# AUTOMATIC SCORING OF SHORT-TEXT ANSWERS USING NATURAL LANGUAGE PROCESSING (NLP)

**Turg‘unova Dilnoza Mamirjonovna**
Sharda University First-year Master of Technology student
DOI:https://doi.org/10.5281/zenodo.24096525

**Abstract.** This article examines the possibilities of automatically scoring short-text answers provided by school pupils and university students in the educational process through the use of Natural Language Processing (NLP) technologies. Traditional assessment methods involve several challenges, including the considerable amount of time required from teachers, the emergence of subjectivity during the assessment process, and variations in the way answers are expressed. NLP technologies make it possible to reduce these challenges, identify the semantic relevance of answers, and accelerate the assessment process. The article analyses the mechanisms of text preprocessing, semantic similarity detection, the BERT architecture, and the use of large language models.

**Keywords:** natural language processing, NLP, automatic scoring, short-text answers, artificial intelligence, machine learning, BERT, semantic analysis.

The development of digital technologies is bringing significant changes to methods of assessing knowledge within the education system. In particular, as the use of e-learning platforms, distance-learning systems, and digital examination tools continues to expand, the issue of assessing learners’ knowledge promptly, consistently, and as objectively as possible is becoming increasingly important.

In traditional testing systems, it is relatively easy to check answers against predetermined options. However, assessing responses to open-ended questions is different. This is because the same idea can be expressed through different words, sentences, and grammatical structures. For example, one student may explain economic growth as an increase in the volume of production, while another may define it as an increase in the value of goods and services produced in a country. Checking these semantically similar answers using simple keywords does not always produce accurate results. For this reason, the use of natural language processing technologies is particularly important. NLP enables computers to process texts in human language, analyse their meaning, identify relationships between words, and classify texts according to specific criteria.

In recent years, the development of transformer architectures such as BERT and large language models has further expanded the possibilities for automatically scoring short-text answers. However, the effectiveness of an automated scoring system depends not only on the model’s technical capabilities but also on the nature of the question, the assessment criteria, and the quality of the training data.

Automatic Short Answer Scoring (ASAS) is the process of analysing a response written by a pupil or student using computer algorithms and assigning it a score according to predetermined criteria. This process is not limited to checking the words contained in the answer. Rather, the system attempts to determine the answer’s relevance to the question, the presence of key concepts, its semantic completeness, and its semantic similarity to the reference answer.

Early automatic scoring systems used methods such as keyword searching, word counting, and classifying answers according to predefined rules. Although these approaches are simple and


© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

computationally convenient, they do not fully take into account the complex characteristics of natural language. For example, consider the following two answers:

Answer 1: Economic growth is an increase in the volume of goods and services produced in a country.

Answer 2: Economic growth refers to an increase in real gross domestic product over time.

Both answers express the concept of economic growth in terms of their meaning. However, their vocabulary and sentence structures differ. A system based on simple keywords may fail to establish a sufficient connection between the second answer and the first. An NLP-based system, by contrast, can identify the meaning of words in context and the semantic relationships between them.

In their article, Bexte, Horbach, and Zesch discuss approaches to the automatic scoring of short-text answers, including rule-based methods, basic machine learning, deep learning, and generative models. The authors emphasise that each approach has its own strengths and weaknesses. Therefore, when developing an automatic scoring system, it is important to consider the characteristics of the answers being assessed rather than simply selecting a single algorithm.

Natural Language Processing technology is used to convert text into a form that can be processed and analysed by a computer. When scoring short answers, this process consists of several consecutive stages. The first stage is text preprocessing. The response entered by a student is initially processed at a technical level. During this stage, extra spaces, unnecessary characters, and certain elements that may interfere with text analysis are standardised. Where necessary, the text is converted to lowercase and divided into sentences and words. When working with Uzbek-language texts, it is necessary to take into account the extensive use of suffixes, case forms, and word-formation features. For example, the words “iqtisodiyot”, “iqtisodiy”, and “iqtisodiyotning” have closely related meanings and grammatical relationships, although their written forms differ. The second stage is converting the text into a digital representation. A computer does not process text directly in the same way as a human does. Therefore, words and sentences are represented as mathematical vectors. This process is called text vectorisation. Traditional approaches use the Bag of Words and TF-IDF methods. TF-IDF helps assess how important a word is to a particular document. However, these methods cannot fully represent the contextual meaning of words. For this reason, modern systems use models that represent words and sentences as context-dependent vectors. The third stage is identifying semantic similarity. This is one of the most important parts of automatic scoring. The system determines the degree of semantic similarity between the student’s answer and the reference answer. Cosine similarity is calculated using the following formula:

**Cosine Similarity = (A · B) / (||A|| × ||B||)**

Here, A is the vector representation of the reference answer; B is the vector representation of the student’s answer; A · B is the dot product of the two vectors; and ||A|| and ||B|| are the lengths of the vectors. This measure is used to determine how close the directions of two texts are in vector space. The result is generally within the range of -1 to +1. In practical text embeddings, the interpretation of the measure depends on the selected model and data. Importantly, a high level of semantic similarity does not always mean that an answer is entirely correct. Therefore, semantic similarity should be used together with other assessment criteria.

The fourth stage is answer assessment and score assignment. The system determines the quality of the answer based on predefined assessment criteria. For example, it takes into account the presence of the main concept, the inclusion of necessary explanations, and whether the response directly addresses the question. As a result, the system assigns a score to the answer and, where necessary, generates explanatory feedback.

Several technological approaches are currently used in NLP-based automatic scoring. Their capabilities and areas of application differ from one another.


© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

In the rule-based scoring method, answers are checked using predefined keywords, phrases, and rules. For example, when answering the question "What is inflation?", the phrase "an increase in the general price level" may be defined as a key criterion. If a student's answer contains expressions corresponding to this concept, the system may assess the answer as correct. The main advantage of this method is that it is simple and fast. Its disadvantage is the limited ability to recognise the same idea when it is expressed using different words.

In machine learning methods, the system is trained using previously assessed answers. The model learns various characteristics of the answers and determines which scoring category a newly submitted text belongs to. Support Vector Machine (SVM), Naive Bayes, and other classification algorithms can be used for this purpose. The effectiveness of this approach depends on the size and quality of the training data and the extent to which it covers different types of answers.

BERT (Bidirectional Encoder Representations from Transformers) is a language model introduced by Google researchers in 2018 and described in detail in a scholarly article in 2019. In their 2019 article, Devlin and co-authors presented BERT as a bidirectional language model that takes into account both the left and right contexts of a text. An important feature of the model is its ability to represent the meaning of words based on the context in which they are used. In short-answer scoring, BERT can identify the semantic relationship between a question and an answer, compare semantically similar answers expressed using different words, identify key concepts in an answer, and assign a score using a trained classification model. A system developed on the basis of BERT can take into account not only exact word-level matches with the reference answer but also the contextual characteristics of texts. At the same time, the model needs to be adapted to a particular subject or language. For example, scoring answers in economics requires a dataset covering economic terms, their definitions, and the relationships between them.

Large language models have created new opportunities for scoring short-text answers. Such models can analyse the content of an answer, take assessment criteria into account, and generate explanatory feedback for students.

In an applied study conducted by Jiang and Bosch, the GPT-4-1106-preview model was tested for scoring student responses in the ASAP-SAS dataset. The authors reported a Quadratic Weighted Kappa score of 0.677 for the best-tuned configuration across 10 questions. The results also demonstrated the model's sensitivity to the assessment criteria and the sample answers provided. Although this result indicates that large language models may be useful in automatic scoring, it does not mean that they perform with the same level of accuracy across all subjects and questions.

When determining the effectiveness of automatic scoring technologies, it is important to establish the extent to which their results correspond with those of human assessors. An article published by Clauser and co-authors in 2024 examined an automated short-answer scoring system using more than 35,000 responses and 71 tasks. In the system used by the authors, agreement with human scores was reported to be in the range of 0.97–0.99 for aggregated results. However, this finding cannot be generalised to all NLP systems or all academic subjects. It is a result obtained under the conditions of a particular scoring system, set of tasks, and dataset.

Bexte, Horbach, and Zesch also analysed the strengths and weaknesses of automatic scoring models and demonstrated the need to consider the linguistic characteristics of the answers being assessed and the purpose of assessment when selecting a system.

The quality of automatic scoring can be determined using the following measures:

- ✓ Accuracy – the proportion of answers correctly classified by the model.


© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

✓ Precision – indicates how many of the answers assigned by the model to a particular category actually belong to that category.

✓ Recall – indicates how many of the answers belonging to a relevant category were identified by the model.

✓ F1-score – the harmonic mean of Precision and Recall.

✓ Quadratic Weighted Kappa (QWK) – used to assess the agreement between ordinal scores assigned by a human assessor and an automated system.

The QWK measure is particularly important when answers are divided into several score levels, such as 0, 1, 2, or 3 points. In addition, it is necessary to examine whether the automated system assigns the same score to the same answer consistently, how it responds to different forms of answers, and whether it can identify cases of incorrect scoring.

Although NLP technologies create important opportunities in the educational process, several challenges remain in their practical implementation.

The first challenge is the complexity of language. In Uzbek in particular, the use of different suffixes, synonyms, and flexible sentence structures may complicate the automatic analysis process. Therefore, dedicated datasets and models adapted to the linguistic characteristics of Uzbek are needed to score short answers in this language.

The second challenge is scoring errors. In some cases, even when a student expresses a correct idea, the model may fail to identify it adequately. Conversely, an answer that is incorrect in terms of meaning may receive a high score because it contains certain keywords or phrases.

The third challenge is dependence on data quality. If a model is trained mainly on answers written in the same format, it may make errors when scoring unusual, brief, or differently worded responses.

The fourth challenge concerns fairness and the protection of personal data. Students' answers and assessment results must be kept confidential, while algorithms must produce fair results for different groups.

The following areas are important for reducing these challenges in the future:

❖ developing an open and high-quality database of questions and answers in Uzbek;

❖ adapting BERT and other transformer models to the Uzbek language;

❖ developing separate assessment criteria for economics, mathematics, law, and other subjects;

❖ regularly comparing the results of human assessors and automated systems;

❖ providing clear explanations alongside assessment results;

❖ making final decisions under human supervision in examinations that involve high levels of responsibility.

These areas will help make NLP-based scoring systems more reliable and practically useful.

In conclusion, NLP-based automatic scoring systems should not be regarded as tools intended to replace teachers entirely, but rather as instruments that help accelerate the assessment process, reduce repetitive tasks, provide students with prompt feedback, and improve the quality of education.

**References:**

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American



© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186. https://doi.org/10.18653/v1/N19-1423

2. Bexte, M., Horbach, A., & Zesch, T. (2024). Strengths and weaknesses of automated scoring of free-text student answers. Informatik Spektrum, 47, 78–86. https://doi.org/10.1007/s00287-024-01573-z
3. Jiang, L., & Bosch, N. (2024). Short answer scoring with GPT-4. Proceedings of the Eleventh ACM Conference on Learning @ Scale, 438–442. https://doi.org/10.1145/3657604.3664685
4. Clauser, B. E., Yaneva, V., Baldwin, P., Ha, L. A., & Mee, J. (2024). Automated scoring of short-answer questions: A progress report. Applied Measurement in Education, 37(3), 209–224. https://doi.org/10.1080/08957347.2024.2386945
5. Ormerod, C., Lottridge, S., Harris, A. E., Patel, M., van Wamelen, P., Kodeswaran, B., Woolf, S., & Young, M. (2023). Automated short answer scoring using an ensemble of neural networks and latent semantic analysis classifiers. International Journal of Artificial Intelligence in Education, 33(3), 467–496. https://doi.org/10.1007/s40593-022-00294-2
6. Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. Educational Measurement: Issues and Practice, 42, 44–58. https://doi.org/10.1111/emip.12544



© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.