Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education

Nils-Jonathan Schaller[1], Yuning Ding[2], Andrea Horbach[2,3], Jennifer Meyer[1], Thorben Jansen[1]
[1]Leibniz Institute for Science and Mathematics Education at the University of Kiel, Germany
[2]CATALPA, FernUniversität in Hagen, Germany
[3]Hildesheim University, Germany

Abstract

Pursuing educational equity, particularly in writing instruction, requires that all students receive fair (i.e., accurate and unbiased) assessment and feedback on their texts. Automated Essay Scoring (AES) algorithms have so far focused on optimizing the mean accuracy of their scores and paid less attention to fair scores for all subgroups, although research shows that students receive unfair scores on their essays in relation to demographic variables, which in turn are related to their writing competence. We add to the literature arguing that AES should also optimize for fairness by presenting insights on the fairness of scoring algorithms on a corpus of learner texts in the German language and introduce the novelty of examining fairness on psychological and demographic differences in addition to demographic differences. We compare shallow learning, deep learning, and large language models with full and skewed subsets of training data to investigate what is needed for fair scoring. The results show that training on a skewed subset of higher and lower cognitive ability students shows no bias but very low accuracy for students outside the training set. Our results highlight the need for specific training data on all relevant user groups, not only for demographic background variables but also for cognitive abilities as psychological student characteristics.

1 Introduction

Educational equity is seen as a foundation for learning with technology (Warschauer et al., 2004), because all students need effective instruction. One of the most effective instructional practices is feedback (Hattie and Timperley, 2007), which can support students in acquiring complex skills like writing (Graham et al., 2015). Automated essay scoring (AES) can be used to provide students with feedback on their writing at scale (Fleckenstein et al., 2023).

The foundation of equity in automated feedback systems is the fairness of the algorithm ((Holstein and Doroudi, 2021), (Pedró et al., 2019)), i.e., the absence of any prejudice or favoritism toward groups of students based on their inherent or acquired characteristics, including their background and their psychological variables((Mehrabi et al., 2019),(Government Equalities Office, 2013)). Algorithmic fairness is widely discussed in various educational contexts from normative (Blodgett et al., 2020; European Commission, Directorate-General for Education, Youth, Sport and Culture, 2022), societal (Baker and Hawn, 2022; Kizilcec and Lee, 2020), or methodological (Mitchell et al., 2021) perspectives, but literature reviews have shown that it is rarely investigated empirically (Li et al., 2023). Specifically in the AES context, only six empirical studies have examined algorithmic fairness, examining differences in algorithmic accuracy and biases for students with different gender, race, and language backgrounds in English-language corpora (Arthur et al., 2021; Baffour et al., 2023; Bridgeman et al., 2009; Litman et al., 2021; Kwako et al., 2022; Yancey et al., 2023). This means that while AES is widely used in education in many countries (Fleckenstein et al., 2023) including non-English speaking countries, it is unclear whether the algorithms used are fair to all groups of students confronted with the results or whether they might disfavor some student gropus. Compounding the problem, the few existing studies have shown that, depending on the algorithms used, students' essays were not scored fairly and disfavored groups related to race/ethnicity, economic status, and English Language Learner status (e.g., Baffour et al. (2023); Litman et al. (2021); Yang et al. (2024)).

So far, previous studies only analyzed fairness in relation to students' demographic variables in corpora with students' essays in English: Extending this research to a corpus on argumentation essays in the German language, we address three main re-

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 210–221

search questions: (1) How fair are AES algorithms for students with different levels of cognitive abilities as psychological characteristics strongly related to writing competence? (Zhang and Zhang, 2023). Addressing this question is linked to the wider equity issue of whether AES systems are likely to widen or narrow the gap between high and low-performing students. (2) How fair are AES algorithms in languages other than English? The question is especially important when automated scoring is based on large language models, mostly trained on English text data. (3) How is the distribution of student characteristics in the training data impacting the mean accuracy and fairness of the prediction?

By answering these questions, our paper makes the following contributions: First, we provide a set of baseline models, including shallow learning, deep learning, and generative large language models (LLM), for the newly released DARIUS corpus, thus enriching the automatic scoring landscape with models for a large German argumentative writing corpus.

Second, we conduct fairness evaluations on our results indicating that none of the models trained on the entirety of training data shows unfair behavior towards specific subgroups.

Finally, to assess the role of the distribution of the training data on algorithmic fairness, we train shallow and deep models with subsets of data from students of low and high cognitive ability, as well as a mixed subset based on low, medium and high cognitive ability, and show that the models are unfair to the groups not included in the training set.

We make all of our code publicly available.[1]

2 Related Work: Fairness in AES Algorithm

According to a literature review by Li et al. (2023), there have been 49 peer-reviewed empirical studies focused on fairness and predictive bias in education since 2010, highlighting the growing academic interest in these issues.

The studies included multiple fairness measures, including the accuracy for the included groups and the mean differences between predicted and annotated scores for each score (e.g., (Litman et al., 2021)). Most of these studies were conducted in contexts other than AES, such as predicting students' course performance or their likelihood of dropping out of a course. To our knowledge, there are only two papers that diagnosed the predictive bias displayed by AES models(Litman et al., 2021; Arthurs and Alvero, 2020), even though the importance of this task has been pointed out as early as in 2012 (Williamson et al., 2012). Litman et al. (2021) evaluated the fairness of shallow and deep learning AES algorithms for essays from the upper elementary level in the English language using three measures: Overall Score Accuracy (OSA), Overall Score Difference (OSD), and Conditional Score Difference (CSD). They found that shallow and deep AES algorithms showed systematically overly positive and negative scoring depending on students' gender, race, and socioeconomic status. Arthurs and Alvero (2020) showed that a shallow learning AES system for college admissions essays based on word vectors favored high-income students over low-income students (see also (Bridgeman et al., 2009) for similar results for essays from the Test of English as a foreign language). Additionally, the authors trained models on only essays from the highest quartile of students in terms of performance, showing that these models are not suitable for students from the other quartiles. Yang et al. (2024) further emphasized that the fairness of AES systems is compromised if such models are used on students or tasks for which they have not been trained.

In addition to the studies included in the literature review, recent studies added an investigation of fairness in Large Language Models scoring essays from a high school context Baffour et al. (2023) in the PERSUADE 2.0 corpus (Crossley et al., 2022). The authors compared the winning entries of the Kaggle Feedback Prize competition.[2] They show differences in the model's accuracy based on demographic factors such as student race/ethnicity, and economic disadvantage. Similar fairness issues based on students' demographic variables were shown for large language models in essays in the English language written by first (Kwako et al., 2023) and second language students (Yancey et al., 2023).

In summary, previous studies on fairness in AES have used shallow learning models, deep learning models, and LLMs and compared whether the accuracy of judgments and systematic over/underrating can be explained by students' demographic vari-

[1] https://github.com/darius-ipn/fairness_AES

[2] https://www.kaggle.com/competitions/feedback-prize-2021

ables. The results showed some fairness problems, which were exacerbated in the studies where the AES was additionally trained only on a homogenous group of students.

3 Data

The DARIUS corpus is a collection of 4,589 annotated argumentative texts written by 1,839 students from German high schools, spread across 114 classes in 33 different schools(Schaller et al., 2024). Essays that were off-topic, shorter than two sentences, empty, or contained names or other data relevant to data protection were removed beforehand. The final dataset consists of essays from two writing assignments focused on socio-scientific issues on the topics *energy* and *automotive*, containing 2,307 and 2,282 essays respectively. Students wrote a draft and revision on one task, followed by an essay on the other task, resulting in up to 3 essays per student. An example text is listed in the Appendix 7. Students also provided demographic data voluntarily, a selection of which is listed in Table 1.

The dataset has been extensively annotated with information about argumentative structure on different levels of granularity. In the present study, we focus specifically on a subset of these annotations, namely *content zone*, *major claim*, *position* and *warrant*. Out of the nine original annotation categories, we selected those as they reflect different parts of an argumentative text, e.g. structure and content, and are annotated on different granularity levels (token level to whole texts). We used the demographic data to measure fairness with respect to gender, profile, school, cognitive ability (KFT), and languages, which are further explained after providing more details on the annotations in Section 3.1. A more extensive description can be found in the original paper (Schaller et al., 2024).

3.1 Annotations

Content zone: This annotation category breaks down the essays into their basic parts: the introduction, the body, and the conclusion. Each section can be as short as one sentence or span several sentences.

Major claim annotation: Central to the argumentative essence of the essays, the Major Claim annotation identifies the pivotal stance taken by the author on the discussed issue. In contrast to similar annotation efforts (Stab and Gurevych, 2014), we also include claims written not only in the opening paragraphs but also within the conclusion, offering a comprehensive view of the argumentative intent. Such claims form the basis for the author's further arguments and the direction of their reasoning.

Position annotation: This annotation extracts the essay's directional stance regarding the thematic issues presented in the writing tasks — whether the argumentation aligns with, diverges from, or remains ambiguous towards the positions debated within the tasks. This annotation is important for understanding the diversity of viewpoints and the critical engagement of students with the socio-scientific topics at hand.

Warrant annotation: A warrant is one out of five argumentative elements annotated in the dataset as part of the Toulmin's Argumentation Pattern (TAP) annotations, following the definitions by Riemeier et al. (2012). TAP describes a structural framework for constructing logical and compelling arguments by including a claim, providing supporting evidence (data), explaining the connection between the claim and data (warrant), and addressing counterarguments (rebuttal). For this study, we focus exemplarily on warrants because the use of warrants indicates already a higher argumentation skill(Osborne et al., 2016). TAP elements are not marked on the sentence level but on the token level, as a TAP sequence can cover a wide range from subordinate clauses to entire paragraphs.

3.2 Demographic and Psychological Data

We consider the following demographic variables:
Grade Grade indicates which grade level the student is in. The dataset was obtained for students between Grade 9 and Grade 12.

Gender The students could indicate their gender. Options were female, male, and diverse.

School The German school system differentiates between different forms of high school.

- Gemeinschaftsschule: non-academic track
- Gymnasium: academic track
- Berufsschule: vocational training

Profile The German school system allows students to choose a profile. The Natural Sciences profile, for example, has a focus on math and science, while the Social Sciences profile can have a focus on politics or ethics.

Languages The students could indicate the language that they speak at home.