(AI)N'T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM

Philipp Hartmann
University of Goettingen, philipp.hartmann@uni-goettingen.de

Nils Holthoff
University of Goettingen, nils.holthoff@stud.uni-goettingen.de

Sebastian Hobert
University of Goettingen, shobert@uni-goettingen.de

Matthias Schumann
University of Goettingen, mschuma1@uni-goettingen.de

(AI)N’T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM

Research Paper

Philipp Hartmann, University of Goettingen, Germany, philipp.hartmann@uni-goettingen.de

Nils Holthoff, University of Goettingen, Germany, nils.holthoff@stud.uni-goettingen.de

Sebastian Hobert, University of Goettingen, Germany, shobert@uni-goettingen.de

Matthias Schumann, University of Goettingen, Germany, mschuma1@uni-goettingen.de

Abstract

Education is increasingly being delivered digitally these days, whether due to the COVID-19 pandemic or the growing popularity of MOOCs. The increasing number of participants poses a challenge for institutions to balance didactical and financial demands, especially for exams. The overall goal of this design science research project is to design and evaluate a semi-automatic machine-learning-based scoring system for essays. We focus on the design of a functional software artifact including the required design principles and an exemplary implementation of an algorithm, the optimization of which is, however, not part of this project and subject for future research. Our results show that such a system is suitable for both scoring essay assignments and documenting these scorings. In addition to the software artifact, we document our results using the work of Gregor et al. (2020). This provides a first step towards a design theory for semi-automatic, machine-learning-based scoring systems for essays.

Keywords: automated essay scoring, machine learning, design principles, education, learning

1 Introduction

The process of digitization has had a strong influence on education over the last years which has been reflected in the increasing number of online courses (Impey and Formanek, 2021). This trend has further been reinforced by the COVID-19 pandemic, as in-class lectures and exams have been suspended and conducted digitally to fight the pandemic (Crawford et al., 2020; Kelly and Columbus, 2020; UNESCO, 2020). In addition to the didactic benefits of multimedia content, e-learning is primarily used to reduce costs (Lai and Liou, 2010; Yusuf and Al-Banawi, 2013). For example, digital learning modules can be reused and allow a larger number of participants due to location-independent use, such as in MOOCs. As a consequence of the growing number of participants, there is also an increasing effort in assessing the taught content at final exams (Balfour, 2013). Whereas in the past, mainly the fully automatically assessable, closed question types (e.g., multiple-choice questions) were used for the examination, the use of essay tasks in e-assessments is also becoming increasingly relevant. From a didactic perspective, essays are more suitable for evaluating competences because through them the understanding of complex relationships, for example the higher taxonomy levels of Bloom et al. (1956), can be tested (Birenbaum et al., 1992; Castellanos-Nieves et al., 2011). While open-ended questions are not difficult to conceptualize, manual scoring proves complex because each response is nearly unique (Attali and Burstein, 2006; Richardson and Clesham, 2021). In particular, when based on superficial features of the answer, such as the use of important terms, students writing generalities and nonsensical content, including the terms sought by the examiners, can cause concentration problems when there are a large number of exams to be scored (Castellanos-Nieves et al., 2011). Therefore, automating the process of scoring essay tasks could significantly reduce the workload for examiners. In addition, the scoring process could be accelerated and the standardization of grading that accompanies automated scoring could lead to greater consistency and fairness in the grades awarded (Richardson and Clesham, 2021; Hung et al., 1993). There are already approaches that deal with automated scoring of essay tasks (Ramesh and Sanampudi, 2021). However, these works mostly address the technical perspective rather than the underlying process of scoring and the examiners' needs. Previous research has shown that trust towards AI-based services, and thus their acceptance or use, is influenced by the relevance of the decision (Ashoori and Weisz, 2019; Lee and See, 2004). Decisions to which users assign a high relevance, such as high stake exams, are often met with reluctance (Ashoori and Weisz, 2019). In addition, human influence on the final decision is often considered important. Thus, users often feel more personally attached if the final decision is made by a human being and the AI-based service merely serves to support the decision (Ashoori and Weisz, 2019). Therefore, in order to gain the users' acceptance, their needs must be considered. In the following, a holistic system for semi-automated essay scoring is considered, taking into account these user requirements. Within the semi-automated system, tasks are accomplished through an appropriate mix of human labor and automated, computerized assistance (Frohm et al., 2008). The scoring system supports the evaluation of answers to essay tasks in the first step by an automated pre-scoring of the answers using an adaptive system and in a second step by helping examiners with the manual post-scoring.

In our research project, we focus on the design of a first functional software artifact that is able to support essay scoring using a semi-automated essay scoring system. To achieve this, we implement a fully-functional user interface artifact and include a first version of a machine-learning-based scoring algorithm. Thus, the aim of this project is to identify core design principles for semi-automated essay scoring systems. Within our research project, we follow a design science research (DSR) approach based on Peffers et al. (2007) and Hevner et al. (2004) to answer the following research questions.

RQ1: How to design a machine-learning-based, semi-automatic scoring system for essays?

RQ2: How do potential users assess the semi-automatic scoring system, supporting the essay-grading process?

Within this research approach, we aim to contribute design knowledge on how to implement semi-automated scoring systems in educational contexts. In the remainder of this paper, we demonstrate and evaluate a first software artifact consisting of a fully-functional user interface and an exemplary machine-learning-based scoring algorithm. The developed software artifact is based on two design-build-evaluate iterations (March and Smith, 1995). While deriving design principles and demonstrating a functional software artifact is the goal of this research project, improving and optimizing the scoring algorithm is subject to future research.

The remainder of this paper is structured as follow: Next, we present related research on semi-automatic scoring systems. Following this, our DSR approach used is first briefly outlined and then applied to the problem in detail. Subsequently, the results are discussed, and the derived design knowledge is summarized based on Gregor et al. (2020).

2 Related Research on Semi-Automatic Essay Scoring

The automatic scoring of essays has been a topic of research for a long time. However, due to technological limitations, these approaches have usually been restricted to single, isolated factors of text composition (e.g., response and word length or grammatical correctness) and the identification of individual terms. For instance, Mitchell et al. (2003) were able to observe an accuracy of semi-automatic scoring of almost 95% for short free-text responses at an early stage. This accuracy was consistent with the accuracy of human examiners but required the creation of complex solution patterns for each task. With the help of an authoring tool, mark scheme templates were created in which syntactic-semantic structures were created by the examiners. The individual parts of the level of expectation were manually broken down into their individual components (nouns, verbs and prepositions). Subsequently, the identified terms were extended by synonyms, which were also to be scored as correct. Due to technical progress as well as the increasing availability of data, approaches for open questions of natural language are feasible nowadays. The use of machine learning enables accurate predictions to be made on the basis of existing empirical values (Mohri et al., 2018). The quality of the scoring depends on the quality and the extent of the empirical values, which, for example, consist of a previous scoring of comparable tasks by human examiners. The approaches that can be taken in this regard differ. Taghipour and Ng (2016) use an approach that works with a long-short-term memory network for evaluation. A Kaggle dataset (Kaggle, 2012) is used, with 60% of the data as the training, 20 % as the development and 20 % as the test dataset. The evaluation takes place on a technical level, with the best model architecture for essay scoring being sought. In contrast, Sharma and Jayagopi (2018) combine two neural networks for transcribing and scoring handwritten responses to essay tasks. The training dataset used 90% of the data and the validation dataset 10%. No differences were observed between AI-based and manual transcription. Another method is the memory network described by Zhao et al. (2017), outperforming a comparable LSTM approach in 7 out of 8 sets. Chen et al. (2010) used an unsupervised automated essay scoring system, which has an adjacent agreement rate of more than 90% and an exact agreement rate of 50%.

In addition, studies have shown that automated essay scoring is highly reliable (Foltz et al., 1999). Thus, individual scoring systems have shown a high correlation between the AI and the examiner scores (Attali and Burstein, 2006; Pearson, 2019). Cohen et al. (2018) also investigated the validity of automated essay scoring, where the "true scores" were determined as the mean across a group of examiners scoring the same task. Compared to a single examiner, the AI was able to achieve comparable results. In contrast, the validity of two or more examiners was significantly higher. From this, they derived that a system as support outperforms careless examiners in particular. Automatic scoring systems are also said to be highly objective. This results primarily from the elimination of the direct influence of human bias, whereby the dependence of the scoring is affected by the scope, quality, and objectivity of the underlying data (Kumar and Boulanger, 2020).

The examples outlined above as well as other identified literature most often focus on the algorithm or model of assessment. However, little attention has been paid to other aspects, such as the process of scoring and the resulting requirements of the examiners. An evaluation of the whole system (instead of only focusing on algorithms) is rarely done. Since the automatic assessment of essays works well, though not perfectly, semi-automatic systems may be a possible solution. For the reasons stated in the introduction, a semi-automatic essay scoring system is developed in this study. The system is at the fourth level of automation, according to the classification of Billings (1997). Here, the system supports the assigned activity, but the user remains in full control of the AI.

3 Research Design

To achieve our research goal of developing and implementing a semi-automatic scoring system for essay answers in exams, we apply a DSR method, as shown in Figure 1. Thereby we intend to (1) develop an artifact for solving the problems listed in the introduction and (2) derive generalizable design principles for the Information Systems (IS) research discipline, according to Gregor et al. (2020). Our research design is based on the DSR process proposed by Peffers et al. (2007), which is suitable when the focus is on developing a practically applicable artifact using a flexible process. In addition, a practical and outcome-oriented evaluation of the artifact will be conducted. Since the present project aims at the design and implementation of an IT artifact in the sense of Hevner et al. (2004), their guidelines are additionally applied. Figure 1 shows an overview of the applied process. In green, the process according to Peffers et al. (2007) with its six phases is shown. The guidelines described by Hevner et al. (2004) are shown in yellow and assigned to the six phases. In blue, equivalent to the phases of the process, the concrete implications for this work are shown.

The first phase deals with identifying the problem by describing it in a comprehensible way based on the user stories and the challenges of the examiners. This is supposed to clarify the added value of our artifact. In the second phase, building on the previous phase, we establish the requirements before deriving the design principles for the planned artifact. Through this, we want to show the potential for improvement compared to manual scoring. The third phase comprises the design and implementation of the artifact. In the first iteration, we focus on the user interface and the presentation of the relevant requirements for the application. In the fourth phase of the DSR process, we demonstrate the application to examiners, and in the subsequent fifth phase, we evaluate the artifact, paying special attention to the user interface and its basic functionalities for the users. The evaluation results are compared with the goals defined in the second phase and, if suitable, implemented in the following. In the second iteration, the evaluation results of the user interface from iteration one are used for improving the system and the scoring mechanism is implemented. The subsequent evaluation of the second iteration focuses on the scoring mechanism. Here we examine the technical suitability of machine learning in scoring essay answers and also address the question of whether it is suitable for out-of-domain data. For the phase of the documentation and communication of the gained knowledge, we use the components of the design principles recommended by Gregor et al. (2020).

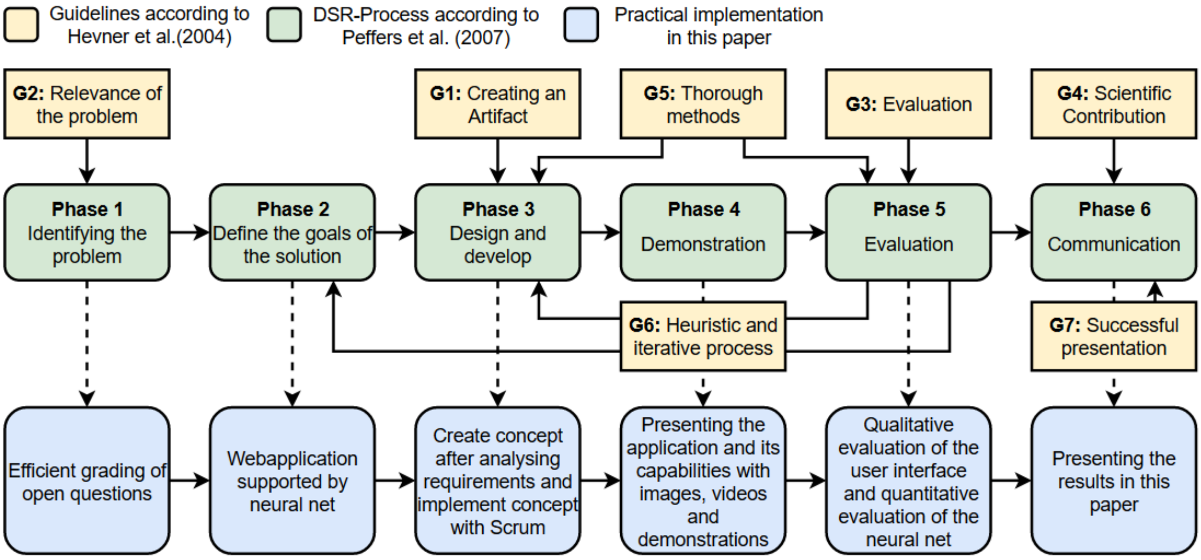

Guidelines according to Hevner et al.(2004)
DSR-Process according to Peffers et al. (2007)
Practical implementation in this paper
G2: Relevance of the problem
G1: Creating an Artifact
G5: Thorough methods
G3: Evaluation
G4: Scientific Contribution
Phase 1 Identifying the problem
Phase 2 Define the goals of the solution
Phase 3 Design and develop
Phase 4 Demonstration
Phase 5 Evaluation
Phase 6 Communication
G6: Heuristic and iterative process
G7: Successful presentation
Efficient grading of open questions
Webapplication supported by neural net
Create concept after analysing requirements and implement concept with Scrum
Presenting the application and its capabilities with images, videos and demonstrations
Qualitative evaluation of the user interface and quantitative evaluation of the neural net
Presenting the results in this paper

Figure 1. Adapted Research Approach (based on Hevner et al. 2004; Peffers et al. 2007)

4 Design and Evaluation

4.1 Specifying the Problem Statement

As stated in the introduction, the challenges of scoring by examiners are addressed in this paper. The goal is to design and implement a machine-learning-based system that automatically performs a pre-scoring of essay tasks and supports examiners in conducting a post-scoring based on the pre-scoring. The main reason for the challenge is an increasing number of students in university teaching and an increasing importance of essay assignments. To overcome these challenges, the system to be designed must provide the examiners with the most accurate pre-scoring possible, which can be comprehended and adapted by the human examiners during the post-scoring process.

Based on the scoring process of essay assignments, we derived three user stories for examiners. First, examiners have to create a level of expectation for the respective essay task before the scoring (U1). This is usually developed as part of the task design and includes possible answers that will be assessed as correct. In addition, the level of expectation should include possible cut scores for individual parts of answers if not all aspects are presented completely and correctly (American Educational Research Association, 2011). The next step in the scoring process is the actual scoring of the essays based on the level of expectation (U2). Finally, the scoring process must be documented for each essay task and each student (U3) to give the exam taker access to the essay score and the interpretation by the examiner (American Educational Research Association, 2011). The documentation includes the identification of the correct components of the level of expectation as well as the respective point allocation. In addition, missing aspects are noted.

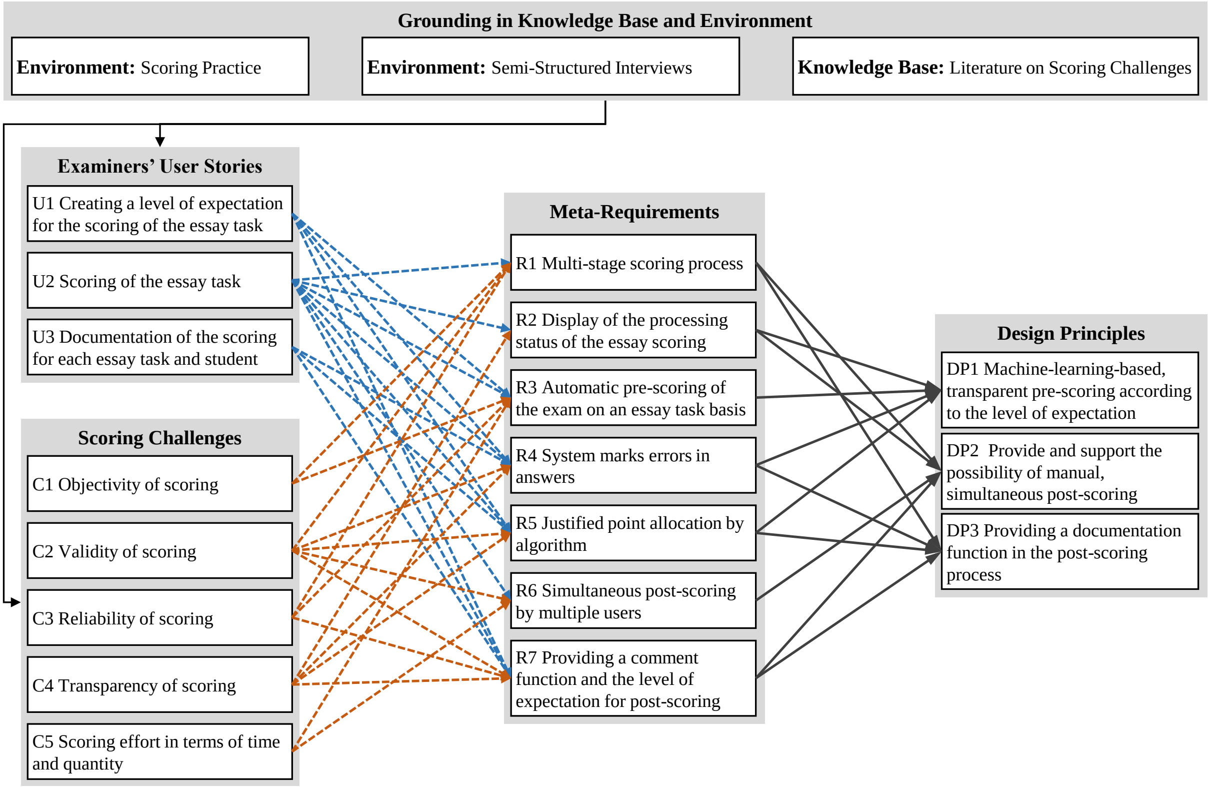

Grounding in Knowledge Base and Environment
Environment: Scoring Practice
Environment: Semi-Structured Interviews
Knowledge Base: Literature on Scoring Challenges
Examiners' User Stories
U1 Creating a level of expectation for the scoring of the essay task
U2 Scoring of the essay task
U3 Documentation of the scoring for each essay task and student
Scoring Challenges
C1 Objectivity of scoring
C2 Validity of scoring
C3 Reliability of scoring
C4 Transparency of scoring
C5 Scoring effort in terms of time and quantity
Meta-Requirements
R1 Multi-stage scoring process
R2 Display of the processing status of the essay scoring
R3 Automatic pre-scoring of the exam on an essay task basis
R4 System marks errors in answers
R5 Justified point allocation by algorithm
R6 Simultaneous post-scoring by multiple users
R7 Providing a comment function and the level of expectation for post-scoring
Design Principles
DP1 Machine-learning-based, transparent pre-scoring according to the level of expectation
DP2 Provide and support the possibility of manual, simultaneous post-scoring
DP3 Providing a documentation function in the post-scoring process

Figure 2. Overview of Requirements and Design Principles

The first three challenges arise from the implementation of the principles of good testing. These include the aspects of objectivity (C1), validity (C2), and reliability (C3) and must also be ensured in the context of scoring (American Educational Research Association, 2011; Hewlett and Kahl-Andresen, 2014). Objectivity (C1) expresses the extent to which a score is derived independently of an examiner's subjective evaluation (Hewlett and Kahl-Andresen, 2014). The American Educational Research Association (2011) also lists the notion of "fairness in testing" in this context. Thus, certain test characteristics may not be comparably difficult for all subgroups (e.g., defined by disability or language) within an exam. However, within a subgroup and its particular test, there should be a fair, objective assessment. Validity (C2) describes the suitability of the measurement for a given goal. A task is intended to test for the presence of a predetermined knowledge or skill (Hewlett and Kahl-Andresen, 2014; American Educational Research Association, 2011). Accordingly, the scoring is intended to provide an accurate assessment of the knowledge and skill being examined. Reliability (C3) describes the consistent replication of scoring. Under comparable conditions, the scoring of examination results should not be random but should show comparable results (Hewlett and Kahl-Andresen, 2014; American Educational Research Association, 2011). In this context, we also talk about the precision of scoring. For the previously mentioned reasons, it is important that the assessment is presented in a transparent and comprehensible way (C4). The last challenge deals with test economy (Hewlett and Kahl-Andresen, 2014). Especially in the case of large numbers of participants, the effort and benefit of the written exam must be taken into account in its design. As the scoring of essays is considered costly, with the costs depending very much on the time needed for correction per essay and the number of examinees, it is rarely used for cost-efficiency reasons (C5).

4.2 Deriving Requirements from Environment and Knowledge Base

In the following, the requirements for the scoring system resulting from the user stories and the scoring challenges are derived. A detailed overview of the relationship between the user stories (U1-U3), the scoring challenges (C1-C5) and the requirements (R1-R7) can be found in Figure 2. This figure also contains the relationship between the requirements and the design principals (DP1-DP3) described in section 4.3.

The first two requirements address the overall structure of the scoring process. A multi-stage scoring process should be implemented, which includes a machine-learning-based pre- as well as a human post-scoring (R1 based on U1, C1-C3). This is intended to make the scoring process objective and, at the same time, to increase validity through human review. The user should be informed about the current processing status of the scoring at any time (R2 based on U2 and C4) to increase the transparency of the procedure. The following three requirements address the machine-learning-based essay scoring. The system should perform the scoring process automatically and on a task basis (R3 based on U1, U2, C1, C3 and C5). During the scoring, the relevant aspects of the scoring should also be made visible in the answer (R4 based on U1-U3, C2 and C4) and the score assignment should be presented (R5 based on U1-U3, C2 and C4). This should improve the transparency of the automation and thus also increase the accuracy in the post-scoring. The possibility of assigning the results of the machine-learning algorithm to the individual text components allows for better human post-scoring. The last two requirements concern human scoring based on the automatic pre-scoring. A major challenge for examiners is the workload of large amounts of essays that need to be scored. Therefore, in addition to the AI-based support, it should also be possible for several examiners to carry out the post-scoring of the essays at the same time (R6 based on U2, C2 and C5). Furthermore, the level of expectation and a comment function should be implemented (R7 based on U1-U3, C1-C4) in order to improve the transparency and accuracy of the scoring as well as ensure the documentation.

4.3 Deriving Design Principles

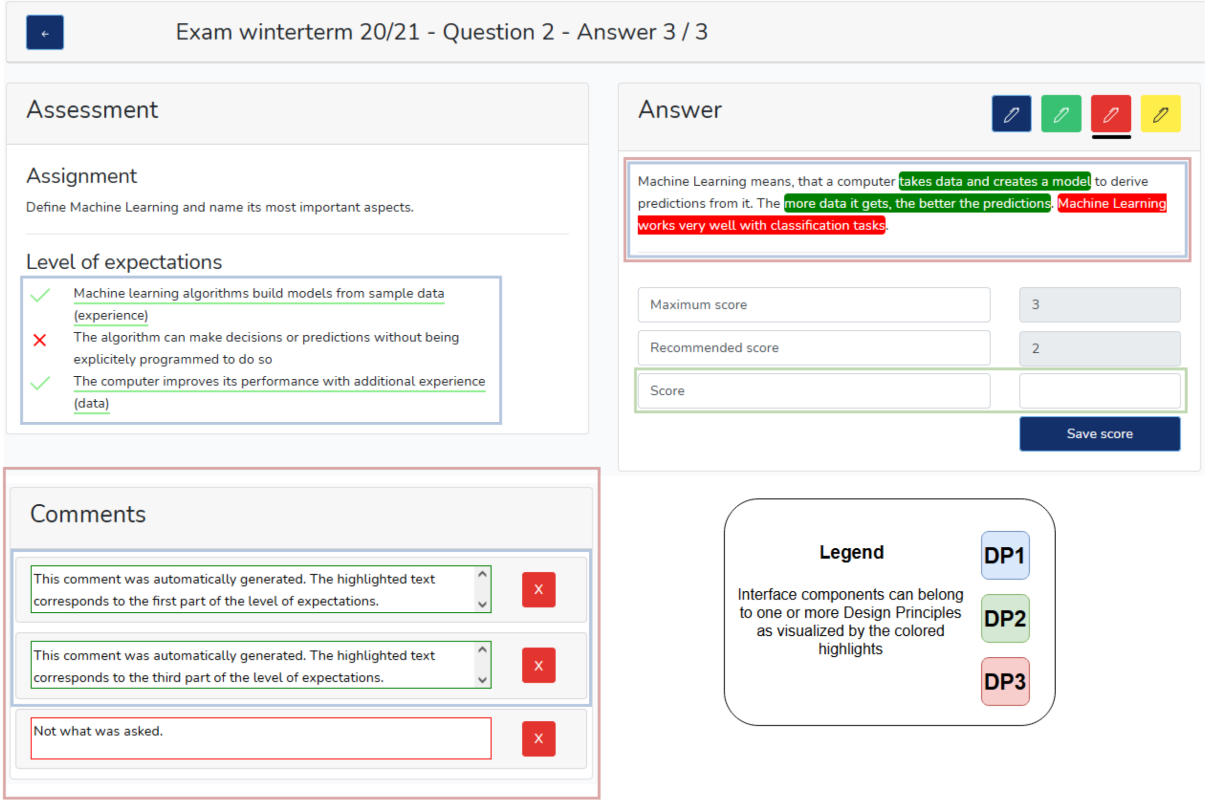
Based on the seven requirements, we derived three design principles. While the first two design principles deal with automatic pre-scoring and human post-scoring, the last design principle describes the documentation of the scoring results.

The first design principle defines that the system should perform an automatic machine-learning-based scoring. This scoring should be comprehensible for the examiners and based on the level of expectation assigned to the respective task (DP1 based on R2-R5). For this purpose, the examiners are given the opportunity to create the essay task and the associated level of expectation in the system. In a next step, the essays of the individual participants are added to the tasks in the system. The essays are then scored using machine learning. The items that are rated as partially or completely correct are highlighted in color. As a second design principle, the system should enable and support the possibility of manual, simultaneous post-scoring of the pre-scored essays (DP2 based on R1, R2, R6 and R7). For this purpose, the human examiners are presented with the level of expectation right next to the essay answer. In addition, it is possible to adjust the scoring from the machine-learning-based pre-scoring. In order to make the results comprehensible, functions for documenting the result of the post-scoring are needed (DP3 based on R1, R4, R5 and R7). Although these also include documentation functions from the AI-based scoring, the final evaluation is carried out by the human examiner. Therefore, the human examiner should have all possibilities to adapt and document the pre-scoring as well.

4.4 First Iteration: Designing the User Interface

To visualize the user interface, first, we implemented mockups as a web-based front-end using HTML, JavaScript and CSS. The user interface is responsible for displaying the application’s data and for receiving and forwarding user input to the server. Due to the nature of the task, the application is primarily designed for use on desktop PCs. However, the use of the Bootstrap framework allows for basic compatibility with mobile devices.

As shown in Figure 3, the front-end is divided into three sections. The *assessment section* includes the assignment and the associated level of expectation. Components of the level of expectation that have been identified by the scoring mechanism or the human scorer in the respective response are marked with a green tick. Components that were not identified are shown with a red cross. The *answer section* includes the individual answers of the participants. Individual text passages can be highlighted in different colors using a highlighter function. Below the answer, the maximum score, the recommended score by the AI, and the final score are displayed. In the *comments section*, comments can be added to the colored highlights.



Exam winterterm 20/21 - Question 2 - Answer 3 / 3

Assessment

Assignment

Define Machine Learning and name its most important aspects.

Level of expectations

✓ Machine learning algorithms build models from sample data (experience)

✗ The algorithm can make decisions or predictions without being explicitely programmed to do so

✓ The computer improves its performance with additional experience (data)

Answer

Machine Learning means, that a computer takes data and creates a model to derive predictions from it. The more data it gets, the better the predictions. Machine Learning works very well with classification tasks.

Maximum score: 3

Recommended score: 2

Score:

Save score

Comments

This comment was automatically generated. The highlighted text corresponds to the first part of the level of expectations.

This comment was automatically generated. The highlighted text corresponds to the third part of the level of expectations.

Not what was asked.

Legend

Interface components can belong to one or more Design Principles as visualized by the colored highlights

Figure 3. *Screenshot of the Manual Scoring Front-end*

4.5 Evaluating the User Interface

The user interface was evaluated using a survey to determine the user experience and completeness of the application versus user expectation. Respondents with experience in the process of scoring essay questions were surveyed. As an introduction, the exemplary use was shown in a four-minute video to represent the dynamics in the process. The individual steps were explained via audio commentary. During the subsequent questioning, screenshots of the application were shown as a reminder so that the respondents could once again intensively deal with the application.

In the first part of the questionnaire, four questions were asked about each of the sub-areas, namely dashboard, course administration, exam administration, scoring interface, and the analysis of the exam results. The questions asked whether the respondents were satisfied with the respective functionality, whether the elements of the user interface were comprehensible and whether the user guidance was satisfactory. Finally, there was the opportunity to formulate further comments and improvement requests. A total of 25 questionnaires were sent out to people who regularly correct essays, of which 13 fully completed responses were received. The adjustments that were made in the second iteration are listed in section 4.6.

The User Experience Questionnaire (UEQ) by Laugwitz et al. (2008) was used to evaluate the user experience. It enables a general assessment of the user experience of the application at a superordinate level and is suitable as a good supplement to the concrete questions of the first part of the evaluation (Schrepp et al., 2014). In comparison to the UEQ benchmark (Schrepp et al., 2014; Schrepp, 2021), the results show that the application is altogether perceived as good to very good by the users. Figure 4 shows the evaluation divided into the six categories of the UEQ. The results for the user interface in four of the six categories are in the top ten percent of the benchmark. The results in the efficiency and stimulation categories are in the top 25% of the benchmark data and can thus be rated as good. Due to the small number of participants in the evaluation, the confidence intervals were also considered. These are at least in the range of above-average results for all categories considered. Only in the evaluation of stimulation a large variance and a lower expression can be observed. We assume that this is due to the nature of the activity under consideration.

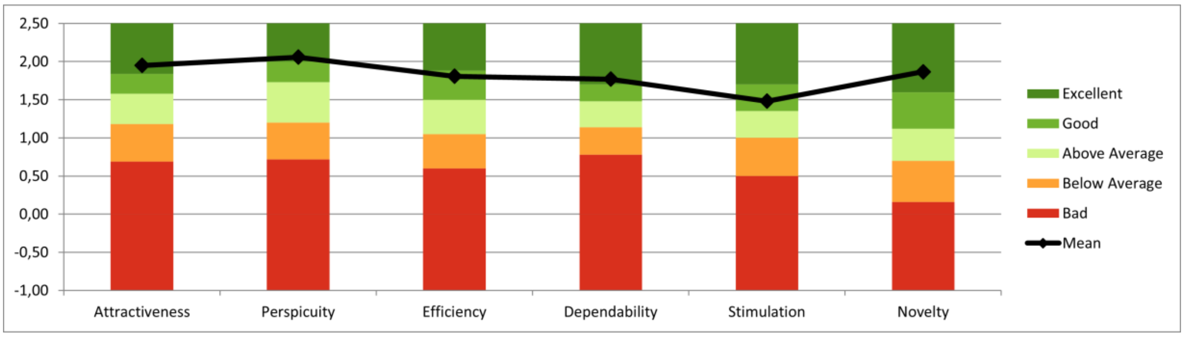


Figure 4. Results of the UEQ (n = 13)

The questionnaire closed with a question about the overall impression of the users, which was predominantly described as positive. Thus, the application was mostly described as helpful. One participant called the application “definitely an improvement because the system does pre-scoring, and you can also quickly and easily approve examiners who are subject to a secure rights concept.”

In particular, potential increases in efficiency were attributed to the overview of the scoring process and the provision of the information needed in each case. One participant mentioned that pre-scoring “is definitely an improvement to the current situation [as] many exams currently have to be written digitally or online. This makes it easier to import students’ solutions into the system without having to digitize them first.” Another participant added that the additional statistical assessment supports the scoring process by saving time otherwise needed for looking at the level of expectation.

4.6 Second Iteration: Revising the User Interface and Implementing the Scoring Mechanism

In the second implementation phase, in line with the DSR process, feedback from the evaluation is used to improve the user interface. The potential improvements identified in Section 4.5, which primarily regard the score assignment in the post-scoring assessment overview, as well as several minor improvements that display additional requested information, have been implemented. Most of the requests addressed the scoring itself.

Regarding DP 2, it was implemented that the scores given, when creating the level of expectation, are also displayed in the scoring interface next to the level of expectation. This should make it easier for examiners to assign (partial) points during post-scoring. In addition, the adjustment of the AI-based pre-scoring has been simplified. Thus, in the post-scoring assessment overview, a part of the level of expectation can now be switched between fulfilled and not fulfilled by clicking on the icon. Hereby, in terms of semi-automatic scoring, the decision of the prototype can be overridden by the human examiner in a quick and uncomplicated way. After switching, the recommendation is adjusted, and the recommended score is automatically corrected by the corresponding amount. To further support the documentation (DP3), it was also implemented that the inserted comments of the multi-level, simultaneous post-scoring are now directly assigned to the authors and that these are identified. In addition, a search function for students and examinations was created. The search is based on the user interface of the exam viewer and serves to make the documentation accessible to the students. Thus, all results of the student with the same student ID number for the selected exam are displayed and the comments can be viewed and changed during and after the scoring process. Additionally, the transfer of a score into a grade by manually specifying a grade delta was implemented. To increase the flexibility of the assignment of grades, the grade deltas can now be assigned manually in addition to the automatic mode. The technical design of the prototype can be divided into three components, the neural network, the application logic, and the user interface.

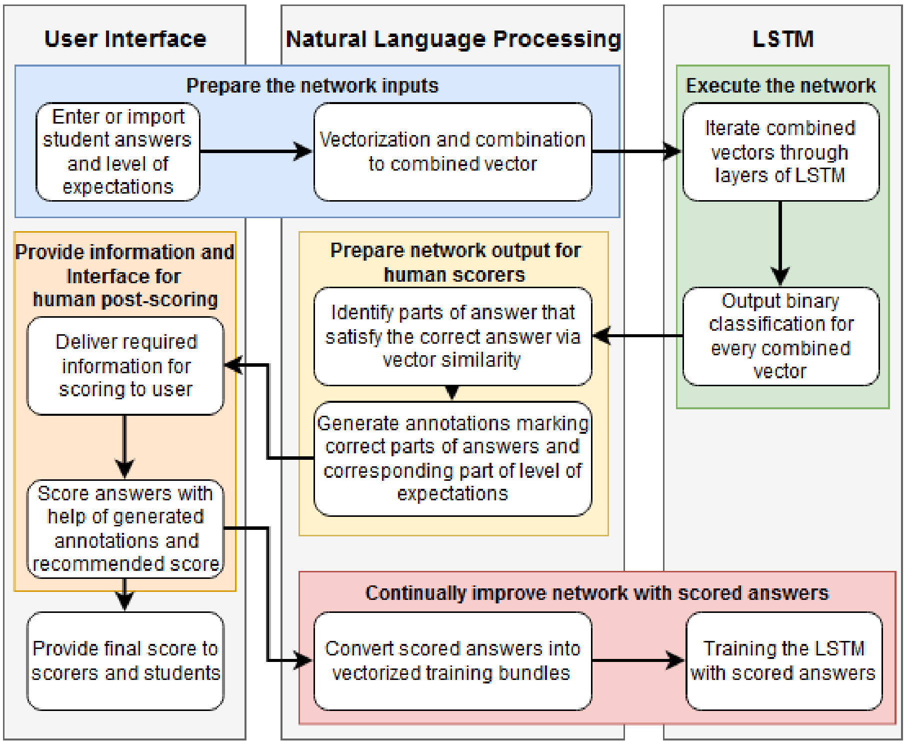

User Interface
Natural Language Processing
LSTM
Prepare the network inputs
Enter or import student answers and level of expectations
Vectorization and combination to combined vector
Execute the network
Iterate combined vectors through layers of LSTM
Output binary classification for every combined vector
Provide information and Interface for human post-scoring
Deliver required information for scoring to user
Score answers with help of generated annotations and recommended score
Provide final score to scorers and students
Prepare network output for human scorers
Identify parts of answer that satisfy the correct answer via vector similarity
Generate annotations marking correct parts of answers and corresponding part of level of expectations
Continually improve network with scored answers
Convert scored answers into vectorized training bundles
Training the LSTM with scored answers

Figure 5. Implementation of the Scoring Mechanism

Furthermore, we implemented the AI-based scoring mechanism as shown in Figure 5. The machine-learning component is responsible for generating the score suggestions of the pre-scoring (DP1) and is separated from the rest of the application logic. A Python server makes the network available to the application via an API. POST requests can be used in the local network to address the neural network and transmit the required data. The PyTorch framework and the Python programming language were used for the implementation. The vectors required for the network were calculated using the spaCy framework, which analyzes the answers to the essay questions and makes the linguistic context and other properties of natural language understandable and usable for the machine-learning component. The evaluation of exam questions in terms of points on a fixed scale poses a classification problem.

Therefore, a neural network (Recurrent Neural Network) is used for the machine learning component. The neural network is trained using past exam scorings consisting of an answer and a score (supervised learning). In addition, the algorithm should be improved by the answers scored in the application (reinforcement learning) and be able to transfer the knowledge of previous scorings to new unknown tasks (transfer learning). The specific neural network becomes a kind of Recurrent Neural Network that maintains the order of information and thus understands contextual information better (Huang and Feng, 2019). Moreover, due to their recursive nature, they can work well with inputs of different sizes and lengths (Chung et al., 2014). Since essay responses are of variable length, an LSTM is chosen as the network for this application. The implementation is done with the open-source framework PyTorch, as it offers a faster implementation as well as shorter training times than comparable frameworks (Cohen et al., 2018; Simmons and Holliday, 2019; Heghedus et al., 2019). A custom word embedding was not used since such an embedding would only represent a known set of words and the generalization to unknown words and topics would be limited.

On the server-side, the application logic is implemented using Laravel. User input is implemented and stored, and database content is retrieved. With the help of the latter, views are created for users.

4.7 Evaluating the scoring mechanism

Since the software artifact is only an improvement on the status quo if the semi-automatic scoring is of sufficient quality, several evaluations were conducted. To train the network, annotated data were needed in which the fulfillment of individual parts of the level of expectation was recognizable in an answer. Since no suitable dataset was found, we built on the Hewlett Foundation's Kaggle dataset (Kaggle, 2012), which partially satisfies the requirements. The dataset contains ten questions with approximately 17,000 responses. For three of these questions, an assignment of awarded points to individual parts of the level of expectation is possible. Annotation was done retrospectively and by hand. Scoring criteria were given for each question, and two expert point ratings were given for each answer. A total of 500 answers for each of the three questions were annotated and used for training and testing.

<u>10-fold Cross-Validation</u>

We performed a 10-fold cross-validation, using one-tenth of each of the 1,500 responses as test dataset while training the network with the remaining data. A separate network was trained and evaluated for each of the ten combinations. Figure 6 shows the quotas of the training and test data of the networks as well as the mean values of the training and test datasets.

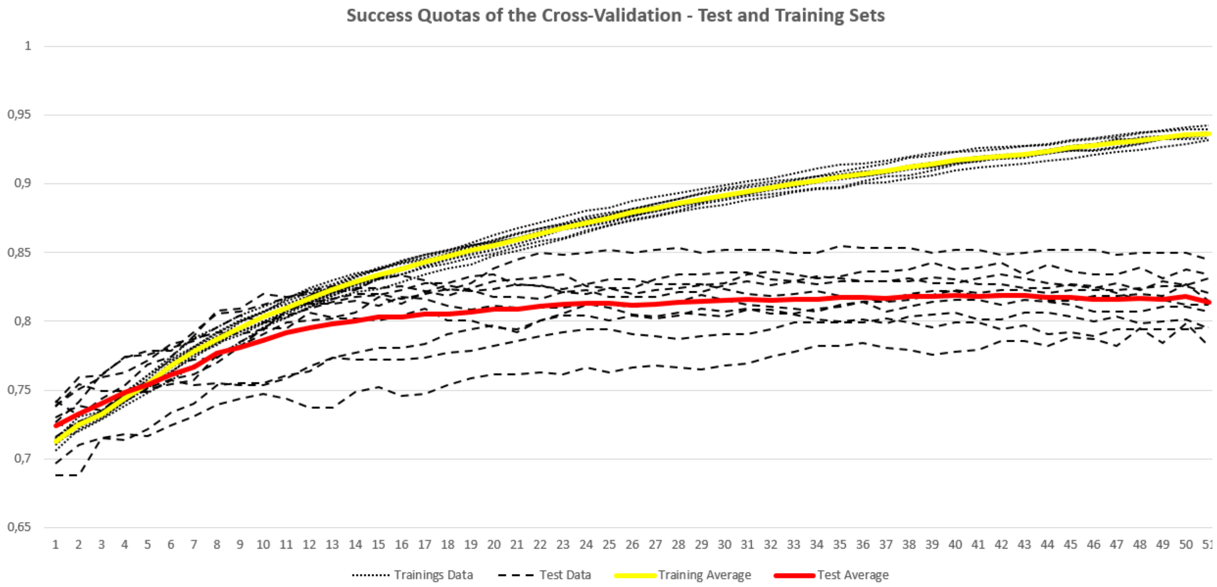


Figure 6. Success Quotas of the Cross-Validation

The results of the training data show that all training quotas are within a very narrow range of maximum two percentage points. For the quotas of the test data, a larger range was shown, with a maximum range of nine percentage points between 76% and 85%. Two outliers were observed here. The observation of the remaining test series showed a range of only five percentage points. For nine out of ten test series, a rate of over 80% was observed. As can be seen in the figure, there is a convergence of the mean value at 81%, which was achieved in most test series between epoch 20 and 23.

Suitability Out-of-Domain Data

Since the trained network will also be used for future unknown essays, its suitability is also evaluated using out-of-domain data. These present a new challenge for the network since both the responses and the level of expectation are unknown. For this purpose, in addition to the existing data, another manually created and comparable dataset was used. This consists of 100 annotated responses to a question with a three-point level of expectation. In this validation, the network achieved a rate of correct assessments of 48%. The result is due to the fact that 48% of the answers in the dataset were incorrect, and all answers were rated to be incorrect by the network. Therefore, the generalization capability of the network can be considered to be strongly limited.

Learning Samples Needed to Adapt to Out-of-Domain Data

In order to increase generalizability, the network already trained with the original data was trained several times with the new dataset of 100 annotated responses. In each case, a portion of the data was used as training dataset and the proportion was incrementally increased. Similarly, the network was trained with this training dataset in an incremental number of epochs. Before calculating the quotas of correct estimates of a combination, the network was reset to its original state. For this rate, the network had to estimate the remaining responses not used as training data. The calculations were performed using a five-fold cross-validation. The results show that a minimum of about 30 learning phases, arbitrarily combined of answers and epochs, are necessary before improvement occurs. For a level of 70%, a minimum of 22 new responses and 19 epochs were required. For a level of 75%, at least 37 answers and 14 epochs were necessary. The number of available answers was more important than the number of epochs. While the quota rose steadily with a constant number of epochs but more answers, it stagnated conversely from a number of 15 to 20 epochs.

4.8 Summarizing and Documenting the Design Knowledge

In the first iteration, the focus was on the front-end and the basic functionalities for examiners. The evaluation results show that the user experience was rated as good to excellent according to the UEQ dimensions. Only minor changes were required, which were implemented in the second iteration. Since these changes only addressed the implementation of DP2 and DP3, but not the design principles themselves, we end the DSR process for these two design principles. The second iteration primarily focused on the implementation and subsequent evaluation of the scoring mechanism. According to DP1, the assessment should be based on a level of expectation and serve as a pre-scoring. Satisfactory results were achieved, particularly in the 10-fold cross-validation. For the use with out-of-domain data only worse results could be observed, so that a training with learning samples was necessary. We were able to show that the machine-learning approach is suitable for scoring essays. The accuracy of the algorithm can be significantly improved by prior training. Since the design of the algorithm was not the first priority in our project, the knowledge gained is sufficient for us to also end the DSR process for the first design principle. The following documentation and communication of the results of the design process is based on the components of the design principles schema according to Gregor et al. (2020). Table 1 shows the derived design principles.

Design Principle Title	**Principle of…**		
	…machine-learning-based pre-scoring (DP 1)	**… manual, simultaneous post-scoring (DP 2)**	**…documenting the post-scoring process (DP 3)**
Aim, and user	To support examiners in the scoring of essays …		
Context	… in exams with a large-scale educational context …		
Mechanism	… provide a machine-learning-based scoring mechanism that is able to generate automated pre-scoring drafts based on transparent evaluation criteria…	… provide an easy-to-use user interface for manual post-scoring that uses the pre-scoring (DP1) as basis to reduce the workload …	… provide a traceable and transparent scoring process by documenting the examiners decisions and the underlying evaluation criteria …
Rationale	… because this can help examiners to reduce the scoring workload while increasing the objectivity of essay scoring.	… because this can help to improve the overall scoring quality, required workload, students' acceptability of exam results, and acts as a manual verification step of the diagnostic quality of the machine-learning-based pre-scoring mechanism.	… because this can increase the transparency of the scoring process, might be mandatory for providing an explainable scoring process, and can be helpful in communicating the scoring results to the students.

Table 1. Documentation of the Design Principles based on Gregor et al. (2020)

In DP1, a machine-learning-based, transparent pre-scoring of essays according to the level of expectation was formulated. Hereby, a more efficient and objective scoring process should be ensured. DP 2 addressed the provision and support of manual, simultaneous post-scoring. The human post-scoring is intended to ensure the diagnostic quality of the whole scoring process. DP3 dealt with documentation in the post-scoring process. This was intended to promote transparency in scoring and is particularly relevant for the communication with students.

5 Discussion and conclusion

The goal of the research project was to derive design principles for a semi-automatic scoring system for essay tasks. In particular, the examiners should be supported in mastering the challenges in the execution of the defined user stories described by the scoring process of the essay tasks. The challenges cover the demands placed on the scoring of an essay by different stakeholders. To achieve this goal, the first step was to derive requirements and design principles based thereon. The implementation in our software artifact was done in an iterative process according to the DSR approach of Hevner et al. (2004) and Peffers et al. (2007). The user interface and the functions were implemented and then evaluated. We showed that potential users were largely satisfied with the functionalities and the user interface and that the artifact can provide additional value for the scoring. Additionally, a technical evaluation of the machine-learning algorithm was carried out, since the added value of the artifact only arises if there is an improvement on the manual scoring process. For the technical evaluation, a modified Kaggle dataset was used. In a first 10-fold cross-validation, it could be shown that on average a convergence of 81% (20 epochs) takes place. For the use with out-of-domain data, no satisfactory results could be achieved at first. A further training was therefore carried out, achieving a success quota of roughly 75% (30 epochs). In addition to the software artifact, we also contributed design knowledge to the scientific knowledge base. The systematic documentation of this knowledge was done in our last step of the DSR process using the structure of Gregor et al. (2020).

The derived design knowledge can not only be applied to the specific use case, but can also be seen as a generalizable basis for comparable software artifacts with different levels of automation and other (semi-) open task types. Thus, for a deviating level of automation, only the share of decisions to be made manually has to be adapted (Frohm et al., 2008). Furthermore, it could be shown that with the help of the machine-learning component, freely formulated answers can be evaluated correctly to a high degree. This can also be transferred to semi-open answers by adapting the respective design of the level of expectation. Furthermore, the design knowledge about the user case can be used in teaching-learning arrangements where essay tasks are used in the context of diagnostic or formative assessments. In this way, feedback can be given to participants immediately after answering an essay task, thus improving error reflection and the learning process. Due to the possibility to transfer our generated design knowledge, our research does not only provide level 1 DSR contribution but also level 2 DSR contribution (Gregor and Hevner, 2013). In addition, we could show that even with a manageable training effort, especially for out-of-domain data, a good result in the pre-scoring could be achieved and examiners can be supported in the scoring process in practice.

The quality of the scoring depends on the volume and quality of the available data. Especially in the search for suitable training data, we have shown that this cannot be taken for granted, even in a data-driven era. Although extensive data are available, there are requirements concerning different key attributes. Since our neural network evaluates the fulfillment of parts of a level of expectations, data must be available that allow a direct correlation between points and parts of the level of expectation. To learn this connection, the network needs data in which it is annotated which part of the level of expectation is fulfilled by the respective answer. In addition, we were able to show in an out-of-domain context that good but only limited use can be made of existing training data for interdisciplinary use. Here, the use of subject-specific networks could be a solution. In addition, the essay task in our scenario had exactly one correct answer. Thus, the prototype is primarily suitable for questions that serve knowledge assessment. Tasks in which a scenario-based subjective evaluation must be carried out also require consideration of the argumentation structures within the decision-making process of the examinees. The added value increases with the number of participants and is probably not suitable for smaller courses. Typical scenarios can be courses in which several hundred students participate or which are repeated identically on a regular basis. Thus, a semi-automatic scoring system can support the execution of the aforementioned user stories especially for large events by reducing the time needed for the scoring and documentation of essay exams. At the same time, it facilitates the fulfillment of the requirements for the scoring of exams by transparently standardizing the scoring and implementing a reduction of human bias. However, the extent to which a system described is actually used depends on other additional factors besides the design principles. For example, potential efficiency benefits can only be realized if both examinees and examiners trust the system and thus use it (Wu et al., 2011).

6 References

American Educational Research Association (2011). *Report and recommendations for the reauthorization of the institute of education sciences.* Washington D.C.: American Educational Research Association.

Ashoori, M. and J. D. Weisz (2019). *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes*. URL: http://arxiv.org/pdf/1912.02675v1 (visited on 03/25/2022).

Attali, Y. and J. Burstein (2006). “Automated Essay Scoring With e-rater® V.2” *Journal of Technology, Learning, and Assessment* 4 (3).

Balfour, S. P. (2013). “Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™” *Research & Practice in Assessment* 8, 40–48.

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Birenbaum, M., K. K. Tatsuoka and Y. Gutvirtz (1992). “Effects of response format on diagnostic assessment of scholastic achievement” *Applied psychological measurement* 16 (4), 353–363.

Bloom, B. S., M. D. Engelhart, E. J. Furst, W. H. Hill and D. R. Krathwohl (1956). *Taxonomy of educational objetives: the classification of educational goals. Handbook I: cognitive domain.* New York, US: David McKay Co Inc.

Castellanos-Nieves, D., J. T. Fernández-Breis, R. Valencia-García, R. Martínez-Béjar and M. Iniesta-Moreno (2011). "Semantic Web Technologies for supporting learning assessment" *Information Sciences* 181 (9), 1517–1537.

Chen, Y. Y., C. L. Liu, C. H. Lee and T. H. Chang (2010). "An unsupervised automated essay-scoring system." *IEEE Intelligent systems* 25 (5), 61–67.

Chung, J., C. Gulcehre, K. Cho and Y. Bengio (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling.* arXiv:1412.3555.

Cohen, Y., E. Levi and A. Ben-Simon (2018). "Validating human and automated scoring of essays against "True" scores" *Applied Measurement in Education* 31 (3), 241–250.

Crawford, J., K. Butler-Henderson, J. Rudolph, B. Malkawi, M. Glowatz, R. Burton, P. Magni and S. Lam (2020). "COVID-19: 20 countries' higher education intra-period digital pedagogy responses" *Journal of Applied Learning & Teaching* 3 (1), 9-28.

Foltz, P. W., D. Laham and T. K. Landauer (1999). "The Intelligent Essay Assessor: Applications to Educational Technology" *EdMedia+ innovate learning*, 939–944.

Frohm, J., V. Lindström, M. Winroth and J. Stahre (2008). "Levels of automation in manufacturing" *Ergonomia - International Journal of Ergonomics and Human Factors* 30 (3), 1–28.

Gregor, S. and A. R. Hevner (2013). "Positioning and presenting design science research for maximum impact" *MIS Quarterly* 37 (2), 337–355.

Gregor, S., L. Kruse and S. Seidel (2020). "Research Perspectives: The Anatomy of a Design Principle" *Journal of the Association for Information Systems* 21 (6), 1622–1652.

Heghedus, C., A. Chakravorty and C. Rong (2019). "Neural network frameworks. comparison on public transportation prediction" *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 842–849.

Hevner, A. R., S. T. March, J. Park and S. Ram (2004). "Design Science in Information Systems Research" *MIS Quarterly* 28 (1), 75–105.

Hewlett, C. and A. Kahl-Andresen (2014). "Prüfungsökonomie statt Prüfungsqualität?" *Berufsbildung in Wissenschaft und Praxis* 14 (3), 6–9.

Huang, J. and Y. Feng (2019). "Optimization of recurrent neural networks on natural language processing" *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, 39–45.

Hung, S.-L., I.-F. Kwok and R. Chan (1993). "Automatic programming assessment" *Computers & Education* 20 (2), 183–190.

Impey, C. and M. Formanek (2021). "MOOCS and 100 Days of COVID: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic" *Social sciences & humanities open* 4 (1).

Kaggle (2012). *The Hewlett Foundation: Short Answer Scoring - Develop a scoring algorithm for student-written short-answer responses.* URL: https://www.kaggle.com/c/asap-sas/overview (visited on 03/25/2022).

Kelly, A. P. and R. Columbus (2020). *College in the Time of Coronavirus: CHALLENGES FACING AMERICAN HIGHER EDUCATION.* American Enterprise Institute for Public Policy Research. URL: https://www.aei.org/wp-content/uploads/2020/07/College-in-the-Time-of-Coronavirus.pdf (visited on 03/25/2022).

Kumar, V. and D. Boulanger (2020). "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value" *Frontiers in Education* 5.

Lai, C.-Y. and W.-C. Liou (2010). "Implementation of e-learning and corporate performance: An empirical investigation" *International Journal of Advanced Corporate Learning* 3 (1), 4–10.

Laugwitz, B., T. Held and M. Schrepp (2008). "Construction and evaluation of a user experience questionnaire" *Symposium of the Austrian HCI and usability engineering group*, 63–76.

Lee, J. D. and K. A. See (2004). "Trust in automation: designing for appropriate reliance" *Human factors* 46 (1), 50–80.

March, S. T. and G. F. Smith (1995). "Design and natural science research on information technology" *Decision support systems* 15 (4), 251–266.

Mitchell, T., N. Aldridge and P. Broomhead (2003). "Computerised marking of short-answer free-text responses" *Manchester IAEA conference*.

Mohri, M., A. Rostamizadeh and A. Talwalkar (2018). *Foundations of Machine Learning.* Second edition. Cambridge, MA: The MIT press.

Pearson (2019). *PTE Academic Automated Scoring White Paper*. *Pearson Test of English Academix: Automated Scoring.* Pearson Education Ltd. URL: https://pearsonpte.com/wp-content/uploads/2018/06/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf (visited on 02/13/2022).

Peffers, K., T. Tuunanen, M. A. Rothenberger and S. Chatterjee (2007). "A Design Science Research Methodology for Information Systems Research" *Journal of Management Information Systems* 24 (3), 45–77.

Ramesh, D. and S. K. Sanampudi (2021). "An automated essay scoring systems: a systematic literature review" *Artificial Intelligence Review,* 1–33.

Richardson, M. and R. Clesham (2021). "Rise of the machines? The evolving role of AI technologies in high-stakes assessment" *London Review of Education* 19 (1), 1–13.

Schrepp, M. (2021). *Data Analysis Tools - Two Excel-Sheets that make the analysis of your results easy*. URL: https://www.ueq-online.org/Material/Data_Analysis_Tools.zip (visited on 03/25/2022).

Schrepp, M., A. Hinderks and J. Thomaschewski (2014). "Applying the user experience questionnaire (UEQ) in different evaluation scenarios" *International Conference of Design, User Experience, and Usability*, 383–392.

Sharma, A. and D. B. Jayagopi (2018 - 2018). "Automated Grading of Handwritten Essays". In: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*: IEEE, pp. 279–284.

Simmons, C. and M. A. Holliday (2019). "A comparison of two popular machine learning frameworks" *Journal of Computing Sciences in Colleges* 35 (4), 20–25.

Taghipour, K. and H. T. Ng (2016). "A neural approach to automated essay scoring" *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1882–1891.

UNESCO (2020). *Global monitoring of school closures caused by COVID-19*. URL: https://en.unesco.org/covid19/educationresponse (visited on 03/25/2022).

Wu, K., Y. Zhao, Q. Zhu, X. Tan and H. Zheng (2011). "A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type" *International Journal of Information Management* 31 (6), 572–581.

Yusuf, N. and N. Al-Banawi (2013). "The impact of changing technology: The case of e-learning" *Contemporary Issues in Education Research (CIER)* 6 (2), 173–180.

Zhao, S., Y. Zhang, X. Xiong, A. Botelho and N. Heffernan (2017). “A memory-augmented neural model for automated grading” *Proceedings of the fourth (2017) ACM Conference on Learning @ Scale*, 189–192.