University of Zurich[UZH]

Department of Psychology - Psychological Methods, Evaluation and Statistics

A Pilot Study on the Use of Transformer Models to Evaluate Open-Ended Response Formats in Educational Assessments.

Rudolf Debelak, Benjamin Wolf

Road Map

- Transformers and Automated Essay Scoring
- A Pilot Study: Automated Scoring of German Essays
- Assessing Validity Using Interpretable Machine Learning
- Outlook: Reliability and Fairness

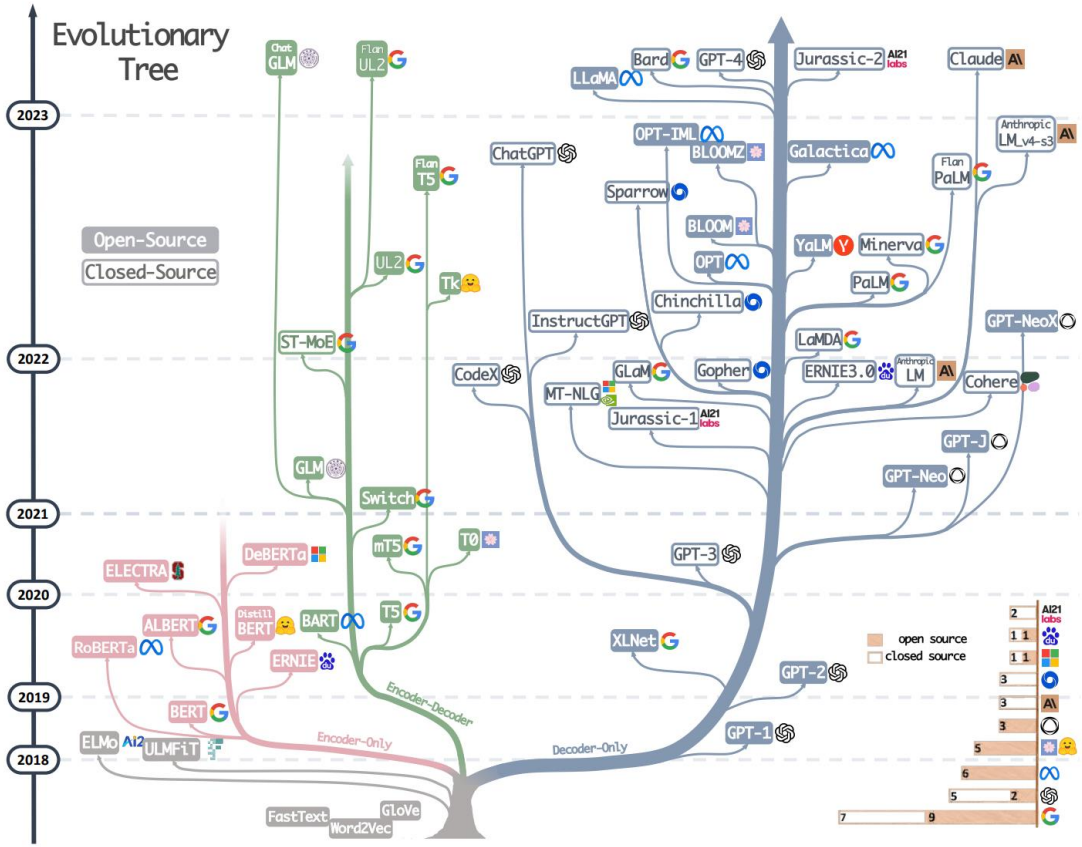
Transformers and Automated Essay Scoring

- The evaluation of written student essays is a classical method for assessing writing skills. They offer the opportunity to assess unique skills that can be hardly assessed by other assessment methods (e.g., multiple choice questions) such as creativity or the ability to address a given topic. (Hussein et al., 2019; Ke & Ng, 2019)
- However, grading student essays takes time and effort.
- There were some attempts to construct automated grading systems. First prototype systems were developed in the late 20th century, but these developments accelerated in the early 21th century. (e.g., Atali & Burstein, 2004)

Transformers and Automated Essay Scoring

- Some very recent approaches suggested the use of pre-trained transformer models for automated essay scoring. (Vaswani et al., 2017)
- Transformer models are a specific architecture of deep learning models that is commonly used with language data.
- On a conceptual level, transformers first process language into numeric output, which is represented as vectors. These vectors can then be used for predication and classification tasks, as well as text generation.
- Almost all transformer models are pre-trained, that is, they have been trained on a large corpus of documents, which allows them to process data in a meaningful way with little additional training on a given task.

Transformers and Automated Essay Scoring



(Yang et al., 2023)

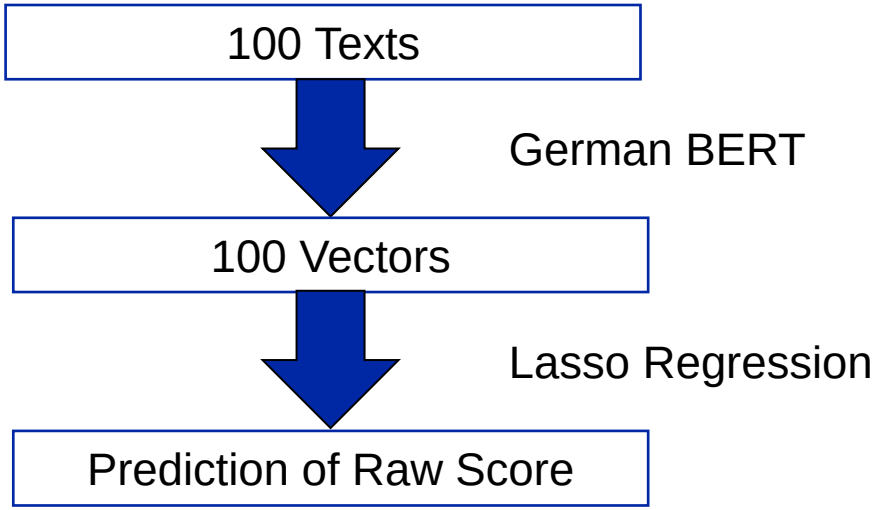
Transformers and Automated Essay Scoring

- This wealth of models leads to the natural question of whether transformer models can be used for automated essay scoring.
- This leads to two families of research questions:
  - How do I train such models so that they can model accurately human raters? This question is related to AI research.
  - How can I check the resulting grade with regard to important psychometric standards, such as validity, fairness and reliability? This question is related to psychological measurement and psychometrics.
- In this presentation, we discuss these problems by investigating an example dataset of German essays in a pilot study.

A Pilot Study

- In this pilot study, we worked with data for 125 German essays from Swiss students around the age of 14. 100 were used as training data set, 25 as test data set.
- The essays were written on one of three topics.
- All essays were graded by the same human rater with regard to style, the correct use of language, creativity, the coherence of the text, and the content. These scores were summarized by an overall rating score.
- The purpose of this study was to predict this rating score from the text.
- To achieve this aim, we used a pre-trained transformer model, namely a German BERT model (https://www.deepset.ai/german-bert), to transform the texts into numerical vectors. Since this leads to a regression problems with many predictors and a relatively small sample size, we use a regularized regression model for the prediction.

A Pilot Study



A Pilot Study

Caveats:

- The sample size is relatively small for training an automated essay scoring model, which makes overfitting likely. More data would be desirable.
- Since all human ratings come from the same rater, we basicly model her assessment. This might include her biases. More raters would be desirable.
- Transformer models are black box models, even when they are combined with regularized regression. This makes it essential to evaluate these models beyond their ability to replicate the human ratings.

A Pilot Study

- The calculations were carried out using the transformers and torch modules in Python. (Wolf et al., 2019; Paszke et al., 2019)
- This module allows to access multiple models for various languages and purposes.
- The Lasso regression was implemented using Scikit-Learn. (Pedregosa et al., 2011)

A Pilot Study

The results:

- Correlation of 0.73 between the predicted and observed sum score in the validation set.
- Correlation of 0.93 between the predicted and observed sum score in the training set.

Possible improvements:

- Apply cross-validation to investigate the stability of these results.

Assessing Validity

”Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. […] Statements about validity should refer to particular interpretations for specified uses. It is incorrect to use the unqualified phrase ‘the validity of the test.’ ” (AERA, APA, & NCME, 2014, p. 11)

There are several sources of validity evidence:

- Test Content
- Response Processes
- Internal Structure (“Construct Validity”)
- Relation to External Variables (“Criterion Validity”)

Assessing Validity

The following sources of validity evidence are not directly related to statistical modeling:

- Test Content
- Response Processes

The following sources are usually handled in a test theoretic framework:

- Internal Structure (“Construct Validity”)
- Relation to External Variables (“Criterion Validity”)

Assessing Validity

Checking the Internal Structure (“Construct Validity”) for an AI framework:

- Application of methods of interpretable machine learning, e.g. LIME and Shapley values, to explain the relationship between individual words or word parts (“tokens”) and the overall output.
  - For instance, spelling mistakes or grammar mistakes should be closely related to a decrease of the overall rating.
  - For some criteria, like overall content, this relationship is less clear, making this approach opaquer.
- Alternatively, one could evaluate the rating for essays for which the correct rating is clear. This could include artificial essays with no errors (positive rating) or typical common errors (negative) rating. The model should be able to discern between such models.

Assessing Validity

Checking the relation to external variables (“Criterion Validity”) for an AI framework:

- Correlation to human ratings
- Correlation to external criteria (e.g., intelligence scores).
- Checking for possible AI biases, such as: Preference of short or long texts, preference of specific topics

This still needs to be investigated for this pilot study.

Outlook: Reliability and Fairness

"The reliability/precision of the scores depends on how much the scores vary across replications of the testing procedure, and analyses of reliability/precision depend on the kinds of variability allowed in the testing procedure (e.g., over tasks, contexts, raters) and the proposed interpretation of the test scores." (AERA, APA, & NCME, 2014, p. 33)

To investigate this standard, one might apply methods that do not consider the specific nature (i.e., text) of the input:

- Correlation of scores for different parts of the text, if the length of the text is unimportant
- Correlation of scores between different texts written by the same student.

Outlook: Reliability and Fairness

One might also apply methods that do consider that the input is text:

- Does the rating change to a large extent if I make small changes to the input?
- Does the rating change in the intended direction? For instance, does the inclusion of a spelling error always decrease the rating?

This approach is related to the idea of checking the robustness of AI models to insignificant changes. In the field of psychometrics, this idea is related to the simulation of measurement errors.

Outlook: Reliability and Fairness

„A test that is fair within the meaning of the Standards reflects the same con struct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct.” (AERA, APA, & NCME, 2014, p. 50)

In psychometrics, this standard is closely related to differential item functioning and measurement invariance. Since we are not interested in the interpretation of model weights in AI models, this specific aspect of having stable weights is less important here.

Outlook: Reliability and Fairness

For checking fairness of an automated essay scoring model, we suggest the following approach:

- Define relevant subgroups of respondents beforehand, including: respondents of different gender, different age groups, students writing on different topics.
- Evaluate the prediction accuracy in each of these groups independently. Is there any group where the model give more inaccurate ratings?
- Is there any group where the model gives on average higher or lower ratings? If so, can these differences be theoretically justified?

Discussion

We discussed the following points:

- Transformers and Automated Essay Scoring
- A Pilot Study: Automated Scoring of German Essays
- Assessing Validity Using Interpretable Machine Learning
- Outlook: Reliability and Fairness

The application of the new approaches for checking validity, reliability and fairness are work in progress and need to be evaluated for the pilot dataset.

References

AERA, APA, & NCME (2014). Standards for Educational and Psychological Testing: National Council on Measurement in Education. Washington DC: American Educational Research Association.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with E-RATER® V.2.0. ETS Research Report Series: i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Hussein, M. A., Hassan, H. A., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5:e208.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In: Kraus, S. (Ed.) Proceedings of the Twenty-Eight International Joint Conference on Artificial Intelligence, pp. 6300–6308.

References

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . ., Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), Advances in neural information processing systems (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pd

References

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. ACM Transactions on Knowledge Discovery from Data. https://doi.org/10.1145/3649506