die hochschul*lehre*
Interdisziplinäre Zeitschrift für Studium und Lehre

die hochschul*lehre* – Jahrgang 8-2022 (12)
Herausgebende des Journals: Svenja Bedenlier, Ivo van den Berk, Jonas Leschke, Marianne Merkt, Peter Salden, Antonia Scholkmann, Angelika Thielsch
Beitrag in der Rubrik Praxisforschung
DOI: 10.3278/HSL2212W
ISSN: 2199-8825 wbv.de/die-hochschullehre

Automatisierte Auswertung von Short-Answer-Aufgaben zur Bestimmung der erworbenen Kompetenzen

Ulrich Bucher

Zusammenfassung

In dem hier vorgestellten Projekt wurde die Genauigkeit der automatisierten Auswertung von Short-Answer-Aufgaben mithilfe von Verfahren des maschinellen Lernens sowie der künstlichen Intelligenz verglichen, wenn das Training der Verfahren einmal auf Basis der Selbstevaluation durch die Studierenden und zum anderen auf einer Expertenbewertung beruhte. Zu diesem Zweck wurde im Bereich des wissenschaftlichen Arbeitens ein Short-Answer-Kompetenz-Test entwickelt, der aus zwölf Aufgaben bestand. Die Antworten zu den Aufgaben wurden zum einen durch die an dem Test teilnehmenden Studierenden selbst bewertet. Zum anderen beurteilte ein Experte, ob die Antworten dem Inhalt einer Musterlösung entsprachen oder nicht. Die Daten wurden dann mithilfe der gleichen Verfahren des maschinellen Lernens sowie der künstlichen Intelligenz automatisiert ausgewertet. Dabei zeigten sich erhebliche Genauigkeitsunterschiede der Verfahren in Abhängigkeit davon, ob diese mit Daten trainiert wurden, die auf einer Selbstevaluation der Studierenden oder der Bewertung durch einen Experten basierten.

Schlüsselwörter: Automatic Short Answer Scoring; Selbstevaluation; maschinelles Lernen; künstliche Intelligenz

Automated evaluation of short-answers to determine acquired competences

Abstract

In the project presented here, the accuracy of the automated evaluation of short-answer tasks using machine learning and artificial intelligence methods was compared when the training of the methods was based on the one hand on self-evaluation by the students and on the other hand on expert evaluation. For this purpose, a short-answer test was developed in the area of scientific work, which consisted of twelve questions. The answers to the questions were evaluated on the one hand by the students taking the test themselves. On the other hand, an expert assessed whether the answers corresponded to the content of a sample solution or not. The data was then automatically evaluated using the same machine learning and artificial intelligence methods. Significant differences in the accuracy of the procedures were found depending on whether they were trained with data based on the students' self-evaluation or the evaluation by an expert.

Keywords: Automatic Short Answer Scoring; Self-Evaluation; Machine Learning; Artificial Intelligence

wbv

1 Einleitung

Die automatisierte Auswertung von Short-Answer-Aufgaben beschäftigt sich damit, die Antworten zu einer Aufgabe automatisiert mit einem Label oder einem Score zu versehen. Die Länge der Antworten geht dabei über einige wenige Sätze meist nicht hinaus. In Abgrenzung zum verwandten Gebiet des Essay-Scorings wird beim Short-Answer-Scoring vorwiegend der Inhalt ausgewertet, ohne dabei etwaige in den Antworten vorhandene Rechtschreib- oder Grammatikfehler zu berücksichtigen (Horbach 2019; Surya, Gayakwad & Nallakaruppan 2019).

Das wachsende Interesse an einer automatisierten Auswertung von Short-Answer-Aufgaben im Bildungsbereich lässt sich leicht nachvollziehen. Die manuelle Bewertung von Antworten ist eine zeitintensive Aufgabe (Horbach, 2019). Eine automatisierte Auswertung verspricht, die Arbeitsbelastung von Dozierenden zu verringern, insbesondere ist der Arbeitsaufwand weitgehend unabhängig von der Anzahl der Studierenden, den Studierenden kann ein unmittelbares Feedback gegeben werden, ein einheitlicher Bewertungsmaßstab sowie ein einheitliches Vorgehen stellen die Konsistenz der Beurteilung sicher und die Dozierenden erhalten einen Überblick über die Lernfortschritte aller Studierenden (vgl. Haley, Thomas, De Roeck und Petre 2007; Pado und Kiefer 2015; Bey und Dillenbourg 2018; Hegarty-Kelly und Mooney 2021). Dem großen Interesse von Dozierenden und Bildungseinrichtungen an einer automatisierten Auswertung von Short-Answer-Aufgaben stehen jedoch die Komplexität sowie der hohe Aufwand bei der Entwicklung und Einführung eines operativen Systems zur automatisierten Auswertung gegenüber (Madnani & Cahill 2018).

Die verschiedenen Verfahren der automatisierten Auswertung von Short-Answer-Aufgaben können in das überwachte Lernen (*Supervised Learning*) und das unüberwachte Lernen (*Unsupervised Learning*) unterteilt werden. Beim unüberwachten Lernen werden die studentischen Antworten mit Musterlösungen verglichen und auf Basis von Ähnlichkeitsmaßen eine Bewertung ermittelt (Sadr & Nazari Solimandarabi, 2019). Die Vielfalt der Sprache, die fehlende Aneignung der Fachsprache in den unteren Semestern sowie der Umstand, dass richtige Antworten nicht immer limitiert und im Vorfeld durch Expertinnen und Experten in einer Musterlösung verfasst werden können, stellen diesen Ansatz jedoch nicht selten vor erhebliche Herausforderungen.

Im Rahmen des vorliegenden Projekts fand daher eine Auseinandersetzung mit dem überwachten Lernen statt. Bei diesem wird gewöhnlich manuell ein Muster erstellt (häufig in Form von Labeln) und die Antworten mithilfe von Verfahren des maschinellen Lernens[1] und insbesondere des Deep Learning[2] analysiert. Zu den gängigen Verfahren zählt dabei, auf Basis von manuell vergebenen Labeln ein neuronales Netz zu trainieren, das ein Muster zwischen den Antworten und den zugeordneten Labels aufdecken soll. Das trainierte Netz wird dann verwendet, um zukünftige Antworten zu kategorisieren. In zahlreichen Studien gelang es, mit dieser Vorgehensweise hohe Trefferquoten zu erzielen, in dem Sinne, dass die Testdaten korrekt den Labeln zugeordnet werden konnten (vgl. Ndukwe, Amadi, Nkomo & Daniel 2020).

Mit dieser Vorgehensweise sind jedoch verschiedene Nachteile verbunden (Lewis Sevcikova, 2018). Ob ein Verständnis stattgefunden hat und die angestrebten Kompetenzen entwickelt wurden, lässt sich auf Basis der Antworten zu einer Aufgabe nur bedingt erkennen. Denn Studierende entwickeln nicht selten ein feines Gespür für die Erwartungen der Dozierenden an die Antworten in einem Test. Entsprechend richten sich die Antworten an der vermuteten Erwartungshaltung und nicht nur am eigenen Verständnis aus. Ob eine Antwort auf einem Verständnis beruht oder

1 Das „maschinelle Lernen“ ist ein Teilgebiet der künstlichen Intelligenz, das darauf abzielt, Systeme in die Lage zu versetzen, Muster in einem Datensatz zu erkennen. Es kann beispielsweise dazu genutzt werden, um Lern-Empfehlungen durch eine Maschine zu bestimmen (u. a. Schöb, S., Kilian, L. Hellmich, C., Brandt, P. & Biel, C. (2019). Adaptives Recommending am Beispiel von EULE. In weiter bilden 4 (2019), S. 22–25. DOI: 10.3278/WBDIE1904W022).

2 Wenn in diesem Beitrag von „Deep Learning“ die Rede ist, dann bezieht sich dieser Begriff nicht auf die Tiefe des Lernens der Studierenden (wie beispielsweise in Riva, N. & Kiehne, B. (2019). Musikpraxis, ohne zu musizieren? Wie Forschendes Lernen in musikwissenschaftlichen Seminaren gelingen kann. die hochschullehre, Jahrgang 5/2019, online unter: www.hochschullehre.org). Vielmehr wird damit eine Methode des maschinellen Lernens bezeichnet, bei der künstliche neuronale Netze zum Einsatz kommen.

nicht, kann daher letztendlich nur durch die Studierenden selbst entschieden werden. Ein zweiter Nachteil der oben genannten Vorgehensweise liegt in dem damit einhergehenden Aufwand. Regelmäßig sind sehr große Datensätze notwendig, um Verfahren des maschinellen Lernens bzw. ein neuronales Netz zu trainieren (Henderson et al., 2017). Insbesondere erfordert die Zuordnung von Labels zu den einzelnen Antworten einen hohen Arbeitsaufwand. Darüber hinaus ist es in methodischer Hinsicht als kritisch zu betrachten, wenn die Labels durch Expertinnen und Experten vergeben werden. Denn auf diese Weise stellt sich die Frage, wie vermieden werden soll, dass ein inneres Schema dieser Expertinnen und Experten Einfluss auf das durch das neuronale Netz erkannte Muster nimmt. Auch wenn bei dem Arbeitsschritt der Zuordnung von Labels zu den Antworten mehrere Expertinnen und Experten parallel die Antworten der Studierenden codieren, wird diese Gefahr nicht ausgeschlossen (da bestimmte Sinnstrukturen durch die Expertinnen und Experten geteilt werden können).

Daher stellte sich in dem hier vorgestellten Projekt die Frage, ob die Selbstevaluation durch die Studierenden mit einer vergleichbar hohen Genauigkeit automatisiert bewertet werden kann, wie dies der Fall ist, wenn die Bewertung durch Expertinnen und Experten vorgenommen wird.[3] Bei einer ähnlich hohen Genauigkeit besitzt die Selbstevaluation durch die Studierenden klare Vorteile gegenüber der Bewertung durch Expertinnen und Experten. Insbesondere kann dadurch der Arbeitsaufwand seitens der Dozierenden deutlich verringert werden. Zudem sind weder umfangreiche Datensätze erforderlich, noch besteht die Notwendigkeit, ein automatisiertes System zu entwickeln. Weiterhin ließe sich damit ein echtes Verständnis überprüfen. Es sind aber insbesondere die didaktischen Vorteile, die eine Selbstevaluation durch die Studierenden gegenüber einer Evaluation durch die Dozierenden als Expertinnen und Experten so attraktiv erscheinen lassen (und Motivatoren für dieses Projekt waren). Insbesondere wird die Fähigkeit der Selbstevaluation als wichtige Voraussetzung betrachtet, um eigenständig Lernfelder zu identifizieren und diese im Sinne eines lebenslangen Lernens in einer volatilen und sich dynamisch entwickelnden Umwelt an die Anforderungen der Arbeitswelt anzupassen (Suwanarak, 2018). Zudem erhalten die Studierenden mit der Selbstevaluation eine aktivere Rolle im Lernprozess sowie mehr Verantwortung für den Lernerfolg (Klenowski, 1995). Die Studierenden können deren Lernfortschritte selbst mit vorgegebenen Kriterien und Standards vergleichen. Eine formative Beurteilung des Lernfortschritts ist durch die Selbstevaluation der Studierenden für die verschiedensten Lernfelder vergleichsweise einfach realisierbar. Auf diese Weise können häufiger formative Beurteilungen in die Lehre integriert werden. Entsprechend wurde in zahlreichen Studien ein Zusammenhang zwischen der Selbstevaluation, dem Lernen sowie dem Lernerfolg gefunden (u. a. Brown & Harris 2013 sowie Elmahdi, Al-Hattami & Fawzi 2018).

Die Selbstevaluation durch die Studierenden wirft jedoch die Frage auf, ob auf diese Weise nicht das Rauschen (*noise*) der Daten anschwillt. So könnte das Rauschen, das bei Textantworten regelmäßig eine erhebliche Herausforderung darstellt, durch die unterschiedliche Genauigkeit und Ehrlichkeit der an einem Test teilnehmenden Studierenden bei der Zuordnung der Labels negativ beeinflusst werden. So zeigte sich in verschiedenen Studien eine Streuung der Genauigkeit der studentischen Selbstevaluationen (Brown, Andrade & Chen 2015). Zudem kann die Durchführung eines formativen Tests, auch wenn dieser nicht in die Notengebung einfließt (wie dies auch in dem vorliegenden Projekt der Fall war), in unterschiedlicher Form ehrlich selbst bewertet werden. Im Extremfall verleitet die Testsituation die Studierenden dazu, eine Täuschung vorzunehmen. Täuschungsversuche von Studierenden sind ein gut untersuchtes Forschungsfeld (Klein et al., 2007). Whitley kommt in einem Review von 36 Studien zu dem Ergebnis, dass durchschnittlich 43,1 % der Studierenden angaben, in einer Prüfung schon einmal geschummelt zu haben (Whitley, 1998). Auch wenn die Zahlen zwischen den verschiedenen Studien erheblich streuen

3 Wenn in diesem Beitrag von Selbstevaluation die Rede ist, dann ist damit die Bewertung der eigenen Antworten durch die Studierenden selbst gemeint. Es wird dabei nicht die eigene Person, sondern die eigene Antwort zu einer Aufgabenstellung bewertet. Die Selbstevaluation wird in Anlehnung an Brown, Andrade & Chen (2015) als Teilgebiet des Self-Assessments betrachtet, das neben der Selbst-Evaluation auch die Deskription umfasst.

(in den von Whitley betrachteten Studien schwankten die Zahlen zwischen 9 und 95 %), sind Täuschungsversuche kein Einzelphänomen. Für die automatisierte Auswertung können diese jedoch eine erhebliche Herausforderung darstellen. Denn wenn beispielsweise von zwei Studierenden, die beide dieselbe falsche Antwort gegeben haben, eine bzw. einer diese als korrekt und die bzw. der andere diese aber als falsch bewertet, dann kann in den Daten möglicherweise kein Muster mehr identifiziert werden, das eine richtige von einer falschen Antwort unterscheidet. Im schlimmsten Fall werden die Verfahren so trainiert, dass eine falsche Antwort automatisch als korrekt klassifiziert wird.

Zu der Gefahr eines Anschwellens des Rauschens kommt hinzu, dass sich die Studierenden möglicherweise besser bewerten, als eine Einschätzung durch einen Dozierenden ausfallen würde. Die Integration der Selbstevaluation in eine Lehrveranstaltung steht und fällt mit der Genauigkeit, mit der die Studierenden diese vornehmen. Nutzt man zur Einschätzung der Genauigkeit die Korrelation der Bewertungen von Studierenden und Dozierenden, dann gelangt man regelmäßig zu einem eher ernüchternden Ergebnis. So fiel die Korrelation zwischen Selbsteinschätzungen und Bewertungen der Dozierenden bei den meistens von Brown und Harris betrachteten Studien lediglich schwach bis moderat positiv aus (Brown & Harris 2013). Nur wenige Studien stellten eine Korrelation fest, bei der r mindestens .60 betrug (Brown und Harris 2013). In vielen Studien bewerteten sich die Schüler:innen bzw. Studierenden besser als deren Lehrer:innen bzw. deren Dozierende (Brown, Andrade & Chen 2015).

2 Methodisches Vorgehen

Um die Kompetenz der Studierenden im Bereich des wissenschaftlichen Arbeitens zu überprüfen, wurde ein Test entwickelt, der aus zwölf Aufgaben bestand. Die Aufgaben beinhalten jeweils einen Ausschnitt aus einer studentischen Arbeit eines vorangegangenen fünften Semesters. Der nachfolgende Abschnitt gibt einen beispielhaften Ausschnitt aus einer studentischen Arbeit wieder.

Beispiel: Ausschnitt einer studentischen Arbeit

„Ziel dieser Arbeit ist es, in Folge der Notwendigkeit einer individuellen Ansprache, die Varianten des Dialogmarketings, sei es mit offline beziehungsweise online Charakter zu ergründend und folgend daraus die aktuelle Entwicklung aufzuführen. Bei dieser Analyse bezüglich des Umgangs mit mobilen Endgeräten wird dabei erörtert, ob sich der Trend in Richtung Mobile Marketing für das Unternehmen zur Investition lohnt. In Folge dessen findet eine tiefere Beleuchtung auf das Thema des Mobile Marketing statt. Grundlegende Definitionen geben dazu einen Überblick in die Thematik; zu fokussierende Zielgruppen sowie die aktuellen Trendbeispiele in Sachen Mobile Marketing sind daher herauszuarbeiten. Durch den direkten Test einer App, die im Bereich des Location Based Service Verwendung findet, wird das Trenddenken Mobile Marketing weiter betrachtet. Der Übergang Analog zur Digital wird ergründet sowie die weiteren Aussichten dieser Thematik beleuchtet, um abschließend eine Handlungsempfehlung und ein Fazit auszusprechen."

Aufgabe der Studierenden war es dann, Schwachstellen in diesen Ausschnitten zu identifizieren und diese zu beschreiben. Erwartungsgemäß lassen sich zu den vorgelegten Ausschnitten zahlreiche Schwachstellen benennen. Die Musterantworten beschränkten sich auf die Schwachstellen, die in der Vorlesung thematisch im Fokus standen und die den angestrebten Kompetenzzielen der Vorlesung entsprachen.

Die zwölf Aufgaben deckten verschiedene Themenfelder des wissenschaftlichen Arbeitens ab. So unter anderem die Forschungsfrage, die Gliederung, die Sprache, die Argumentation, die Literaturrecherche sowie den Fußnotenapparat. Jede Aufgabe adressierte ein anderes Themenfeld, da-

mit das Antwortverhalten zu einer Frage nicht durch die vorangegangenen Aufgaben beeinflusst wird.

Die Datenerhebung fand über einen fünfjährigen Zeitraum zwischen 2015 und 2020 in verschiedenen Kursen des vierten sowie des ersten Semesters an der DHBW Stuttgart statt. Die Ergebnisse des Tests gingen nicht in die Notengebung ein. Der Test wurde auf freiwilliger Basis durchgeführt. Dass die Anzahl der Test-Teilnehmenden der Anzahl der an den Lehrveranstaltungen anwesenden Studierenden entspricht, ist insbesondere den Umständen geschuldet, dass die Tests während der Lehrveranstaltung durchgeführt wurden und das Testformat ein hohes Interesse seitens der Studierenden hervorrief. Es ist jedoch nicht auszuschließen, dass einzelne Studierende mehrfach an dem Test teilgenommen haben und zufällig dieselbe Anzahl an Studierenden die Teilnahme verweigert hat. Die Beobachtungen des Autors während der Tests sowie die Inhalte der Antworten (Duplikate etc.) liefern jedoch keinen Grund zu dieser Annahme.

Damit die Selbstevaluation durch die Studierenden nicht zu einem zusätzlichen Rauschen sowie zu einem Genauigkeitsverlust der Daten führt, wurde im Vorfeld der Tests an deren Ehrlichkeit appelliert. Die teilnehmenden Studierenden wurden gebeten, eine ehrliche und korrekte Einschätzung abzugeben, ob ihre Antworten den Inhalten der Musterantworten entsprechen, die ihnen im Anschluss an eine Aufgabe gezeigt wurden. Zudem erfolgte der Hinweis, dass der Kompetenztest vollständig anonym erfolgt. Die anonyme Durchführung des Tests folgte der Empfehlung von Brown, Andrade und Chen, die darin einen positiven Einfluss auf die Genauigkeit der studentischen Selbstevaluationen sehen (Brown, Andrade & Chen 2015). Unterstrichen wurde dies dadurch, dass sich alle Studierende mit derselben Kennung zu dem Test einloggten. Der Test wurde online auf der Plattform www.m-lernen.de durchgeführt.

Einen Überblick über das methodische Vorgehen bei der Datenauswertung gibt die Abbildung 1. In einem ersten Schritt wurden in dem Datensatz alle Zeichen bereinigt, die keine alphanumerische Ausprägung hatten oder ein Satzzeichen darstellten. Im zweiten Schritt, dem *Preprocessing* der Daten, wurde ein Corpus gebildet. Außerdem wurden die Daten in einzelne Sätze bzw. Wörter zerlegt. Anschließend wurden die Daten in Trainings- und Testdaten gesplittet (70 % bzw. 30 %). Die Zuordnung eines Datensatzes erfolgte nach dem Zufallsprinzip. Im Rahmen des *Feature Extraction* wurden für alle Datensätze drei Eigenschaften ausgewertet: 1. die Länge der Antworten (da falsche Antworten tendenziell kürzer ausfallen als richtige) sowie 2. ob eine zustimmende bzw. 3. ablehnende Haltung zu den vorgelegten Beispielen in den studentischen Antworten ersichtlich ist. Zudem wurde auf Basis der Trainingsdaten ein Bag-of-Words-Modell erstellt. Auf Grundlage dieses Bag-of-Words-Modells wurden fünf verschiedene Modelle des maschinellen Lernens entwickelt, um die Labels der Datensätze vorherzusagen: Multinomial Naïve Bayes, logistische Regression, Support Vector Machines in zwei verschiedenen Ausprägungen (als SGDClassifier und LinearSVC) sowie ein Random Forest Classifier. Die fünf entwickelten Modelle wurden dann genutzt, um für jeden Datensatz sowohl des Trainingssets als auch des Testsets eine Vorhersage der Klasse (richtig/falsch) zu treffen.

Gemeinsam mit der Länge der Antworten sowie der Merkmale Zustimmung/Ablehnung flossen die Vorhersagen der oben beschriebenen fünf Methoden in die Modellierung eines neuronalen Netzes ein. Zur Modellierung wurde das im dritten Arbeitsschritt erstellte Trainingsset genutzt. Damit kein Overfitting stattfindet und das neuronale Netz besser aus den Daten generalisiert wird, wurde die Länge der Antworten in eine kategoriale Variable mit vier Klassen überführt. Das neuronale Netz wurde als mehrschichtiges Perzeptron mit einer verborgenen Schicht angelegt. Das Training erfolgte als Batch-Training. Als Aktivierungsfunktion wurde Softmax und als Fehlerfunktion die Kreuzentropie gewählt. Das Training wurde automatisch abgebrochen, wenn in einem aufeinanderfolgenden Schritt keine Verringerung des Fehlers stattfand. Die Anwendung des neuronalen Netzes auf das Trainings- und Testset resultierte in einer finalen Vorhersage der Klasse.

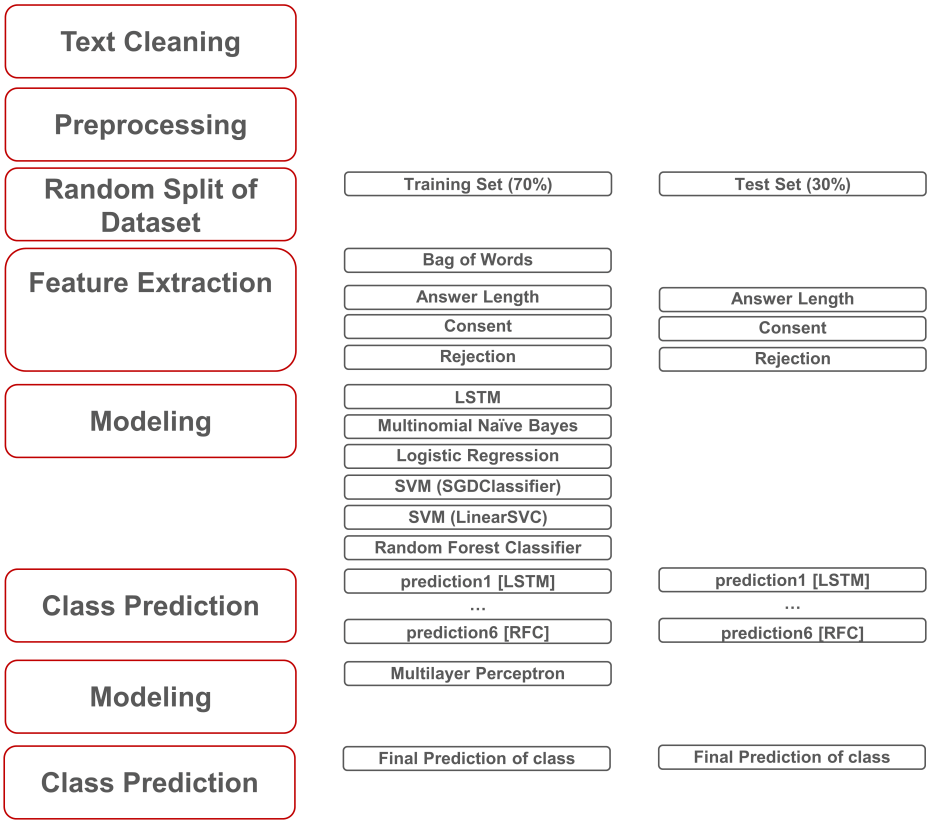

Abbildung 1: Methodisches Vorgehen bei der Datenauswertung (Quelle: Eigene Darstellung)

3 Beschreibung des Datensatzes

Insgesamt haben 290 Studierende an dem Test zur Messung der Kompetenz teilgenommen. Über die 12 Aufgaben hinweg gingen 3208 Antworten in die Auswertung ein. Die meisten Antworten fielen dabei vergleichsweise kurz aus. Im Durchschnitt bestanden die Antworten aus 7,39 Wörtern (sd = 7,656).

Tabelle 1: Antwortverhalten der Teilnehmenden über die Aufgaben hinweg

Aufgabe	1	2	3	4	5	6	7	8	9	10	11	12
n	290	274	276	278	275	271	262	263	262	258	253	246
Mittelwert	11,44	11,68	10,91	8,14	6,86	6,44	6,11	6,44	4,50	6,65	3,09	5,38
Std.-Abw.	9,312	8,752	9,125	8,506	6,740	6,385	6,973	5,237	5,831	7,295	4,030	6,331
richtig	174	112	140	112	162	197	202	234	91	110	108	129
falsch	116	162	136	166	113	74	60	29	171	148	145	117
richtig (%)	60	40,9	50,7	59,7	58,9	72,7	77,1	89,0	34,7	57,4	57,3	52,4
falsch (%)	40	59,1	49,3	40,3	41,1	27,3	22,9	11,0	65,3	42,6	42,7	47,6

Dabei kann festgestellt werden, dass die Antworten über den Test hinweg zunehmend kürzer ausfielen. Während die ersten drei Aufgaben noch vergleichsweise umfangreich beantwortet wurden, nimmt die Länge der Antworten ab der vierten Aufgabe deutlich ab.

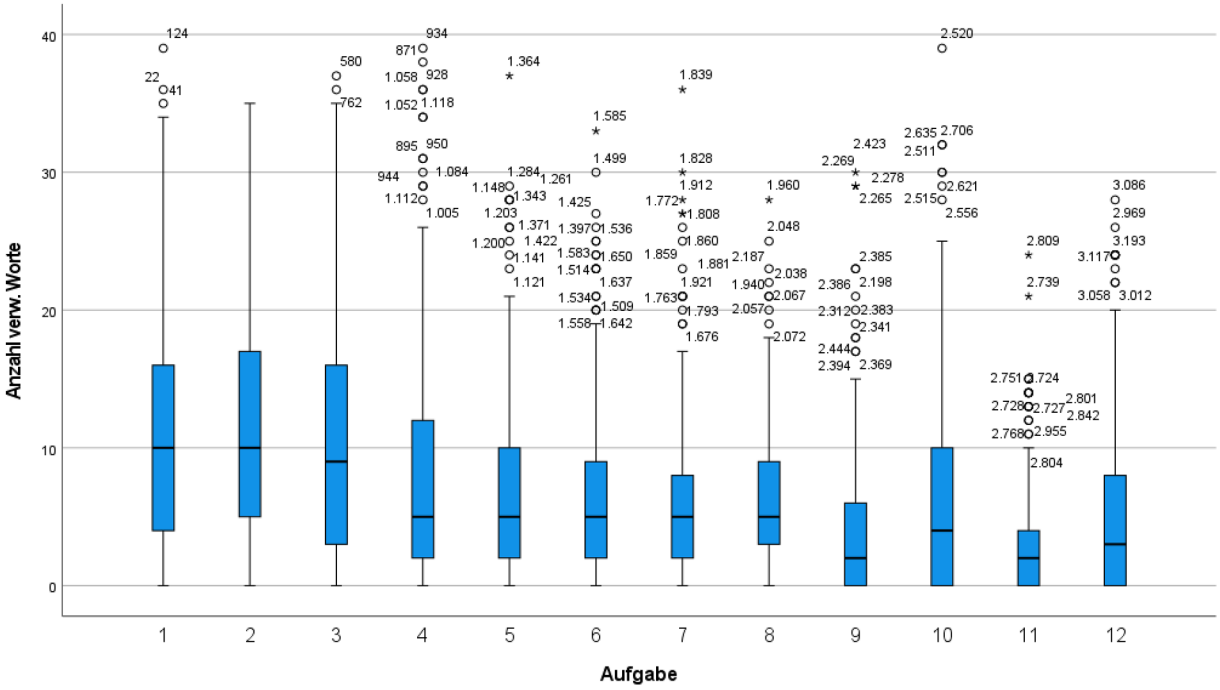

Abbildung 2: Länge der Antworten im Verlauf des Tests (Quelle: Eigene Darstellung)

Zur Prüfung, ob zwischen den beiden Gruppen Mittelwertunterschiede bezüglich der Länge der Antworten existieren, wurde ein T-Test durchgeführt. Dabei zeigten sich signifikante Unterschiede zwischen den Gruppen (Signifikanz 2-seitig: 0,007). Während bei den als richtig klassifizierten Antworten durchschnittlich 7,72 Wörter verwendet wurden (n = 1.771, Std.-Abweichung: 7,615), fielen die als falsch klassifizierten Antworten mit 6,99 durchschnittlich verwendeten Wörtern signifikant kürzer aus (n = 1.437; Std.-Abweichung: 7,690).

4 Ergebnisse

Um die Übereinstimmung zwischen der studentischen Selbstevaluation und der Expertenbewertung zu bestimmen, wurde Cohens Kappa ermittelt. Über alle Aufgaben hinweg fällt die Übereinstimmung zwischen der Selbstevaluation der Studierenden sowie der Expertenbewertung nur schwach aus (Cohens Kappa: 0,306; bei einem Grad der Übereinstimmung zwischen 0,1 und ≤ 0,4 sprechen Wirtz und Caspar von einer schwachen Übereinstimmung, zwischen 0,4 und ≤ 0,6 von einer deutlichen Übereinstimmung, Wirtz & Caspar 2002).[4]

Betrachtet man die einzelnen Aufgaben, dann zeigen sich deutliche Unterschiede bezüglich der Übereinstimmung. Insbesondere bei den Fragen 2 bis 5 fällt diese nur schwach aus. Lediglich die Aufgabe 7 kann eine deutliche Übereinstimmung zwischen der Selbstevaluation der Studierenden sowie der Expertenbewertung aufweisen.

4 Autor und Experte sind in der vorliegenden Studie ein und dieselbe Person.

Tabelle 2: Übereinstimmung zwischen der Selbstevaluation der Studierenden sowie der Expertenbewertung

Aufgabe	1	2	3	4	5	6	7	8	9	10	11	12
Cohens Kappa	,309	,133	,156	,137	,124	,211	,423	,252	,227	,339	,165	,245

Die wesentliche Ursache für die schwache Übereinstimmung der Bewertungen liegt darin, dass deutlich mehr Antworten, welche der Experte als falsch beurteilte, durch die Studierenden als korrekt klassifiziert wurden. Vergleichsweise selten bewerten die Studierenden deren Antworten als falsch, während der Experte diese als richtig kategorisiert (n = 140). Auch wenn die Fallzahlen gering sein mögen, macht dies deutlich, dass eine Bewertung der Antwort als korrekt durch Expertinnen und Experten nicht immer mit einem Verständnis auf studentischer Seite einhergehen muss.

Tabelle 3: Gegenüberstellung Selbstevaluation/Bewertung durch Expertinnen und Experten

		Bewertung durch Expertinnen und Experten		
		falsch	**richtig**	**Gesamt**
Selbstevaluation	falsch	1294	140	1434
	richtig	1024	746	1770
	Gesamt	2318	886	3204

Die Unterschiede zwischen der Selbstevaluation durch die Studierenden und der Expertenbewertung resultieren unter anderem daher, dass der Experte bei einer unbeantworteten Aufgabe konsistent diese Aufgabe als falsch gewertet hat, während 42,9 % der Selbstevaluationen diese als korrekt klassifizierten. Betrachtet man das Kategorisierungsverhalten der Studierenden über die verschiedenen Aufgaben hinweg, dann fällt auf, dass insbesondere ab Frage 9 ein deutlich höherer Anteil der Aufgaben unbeantwortet blieb. Eine zweite Auffälligkeit besteht darin, dass die absolute Zahl der Studierenden, die eine unbeantwortete Aufgabe als richtig kategorisiert haben, ab Aufgabe 3 weitgehend konstant bleibt (und um einen Wert von 22 Studierenden schwankt), während die Zahl der Studierenden, die eine unbeantwortete Frage als falsch kategorisiert haben, deutlich schwankt.

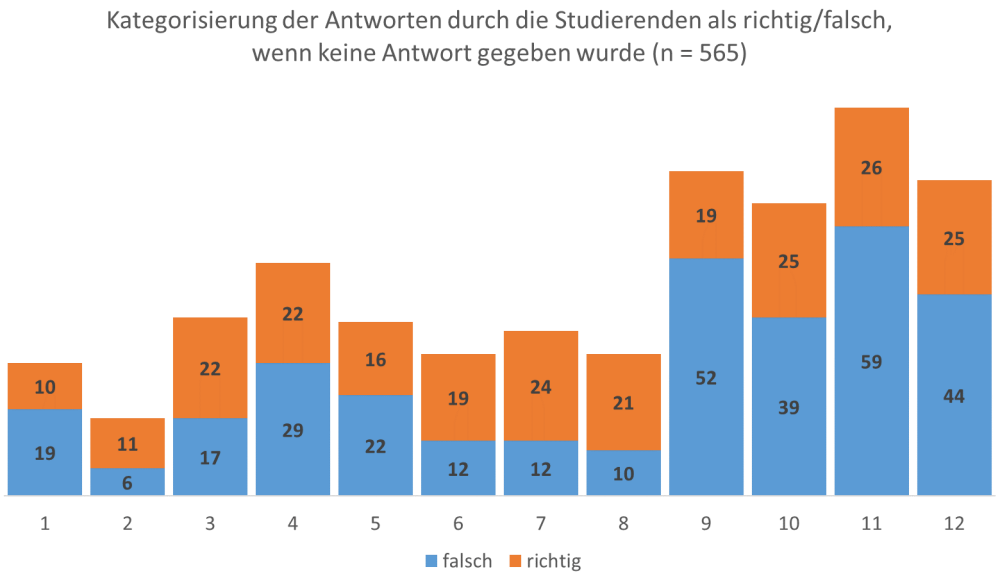

Abbildung 3: Kategorisierung der Antworten durch die Studierenden (eigene Darstellung)

Eine Tendenz, die Aufgaben unbeantwortet zu lassen und diese dennoch als richtig zu kategorisieren, findet sich bei den Studierenden nur vereinzelt. Lediglich 21 der 290 Studierenden haben vier oder mehr Aufgaben unbeantwortet gelassen und diese dennoch als richtig charakterisiert.

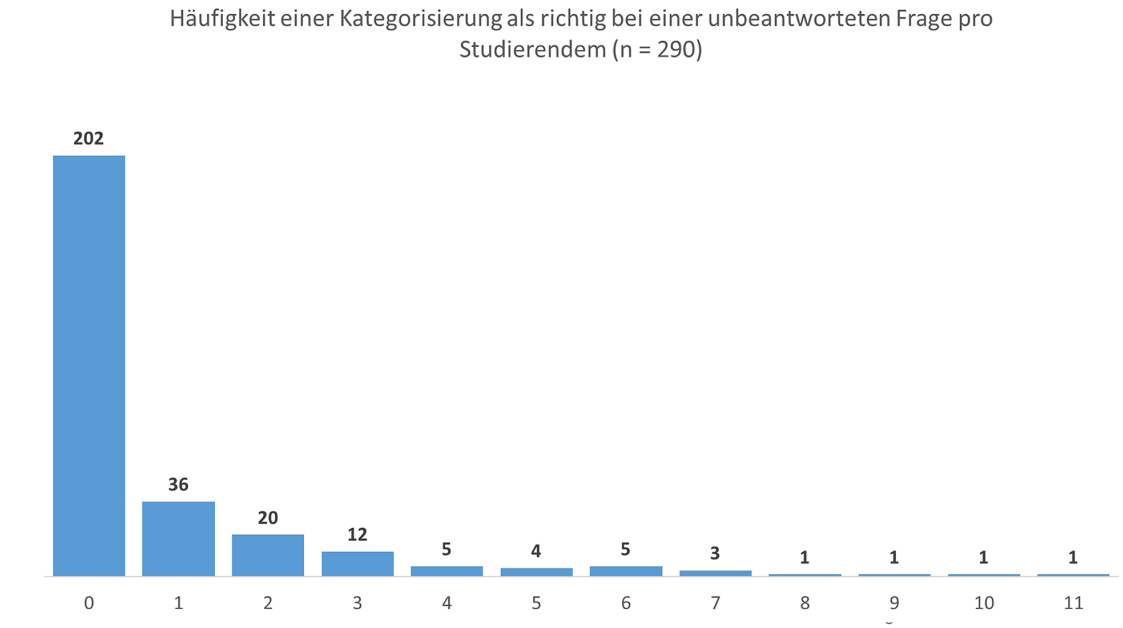

Abbildung 4: Kategorisierungsverhalten bei unbeantworteten Fragen (eigene Darstellung)

Zur Bestimmung der Eignung einer Selbstevaluation im Vergleich zu einer Bewertung durch Expertinnen und Experten wurde als Bewertungsmaßstab die Anzahl der korrekt klassifizierten Trainingsdaten verwendet. Der nachfolgende Boxplot zeigt erhebliche Unterschiede zwischen den beiden Bewertungsverfahren. Während bei der Selbstevaluation im Durchschnitt über alle zwölf Aufgaben hinweg etwas mehr als 80 Prozent der Antworten des Testsets (n = 966) korrekt klassifiziert wurden, waren dies bei der Expertenbewertung über 95 Prozent.

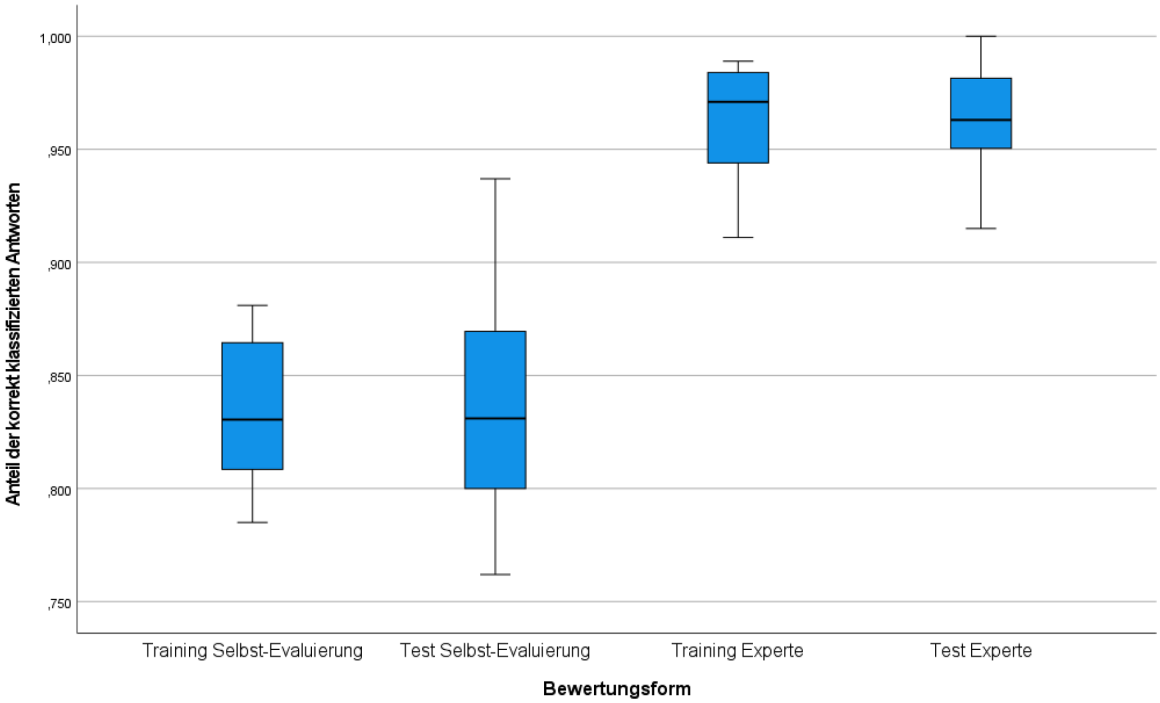

Abbildung 5: Boxplots nach Bewertungsform und Datenset (eigene Darstellung)

Die Ergebnisse machen damit deutlich, dass die Vorteile einer Selbstevaluation durch die Studierenden mit einem erheblichen Genauigkeitsverlust bei einer automatisierten Auswertung erkauft werden. Einem Experten gelingt es deutlich besser, einem einheitlichen Bewertungsschema zu folgen, als dies bei einer Selbstevaluation durch die Studierenden der Fall ist.

5 Implikationen

Bei der automatisierten Auswertung von Short-Answer-Aufgaben wird gewöhnlich nach Mustern gesucht, die von den Antworten auf die Bewertungen schlussfolgern lassen. Diese von sich aus schon sehr anspruchsvolle Aufgabe wird durch Ungenauigkeiten in den Bewertungen wesentlich erschwert. In der Folge leidet die Qualität der automatisierten Bewertung der Antworten. Da in der Lehrpraxis meist ein hoher Anspruch an die korrekte Kategorisierung der Antworten besteht (man denke nur an die ansonsten entstehenden Konfusionen bei den Lernenden), ergibt sich ein hoher Qualitätsanspruch an die Bewertungen, die zur Aufdeckung der oben genannten Muster verwendet werden, und stellt sich die Frage, ob eine Selbstevaluation der geeignete Weg zu deren Erstellung ist.

Dies führt wiederum zur Frage, von welchen Einflussfaktoren eine ehrliche/unehrliche Selbst-Evaluation abhängig ist und wie diese generell im positiven Sinne beeinflusst werden kann. In der Literatur wird vielfach ein Zusammenhang zwischen der Ehrlichkeit/Unehrlichkeit und dem aus der jeweiligen Verhaltensweise gezogenen Nutzen gemacht (Becker, 2000). Unehrliches Verhalten wird demnach begünstigt, wenn dieses in einem Nutzen resultiert.

Die Studie zeigt auf, dass Unehrlichkeit sich nicht auf Einzelfälle beschränkt und als Phänomen auch dann bedeutsam ist, wenn ein externer Nutzen nicht vorhanden ist. Denn weder wurde der Test bewertet noch konnte das Ansehen der eigenen Person bei dem Dozierenden gesteigert werden. Die Teilnahme unter einer einheitlichen Kennung und in vollständig anonymer Form sowie der Hinweis darauf, dass der Test bzw. dessen Ergebnisse einzig und allein den Teilnehmenden dienen, hat diesen Umstand verdeutlicht.

Ist ein externer Nutzen nicht vorhanden, dann kann der Nutzen nur intern vorliegen oder die Teilnehmenden haben einen externen Nutzen erwartet, konnten diesen nicht ausschließen oder wollten den Dozierenden nicht enttäuschen. Alle genannten Punkte liefern nur vergleichsweise schwache Gründe für ein unehrliches Verhalten. Insofern erscheint auch dann eine Auseinandersetzung mit unehrlichem Verhalten notwendig zu sein, wenn der externe Nutzen dieser Verhaltensweise nur gering ist. Dies gilt insbesondere vor dem Hintergrund, dass die Lernenden einen Nachteil durch ein unehrliches Verhalten hatten, indem Lernfelder in diesem Fall nur unvollständig aufgedeckt wurden.

Ist der externe Nutzen eines unehrlichen Verhaltens wie im vorliegenden Fall nicht vorhanden bzw. vergleichsweise gering, dann rücken interne Belohnungen, wie die Aufrechterhaltung eines positiven Selbstkonzeptes, in den Vordergrund. Aus Sicht von Mazar et al. werden der Unehrlichkeit durch das Selbstkonzept gewisse Grenzen gesetzt (Mazar et al., 2008). Demnach ist der Mensch nur bis zu dem Grad unehrlich, wie dies mit seinem Selbstkonzept vereinbar ist, was wiederum u. a. vom Kontext, der Tätigkeit sowie der Interpretation des eigenen Verhaltens abhängig ist (Mazar et al., 2008).

Legt man diese Sichtweise zugrunde, dann leiten sich daraus verschiedene Ansatzpunkte für die Beeinflussung eines ehrlichen Verhaltens in einer Selbst-Evaluation ab. Dazu zählt insbesondere die Aufrechterhaltung eines positiven Selbstkonzeptes, beispielsweise indem die ersten Fragen einen niedrigen Schwierigkeitsgrad aufweisen, an die Ehrlichkeit der Lernenden appelliert wird oder statt der dichotomen Kategorisierung der Antworten in richtig/falsch weniger offensive Labels verwendet werden. Auch die Belohnung eines ehrlichen Verhaltens für den Fall, dass in der Selbst-Evaluation Defizite eingestanden werden, begünstigt die Aufrechterhaltung eines positiven Selbstkonzeptes. Zudem erscheint es wichtig, den Kontext und damit die Rahmenbedingungen so

zu gestalten, dass eine ehrliche Selbst-Evaluation gefördert wird. Die im vorliegenden Projekt genutzte Selbst-Evaluation auf Basis einer Musterlösung scheint ein unehrliches Verhalten vergleichsweise einfach zu machen. Es benötigt nicht viel Fantasie, um sich eine Interpretation zurechtzulegen, bei der ein Konflikt zwischen dem unehrlichen Verhalten und dem Selbstkonzept (als ehrliche Person) vermieden wird; beispielsweise indem postuliert wird, dass eine Aussage doch so gemeint war, dass einzelne Aspekte der Musterantwort entsprachen, man die Lösung im Kopf hatte und nur nicht verschriftlicht hat oder dass die Antwort inhaltlich korrekt war (auch wenn sie nicht dem Erwartungshorizont zu einer Aufgabe entsprach). Vor allem gilt es jedoch, die Fähigkeit der Lernenden zur Selbst-Evaluation zu stärken, denn letztlich profitieren diese von einer ehrlichen Bewertung am meisten.

6 Limitationen

Das hier vorgestellte Projekt tendiert dazu, die Genauigkeit des automatisierten Scorings zu überschätzen, wenn dieses auf Basis der Expertenbewertung erfolgte. Dafür sind verschiedene Ursachen verantwortlich. So wurden die Antworten lediglich von einem einzigen Experten bewertet. Daher kann nicht ausgeschlossen werden, dass andere Expertinnen und Experten zu einer anderen Bewertung der Antworten gelangen würden. Dies gilt insbesondere für den Fall, dass die Bewertung nicht anhand einer Musterlösung stattfindet. Zudem wurden aufgrund der Größe des Stichprobenumfangs in dem vorliegenden Projekt lediglich ein Trainingsset und ein Testset erstellt. Da keine Unterteilung der Daten in Trainings-Set, Validations-Set und Holdout-Set stattgefunden hat, kann bei einer zukünftigen Anwendung die Performance möglicherweise deutlich sinken.

7 Fazit

Die automatisierte Auswertung von Short-Answer-Aufgaben ist ein Feld, auf dem derzeit ein hoher Aufwand betrieben werden muss, um selbst bei eng definierten Aufgaben die Antworten mit einer hohen Genauigkeit zu bewerten. Daher existiert ein großes Interesse der Praxis daran, den Aufwand der Entwicklung und Einführung von operativen Systemen in diesem Bereich zu verringern. Ein Ansatz, der im Rahmen dieses Projekts geprüft wurde, bestand darin, die Antworten durch die Studierenden anhand von bereitgestellten Musterlösungen selbst evaluieren zu lassen. Die Ergebnisse des Projektes machen deutlich, dass diese Vorgehensweise zu einem zusätzlichen Rauschen in den Daten führt, worunter wiederum die Genauigkeit der Bewertungen leidet. Zudem bewerten die Studierenden deren Antworten gegenüber dem Experten erheblich besser. Für eine automatisierte Auswertung der Short-Answer-Aufgaben kommen die durch die Selbstevaluation erzeugten Daten daher nur sehr bedingt infrage.

Betrachtet man die Selbstevaluation nur unter dem Aspekt der Genauigkeit einer automatisierten Auswertung, dann erscheint diese Vorgehensweise der klassischen Bewertung durch Expertinnen und Experten deutlich unterlegen zu sein. Die geringere Genauigkeit bedeutet jedoch nicht, dass auf diese verzichtet werden sollte. Denn in der Praxis findet sich nicht selten eine Kluft zwischen der Wahrnehmung der Lernenden und der tatsächlichen Performance (Osterhage et al., 2019). Die Kombination aus Selbstevaluation und automatisierter Auswertung bietet damit die Chance für die Lernenden zur Selbstreflexion. Zu diesem Zweck sollte die Auswertung der Antworten auf einer detaillierteren Ebene stattfinden, als dies in dem durchgeführten Projekt der Fall war. So könnte die Selbstreflexion insbesondere dann profitieren, wenn sehr viel konkreter aufgezeigt würde, wo Überschneidungsflächen zwischen den Antworten und der Musterlösung vorhanden sind und welche Teile der Musterlösung sich in den Antworten nicht wiedergefunden haben.

Literatur

Becker, G. S. (2000). Crime and Punishment: an Economic Approach. In N. G. Fielding, A. Clarke & R. Witt (Hrsg.), *The Economic Dimensions of Crime* (p. 13–68). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-62853-7_2

Bey, A., Jermann, P. & Dillenbourg, P. (2018). A Comparison between Two Automatic Assessment Approaches for Programming. An Empirical Study on MOOCs. *Journal of Educational Technology & Society, 21 (2)*, 259–272. Online unter: http://www.jstor.org/stable/26388406 [24.10.2021].

Brown, G. T. L., Andrade, H. L. & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22 (4)*, 444–457.

Brown, G. T. L. & Harris, L. R. (2013). Student Self-Assessment. In J. H. McMillan (Hrsg.), *The SAGE Handbook of Research on Classroom Assessment*, 367–393.

Elmahdi, I., Al-Hattami, A. & Fawzi, H. (2018). Using Technology for Formative Assessment to Improve Students' Learning. *Turkish Online Journal of Educational Technology - TOJET, 2 (2018)*, 182–188.

Haley, D. T., Thomas, P., De Roeck, A. & Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML. *ACM 9th Proceedings of the Ninth Australasian Conference on Computing Education, 66*, 35–42.

Hegarty-Kelly, E. & Mooney, A. (2021). Analysis of an automatic grading system within first year Computer Science programming modules. *CEP '21: Computing Education Practice*, 17–20.

Henderson, M., Al-Rfou, R., Strope, B., Yun-hsuan, S., Lukacs, L., Guo, R. et al. (2017). Efficient Natural Language Response Suggestion for Smart Reply. Online unter: https://arxiv.org/pdf/1705.00652 [24.10.2021].

Horbach, A. (2019). *Analyzing Short-Answer Questions and their Automatic Scoring. Studies on Semantic Relations in Reading Comprehension and the Reduction of Human Annotation Effort*. Dissertation. Universität des Saarlandes, Saarbrücken. Online unter: https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/27666/4/Dissertation_UdS_Horbach_kv.pdf [24.10.2021].

Klein, H. A., Levenburg, N. M., McKendall, M. & Mothersell, W. (2007). Cheating During the College Years: How do Business School Students Compare? *Journal of Business Ethics,* 72 (2), 197–206. https://doi.org/10.1007/s10551-006-9165-7

Klenowski, V. (1995). Student Self-evaluation Processes in Student-centred Teaching and Learning Contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice, 2 (2)*, 145–163. https://doi.org/10.1080/0969594950020203

Lewis Sevcikova, B. (2018). Human versus Automated Essay Scoring: A Critical Review. *Arab World English Journal (AWEJ), 9 (2)*, 157–174. https://doi.org/10.24093/awej/vol9no2.11

Madnani, N. & Cahill, A. (2018). Automated scoring: Beyond natural language processing. *Proceedings of the 27th International Conference on Computational Linguistics*, 1099–1109.

Mazar, N., Amir, O. & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research, 45(6)*, 633–644. https://doi.org/10.1509/jmkr.45.6.633

Ndukwe, I. G., Amadi, C. E., Nkomo, L. M. & Daniel, B. K. (2020). Automatic Grading System Using Sentence-BERT Network. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin & E. Millán (Hrsg.), *Artificial Intelligence in Education, Bd. 12164*. Cham: Springer International Publishing, 224–227.

Osterhage, J. L., Usher, E. L., Douin, T. A. & Bailey, W. M. (2019). Opportunities for Self-Evaluation Increase Student Calibration in an Introductory Biology Course. *CBE life sciences education, 18 (2)*, ar16. https://doi.org/10.1187/cbe.18-10-0202

Pado, U. & Kiefer, C. (2015). Short answer grading: When sorting helps and when it doesn't. *Proceedings of the Nodalida-2015 workshop, Bd. 114*, 42–50. Online unter: https://www.aclweb.org/anthology/W15–1905.pdf [24.10.2021].

Sadr, H. & Nazari Solimandarabi, M. (2019). Presentation of an Efficient Automatic Short Answer Grading Model Based on Combination of Pseudo Relevance Feedback and Semantic Relatedness Measures. *Journal of Advances in Computer Research (JACR),* 10 (2), 17–30. Online unter: http://jacr.iausari.ac.ir/article_663355.html [24.10.2021].

Surya, K., Gayakwad, E. & Nallakaruppan, M. K. (2019). Deep learning for Short Answer Scoring. *International Journal of Recent Technology and Engineering (IJRTE),* 7 (6), 1712–1715.

Suwanarak, K. (2018). Self-Evaluation of Thai Adult Learners in English Writing Practice. *3L, 24 (2)*, 95–111. https://doi.org/10.17576/3L-2018-2402-08

Whitley, B. E. (1998). Factors associated with cheating among college students: a review. *Research in Higher Education,* 39 (3), 235–274. https://doi.org/10.1023/A:1018724900565

Wirtz, M. A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe Verl. für Psychologie.

Autor

Prof. Dr. Ulrich Bucher, DHBW Stuttgart, Studienzentrum Dienstleistungsmanagement, Stuttgart, Deutschland, E-Mail: ulrich.bucher@dhbw-stuttgart.de

Zitiervorschlag: Bucher, U. (2022). Automatisierte Auswertung von Short-Answer-Aufgaben zur Bestimmung der erworbenen Kompetenzen. *die hochschullehre*, Jahrgang 8/2022. DOI: 10.3278/HSL2212W. Online unter: wbv.de/die-hochschullehre