СЕКЦІЯ 3
КОМП'ЮТЕРНІ НАУКИ ТА ПРОГРАМНА ІНЖЕНЕРІЯ

UDC 519.8

ANALYZE AND EVALUATE STUDENT RESPONSES TO OPEN-ENDED QUESTIONS USING NATURAL LANGUAGE PROCESSING (NLP)

Viktoriia Astafieva, student, vikaastafyevaa@gmail.com, State University of Trade and Economics
Selivanova Anna, senior lecturer, ann.selivanova1@gmail.com, State University of Trade and Economics

Using Natural Language Processing (NLP) to analyse and evaluate student responses to open-ended questions marks a significant advancement in educational technology. Lots of examination testing require manual evaluation and rating, which is time-consuming and potentially error-prone. Natural language processing (NLP) should be used to automatically analyse free-text answers to support the review process [1].

Several studies have demonstrated the efficacy and challenges of using NLP in educational settings. For example, research on automated essay scoring systems has shown that they can provide scores comparable to human graders in terms of reliability and validity. However, these studies also highlight the importance of combining NLP tools with human oversight to ensure fair and accurate assessments [2],[3].

The use of good NLP models including: ELECTRA-small, RoBERTa-base, XLNet-base, and ALBERT-base-v2 for automated formative assessment will substantially improve education technology implementation, design and use in the context of online education and large cohorts [3].

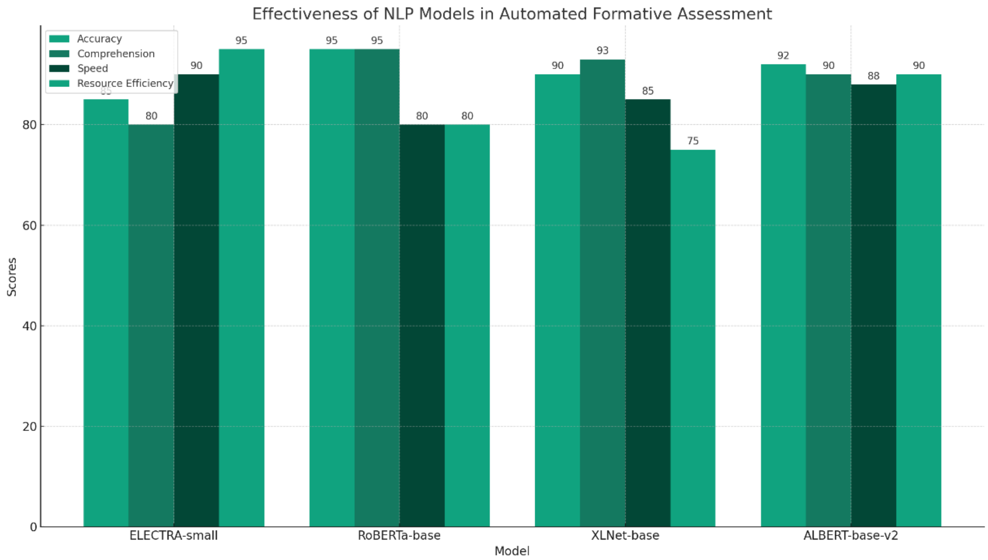

Figure 1 – Result of NLP models efficacy measure using Accuracy, Comprehension, Speed and Resource Efficiency

The Cunningham-Nelson's (2019) automated assessment techniques, utilizing transformer-based NLP models to analyse students' responses for conceptual understanding is an outstanding example. It built on six adapted concept inventory questions to develop a model that assesses four pointers within a response, determining the student's conceptual grasp. The research results show NLP capability aids in evaluating students' understanding, using accuracy and ROC curve metrics to measure performance. The pointers allowed a nuanced analysis of students' conceptual understanding, including assessing their confidence in their responses [3].

It's also important to note that two "impossible" scenarios in the context of assessing student responses using NLP are:

1) The model incorrectly identifies a response as having correct reasoning without the mention of the correct concept;

2) The model finds correct reasoning despite the student choosing the wrong multiple-choice answer.

Both scenarios are likely due to misclassifications by the NLP system. Another significant point is to make sure that the NLP model focuses on free-text validity and confidence-in-response pointers, emphasizing the selection of transformer-based NLP models based on their performance in the GLUE benchmark for this purpose.

During the preprocessing steps for NLP model training, emphasizing minimal changes to maintain students' response semantics. This includes lower-casing, punctuation removal, spell checking, and removing duplicates and single-word responses to ensure data quality. Using the simple transformers library for optimal tokenization and optimizing model training parameters through early stopping algorithms could be beneficial in this case. Additionally, ensemble modeling by combining the strengths of individual models through majority voting help to enhance accuracy.

The effectiveness of an NLP engine that uses word lists, rule-based synonyms, and decision-tree learning was evaluated. The use of final scores as performance metrics demonstrated the potential for NLP models to enhance manual evaluation processes and aid in the creation of automated review systems. NLP is able to perform better with structured questions and answers, which require less effort for accurate interpretation.

The NLP models could effectively assess students' conceptual understanding from their textual responses, achieving over 95% accuracy. The approach of leveraging ensemble models and minimal preprocessing, aligns with human evaluative standards, offering a scalable, automated formative assessment tool suitable for online and flexible educational settings. Despite limitations, such as dependency on data set size and content specificity, this method provides a significant step forward in applying NLP to educational assessment, promising enhanced learning and teaching experiences [4].

Conclusion. By using the NLP approach, free-text responses can be effectively automated for analysis and grading, which can streamline the laborious and error-prone manual review process. Structured question-answer formats result in a better performance of the NLP engine, which reduces processing effort. Open-ended questions are useful for assessing critical skills, but their evaluation is resource-intensive, which highlights the potential benefits of NLP in improving educational assessments.

References

1. Khurana, D., Koli, A., Khatter, K. and Singh, S. (2022). Natural Language processing: State of the art, Current Trends and Challenges. *Multimedia Tools and Applications*, [online] 82(3), pp.3713–3744.
2. Raina, V. and Krishnamurthy, S. (2021). Natural Language Processing. *Building an Effective Data Science Practice*, pp.63–73.
3. Somers, R., Cunningham-Nelson, S. and Boles, W. (2021). Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australasian Journal of Educational Technology*, 37(5), pp.98–115.
4. Stoehr, F., Benedikt Kämpgen, Lukas Müller, Laura Oleaga Zufiría, Junquero, V., Merino, C., Mildenberger, P. and Kloeckner, R. (2023). Natural language processing for automatic evaluation of free-text answers — a feasibility study based on the European Diploma in Radiology examination. *Insights into Imaging*, 14(1).