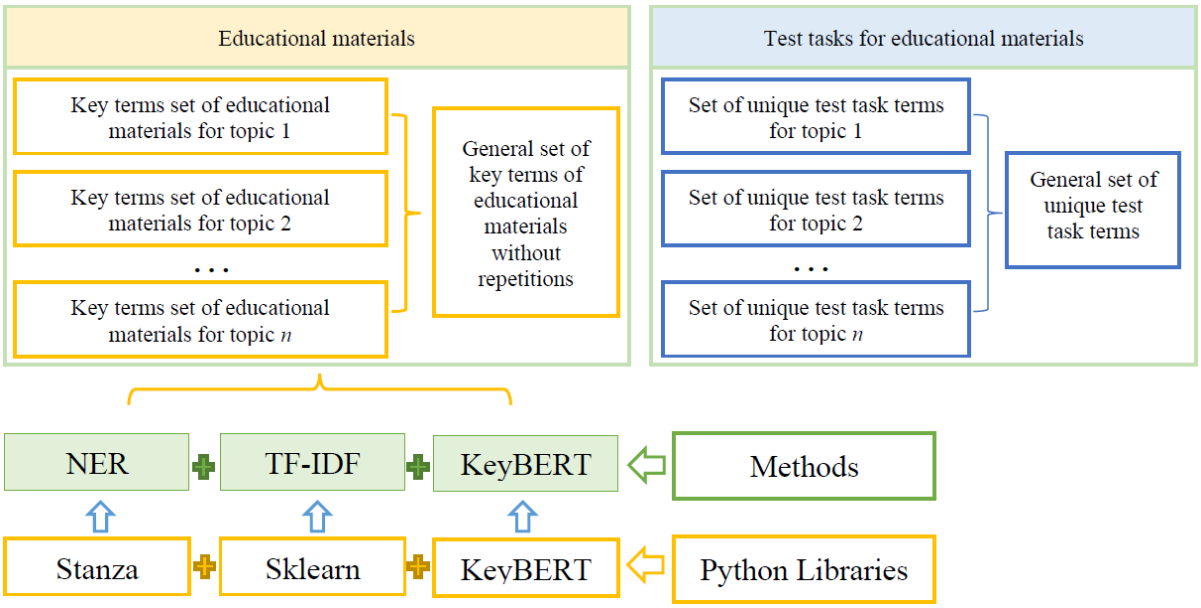

Figure 3: Illustration of stage 2 implementation of the method for key terms recognition.

Stage 3 is responsible for key terms filtering for educational materials and test tasks. Filtering occurs in several stages:

- lemmatization of key terms set of educational and test materials by topic and general entities;
- filtering of received lemmatized sets through the common words blacklist;
- duplicates removal.

In stage 4, the correspondence of test tasks to the semantic component of educational materials is evaluated. The evaluation is calculated as the ratio of key term numbers for each topic separately and for the entire STEM-discipline.

The output data of method: percentage score of test material coverage of educational materials, percentage scores of coverage of test materials for each topic, key terms set for each topic of educational materials, general set of key terms of STEM-discipline and the set of key terms not covered by existing tests – general and for each topic.

The use of different approaches to key terms recognition allows for comprehensive consideration of various aspects of educational materials, from basic named entities to contextually important concepts [27]. Thus, implementing the method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP not only improves the content relevance of tests but also increases the overall efficiency of the knowledge assessment process. Furthermore, this approach can be adapted to different disciplines and languages, which expands its potential for use in international educational environments. It also contributes to generating more accurate feedback for educational materials developers, which stimulates the improvement of their structure and content.

4. Experiment

Software in the form of web application was created to test the method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP. The software development was carried out using the Flask microframework [28], the Python programming language [29], the Stanza, Sklearn, and KeyBERT libraries.

The appearance of the experimental web application for studying the developed method is shown in figure 4.

In the case shown figure 4, an analysis of tests for the STEM discipline "Algorithmization and Programming" was performed for one of 17 topics, each of which provides tests to confirm the knowledge level of educational material. In accordance with the rubrication system of educational material, the method is similarly used for each of educational material topics separately.

In total, the research was conducted on 8 Ukrainian disciplines that integrate knowledge of mathematics, computer science, and engineering and belong to STEM-disciplines. The total content of these disciplines is 129 topics and 129 tests corresponding to these topics. The purpose of the experiment is to check the degree of correspondence between test tasks and educational materials based on key terms obtained from semantic analysis, as well as to compare the results obtained automatically with the results obtained by experts.

5. Results and discussion

The PCA dimensionality reduction method [30] was used to visualise the results, which allowed to project multidimensional semantic vectors into a two-dimensional space. This provided the opportunity to identify the structural distribution of terms in the context of individual topics and the discipline as whole. In the resulting graph (figure 5), each point represented a term, and its position in two-dimensional space was determined based on semantic proximity to other terms. Different topics were displayed as clusters, coloured in the corresponding colours.

Based on the presented graph, the visualization results demonstrate the terms distribution in two-dimensional space, which is the result of implementing the method for automated evaluation of test tasks correspondence to semantic structure of STEM-disciplines educational materials using NLP.

The graph confirms that the method provides systematization and clustering of terms by topics, reflecting the degree of coverage of key terms by tests and educational materials. Different topics have

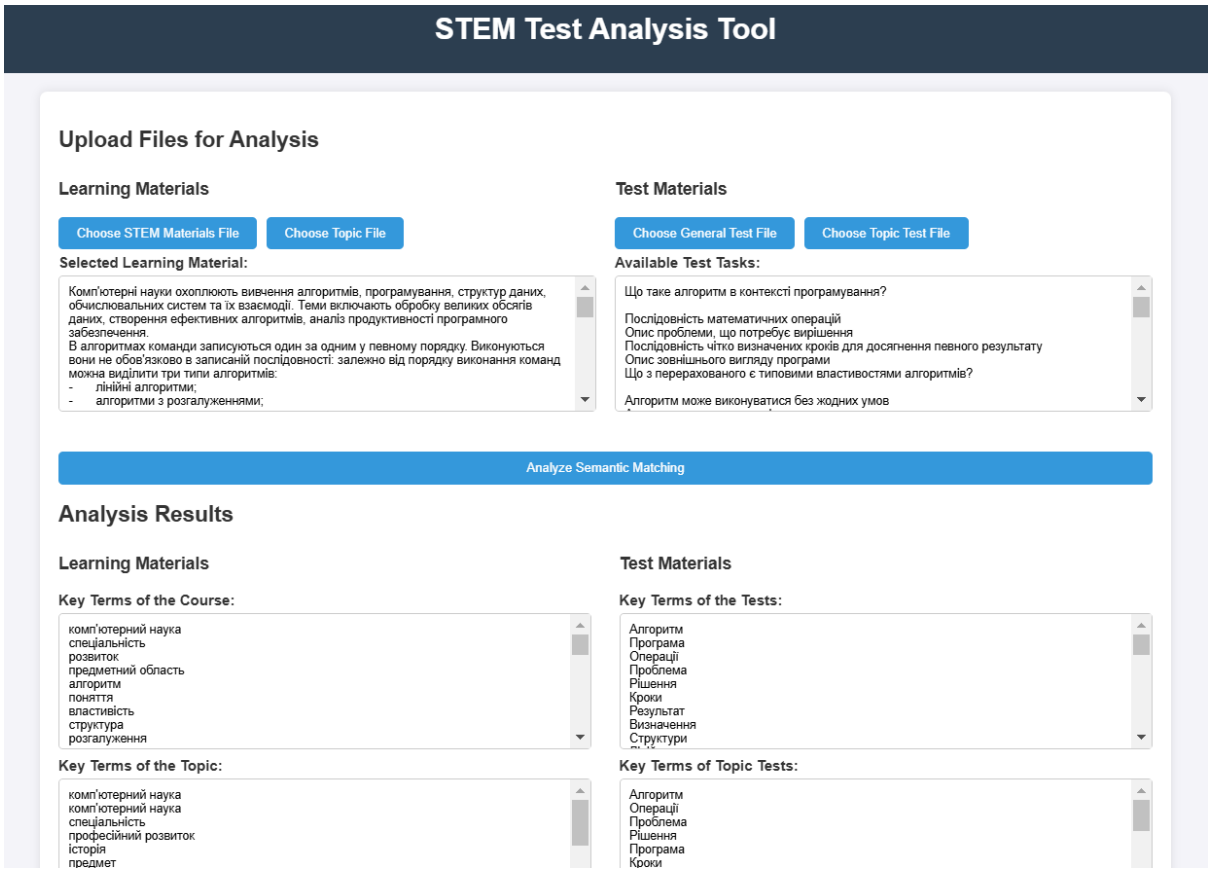

Figure 4: Developed software for analyzing the compliance of test tasks with the semantic structure of STEM-disciplines educational materials.

different levels of cluster density: this indicates heterogeneity of coverage and semantic variability of terms within the educational material. More compact clusters indicate consistency between test tasks and educational materials for specific topics, while more scattered groups reveal possible shortcomings in compliance.

The systematic nature of the method is confirmed by the possibility of assessing both the overall coverage of educational materials by tests and the coverage for each topic separately. The applied NLP approaches, including the use of named entities, TF-IDF and contextually-oriented KeyBERT models, allow for multidimensional analysis of key terms, taking into account both their overall significance and specific context.

The percentage of example coverage of educational materials semantic structure by test tasks is given in table 1. The average coverage rate obtained is 84.8%, indicating a high correspondence of test tasks to educational material. The distribution of coverage by topic shows little variability, which emphasizes the balanced development of tests for different topics.

Figure 5 and table 1 demonstrate a high correlation between the spatial distribution of points and the percentage coverage of educational materials by test tasks.

In the graph, topics with a high percentage of test tasks coverage (e.g., Topic 1 with 90.1% or Topic 13 with 87.4%) are located closer to the cluster centers or occupy higher positions on the axis of coverage by educational materials. Topics with lower coverage (e.g., Topic 4 with 80.3% or Topic 12 with 81.1%) are distributed on the periphery of the clusters or closer to the bottom of the graph.

This distribution confirms that the percentage coverage in the table corresponds to the visual clustering in the graph. Points belonging to topics with similar percentages of coverage also form geographically close clusters, which indicates the consistency of the automated assessment results. Thus, the graph successfully reflects the content of the table, emphasizing the connection between semantic coverage by test tasks and the distribution by topics.

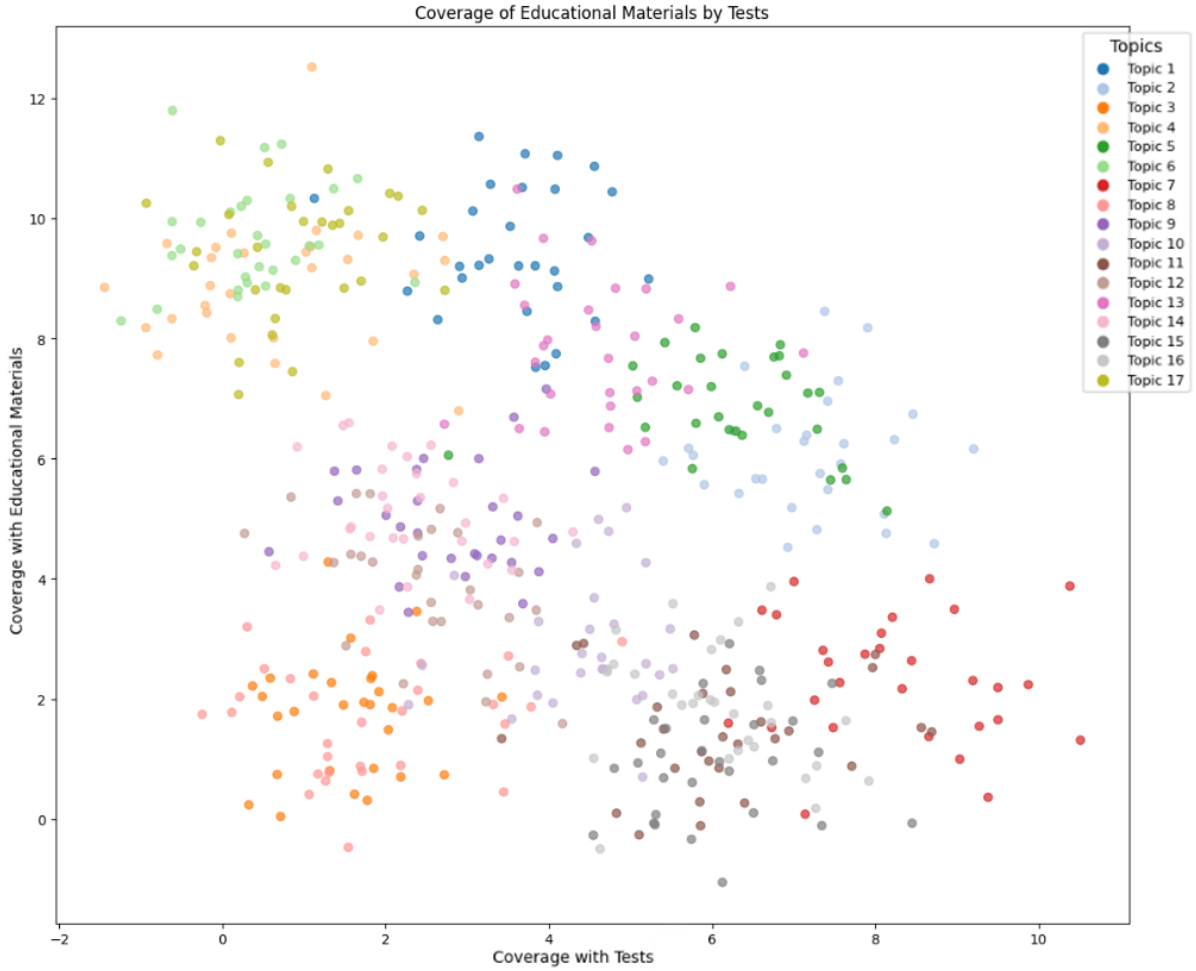

Figure 5: Visual analytics example for covering educational material with test tasks.

Table 1
Results of evaluation of the coverage of educational materials' semantic structure by test tasks.

Topic	Coverage, %	Topic	Coverage, %
Topic 1	90.1	Topic 10	86.1
Topic 2	85.4	Topic 11	83.2
Topic 3	87.7	Topic 12	81.1
Topic 4	80.3	Topic 13	87.4
Topic 5	84.2	Topic 14	85.3
Topic 6	87.2	Topic 15	83.7
Topic 7	82.7	Topic 16	82.6
Topic 8	83.1	Topic 17	86.2
Topic 9	84.9		

The task of evaluation of test tasks correspondence to semantic structure of STEM-disciplines is subject-oriented; therefore, experts were involved for its verification. For each STEM-discipline, the experts were the author of corresponding educational course and three teachers of other STEM-disciplines, which ensured the averaged solutions objectivity. The task of each expert was to assess the degree of correspondence between test tasks and each of topics of each educational material, for further comparison of these estimates with the estimates obtained automatically through the applied use of developed method. As result of comparing the averaged estimates of experts with the estimates obtained automatically, it was found that the developed method allows evaluating the correspondence

of test tasks to the semantic structure of educational materials in STEM-disciplines with average accuracy of 94.6%. Within individual topics of STEM-disciplines, the minimum accuracy was 71.8%, the maximum accuracy was 97.4%; within individual STEM-disciplines, the minimum accuracy was 85.5%, the maximum accuracy was 96.1%.

The proposed method's limitation is that it is designed to work with test tasks that contain text content. At the same time, the method provides content analysis of test tasks of various types: logical choice, choosing one correct answer from several, choosing several correct answers, establishing correspondence, determining the correct sequence, and short answer input.

6. Conclusion

A method for automated evaluation of test tasks that correspond to semantic structure of STEM-discipline educational materials using NLP was proposed. The use of various approaches to selecting key terms, such as recognition for named entities, context analysis, and use of lemmatization methods, allows for comprehensive consideration of various aspects of educational materials, which ensures high level of correspondence of tests to educational goals. This method not only improves the content correspondence of tests, but also significantly reduces the resources required to check test tasks, making this process more automated and convenient for use in educational systems. In addition, the method's adaptability to different disciplines and languages opens up opportunities for its use in international educational environments. This not only contributes to improving the testing quality, but also stimulates the improvement of structure and content of educational materials, which, in turn, increases the overall quality of education.

The research contributes to achievement of several Sustainable Development Goals set by the United Nations, in particular SDG No. 4 – ensuring comprehensive, inclusive and equitable education, as well as SDG No. 9, SDG No. 10 and SDG No. 12. The use of modern NLP technologies in the field of education allows creating intelligent systems that adapt the educational process to the needs of different groups of students, reduce inequality in education, and contribute to the development of sustainable infrastructure and innovations in educational. Automation of test tasks scoring not only improves the quality of education but also allows reducing resource consumption, introducing more environmentally sustainable approach to the creation and use of educational materials. The results of the conducted experimental research of developed method for automated evaluation of test tasks correspondence to semantic structure of educational materials of number of STEM-disciplines showed the correlation of the results of method and conducted cluster analysis with visual interpretation. The created method has limitations, educational materials should not have abbreviations and abbreviations, and currently works with test tasks of logical type and single choice.

As result of comparing the averaged estimates of experts with the estimates obtained automatically, it was found that the developed method allows evaluating the correspondence of test tasks to the semantic structure of educational materials in STEM-disciplines with average accuracy of 94.6%; within individual topics of STEM-disciplines the minimum accuracy was 71.8%, and the maximum accuracy was 97.4%.

Further research will be aimed at ensuring the analysis of test tasks of other types, as well as at taking into account generalizations of noun-named entities for correct processing of synonymous constructions, abbreviations and reductions. Promising direction for further research is the semantic analysis of answers texts entered by students in response to test tasks, such as essays, theses and abstracts. The separate direction of continuing research is the automated construction of semantic trees of test tasks for adaptive testing scenarios. It is also advisable to expand the parameters of analysis of correspondence of test tasks to semantic component of educational materials in STEM-disciplines, in particular by determining the completeness of coverage of educational material content by test tasks and determining the coverage of educational material content by test tasks of different types.

Author Contributions

Conceptualization – Olexander Barmak; methodology – Olexander Mazurets; formulation of tasks analysis – Maryna Molchanova and Iurii Krak; software – Maryna Molchanova and Olexander Mazurets; writing – original draft – Maryna Molchanova and Olena Sobko; analysis of results – Olexander Mazurets and Maryna Molchanova; visualization – Iurii Krak and Olena Sobko; reviewing and editing – Daryna Hardysh and Olexander Mazurets. All authors have read and agreed to the published version of the manuscript.

Funding

This study did not receive any funding.

Data Availability Statement

No new data were created or analysed during this study. Data sharing is not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

The research was verified and evaluated in actual conditions with the help of the Faculty of Physics and Mathematics of the Ternopil Volodymyr Hnatiuk National Pedagogical University. Thanks to the university's support, the authors' team had access to the necessary software, which significantly increased research efficiency. The authors express their gratitude to Sehriy Semerikov and Tetiana Vakaliuk to scientific and organizational support of 4th Yurii Ramskyi STE(A)M Workshop co-located with XVII International Conference on Mathematics, Science and Technology Education.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] N. Zaki, S. Turaev, K. Shuaib, A. Krishnan, E. Mohamed, Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation, Education and Information Technologies 28 (2023) 16723–16742. doi:10.1007/s10639-023-11877-4.

[2] L. K. Kalyani, The role of technology in education: Enhancing learning outcomes and 21st century skills, International journal of scientific research in modern science and technology 3 (2024) 05–10.

[3] C. W. Tan, K. Y. Lim, Revolutionizing Formative Assessment in STEM Fields: Leveraging AI and NLP Techniques, in: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2023, pp. 1357–1364. doi:10.1109/APSIPAASC58517.2023.10317226.

[4] T. Kubiszyn, G. D. Borich, Educational testing and measurement, John Wiley & Sons, 2024.

[5] P. Basu, S. S. Mohanty, Developing Multilingual Glossaries for STEM Terminology Using AI-NLP, in: Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages, Springer, 2024, pp. 115–122. doi:doi.org/10.1007/978-981-97-1987-7_9.

[6] A. Singh, A. Kanaujia, V. K. Singh, R. Vinuesa, Artificial intelligence for Sustainable Development Goals: Bibliometric patterns and concept evolution trajectories, Sustainable Development 32 (2024) 724–754. doi:doi.org/10.1002/sd.2706.

[7] M. van Geel, T. Keuning, K. Meutstege, J. de Vries, A. Visscher, C. Wolterinck, K. Schildkamp, C. Poortman, Adapting Teaching to Students' Needs: What Does It Require from Teachers?, in: Effective Teaching Around the World: Theoretical, Empirical, Methodological and Practical Insights, Springer International Publishing Cham, 2023, pp. 723–736.

[8] F. M. Reimers, The sustainable development goals and education, achievements and opportunities, International Journal of Educational Development 104 (2024) 102965. doi:doi.org/10.1016/j.ijedudev.2023.102965.

[9] H. A. Younis, N. I. R. Ruhaiyem, W. Ghaban, N. A. Gazem, M. Nasser, A systematic literature review on the applications of robots and natural language processing in education, Electronics 12 (2023) 2864.

[10] M. Ouahi, S. Khoulji, M. L. Kerkeb, Analysis of Deep Learning Development Platforms and Their Applications in Sustainable Development within the Education Sector, in: E3S Web of Conferences, volume 477, EDP Sciences, 2024, p. 00098. doi:doi.org/10.1051/e3sconf/202447700098.

[11] S. Uda, B. Basrowi, Environmental education using SARITHA-Apps to enhance environmentally friendly supply chain efficiency and foster environmental knowledge towards sustainability, Uncertain Supply Chain Management 12 (2024) 359–372. doi:10.5267/j.uscm.2023.9.015.

[12] H. M. Ahmed, S. E. Sorour, Classification-driven intelligent system for automated evaluation of higher education exam paper quality, Education and Information Technologies (2024) 1–27. doi:doi.org/10.1007/s10639-024-12555-9.

[13] F. M. V. Falcão, D. S. Pereira, J. M. Pêgo, P. Costa, Progress is impossible without change: implementing automatic item generation in medical knowledge progress testing, Education and Information Technologies 29 (2024) 4505–4530. doi:doi.org/10.1007/s10639-023-12014-x.

[14] Y. V. Krak, O. Barmak, O. Mazurets, The practice investigation of the information technology efficiency for automated definition of terms in the semantic content of educational materials, Problems in programming (2016) 237–245. doi:doi.org/10.15407/pp2016.02-03.237.

[15] O. V. Barmak, O. V. Mazurets, I. V. Krak, A. I. Kulias, L. E. Azarova, K. Gromaszek, S. Smailova, Information technology for creation of semantic structure of educational materials, in: Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, volume 11176, SPIE, 2019, pp. 616–626. doi:doi.org/10.1117/12.2537064.

[16] L. Kaldaras, K. Haudek, J. Krajcik, Employing automatic analysis tools aligned to learning progressions to assess knowledge application and support learning in STEM, International Journal of STEM Education 11 (2024) 57. doi:doi.org/10.1186/s40594-024-00516-0.

[17] I. Katerynchuk, O. Komarnytska, A. Balendr, The Use of Artificial Intelligence Models in the Automated Knowledge Assessment System, in: IEEE lnternational Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, Springer, 2024, pp. 274–288. doi:doi.org/10.1007/978-3-031-61221-3_13.

[18] F.-Y. Hsu, H.-M. Lee, T.-H. Chang, Y.-T. Sung, Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques, Information Processing & Management 54 (2018) 969–984. doi:doi.org/10.1016/j.ipm.2018.06.007.

[19] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education (2023) 1–53. doi:doi.org/10.1007/s40593-023-00362-1.

[20] V. Albano, D. Firmani, L. Laura, J. G. Mathew, A. L. Paoletti, I. Torrente, NLP-Based Management of Large Multiple-Choice Test Item Repositories., Journal of Learning Analytics 10 (2023) 28–44.

[21] A. A. Rezigalla, A. M. E. S. A. Eleragi, A. B. Elhussein, J. Alfaifi, M. A. ALGhamdi, A. Y. Al Ameer, A. I. O. Yahia, O. A. Mohammed, M. I. E. Adam, Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items, BMC Medical Education 24 (2024) 445. doi:doi.org/10.1186/s12909-024-05433-y.

[22] C. P. Chai, Comparison of text preprocessing methods, Natural Language Engineering 29 (2023)

509–553. doi:doi:10.1017/S1351324922000213.
[23] Stanza, Stanza – A Python NLP Package for Many Human Languages , https://stanfordnlp.github.io/stanza/, 2025. Accessed: 17 January 2025.
[24] S. Jain, S. K. Jain, S. Vasal, An Effective TF-IDF Model to Improve the Text Classification Performance, in: 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2024, pp. 1–4. doi:doi.org/10.48550/arXiv.2308.04037.
[25] scikit learn, scikit-learn Machine Learning in Python, https://scikit-learn.org/stable/, 2025. Accessed: 17 January 2025.
[26] KeyBERT, A minimal method for keyword extraction with BERT , https://maartengr.github.io/KeyBERT/api/keybert.html, 2025. Accessed: 17 January 2025.
[27] O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. V. Barmak, I. V. Krak, Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network, in: International Conference on Computational Linguistics and Intelligent Systems, 2023. URL: https://api.semanticscholar.org/CorpusID:258688336.
[28] Flask, Welcome to Flask — Flask Documentation (3.1.x), https://flask.palletsprojects.com/en/stable/, 2025. Accessed: 17 January 2025.
[29] Python, Welcome to Python.org , https://www.python.org//, 2025. Accessed: 17 January 2025.
[30] J. P. Bharadiya, A tutorial on principal component analysis for dimensionality reduction in machine learning, International Journal of Innovative Science and Research Technology 8 (2023) 2028–2032.