Students’ Trust in Automated Grading Through Explainable AI Visualizations.

VLAD-GABRIEL STOIAN, University of Twente, The Netherlands

Currently, Artificial Intelligence (AI) has a big impact on people’s life. This technology evolved more and more every year, and now it is used worldwide, across most of the sectors. For example, its utilizations can be seen in transportation, logistics, medicine, military and also education. At present, different types of AI-based tools are developed for enhancing students’ experience in higher education. But, due to the increasing complexity of algorithms, lack of transparency can be observed, especially in AI-based tools used for summative assessments. Thus, it is believed that Explainable AI might play an important role in addressing this issue. The aim of this research is to discover what Explainable AI methods enhance the transparency of results in an AI-based grading tool and which of them are the most trusted by students. This will be done through literature review and interviews with university students. Furthermore, a clear overview of the selected Explainable AI methods will be provided with mock-ups. They will also be used to illustrate examples in the interviews. This research provides new insights for a major project at the University of Twente (an AI-grading tool).

Additional Key Words and Phrases: Explainable AI, Visualizations, Explanations, Automated Grading, Higher Education, Trust, Transparency

1 INTRODUCTION

In higher-education, the assessment of students is one of the most critical, important and also challenging processes for examiners. This process can be defined as “the mechanics or steps required to effectuate a judgement” [28]. Furthermore, this process is divided into two types: the process of formative assessments and the process of summative assessments. The latter one is the focus of this research. It can be defined as “a judgement which encapsulates all the evidence up to a given point” [28].

Currently, the number of enrolled students in universities is exponentially increasing every year and, for teachers, time that is left for assessing students becomes more limited. Due to these aspects, the examiners try to make use of the latest Artificial Intelligence (AI) technologies to automate the summative assessment process and make it more efficient. This is the main reason that justifies the development and implementation of AI-based grading tools.

However, the adoption of assistive AI systems may be limited by the distrust of humans in the predictions, due to the lack of transparency [25]. One of the emerging methods that addresses this issue and also helps to increase stakeholders’ trust in AI-based systems is Explainable AI (XAI), which produces visualizations and explanations of the decisions made by the AI-system [15].

This research aims to discover what types of visualizations and explanations are suitable for an AI-based grading system and which of them students find the most trustworthy.

TScIT 39, July 7, 2023, Enschede, The Netherlands

2 PROBLEM STATEMENT AND RESEARCH QUESTIONS

An AI-based grading tool brings plenty of benefits to examiners and students. Firstly, this type of tool can significantly reduce the time spent in grading. A trained machine learning model can review answer sheets in 90% less time than teachers [14]. Secondly, it can provide unbiased and high quality feedback to students.

However, due to the black-box structure of most systems, it is hard for the end user to interpret their autonomous decisions. This interpretation is very important for a student who wants to know how the grading has been done and what the feedback they got is based upon. In this case, explanations are essential to understand and trust the system [12].

Based on the problem statement, a main research question is defined and further analyzed in this research:

RQ1: What Explainable AI methods are suitable to enhance transparency of results in an AI-based grading tool in higher-education?

Considering the importance of transparency and feedback for students, their input is necessary for creating a more accurate overview. So, another sub-question is defined:

RQ2: What methods do students find the most trustworthy?

3 RELATED WORK

In order to gather relevant literature for this research, a systematic search has been performed on the following domains: Google Scholar, Semantic Scholar, Scopus, Web of Science, IEEE Xplore. Considering that the first two domains do not make use of query searching, the results are quite broad. However, these tools are useful for creating a clear picture of relevant literature in this specific area and for collecting some starting papers. Furthermore, in-depth query searching has been performed on the last three domains. Also, the useful references from some papers have been analyzed as well. The search has been performed using the following keywords: “Explainable AI”, “Visualizations”, “Explanations”, “AI tools”, “Higher Education”, ”Grading”, “Students/Undergraduates”. The results are grouped in different research streams.

The first stream contains literature about general applicability of Explainable AI and how it is linked with different AI-systems. Research indicates that an AI-system can make decisions based on a different set of techniques such as: Machine Learning (ML), Natural Language Processing (NLP), Deep Neural Networks (DNN). The “black-box” aspect makes the interpretation of results almost impossible, leaving room for methods to solve this issue. Explainable AI represents a set of tools and methods that allows end-users to understand and trust the AI-algorithms and how they performed the decision-making by adding layers of accuracy, fairness and transparency [6, 12, 22, 25, 30].

The second stream consists of literature about direct applicability of Explainable AI methods in the educational sector. Given the importance of feedback in students' learning cycle [13], a framework for Explainable AI in education has been already established by research in this field. The XAI-ED framework consists of six different key elements: stakeholders, benefits, approaches for showing explanations, used classes of AI-models, human-centered UI and pitfalls [15]. Research shows that this framework facilitates the implementation of XAI in AI-based tools across this sector [3, 15]. However, the focus is on a general picture of different types of AI-based tools that facilitate both processes of formative and summative assessments [3, 9, 14, 15]. Furthermore, there is a lack of input from students regarding which of these Explainable AI methods may address the issue of transparency, especially in an AI-tool for summative assessments.

Thus, this research will fill in the gap in literature related to applicable Explainable AI methods in an AI-based grading tool and students' perspectives about the extent to which the selected methods can enhance the transparency of the tool.

4 METHODOLOGY AND APPROACH

This section gives an overview of the methodology that is used in conducting this research and the general approach.

In order to perform this study, a qualitative approach has been chosen. This type of research aims to develop an understanding of humans' perspectives by gathering in-depth knowledge [11]. Thus, relevant data regarding students' beliefs and opinions about different Explainable AI methods are gathered and further analyzed.

4.1 Performing Literature Review

The base of the research is the literature review. An extensive literature review is performed in order to gather relevant knowledge about different types of Explainable AI methods that can be incorporated in an AI-based grading tool used in higher-education with the purpose of enhancing transparency for students. This review helps in answering the first research question by analyzing the selected techniques. Based on this knowledge, mock-ups that illustrate the selected techniques are created. These mock-ups exemplify students' point of view in an AI-based grading tool.

4.2 Conducting Interviews

For the second part of this research, semi-structured interviews with students are conducted. This type of interviews "are characterized by open-ended questions and the use of an interview guide in which the broad areas of interest are defined" [2]. Also, semi-structured interviews do not restrict students' freedom in answering questions. This method helps the researcher in gathering in-depth knowledge and insights about students' opinions and beliefs. The questions are built based on the analyzed literature and the mock-ups created in the previous step. Nevertheless, the participants are able to choose how the interviews are conducted: physically or online. If a participant preferred an online interview, a safe platform has been used (MS Teams). After conducting the interviews, the gathered data is analyzed. Lastly, conclusions that help answering the second research question are drawn. In order to ensure the ethical aspect of the interviews, a request containing all of the relevant details about the chosen approach to conduct the interviews is approved by The Ethics Committee of University of Twente. The participants are verbally informed about the study beforehand but they also receive a written overview of the study in order to ensure full transparency and avoid misunderstandings. Furthermore, participants had to sign a consent form if they wanted to take part in this research. Their consent can be withdrawn at any time and their data will be deleted and never used anymore. Also, their answers are anonymous.

5 LITERATURE REVIEW

In this section, an answer to the main research question is found by analyzing relevant literature about the purpose of Explainable AI (Section 5.1), its importance in the educational sector (Section 5.2) and literature about different Explainable AI techniques (Section 5.3). Based on this, specific techniques that can be applied to an AI-based grading tool are illustrated by mock-ups and further used during this research.

5.1 Explainable AI and Transparency

The field of Artificial Intelligence became more and more complex. This increase in the complexity of the algorithms brings to light the necessity of explanations and the enhancement of transparency. Based on research in this field, Explainable AI can address this issue. This solution provide explanations, interpretations and insights about how the AI algorithm has performed the computations [3, 6, 9]. Researchers have proved that explanations are useful for "opening the black box" [24], proof that has further implications and advantages for the stakeholders: improving the understanding of the algorithm, increasing trust, enhancing the acceptance of the outcomes, facilitating the decision-making process based on the outcomes [3]. Moreover, there is an overlap in the literature about the general goals of Explainable AI. Fiok states that the general goals are: transparency, causality, privacy, fairness, trust, usability and reliability [9]. Even if some elements differ from research to research, the general concept is the same.

However, not all of the AI-systems need the same level of explanation for the stakeholders to be able to understand the outcomes. The inner complexity of the systems is different, depending on which type of algorithm or method is used. Research shows that "explainability of a machine learning model is usually inverse to its prediction accuracy- the higher the prediction accuracy, the lower the model explainability" [30]. For example, decision trees and support vectors machines have a high level of explainability, but they are lacking accuracy in results. In the other extreme, deep learning methods such as deep neural networks (DNN) achieve a high-level of accuracy but the ""black-box" aspect does not offer room for any interpretations [30]. Arguably, the term "stakeholders" is too broad in this context to remove the necessity of explanations even for the models that have a high degree of explainability. Developers or technical persons are able to understand the reasoning behind the algorithm but the focus of this research is on students who may have not any knowledge in the field of Artificial Intelligence. For them, transparency is a key factor during the summative assessment process. However, the importance of Explainable AI in education is analyzed in Section 5.2.

Moreover, the concept of Explainable AI can be divided in two categories: transparency design and post-hoc explanations [19, 30]. Transparency design, as the name suggests, helps in "understanding how the mechanism by which the model works" [19] at different levels: entire model (simulatability), individual components (decomposability), and the level of the trained algorithm (algorithmic transparency) [19]. These techniques reveal the functionality of the algorithm from developers' perspective [30]. In contrast to this approach, post-hoc interpretability offers valuable information to the end-users by text explanations, visualizations, local explanations, explanations by examples, even if it does not entirely elucidate how a model works [19]. In this way, "opaque models can be interpreted after the fact, without sacrificing predictive performance" [19]. However, even if transparency design is better known to offer meaning for technical stakeholders, it should not be neglected because, from students' perspective, these techniques can be considered a proof for the way they have been graded by the AI-based grading tool, as well as feedback providers.

5.2 Explainable AI in Education

Education is one of the sectors in which extra care should be considered when designing and implementing AI-based tools. The need for intelligent tools that can assist or even replace some of the human processes performed in higher-education has risen due to the increasing number of students and time-constraints [14]. However, adopting AI-based tools is not an easy task to do because it should meet some ethical requirements in order to be accepted and trusted by the population, especially if these tools are built for summative assessment processes. Besides ensuring that ethical requirements are met, research shows that feedback is one of the most important factors for enhancing students' learning process [13, 28]. Feedback can be defined as "information provided by an agent regarding aspects of one's performance or understanding" [13]. Besides its main purposes (development of domain knowledge, skills and a sense of being), feedback is "seen as a relational process through which teachers may encourage positive motivation and help learners build confidence and self-esteem" [15]. Nevertheless, if the human-agent is replaced by a computer-based agent, the transmission of similar feedback should be facilitated in order to ensure at least the same level of transparency. Research is criticizing the automation in education due to the lack of feedback that can be discouraging for some individuals [8, 27]. In this regard, Explainable AI can facilitate the implementation and acceptance of automated grading tools in higher education.

However, for taking into consideration all of the human needs, preferences, ways of learning and teaching, research suggests that Explainable AI in education "draw insights and best practices from the fields of AI, Human-Computer Interaction and the interdisciplinary and emerging field of Human-Centred AI" [15]. For this purpose, [15] established the XAI-ED framework. This framework aims to contribute to the development of effective Explainable AI educational systems by considering the following six dimensions: stakeholders, benefits, approaches for showing explanations, used classes of AI-models, human-centered UI and pitfalls. Thus, the XAI-ED framework serves as guidance tool during this research. Therefore, Section 5.3 analyzes different techniques for showing explanations that can be implemented for specific AI-models.

5.3 Explainable AI Methods

In this section, three Explainable AI methods are analyzed with the goal of establishing if they are suitable for enhancing the transparency of an AI-based grading tool used in higher education.

5.3.1 *Confidence Measures.*

Confidence measures can be defined as "measures that provide an expectation that an advice will prove to be correct" [29]. Thus, the scope of these measurements is to provide the end-user with relevant information about how accurate the results are. To facilitate the transparent implementation of accurate confidence measures, the Interpretable Confidence Measure (ICM) framework has been established by [29]. This framework assumes "that a confidence measure should be: accurate, able to explain a single confidence value, use a transparent algorithm and provide confidence values that are predictable to humans" [29]. In the field of Machine Learning, many confidence measures are used (confusion metrics, prediction score, rescaling, probability, voting) but they do not meet the requirements of the framework. This is due to the fact that "the purpose of these measures is to convey performance of a Machine Learning model to a developer, not the confidence of the system in an advice to a user" [29]. Thus they do not belong to the category of post-hoc interpretability due to their model-agnostic approach. Confusion metrics is the only measure that is built on a system-agnostic approach but it also lacks of accuracy and explainability [29].

However, the ICM framework "relies on a system-agnostic approach and performs a regression analysis with the correctness of and advice as the regressor. It does so based on case-based reasoning" [29]. Considering the fact that case-based reasoning has the k-Nearest Neighbours algorithm as basis [10], this method can easily be implemented in an AI-Grading tool. It can compare the students' answers with the k most similar answers from the data set and "assign the case with a weighted aggregation of the neighbour's labels" [29]. Then, the confidence measure resulted from the regression analysis will be displayed, leading to the enhancement of transparency of the algorithm. The students will be able to understand the accuracy of the algorithm that automatically graded their exam. Furthermore, research shows that case-based learning methods also allow for example-based explanations [7], which can enhance the transparency of the tool even more. Thus, on top of the confidence level of the algorithm, a similar answer can be displayed and the student will be able to understand the grade by means of comparison.

5.3.2 *Local Interpretable Model-Agnostic Explanations.*

As discussed previously, trusting a prediction is a real problem in the field of Artificial Intelligence. However, [23] proposes a solution that can address this issue. Local Interpretable Model-Agnostic Explanations (LIME) is "an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model" [23]. To make the definition clearer, the researcher explains what exactly "explaining a prediction" means in this context. It means "presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components and the model's prediction" [23]. This "understanding" is represented by a list of explanations that reflect the contribution of the features to the prediction as output. Furthermore, this method is based on some principles. Firstly, the explanations should be interpretable in order to "provide qualitative understanding between the input variables and the response" [23]. Secondly, the explanation should be "at least locally faithful, i.e. it must correspond to how the model behaves in the vicinity of the instance being predicted" [23]. Lastly, "an explainer should be able to explain any model, and thus be model-agnostic" [23]. However, the feature importance scores produced by LIME do not provide enough understanding for students so, the transparency is not enhanced. In order to display these scores in a user-friendly manner, a visualization technique is needed. Saliency "has been primarily used to visualize the importance scores of different types of elements in XAI learning systems, such as highlighting words in input text" [4]. Thus, these two techniques can be used in junction to highlight words in students' answers. The relevance of the words can vary and it can be represented by different tones of color. In this way, the students can understand which words in their answer are more relevant and which words are less relevant compared with the correct answer that has been used in the data set. This can also be viewed as a justification of the grade for that specific question.

5.3.3 Concept Activation Vectors. Another method that can be used for enhancing the transparency and interpretation of deep learning models are Concept Activation Vectors (CAV). A concept activation vector (CAV) "provides an interpretation of a neural net's internal state in terms of human-friendly concepts" [16]. Thus, "a CAV for a concept is simply a vector in the direction of the values (e.g. activations) of that concept's set of examples" [16]. This method is also based on relevance scores. Using each individual score that resulted from breaking down the text into words, a high-dimensional vector is created. This vector is useful for displaying which pieces of information are "activated" based on the scores. So, this method can be used to highlight the "activated" words in students' answers. They will be able to understand which are the key terms that are also found in the correct answer provided to the model and how these key terms contributed to the number of points received for answering a specific question.

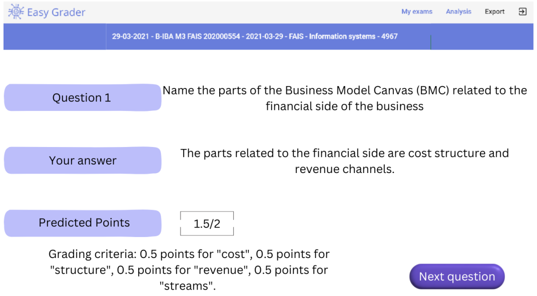
6 ILLUSTRATIONS

In this section, the Explainable AI methods which have been analyzed during the literature review are illustrated by mock-ups. They have been designed using the theme of the Easy Grader software. This software is currently under development and testing by researchers at the University of Twente. These mock-ups help in visualizing how the methods are used and displayed in this specific software. The XAI-ED framework that has been discussed in Section 5.2 was taken into consideration. For this research, the main stakeholders are students. The explanations should be intuitive for their level of understanding and the user interface is human-centered.

All of the mock-ups have the same question as example. The question is: "Name the parts of the Business Model Canvas (BMC) related to the financial side of the business". The correct answer is: "The parts related to the financial side are cost structure and revenue streams" [21]. However, the examples provided as students' answers are not entirely correct, having half a point deduction. The purpose of this mistake in the answer is to illustrate how the Explainable AI methods cope with this situation.

6.1 Answer Key

Figure 1 illustrates the output of the software without any Explainable AI method added. This can be considered as an extra mock-up that is useful during the interview phase because it illustrates the current way of working. However, the only difference is that the algorithm is grading the question whilst the teacher provides the grading criteria. This is a general scheme of how this specific question is graded for everyone sitting the exam. The score can be seen in the box below the answer. This specific deduction of points is due to using the word "channels" instead of "streams". Otherwise, the student would have received the maximum number of points because the value of each key word is half a point.



29-03-2021 - B-IBA M3 FAIS 202000554 - 2021-03-29 - FAIS - Information systems - 4967

Question 1: Name the parts of the Business Model Canvas (BMC) related to the financial side of the business

Your answer: The parts related to the financial side are cost structure and revenue channels.

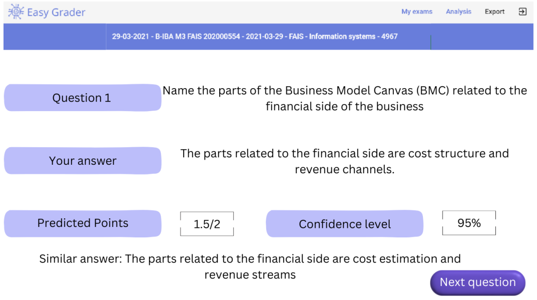
Predicted Points: 1.5/2

Grading criteria: 0.5 points for "cost", 0.5 points for "structure", 0.5 points for "revenue", 0.5 points for "streams".

Fig. 1. Answer Key

6.2 Confidence Level and Alternative Answer

Figure 2 illustrates the output of the software using the method discussed in Section 5.3.1. Besides the score that was predicted by the algorithm, the computed confidence level and an alternative answer are provided.



29-03-2021 - B-IBA M3 FAIS 202000554 - 2021-03-29 - FAIS - Information systems - 4967

Question 1: Name the parts of the Business Model Canvas (BMC) related to the financial side of the business

Your answer: The parts related to the financial side are cost structure and revenue channels.

Predicted Points: 1.5/2

Confidence level: 95%

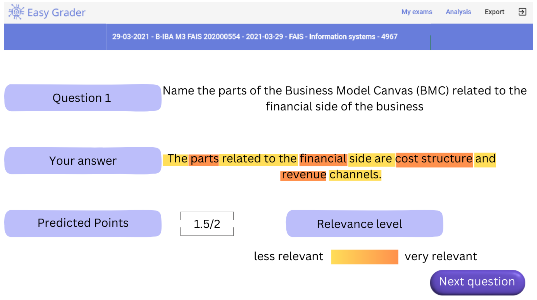
Similar answer: The parts related to the financial side are cost estimation and revenue streams

Fig. 2. Confidence Level and Alternative Answer

The confidence level gives an assurance to students regarding algorithm's reliability. In this case, students can be sure that the algorithm is 95% reliable for this specific question. This high percentage is the confirmation of a correct behavior. Furthermore, an alternative answer is provided. However, the answer is not a correct one. It is an answer that is represented by the same amount of points. By this means, students will be able to understand why they got that specific deduction of points while comparing both answers.

6.3 Highlighting Based on Relevance

Figure 3 illustrates the output of the software using the method discussed in Section 5.3.2. Each word in students' answer is highlighted in order to display the relevance of each instance.



29-03-2021 - B-IBA M3 FAIS 202000554 - 2021-03-29 - FAIS - Information systems - 4967

Name the parts of the Business Model Canvas (BMC) related to the financial side of the business

The parts related to the financial side are cost structure and revenue channels.

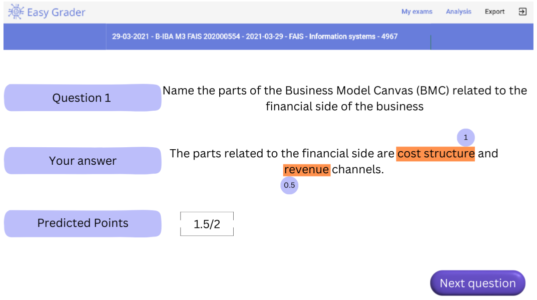
1.5/2

Fig. 3. Relevance Highlighting

As it can be observed, the nuance of the color differs by the level of relevance. The words which are highlighted with orange are detected key words that are linked with the correct answer. The word "channels" is wrong in this context (however, it is marked as less relevant because it is not considered a correct key word for this question). So, this method is a visual proof and justification for a specific grade.

6.4 Words Activation

Figure 4 illustrates the output of the software using the method discussed in Section 5.3.3. The "activated" words are highlighted in students' answers.



29-03-2021 - B-IBA M3 FAIS 202000554 - 2021-03-29 - FAIS - Information systems - 4967

Name the parts of the Business Model Canvas (BMC) related to the financial side of the business

The parts related to the financial side are cost structure and revenue channels.

1.5/2

Fig. 4. Highlighting the Activated Words

Besides highlighting the key terms, the points that contribute to the final grade are also displayed. This can be done due to the fact that those "activated" key words achieved the highest relevance score. However, it can be observed that "channels" is not highlighted, meaning that the word is wrong in that context. In this way, students have a clear overview of the correct key words and how they contributed to the achieved number of points.

7 CONCEPTUALIZATION OF TRUST

Trust is an abstract concept that has been extensively analyzed over the years. There are many models and frameworks that divide this construct in multiple sub-constructs in order to create a clear conceptualization [1, 17, 18, 20, 26]. However, these different models conceptualize trust in a broad range of complex scenarios, implying interpersonal trust to some extent, whilst this research aims to capture students' level of trust in a very specific scenario (trust towards different Explainable AI methods implemented in an AI-based grading tool). Thus, some elements from these studies are extracted and further used during this research. The selected concepts act as categories based on which comparisons are made. Moreover, taking into consideration that the trustee is represented by a specific technology, some elements from the Technology Acceptance Model (TAM) developed by Davis are used as well. So, the following dimensions are selected for being analyzed: perceived usefulness, perceived ease of use and perceived trustworthiness.

Perceived usefulness can be defined as "the degree to which a person believes that using a particular system would enhance his/her performance" [5]. "A system high in perceived usefulness [. . .] is one for which a user believes in the existence of a positive use-performance relationship" [5]. This theme is explored in order to understand students' opinions about advantages of the selected Explainable AI method and to what extent, useful feedback that can enhance the learning process is facilitated. Moreover, disadvantages are also being discussed within this theme. In this way, the method with the highest level of perceived usefulness is discovered along with its disadvantages.

Perceived ease of use can be defined as "the degree to which a person believes that using a particular system would be free of effort"[5]. Davis claims that "[. . .] an application perceived to be easier to use than another is more likely to be accepted by users" [5]. In this case, "easier to use" is connected with how understandable and intuitive each Explainable AI method is in students' perspective. Thus, exploring this theme helps in finding which of the methods facilitates a clear understanding for students about the inner-working of the algorithm. Moreover, this connection of concepts leads towards findings about an important aspect that has been discussed in the previous sections: enhancement of transparency.

Perceived trustworthiness can be defined as "trustor's perception of the trustee's competence, benevolence and integrity" [1]. This definition is aligned with McKnight's conceptualization of trusting beliefs. He claims that the same three dimensions build a strong belief of a trustor towards a trustee [20]. Thus, by exploring these three dimensions, meaningful insights will be found about students' perceived trustworthiness towards the selected methods. Also, the method which is the most trusted by them is discovered. However, the dimensions are adapted in order to fit the context of this research where the trustee is not a human-being. Competence is explored with the purpose of finding how well the Explainable AI methods explain the grading process from students' perspective. Integrity is investigated with the aim of finding to what extent the methods provide a complete and detailed explanation of the grading process that satisfy the students' need for justification. Furthermore, due to the limitations of this research, the students are not able to interact with a real life system and they have to rely on the created mock-ups (Section 6). Thus, benevolence of the general concept of AI-based grading tools is explored. However, this is a positive aspect for this research because a high-level overview of students' opinions towards the concept of AI-based grading is created.

8 RESULTS

In this chapter, the results of this study are discussed. The first research question was answered by conducting a literature review (Section 5). The field of Explainable AI has been explored and a deep understanding of the concept has been developed. Then, a connection between this concept and the educational sector has been made. It has been found that it is necessary to adapt the AI-based tools used in this sector in order to ensure the ethical behavior of the algorithm, display the correct behavior in order to improve the stakeholders' level of trust and facilitate feedback. Thus, different Explainable AI methods help to solve this issues. However, their implementation should be made carefully because many human-related aspects combined with learning and teaching methodologies have to be considered. This is the main purpose for the establishment of the XAI-ED framework. Lastly, three common Explainable AI methods have been discussed. Their functionalities allow them to be applied to an AI-based grading tool used in higher education. Furthermore, they can be adapted in different ways, but the chosen way of implementation has been discussed (Section 5.3) and illustrated by mock-ups (Section 6). So, interviews with students are conducted in order to fully understand whether they find the selected methods suitable as well and understand which of them they find the most trustworthy.

8.1 Population Overview

The target population of this study consists of students. However, only students from University of Twente were interviewed because of two reasons. Firstly, this study helps in the development of the EasyGrader tool which is an internal product developed by researchers at the University of Twente. Secondly, most of the students at this university are familiar with the actual platform that is used for sitting exams (Remindo). This is an important factor that was taken into consideration because the current way of working is the basis for the transition to an AI-based grading tool whose transparency can be enhanced by the presented Explainable AI methods. A clear understanding of how the current process of grading works facilitates a better understanding of the proposed methods by means of comparison. The total number of interviewees is twelve. The distribution of the study programmes is: five students from Technical Computer Science, four students from Business Information Technology, one student from Educational Science & Technology, one student from Civil Engineering & Management and one student from Management, Society & Technology.

As it can be observed, over half of the population is currently studying Technical Computer Science and Business Information Technology. The fact that these students have background knowledge about the topic is neither an advantage, nor a disadvantage because the study does not require technical expertise. The only important prerequisite is to be familiar with the Remindo platform. All of the interviewed students have sat at least an exam using this platform and they have participated to at least an exam review where they could check how they have been graded.

Moreover, the age range of participants lies between 19 and 24 years old. This population includes students that are currently studying in their first year of bachelor's degree as well as students that are currently studying in the second year of masters' degree. Thus, the personal experiences are diverse.

8.2 Perceived Usefulness

By exploring this theme, meaningful insights have been found regarding the degree to which the methods provide valuable feedback for students that can enhance their learning process for further examinations. The most preferred method among students has been discovered alongside with its disadvantages. An overview of the results can be seen in Figure 5.

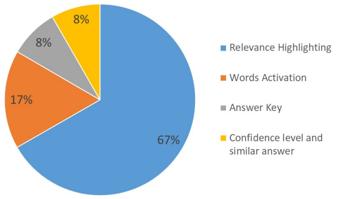


Fig. 5. Perceived Usefulness

As it can be seen, two thirds of the interviewees considered that the concept of relevance highlighting (Section 6.3) is the most useful in terms of receiving feedback with the purpose of learning enhancement. Even if there is no written feedback provided, the participants mentioned that visualizing different tones of colours that are representing the relevance level of each word found in their answer helps in understanding which were the key words that matter in answering that specific question. Furthermore, in case of failing the exam, they will be able to focus more on the right parts and clearly understand what knowledge they are lacking. Also, taking into consideration the answers that are longer and more complex, the difference in the level of relevance leaves room for interpretations. Besides the identified key words that are the most important and mandatory to have in the answer in order to be considered correct, the words with a lower degree of relevance (but not the lowest) can spark students' interest and make them explore the learning materials more in depth. They will be wondering why that specific word matters to some extent. Here, a further connection has been made with the expectations of the teachers. As long as students can observe the difference among the relevance of the words, they can understand what expectations teachers' had from students. Depending on the question, it might be enough to list a number of concepts and the answer will be considered correct, but sometimes a more complex answer that require in-depth explanations is expected. Nevertheless, disadvantages were also discovered alongside with the advantages. The most mentioned disadvantage is that the grading system is unclear. Students cannot really observe to what extent each key word contributed to the grade. No further information about the points is provided. This would be even more confusing for complex answers were words have a different level of relevance. In participants' opinions, this would be the most useful addition to this specific method. Another disadvantage that has been discovered is the lack of correct answers in displaying the results. Besides the fact that the point system is unclear, a correct answer is not provided. This addition is useful especially if you do not know how to answer a question or if the answer is completely wrong. In this case, there is nothing to be highlighted if the answer is missing or everything would be highlighted as "less relevant". By having a correct answer provided, the learning experience is even more enhanced if students' current context-specific knowledge does not exist.

8.3 Perceived Ease of Use

Exploration of this theme leads to discoveries in two slightly different directions. Firstly, the method that is the easiest to be understood by the students is discovered. Secondly, the method that provides the clearest picture about the inner-working of the algorithm is discovered as well. Thus, this is the method that works the best for enhancing the transparency of the algorithm in students' opinions. However, the findings are surprising and a single method fulfills the needs of both sub-categories. An overview of the results can be seen in Figure 6.

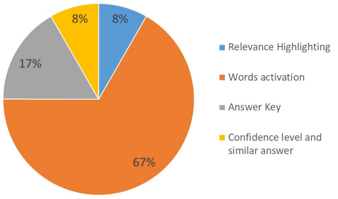


Fig. 6. Perceived Ease of Use

As it can be observed, two thirds of the interviewees considered that words activation (Section 6.4) is the clearest and easiest method to be understood among all of them. It facilitates a clear understanding by highlighting only the key words that contributed to the grade and also by providing and displaying the number of given points. In this way, students can understand the grading process and the expected words from someone's answer. One of the participants emphasized that they "love the simplistic but still efficient way of displaying the results by this method because it provides everything that a student needs to acknowledge and accept a specific result". Furthermore, the participants who chose this method as the easiest to be understood also chose it as the one which provides the clearest picture about the inner working of the algorithm. This is due to the fact that by highlighting the key words and displaying the specific number of points, they can understand "what was the algorithm looking for in the answer" and "how the algorithm performed the grading based on what it has been searching within the answer". These elements play a role in enhancing the transparency of the AI-based tool that was used for grading their exams. Moreover, taking into consideration that the methods were chosen and visually illustrated having the goals of the XAI-ED framework in mind, the positive results within this category prove the importance of taking the main stakeholders into consideration, choosing the right methods based on the contextual needs and having a human-centered design. Furthermore, students had some interesting remarks about the degree to which this method is enhancing the transparency of the tool. A useful addition to the concept of words activation would be the confidence level of the algorithm (Section 6.2). It adds an extra layer of transparency because the confidence level is a proof of the correct behavior of the algorithm. The measurement tells how accurate the algorithm was in grading that specific question. One of the interviewees mentioned that "by combining these two methods, there is no room left for interpretations about how the grading has been performed by the system".

8.4 Perceived Trustworthiness

This broader theme is explored with the aim of finding which of the presented methods is trusted more by students and the reasons behind. However, an extra step is needed in order to understand to what extent students trust an automated tool that is used for grading their questions at an exam. This is necessary for having a high-level overview of the concept and find out which is the acceptance rate of such a tool among students. The results clearly show that students are willing to adapt themselves to this new change but they mention some extra conditions that should be in place. 84% of the students claimed that an AI-based tool used for grading can be objective and can bring plenty of advantages to the grading process. The advantage that was mentioned the most is that this type of tool can significantly reduce the waiting times for receiving a grade after sitting an exam. This factor helps them to reduce the level of stress related to that specific examination. Another mentioned advantage is the reduction of human-bias. The AI-tool can maintain the same level of objectivity during the whole process while a teacher can be negatively influenced by external factors such as tiredness, personal problems, increased workload. However, the algorithm has to be well trained, the correct behavior should be proved and it should not have any biases in the data set. Also, they mentioned that an automated tool would work effectively for grading short open-ended questions whose answers are mainly based on keywords. For questions that require a more complex answer, the tool is prone to errors in trying to recognize different types of phrasings that might be correct even if they do not include some expected keywords.

Consequently, the trust towards Explainable AI methods is further analyzed. Based on the previously made conceptualization (Section 7), trustworthiness is explored within two dimensions: competence and integrity. An overview of the perceived competence can be observed in Figure 7. 59% of the students consider that the concept of words activation is the most suitable method for explaining how the grading has been done for them personally. The results are not surprising considering the fact that even a higher percentage considered this method as the most intuitive one (Section 8.3). Students mentioned that this method "is simplistic, straightforward and self-explanatory" and it would meet their personal expectations

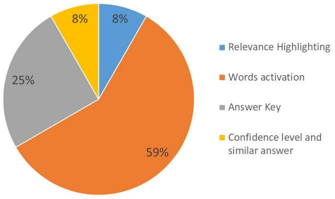


Fig. 7. Perceived Competence

and requirements. Besides the traditional method of providing the answer key (Section 6.1), this is the only method that displays the number of points given for each specific key word. One of the participants mentioned that "it is very clear how the grading has been performed and how the points have been divided among the correct key words". However, the statistics are different for integrity. An overview can be observed in Figure 8. As it can be seen, half of the interviewed students still consider that the need for justification is satisfied by having the grading criteria provided (Section 6.1). One student made an interesting comparison, claiming that "even if the words activation displays the number of points given for each key word included in the answer, a complete grading criteria lists all of the relevant key words and their contribution to the grade." So, it seems that the need of justification is entirely fulfilled by displaying the list of all the correct key words due to the fact that the other methods do not explicitly state what an entirely correct answer would be. In this way, there is no room left for interpretations about what is missing or what is wrong in the answer, and the justification of the grade is complete. Lastly, students were asked which of these methods they find the most trustworthy and which of them they would rely on if would be implemented in real-life. Still, 67% trust the words activation the most because it provides "almost everything that a student would want to know about the grade". This individual question is also proven by the results from previous categories where this concept was the main discussion point and most popular option among the students that were interviewed. However, the grading criteria is a necessary addition that helps in fulfilling all students' needs and achieve a maximum trust level of the system.

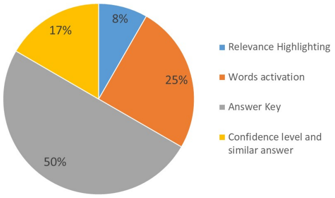


Fig. 8. Perceived Integrity

9 CONCLUSION AND DISCUSSION

This section concludes the discoveries of this research and discuss further areas of exploration.

9.1 Conclusion

To sum up, the purpose of this research was to discover how different Explainable AI methods can enhance the transparency of an AI-based grading tool used for grading exams in higher education and check which of them do students find the most trustworthy. During the literature review phase, the field of Explainable AI has been explored as well as the importance of enhancing the transparency of AI-based tools used in the educational sector by Explainable AI methods. Lastly, three different Explainable AI methods have been analyzed and they have been found suitable to be applied to an AI-based grading tool. They were illustrated by mock-ups in order to create a real-life example and an overview of their potential implementation. Furthermore, a qualitative approach has been chosen in order to find the level of students' trust towards these methods. However, the concept of "trust" required a conceptualization in order to catch the true meaning of the construct from multiple perspectives. The target population consists of bachelor and master students from University of Twente. They helped in discovering which of the methods fits the best in three different categories. The findings shows that highlighting the words in students' answers based on the level of relevance can enhance the learning process, displaying the activated words is the most intuitive and easy to understand as well as the one which provides the most insights about the inner working of the algorithm and last but not least, displaying the activated words is the method that students trust the most in terms of personal preference of how the grading has been done and also the answer key as the method that satisfies the personal need of grade justification. Furthermore, useful combinations of methods were discovered that have a bigger impact than the standalone methods. For example, the words activation can be jointly implemented with the confidence level of the algorithm in order to provide a more detailed picture about the inner working of the algorithm. Also, providing the grading criteria alongside with the concept of words activation builds an unambiguous concept that entirely fulfills the need of justification from students' perspective.

9.2 Further Research

During this research, further areas of exploration and work have been discovered. This research represents one of the first taken steps in developing a tool based on the discussed concept of AI-based grading. Thus, it adds value to the preliminary analysis of the concept that is still in progress. After completing this phase, testing with a real-life system can be started. A real-life implementation of the software as well as Explainable AI methods would lead to more accurate results by gathering higher-quality data. Thus, the interviews can be transformed into controlled experiments that help in capturing unbiased knowledge. Secondly, this concept of automated grading that has been discussed within this research focuses on grading short open-ended questions and the Explainable AI methods have been adapted for this. After a successful implementation of the software for this specific case, it can be expanded to meet other types of requirements such as answering code-related questions through writing lines of code or analyzing models, diagrams and drawings.

REFERENCES

[1] Gene M. Alarcon, Joseph B. Lyons, James C. Christensen, Samantha L. Klosterman, Margaret A. Bowers, Tyler J. Ryan, Sarah A. Jessup, and Kevin T. Wynne. 2018. The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads. *Behavior Research Methods* 50 (10 2018), 1906–1920. Issue 5. https://doi.org/10.3758/s13428-017-0959-6

[2] Loraine Busetto, Wolfgang Wick, and Christoph Gumbinger. 2020. How to use and assess qualitative research methods. *Neurological Research and Practice* 2 (5 2020). Issue 1. https://doi.org/10.1186/s42466-020-00059-z

[3] Rianne Conijn, Patricia Kahr, and Chris Snijders. 2023. The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *Journal of Learning Analytics* 10 (3 2023), 37–53. Issue 1. https://doi.org/10.18608/jla.2023.7801

[4] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. (10 2020). http://arxiv.org/abs/2010.00711

[5] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[6] Derek Doran, Sarah Schulz, and Tarek R. Besold. 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. (10 2017). http://arxiv.org/abs/1710.00794

[7] Dónal Doyle, Alexey Tsymbal, and Pádraig Cunningham. 2003. A review of explanation and explanation in case-based reasoning. (2003).

[8] Rebecca Ferguson. 2019. Ethical challenges for learning analytics. *Journal of Learning Analytics* 6 (2019), 25–30. Issue 3. https://doi.org/10.18608/jla.2019.63.5

[9] Krzysztof Fiok, Farzad V. Farahani, Waldemar Karwowski, and Tareq Ahram. 2022. Explainable artificial intelligence for education and training. *Journal of Defense Modeling and Simulation* 19 (4 2022), 133–144. Issue 2. https://doi.org/10.1177/15485129211028651

[10] Evelyn Fix. 1985. *Discriminatory analysis: nonparametric discrimination, consistency properties*. Vol. 1. USAF school of Aviation Medicine.

[11] Ellie Fossey, Carol Harvey, Fiona McDermott, and Larry Davidson. 2002. Understanding and evaluating qualitative research. *Australian & New Zealand journal of psychiatry* 36, 6 (2002), 717–732.

[12] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang Zhong Yang. 2019. XAI-Explainable artificial intelligence. *Science Robotics* 4 (12 2019). Issue 37. https://doi.org/10.1126/scirobotics.aay7120

[13] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research* 77 (2007), 81–112. Issue 1. https://doi.org/10.3102/003465430298487

[14] Kutubuddin Sayyad Liyakat Kazi, Macha Babitha, C Sushama, Vijaya Kumar Gudivada, and Srinivasa Rao Bandaru. 2022. Trends of Artificial Intelligence for Online Exams in Education. *Article in International Journal of Early Childhood Special Education* 14 (2022), 2457–2463. https://doi.org/10.9756/INT-JECSE/V14I1.290

[15] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence* 3 (1 2022). https://doi.org/10.1016/j.caeai.2022.100074

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).

[17] Matthew KO Lee and Efraim Turban. 2001. A trust model for consumer internet shopping. *International Journal of electronic commerce* 6, 1 (2001), 75–91.

[18] Roy J. Lewicki, Edward C. Tomlinson, and Nicole Gillespie. 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management* 32 (12 2006), 991–1022. Issue 6. https://doi.org/10.1177/0149206306294405

[19] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[20] D Harrison McKnight and Norman L Chervany. 2000. What is trust? A conceptual analysis and an interdisciplinary model. (2000).

[21] Alexander Osterwalder and Yves Pigneur. 2010. *Business model generation: a handbook for visionaries, game changers, and challengers*. Vol. 1. John Wiley & Sons.

[22] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29 (9 2019), 441–459. Issue 3. https://doi.org/10.1007/s11023-019-09502-w

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016, 1135–1144. https://doi.org/10.1145/2939672.2939778

[24] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[25] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29 (10 2020), 260–278. Issue 4. https://doi.org/10.1080/12460125.2020.1819094

[26] F David Schoorman, Roger C Mayer, and James H Davis. 2007. AN INTEGRATIVE MODEL OF ORGANIZATIONAL TRUST: PAST, PRESENT, AND FUTURE. , 344-354 pages. Issue 2.

[27] Neil Selwyn. 2019. What's the problem with learning analytics? *Journal of Learning Analytics* 6 (2019), 11–19. Issue 3. https://doi.org/10.18608/jla.2019.63.3

[28] Maddalena Taras. 2005. Assessment - Summative and formative - Some theoretical reflections. *British Journal of Educational Studies* 53 (12 2005), 466–478. Issue 4. https://doi.org/10.1111/j.1467-8527.2005.00307.x

[29] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. 2020. Interpretable confidence measures for decision support systems. *International Journal of Human Computer Studies* 144 (12 2020). https://doi.org/10.1016/j.ijhcs.2020.102493

[30] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11839 LNAI, 563–574. https://doi.org/10.1007/978-3-030-32236-6_51

A APPENDIX - INTERVIEW QUESTIONS

- What is your study?
- What is your age?
- Have you sat at least an exam on Remindo?
- Have you participated to at least an exam review on Remindo?

Perceived Usefulness:

- Which of the Explainable AI methods displays useful feedback that can enhance your learning process and why?
- Can you find at least a disadvantages of this method?

Perceived Ease of Use:

- Which of the Explainable AI methods do you find the most intuitive (easy to understand) and why?
- Which of the Explainable AI methods creates the most transparent picture of how the algorithm has graded your question and why?

Perceived Trustworthiness:

- What is your opinion about the objectivity and impartiality of an AI-tool used for grading your questions at the exams?
- In your perspective, which of the Explainable AI methods works the best for explaining how the grading has been done by the AI-system and why?
- In your perspective, which of the Explainable AI methods satisfies your need for justification of how a question has been graded?
- Having in mind everything that has been discussed, which of the Explainable AI methods do you find the most trustworthy?