DOI: 10.12700/AIS.2025.040

Educational material quality evaluation using text analysis in Python

Olga Shvets
Faculty of Digital Technologies
Novosibirsk State University of Economy and Management
Novosibirsk, Russia
0000-0001-5710-9056

***Abstract*— The study aims to evaluate educational material using modern text processing and analysis methods. The work helps determine how easily students understand the text, its structure, the use of difficult-to-understand terms, and more. Analyzing the educational material facilitates decisions on the feasibility of publishing new student textbooks.**

Keywords*— *text quality evaluation, text processing, educational material analysis methods

I. Introduction

Digital transformation is gradually permeating virtually every aspect of our lives. Significant changes leading to the transition from third-generation industry to Industry 4.0 involve the use of advanced technologies such as the Internet of Things, artificial intelligence, big data, automation, and the digitalization of various processes.

II. Modern Condtions in Education Development

A. New Challenges in Education

Industry 4.0 and the digitalization of higher education offer numerous opportunities for the development and improvement of educational processes [1 - 4]. They stimulate changes in approaches to teaching, opening up new formats and resources for student learning. However, the key is the ability to adapt to a rapidly changing world and utilize advanced technologies to achieve better results. To harness this potential, universities are shifting their focus to student participation and independent research. These processes are constantly interacting with each other and with the environment. It creates new challenges, such as the development of relevant curricula that meet student needs and the demands of the modern labour market [5 - 9]. The digitalization of education also raises issues of data security and privacy, the need to develop information literacy, and copyright compliance.

In these conditions a huge amount of new methodological material in various formats is emerging. The problem of assessing the quality of methodological materials has long been relevant. There are well-known methods for determining the quality of such materials. These are primarily expert assessments that take into account the inclusion of topics in methodological materials that meet the requirements of the Federal State Educational Standard (FSES) and educational programs. Educational and methodological literature is created by people, and their views may be biased, perhaps focusing on only one scientific school, or may become outdated, etc.

Currently, there is no clear concept for determining the impact of the quality of methodological materials on the development of competencies and, consequently, on the quality of education especially in Russia.

So, education quality assessment also requires revision and a new approach, taking these challenges into account.

B. New Concept of Education

One of the main challenges in assessing the quality of higher education is the changing concept of education itself. Today, education is no longer limited to simply imparting knowledge and skills. It must shape individuals' personalities, develop critical thinking, and foster social and communication skills.

Tools such as cloud services, video conferencing, and digital libraries have become available, enabling remote learning and work, reducing the need for students and faculty to be present in a single physical location. Furthermore, digital technologies offer new learning formats based on interactivity, individualized approaches, and autonomy student work.

There are various concepts for assessing the quality of education, but there is currently no review of the relevant approaches that address the impact of the quality of methodological materials on the development of competencies and, accordingly, on the quality of education in the context of digital transformation.

III. Natural Language Processing

Analysing educational material requires the use of modern models and technologies. This study requires processing and analysing large volumes of text data, which requires the use of NLP. Numerous text generation tools, such as OpenAI's ChatGPT, rely on NLP. Natural language processing models consist of two components: training data and specialized algorithms [10, 11]. The process of analyzing teaching materials involves several stages:

A. The first stage

First, the data must be cleaned of unnecessary or duplicate information [12, 13, 14]. The more contaminated the data, the more difficult it will be for the model to understand what is important and what is not. Therefore, it is necessary to remove duplicates, convert strings to the same case, and remove unnecessary characters.

B. Tokenization

Next is a tokenization step. To enable the model to process text at the semantic level, the cleaned data was broken down into individual units—tokens [12, 13, 14]. In this study, tokens are represented by words. In the future, we plan to use other token types: symbols, phrases, and other elements, depending

on the task and context. An example of tokenization is dividing text into individual words:

"Education is important" → [Education, is, important]

Tokenization allows to transform text into a structured representation used for further analysis or processing. Russian and many other languages have suffixes that change the form of a word but not its meaning. To avoid confusing the program, words had to be converted to their dictionary form—a lemma—by performing lemmatization.

Stemming is also a similar process where the root of a word is extracted. Lemmatization and stemming improve the efficiency of text processing by reducing the number of unique tokens. The model will run faster, and less memory will be required to store words.

C. . Data labeling and creating a dataset

The next step is data labelling. Each document, text fragment, or word—that is, each token—must be assigned a label that describes the object in question [12, 13, 14]. The format and content of the label depend on the task at hand.

Another step is creating a dataset. Before training the model, the labelled data must be transformed into a dataset—that is, structured. The dataset is a table of tokens with their relevant features and labels. For text data in this study, it is stored in CSV format.

Word embedding is also possible for data presentation.

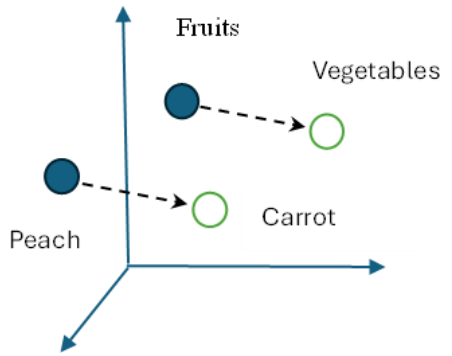

Fig. 1. Example of word embadding

D. Data Vectorization

Machine learning models can't handle text data; they only understand numbers. Therefore, data scientists need to translate tokens into a set of numerical values.

There are different approaches to this.

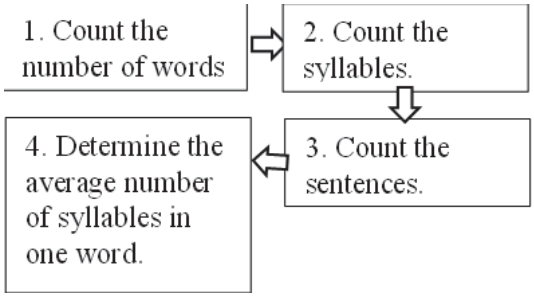

Fig. 2. Algorithm

The two most popular are bag-of-words and N-grams.

Bag-of-words simply encodes tokens into numbers, accounting for their quantity but ignoring context and specific order.

When using N-grams, words are encoded not one at a time, but two or three at a time. This preserves the structure of sentences and their context.

Notes.

Individual words, words with an apostrophe, abbreviations, numbers, symbols, and combinations of symbols are counted as units.

Count syllables in words, and count abbreviations, numbers, symbols and their combinations as monosyllabic words.

Each complete speech phrase separated by a period, colon, semicolon, question mark, exclamation mark, or dash is considered one sentence.

IV. Web Service Development

A web service based on large-scale language models (LLM) was developed for the automatic analysis of educational publications using artificial intelligence. The system executes comprehensive processing of text materials, including content complexity assessment, structural analysis, logical consistency checking, and identification of semantic relationships between thematic blocks.

The service allows to download educational materials in PDF format and receive a detailed content analysis.

A. Technologies:

- Backend: Quart (asynchronous version of Flask).
- Interface: HTML/CSS.
- ML methods: textstat, TF-IDF, BART, BERT.

B. Requirements:

- OS: Windows 10/11 (64-bit).
- RAM: 8 GB (16 GB recommended for large texts).
- Python: not less than 3.9.
- Video card: NVIDIA (preferred for BERT acceleration).
- Analysis file format: PDF.
- Maximum file size: up to 10 MB.
- Processing time for one file: approximately 1.5 - 2 minutes.

C. Web Service Exploitation

The web service exploitation is simple:

Download the PDF file of the textbook using the download button on the main page, → Click the "Check" button. → Wait 1.5 to 2 minutes for the service to process the file. → Once the analysis is complete, we will receive a report with the results.

The page with results is presented on Fig. 3.

The first block is readability in percentage.

Fig. 3. The page with results

The second is coherence (from the Latin cohaerens - "being in connection") – in linguistics, "the integrity of a text, consisting of the logical-semantic, grammatical and stylistic correlation and interdependence of its constituent elements (words, sentences, etc.)"; one of the defining characteristics of a text/discourse and one of the necessary conditions of textuality.

The third one checks whether the text structure complies with standards.

The fourth block allows to determine a text style.

There are the details of the text analysis in fifth block: Flash index, number of sentences, number of words, number of syllables and text complexity.

The service selects keywords automatically (sixth block).

After the analysis, a set of recommendations (seventh block) for improving the text of the educational material is formed.

CONCLUSION

With the growing volume of educational literature from Russian authors, the developed software reduces the workload of staff involved in selecting materials for students. This web service helps automate the process of analysing educational materials at the university. While existing text analysis tools exist in European languages, the development takes into account such specific features as working with Russian-language text, which differs in structure and word formation. This service is also integrated into the university's system for accepting educational materials for publication, making the web service relevant and in demand. Future plans include expanding the service's functionality, refining its algorithms, and incorporating fuzzy models to improve analysis.

REFERENCES

[1] A. R. Malik, Y. Pratiwi, K. Andajani, I. W. Numertayasa, S. Suharti, A. Darwis, "Exploring Artificial Intelligence in Academic Essay: Higher Education Student's Perspective", International Journal of Educational Research Open, vol. 5, 2023, (https://www.sciencedirect.com/science/article/pii/S2666374023000717)

[2] A. Monken, W. Ampeh, F. Haberkorn, U. Krishnaswamy, F. A. Batarseh, "Assuring AI methods for economic policymaking", Academic Press, 2023, pp. 371-427, (https://www.sciencedirect.com/science/article/pii/B9780323919197000251)

[3] C. Cacciuttolo., Y. Vásquez, D. Cano., F. Valenzuela, "Research Thesis for Undergraduate Engineering Programs in the Digitalization Era: Learning Strategies and Responsible Research Conduct Road to a University Education 4.0 Paradigm", Sustainability. 2023; 15(14):1120:10-35. (https://doi.org/10.3390/su151411206)

[4] Ch. Zhai, S. Wibowo, "A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university", Computers and Education: Artificial Intelligence, vol. 4, 2023, (https://www.sciencedirect.com/science/article/pii/S2666920X23000139)

[5] D. Hampus, "Graphic design, Already Intelligent? Current possibilities of generative AI applications in graphic design", Degree: Faculty of Technology and Society (TS), Malmö University, 2023, (http://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-62512)

[6] M. Gheisari, B. Becerik-Gerber, C.S. Dossick. "Emerging learning technologies for future of work and education in engineering [Internet]", Advanced Engineering Informatics. 2022; (https://www.sciencedirect.com/science/article/abs/pii/S1474034622002336.).

[7] R. Kohli, S. Johnson, "Digital Transformation in Latecomer Industries: CIO and CEO", Leadership Lessons from Encana Oil & Gas, USA, Inc, 17, 2011.

[8] A. Kukulska-Hulme, C. Bossu, K. Charitonos , T. Coughlan, A. Deacon, N. Deane, R. Ferguson, C. Herodotou, C-W. Huang, T. Mayisela, I. Rets, J. Sargent, E. Scanlon, J. Small, S. Walji, M. Weller, D. Whitelock, "Innovating Pedagogy 2023: Open University Innovation Report 11", Milton Keynes: The Open University, 2023.

[9] M.C. Laupichler, A. Aster, J. Schirch, T. Raupach, "Artificial intelligence literacy in higher and adult education: A scoping literature review", Computers and Education: Artificial Intelligence, 2022; vol. 3.

[10] W.M. Lim, A. Gunasekara, J.L. Pallant, J.I. Pallant, E. Pechenkina, "Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators", The International Journal of Management Education, 2023; 21 (2), 100790.

[11] M. Caglar Sahin, A. Sozer, P. Kuzucu, T. Turkmen, M. Buke Sahin, E. Sozer, O. Y. Tufek, K. Nernekli, H. Emmez, E. Celtikci, "Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams", Computers in Biology and Medicine, vol. 169, 2024, 107807. (https://www.sciencedirect.com/science/article/pii/S0010482523012726)

[12] M.V. Pashkov, V.M. Pashkova, "Problems and Risks of Digitalization in Higher Education", Higher Education in Russia, 2022; 31; vol. 3, pp. 40–57.

[13] S. Kumar, P. Rao, Sh. Singhania, Sh. Verma, M. Kheterpal, "Will artificial intelligence drive the advancements in higher education? A tri-phased exploration", Technological Forecasting and Social Change, vol. 201, 2024, 123258, (https://www.sciencedirect.com/science/article/pii/S0040162524000544)

[14] L.O. Tedeschi, "Review: The prevailing mathematical modeling classifications and paradigms to support the advancement of sustainable animal production", Animal, vol. 17, Supp. 5, 2023, 100813, https://doi.org/10.1016/j.animal.2023.100813. (https://www.sciencedirect.com/science/article/pii/S175173112300109X)