Higher education and in-service teachers II. (chair: Edit Katalin Molnár)

T-15

UTILIZING AI FOR ANALYZING WRITTEN ASSESSMENT REPORTS IN HIGHER EDUCATION

Anthea Moravánszky

Doctoral School of Education, University of Szeged

Keywords: Artificial Intelligence (AI); Assessment Report; Natural Language Processing (NLP)

This qualitative study explored written assessments of bachelor theses with the intention of uncovering discrepancies between assessors` feedback and awarded grades using AI for in-depth textual analysis. The study analyzed a dataset of 214 assessment reports from 107 bachelor theses written by 40 supervisors and co-supervisors between 2020 and 2022. In the context of this study, bachelor theses underwent evaluation using a grading rubric. This rubric included criteria such as topic delimitation, research questions, methods, structure, language, references, and citations. The grading scale ranged from 6 (excellent) to 1 (insufficient), with scores below 4 indicating a specific area failure. In addition to individual aspect grading, the assessors (supervisor/co-supervisor) assigned each an overall impression grade and provided a summary of their evaluation in the form of a written assessment report. The study employed a qualitative approach to examine the language, specific phrases, and text length in the assessors` feedback. Its primary objective was to uncover latent patterns or inconsistencies between the feedback and the grades awarded. This research also included calculating interrater reliabilities, containing the grades of the human assessors and the AI. The research aimed to explore the grade assigned by AI (ChatGPT 4) based on its analysis of the assessment reports. It also aimed to compare this AI-assigned grade with the grades provided by human assessors in their written assessment reports. Longo`s human-in-the-loop approach was applied by utilizing AI assessment as the initial step, followed by random re-assessment by a human assessor who had not been previously involved in the assessment process. The main finding of this study revealed that there was often no noteworthy difference in assessments falling within the excellent grading range (5–6) when comparing AI`s evaluations to those of human assessors. However, discrepancies became more apparent between 4–5, where assessors struggled to express and balance credit and objection in their feedback. In such cases, AI tended to interpret a lower grade due to specific formulations that intensified the negative tone in the assessment reports. This study contributes to the broader discourse on written assessment reports by highlighting potential disparities in representing the actual grades. It provides valuable insights that could be the foundation for creating an AI model to assist assessors in writing assessment reports. Furthermore, the findings could be used to develop training programs to help assessors align their assessment reports more closely with their grading decisions. The research into AI-driven analysis of thesis grading underscores AI`s potential to enhance the accuracy of academic assessment reports while deepening our understanding of the challenges assessors encounter when conveying nuanced evaluations in their written reports and potential solutions to address them.