AI Fairness in education

Unfairness Risk Assessment

The deployment of AI in education raises critical issues of fairness and bias AI systems can inadvertently perpetuate and amplify existing biases present in the data they are trained on, leading to unfair outcomes, particularly for historically marginalized groups. For instance, AI-powered essay grading systems may reflect the biases of the data they are trained on, which could include the subjective preferences of human graders. This could result in unfair assessments for students whose writing style or cultural background differs from the norm established by the training data. Ethical considerations also come into play when discussing the transparency and accountability of AI systems in education. The "black box" nature of AI algorithms can make it challenging to understand and contest their decisions, raising concerns about the ethical implications of their use in educational evaluations. The strength assessment is based on natural language processing and assess student work, automate grading and feedback, and facilitate tailored learning experiences. The AI-driven assessment tool under analysis provides consistent, objective, and efficient evaluations of student performance while also enabling personalized feedback, identifying knowledge gaps, and informing instructional strategies It aim to reduce the workload for educators and offer scalable assessment solutions, particularly in large-scale online education settings. The score assessment under analysis applies uniform criteria to all student submissions, which can help reduce human biases and inconsistencies in grading. This is one of the main advantages of this Tool. Despite this advantage, inherent biases in AI assessment systems are a source of concern. The evaluations and feedback produced by the system may reinforce unjust treatment or prejudices against particular demographic groups if the training data used to create it is biased or lacks variety.

Biases mitigation mechanisms

To detect and mitigate bias a set of mechanisms have been implemented in the strength assessment under analysis in order to counteract the following types of biases:

- Data bias – different types of data biases have been considered: selection bias, sampling bias, labelling bias, temporal bias, aggregation bias, historical bias, measurement bias, confirmation bias, proxy bias, cultural bias, under-representation bias.
- User interaction bias to mitigate the adaptation of the AI behaviour based on user feedback potentially reinforcing and amplifying existing biases.

The mitigation measures belong to the following classes:

- Pre-processing fairness modifying training data to balance group representation;
- In processing technique that modified learning algorithms to integrate fairness;
- Post processing to adjust model predictions to align with fairness goals.

Fairness metrics

The following fairness metric has been considered: Disparate Impact. Disparate Impact. One of the most widely used notions of fairness is based on the legal concept of disparate impact which occurs whenever a neutral practice negatively impacts a protected group. The principle of disparate impact can be extended to ML models by considering their output with respect to protected attributes To quantify the disparate impact in regression and classification, a fairness indicator has been considered: Disparate Impact Discrimination Index (DIDI). The higher the DIDI, the more disproportionate the model output is with respect to protected attributes, and the more it suffers from disparate impact.