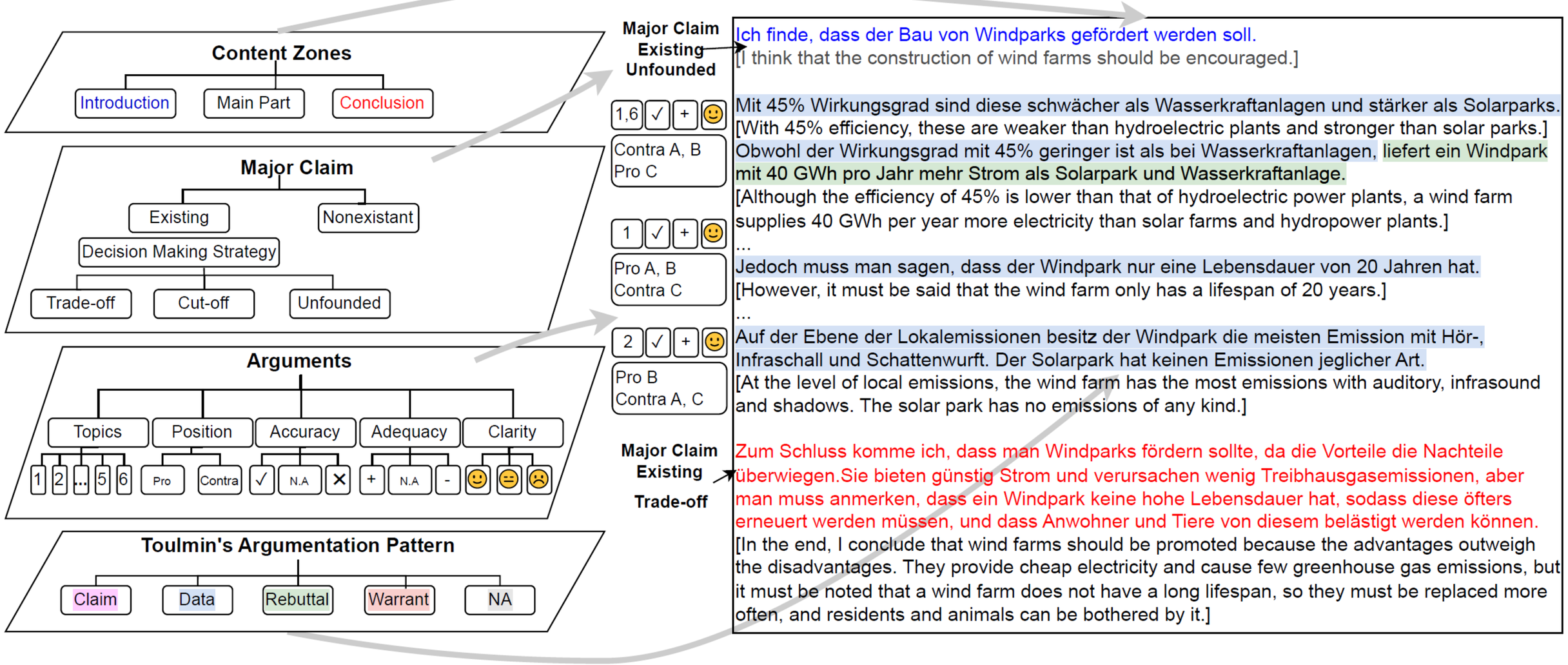
Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education

DARIUS
Digital Argumentation Instruction for Science

Nils-Jonathan Schaller[1], Yuning Ding[2], Andrea Horbach[2,3], Jennifer Meyer[1], Thorben Jansen[1]



Research Questions and Method

1. How fair are AES algorithms for students with different levels of cognitive abilities as psychological characteristics strongly relate to writing competence?
2. How fair are AES algorithms in languages other than English?
3. How is the distribution of student characteristics in the training data impacting the mean accuracy and fairness of the prediction?

1. Trained models on the complete dataset for each annotation.
2. Trained models on subsets of the training data based on cognitive abilities:
   - Lower quartile of KFT
   - Upper quartile of KFT
   - Mixed (quartile of each quartile) KFT
3. Evaluated the fairness of all models on a number of subgroups.

Fairness Measures and Corpus

- **overall score accuracy** (osa): Model score vs human score across groups. $(S - H)^2$
- **overall score difference** (osd): How and whether some groups tend to score consistently higher or lower due to the model's predictions. (S-H)
- **conditional score difference** (csd): measures whether an AES scores different groups with different scores, even when having the same human score.
- Scores over 0.01 suggest unfairness. (Williamson et al. 2012)

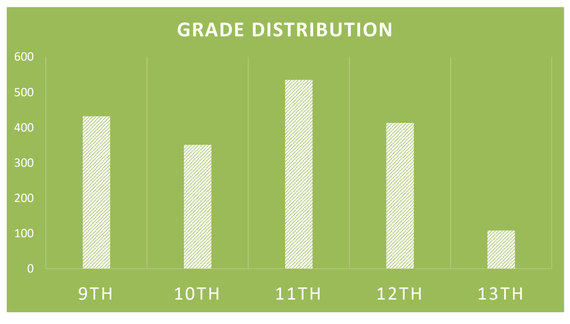
4589 annotated essays from 1839 student.

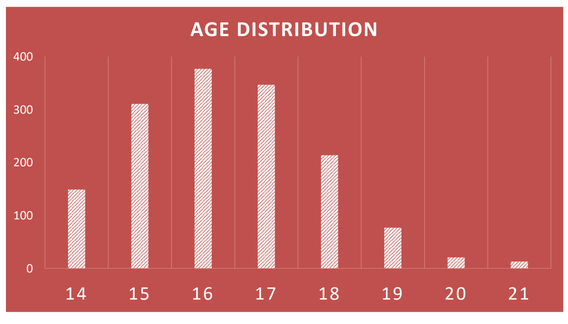
Corpus Available for Download:

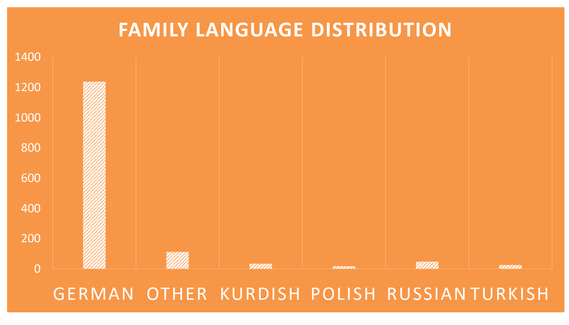
The DARIUS Corpus was created to develop automated tools for assessing and enhancing German high school students' argumentation skills. It consist of 4.589 essays on "energy" and "automotive" topics from 1.839 students across 33 schools. Students advocated for one of three options on power plants or car engines, produced drafts, received feedback, revised their work, and tackled a transfer task.

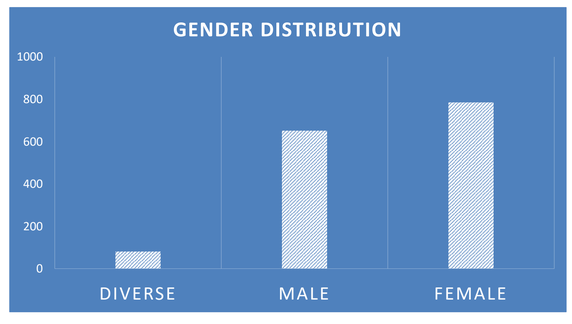
Annotation and Metadata Selection

Content Zone	Major Claim	Position	Toulmin's Argumentation Pattern
Sequences of **introduction**; **main part**; **conclusion**.	Sequence of author's stance on a topic.	Argument stance. **Pro**; **Contra**; **Unclear**	Sequences of argument parts. **Claim**; **Data**; **Warrant**; **Rebuttal**.

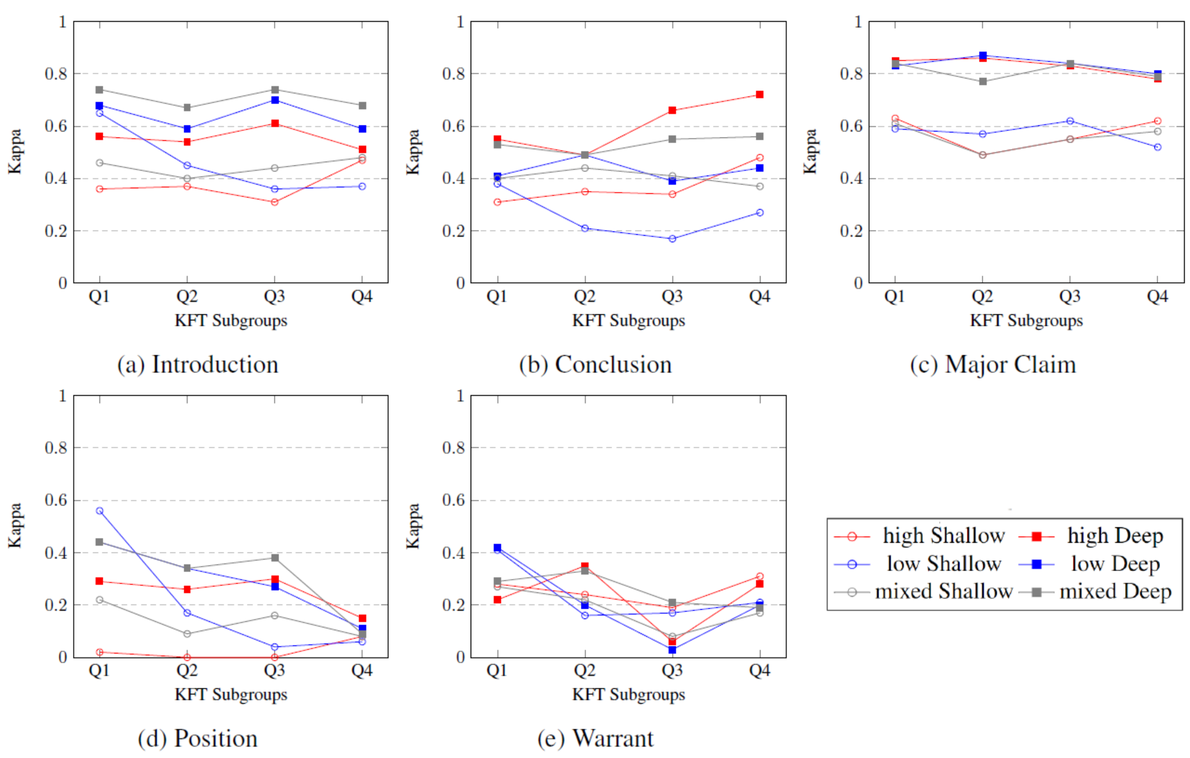








Results



Fairness and Performance Evaluated.

- **Performance**:
  - Task-specific deep learning performs best.
  - *Considerations: better prompts, other versions etc.*
- **Fairness**:
  - Training data from only one student group = lower performance on other groups.
  - OSA, OSD, CSD: No values over the threshold of .01.
  - Performance and Fairness should both be evaluated.
  - *Considerations: groups to homogenous? Some groups too small?*

Training Data	Model	Intro-duction	Conclusion	Major Claim	Position	Warrant
All	Shallow	0.63	0.55	0.68	0.41	0.43
All	Deep	0.81	0.7	0.88	0.44	0.44
-	LLM	0.60	0.68	0.75	0.32	0
KFT high	Shallow	0.38	0.39	0.57	0.02	0.26
KFT high	Deep	0.56	0.62	0.83	0.29	0.23
KFT low	Shallow	0.47	0.25	0.58	0.37	0.23
KFT low	Deep	0.65	0.44	0.84	0.34	0.2
KFT mixed	Shallow	0.46	0.42	0.56	0.16	0.17
KFT mixed	Deep	0.71	0.54	0.81	0.37	0.25

[1]Leibniz Institute for Science and Mathematics Education at the University of Kiel; [2]CATALPA, FernUniversität in Hagen, Germany; [3]Hildesheim University, Germany

Contact: schaller@leibniz-ipn.de

This work was supported by the Deutsche Telekom Stiftung and partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.