Artikel erschienen in:

Jolanda Hermanns (Hrsg.)

PSI-Potsdam
Ergebnisbericht zu den Aktivitäten im Rahmen der Qualitätsoffensive Lehrerbildung (2019 – 2023)

(Potsdamer Beiträge zur Lehrkräftebildung und Bildungsforschung ; 3)

2023 – 393 S.
ISBN 978-3-86956-568-2
DOI https://doi.org/10.25932/publishup-60187

Empfohlene Zitation:
Peter Wulff; Lukas Mientus; Anna Nowak; Andreas Borowski: KI-basierte Auswertung von schriftlichen Unterrichtsreflexionen im Fach Physik und automatisierte Rückmeldung, In: Jolanda Hermanns (Hrsg.): PSI-Potsdam. Ergebnisbericht zu den Aktivitäten im Rahmen der Qualitätsoffensive Lehrerbildung (2019–2023) (Potsdamer Beiträge zur Lehrkräftebildung und Bildungsforschung 3), Potsdam, Universitätsverlag Potsdam, 2023, S. 103–115.
DOI https://doi.org/10.25932/publishup-61636

KI-basierte Auswertung von schriftlichen Unterrichtsreflexionen im Fach Physik und automatisierte Rückmeldung

Peter Wulff[1], *Lukas Mientus*[2], *Anna Nowak*[3] & *Andreas Borowski*[4]

[1] Pädagogische Hochschule Heidelberg, 0000-0002-5471-7977
[2] Universität Potsdam, 0000-0001-5344-4770
[3] Universität Potsdam, 0000-0002-6890-3463
[4] Universität Potsdam, 0000-0002-9502-0420

ZUSAMMENFASSUNG: Für die Entwicklung professioneller Handlungskompetenzen angehender Lehrkräfte stellt die Unterrichtsreflexion ein wichtiges Instrument dar, um Theoriewissen und Praxiserfahrungen in Beziehung zu setzen. Die Auswertung von Unterrichtsreflexionen und eine entsprechende Rückmeldung stellt Forschende und Dozierende allerdings vor praktische wie theoretische Herausforderungen. Im Kontext der Forschung zu Künstlicher Intelligenz (KI) entwickelte Methoden bieten hier neue Potenziale. Der Beitrag stellt überblicksartig zwei Teilstudien vor, die mit Hilfe von KI-Methoden wie dem maschinellen Lernen untersuchen, inwieweit eine Auswertung von Unterrichtsreflexionen angehender Physiklehrkräfte auf Basis eines theoretisch abgeleiteten Reflexionsmodells und die automatisierte Rückmeldung hierzu möglich sind. Dabei wurden unterschiedliche Ansätze des maschinellen Lernens verwendet, um modellbasierte Klassifikation und Exploration von Themen in Unterrichtsreflexionen umzusetzen. Die Genauigkeit der Ergebnisse wurde vor allem durch sog. Große Sprachmodelle gesteigert, die auch den Transfer auf andere Standorte und Fächer ermöglichen. Für die fachdidaktische Forschung bedeuten sie jedoch wiederum neue Herausforderungen, wie etwa systematische Verzerrungen und Intransparenz von Entscheidungen. Dennoch empfehlen wir, die Potenziale der KI-basierten Methoden gründlicher zu erforschen und konsequent in der Praxis (etwa in Form von Webanwendungen) zu implementieren.

KEYWORDS: Künstliche Intelligenz, Maschinelles Lernen, Natural Language Processing, Reflexion, Professionalisierung

ABSTRACT:[1] For the development of professional competencies in pre-service teachers, reflection on teaching experiences is proposed as an important tool to link theoretical knowledge and practice. However, evaluating reflections and providing appropriate feedback poses challenges of both theoretical and practical nature to researchers and educators. Methods associated with artificial intelligence research offer new potentials to discover patterns in complex datasets like reflections, as well as to evaluate these automatically and create feedback. In this article, we provide an overview of two sub-studies that investigate, using artificial intelligence methods such as machine learning, to what extent an evaluation of reflections of pre-service physics teachers based on a theoretically derived reflection model and automated feedback are possible. Across the sub-studies, different machine learning approaches were used to implement model-based classification and exploration of topics in reflections. Large language models in particular increase the accuracy of the results and allow for transfer to other locations and disciplines. However, entirely new challenges arise for educational research in relation to large language models, such as systematic biases and lack of transparency in decisions. Despite these uncertainties, we recommend further exploring the potentials of artificial intelligence-based methods and implementing them consistently in practice (for example, in the form of web applications).

KEYWORDS: Artificial intelligence, machine learning, natural language processing, reflexion, professionalization

1 PROFESSIONALISIERUNG ANGEHENDER (PHYSIK-)LEHRKRÄFTE

Für die Entwicklung professioneller Handlungskompetenzen in den verschiedenen Phasen der Lehrkräfteausbildung wird die Unterrichtsreflexion als wichtiges Instrument betrachtet, um Theoriewissen und Praxis miteinander zu verzahnen (Carlson et al., 2019; Darling-Hammond, 2012; Korthagen & Kessels, 1999). Dabei analysieren angehende Lehrkräfte eigenen oder fremden Unterricht und verfassen unter Anleitung Reflexionstexte, die dann modellbasiert ausgewertet werden. Oftmals sind die Unterrichtsreflexionen zu Beginn eher beschreibend und inhaltlich oberflächlich (Mena-Marcos et al., 2013; Sorge et al., 2018). Um angehende Lehrkräfte im Reflexionsprozess zu unterstützen, können Rückmeldungen zu den Unterrichtsreflexionen gegeben werden (Lai & Calandra, 2007). Allerdings fehlen oftmals die Ressourcen auf Seiten der Forschenden und Dozierenden, um die Unterrichtsreflexionen systematisch, skalierbar und reliabel

1 Wurde unter Zuhilfenahme der generativen KI ChatGPT Version 4 (12. Mai 2023) erstellt, mit dem Prompt: „Übersetze den folgenden Text ins Englische". Verfügbar unter: https://chat.openai.com/?model=gpt-4. Einige Anpassungen auf Wortebene („aspiring teachers" → „pre-service teachers"; „class reflection" → „reflection"; „subjects" → „disciplines") sowie Satzebene wurden vorgenommen.

auszuwerten. Fortschritte im Bereich der Forschung zu Künstlicher Intelligenz (KI) können hier Abhilfe schaffen, indem sie sowohl Forschenden neue, evidenzbasierte Analysen von Unterrichtsreflexionen ermöglichen als auch Dozierende entlasten und angehenden Lehrkräften regelmäßig und angemessen eine Rückmeldung zu deren Unterrichtsreflexionen erstellen (Yeadon et al., 2023).

In diesem Beitrag berichten wir von einem Projekt im Bereich der Physikdidaktik der Universität Potsdam im Rahmen der Qualitätsoffensive Lehrerbildung (Schwerpunkt 2 – Schulpraktische Studien, PSI Potsdam), das zum Ziel hatte, KI-Methoden zur automatisierten Auswertung schriftlicher Unterrichtsreflexionen angehender Physiklehrkräfte inklusive der Erstellung entsprechender Rückmeldungen zu erproben. Die Reflexion des eigenen Unterrichts wird als Bedingungsfaktor für die professionelle Entwicklung von Lehrkräften gesehen (Darling-Hammond, 2012; Korthagen, 2005; Mientus et al., 2022). Die kontinuierliche Reflexion über die Berufslaufbahn hinweg kann dabei helfen, die eigene Praxis auf Basis theoretischen Wissens zu verstehen und weiterzuentwickeln.

2 MASCHINELLES LERNEN ALS AUSWERTUNGSVERFAHREN IN DER NATURWISSENSCHAFTSDIDAKTIK

In der fachdidaktischen Forschung werden KI-Methoden bereits vielfach verwendet, beispielsweise um offene Antwortformate wie Essays automatisiert auszuwerten (Zhai et al., 2020). Dabei kommen unterschiedliche datenzentrierte Auswerteverfahren zum Einsatz, die zumeist auf Methoden des maschinellen Lernens (ML) sowie der natürlichen, computerbasierten Sprachverarbeitung (natural language processing, NLP) basieren. ML ist ein induktiver Ansatz des computerbasierten Problemlösens. Hierbei können Verfahren, die auf Basis kodierter Datensätze eine Zuordnung lernen und die Zuordnung dann an ungesehenen Daten vornehmen können (sog. supervised ML), von solchen unterschieden werden, die in komplexen Daten Muster erkennen können, ohne dass kodierte Datensätze bereits vorliegen müssen (sog. unsupervised ML) (Marsland, 2015). NLP bezeichnet die strukturierte und systematische (zumeist computerbasierte) Auswertung von Sprachdaten, oft unter Zuhilfenahme von ML-Verfahren.

In der naturwissenschaftsdidaktischen Forschung konnten ML und NLP bereits gewinnbringend eingesetzt werden. Anwendungen in den Naturwissenschaftsdidaktiken umfassen die automatisierte und reliable Kodierung von Konzepten (Vorstellungen zu Evolution, Überlegungen zur Thermodynamik oder Modellvorstellungen) in offenen Antworten (Donnelly et al., 2015; Krüger & Krell, 2020; Nehm & Härtig, 2012). Ebenso konnten explorativ Themen in Physik-Konferenzabstrakten identifiziert werden und deren Variation über einen be-

stimmten Zeitraum (Odden et al., 2020) oder Erklärungsansätze zu den Jahreszeiten in Interviewtranskripten (Sherin, 2013). Oft greifen diese Arbeiten noch auf vergleichsweise einfache ML-Algorithmen zurück. Des Weiteren bezogen sich diese Arbeiten zumeist auf überschaubare offene Textantworten von Schüler:innen (Zhai et al., 2020). In den letzten Jahren sind insbesondere sog. Große Sprachmodelle (GSM) entwickelt worden, auf deren Basis die bisherigen Analysen in Bezug auf Sprachdaten noch weiterentwickelt werden können.

Insbesondere für die Analyse von Essays oder Berichten – wie etwa Unterrichtsreflexionen – bieten ML und NLP interessante Potentiale (Buckingham Shum et al., 2017). Beispielsweise könnten durch ML und NLP Textformen wie Reflexionstexte in der Naturwissenschaftsdidaktik häufiger und individualisierter eingesetzt werden, da angehende Lehrkräfte kontinuierlich und instantan eine Rückmeldung zu ihren Texten erhalten können, was sonst eher selten und nur mit vergleichsweise hohem Ressourcenaufwand stattfinden kann.

3 KI-BASIERTE AUSWERTUNG VON SCHRIFTLICHEN UNTERRICHTSREFLEXIONEN UND AUTOMATISIERTE RÜCKMELDUNG

Wie für die Analyse von Texten allgemein gilt auch für die Analyse von Unterrichtsreflexionen, dass ganzheitliche Bewertungsmaße (etwa eine Gesamtnote) oft nur unzureichend die Vielschichtigkeit dieser Lernprodukte berücksichtigen können (Poldner et al., 2014; Wang et al., 2008). Vielmehr sind analytische Ansätze notwendig, die einzelne Aspekte der Texte separat berücksichtigen. Automatisiertes Essay-Scoring berücksichtigt beispielsweise Aspekte wie die Kohärenz, den verwendeten Wortschatz oder die Satzlänge als wichtige Maße für die Qualität eines Textes. Um ML und NLP auf Unterrichtsreflexionen anzuwenden, ist es zunächst notwendig zu definieren, welche Merkmale eine gute Unterrichtsreflexion ausmachen. Neben der Qualität einer tiefen inhaltlichen, theoriebasierten Auseinandersetzung mit den eigenen Erfahrungen ist zunächst die Strukturierung des Reflexionsprozesses im Text wichtig. Unterrichtsreflexionen sollen nicht bloße Schilderungen von Erfahrungen und Bewertungen dieser Erfahrungen sein, sondern auch alternative Handlungsoptionen abwägen und Konsequenzen für die eigene Praxis ableiten (Hatton & Smith, 1995; Nowak et al., 2019). Nowak et al. (2019) und andere Forschende (Poldner et al., 2014; Ullmann, 2019; von Aufschnaiter et al., 2019) unterscheiden verschiedene Elemente, die eine vollständige (Unterrichts-)Reflexion enthalten sollte: Eine Darstellung der *Rahmenbedingungen* der Unterrichtsstunde, wie das Lernziel und zu vermittelnde Kompetenzen, eine *Beschreibung* von beobachteten Situationen und Er-

eignissen, eine *Bewertung* eben dieser Ereignisse und mögliche *Alternativen* für Handlungen. Schließlich sollen *Konsequenzen* für die eigene professionelle Entwicklung oder die Weiterentwicklung des Unterrichts abgeleitet werden, da Unterrichtsreflexion, im Gegensatz zur Analyse von Unterricht, stets auch die professionelle Entwicklung der angehenden Lehrkraft zum Thema machen soll (von Aufschnaiter et al., 2019).

4 KI-BASIERTE AUSWERTUNG VON UNTERRICHTSREFLEXIONEN VON PHYSIK-LEHRAMTSSTUDIERENDEN

4.1 Forschungsfragen

In der Physikdidaktik der Universität Potsdam haben wir uns zum Ziel gesetzt, basierend auf dem Reflexionsmodell von Nowak et al. (2019) verschiedene KI-Algorithmen anzuwenden, um Unterrichtsreflexionen angehender Physik-Lehramtsstudierender automatisiert auszuwerten. Auf Basis der unterschiedlichen Ansätze des ML (supervised und unsupervised) wurden folgende Forschungsfragen im Zusammenhang unserer KI-basierten Analysen der Unterrichtsreflexionen beantwortet:

1. Inwieweit kann mittels ML die Struktur im Sinne der Reflexionselemente automatisiert erfasst werden?
2. Auf welche Weise können Themen in den Unterrichtsreflexionen mit Hilfe von ML und NLP identifiziert werden?

4.2 Erhebung von Unterrichtsreflexionen

Die Stichprobe sowie die Instruktion wurden in Vorarbeiten beschrieben (Nowak et al., 2019). Zusätzlich zu diesen Daten haben wir kontinuierlich im Praxissemester Physik Selbstreflexionen zum Unterricht der angehenden Physiklehrkräfte erhoben. Darüber hinaus haben wir eine Unterrichtsvignette erstellt, die als Reflexionsanlass für eine Fremdreflexion in unterschiedlichen Seminaren im Physik-Lehramtsstudium eingesetzt wurde (Wulff, Buschhüter et al., 2022). Ausgehend von diesen Arbeiten haben wir mittlerweile einen Korpus an Unterrichtsreflexionen erfasst, der über 500 Unterrichtsreflexionen (auch anderer Standorte sowie Fächer) umfasst.

4.3 Klassifikation der Unterrichtsreflexionen mit supervised ML

Um Forschungsfrage 1 zu beantworten, haben wir in einer ersten Teilstudie zunächst die Unterrichtsreflexionen in elementare Kodiereinheiten (hier: Sätze) segmentiert und diese nach den Elementen im Rahmenmodell für Reflexion von Nowak et al. (2019) kodiert (Wulff et al., 2020). Die Elemente waren *Rahmenbedingungen, Beschreibung, Bewertung, Alternativen* und *Konsequenzen* (siehe Abschn. 3). Um sicherzustellen, dass die Elemente von unterschiedlichen menschlichen Kodierenden in gleicher Weise kodiert werden, haben wir unterschiedliche, qualifizierte Personen gebeten jeweils Teildatensätze zu kodieren und deren Übereinstimmung bestimmt. Es zeigte sich, dass eine substanzielle Übereinstimmung erreicht werden kann (Cohens Kappa). Anschließend wurden systematisch verschiedene etablierte ML-Algorithmen (Multinomial Bayes, Logistic Regression, Decision Tree, siehe: Wulff et al., 2020) angewendet, um die Zuordnung der Sätze (Größe des Trainingsdatensatzes) zu den Reflexionselementen zu lernen. Supervised ML funktioniert in der Weise, dass adaptiv bestimmte Parameter in den Modellen angepasst wurden, die im Trainingsdatensatz zur höchsten Klassifikationsgüte zwischen Mensch und Maschine führen. Das genaueste Modell wurde abschließend auf einem im Training nicht präsentierten Datensatz (dem Testdatensatz) getestet, um die Generalisierbarkeit abzuschätzen. Obwohl die ML-Modelle auf dem Validierungsdatensatz hinreichend gut funktionierten (Genauigkeit etwa 71 %), war ein substantieller Übereinstimmungsverlust für die Testdaten zu verzeichnen (Genauigkeit etwa 58 %).

Fortschritte im Bereich der KI-Forschung legen nahe, dass tiefe künstliche neuronale Netze besser generalisieren können als einfache ML-Algorithmen. Zumindest korrespondiert die Tiefe dieser Netze mit der Fähigkeit komplexe Funktionen zu approximieren (Eldan & Shamir, 2016). Insbesondere im Bereich der Sprachverarbeitung (NLP) konnten künstliche neuronale Netze als Sprachmodelle eingesetzt werden, die dann sprachbezogene Aufgaben wie Übersetzung, Zusammenfassung oder Klassifikation lösen konnten, ohne spezifisch auf diese Aufgaben hin trainiert worden zu sein (Devlin et al., 2018; Lewkowycz et al., 2022). Aus diesen Gründen wurden dann tiefe neuronale Netze verwendet, um die Unterrichtsreflexionen zu klassifizieren. Es konnte gezeigt werden, dass die Klassifikation mithilfe von Sprachmodellen zu einer Genauigkeit von etwa 83 % führt und ebenso besser auf die ungesehenen Testdaten generalisiert (ebenso 81 %), und zwar bereits bei einem vergleichsweise kleine Trainingsdatensatz (Wulff, Mientus et al., 2022). Die trainierten ML-Modelle konnten sogar in anderen Fachbereichen reliabel eingesetzt werden (Wulff et al., 2023).

Die einst trainierten Modelle können nunmehr verwendet werden, um Unterrichtsreflexionen der Lehramtsstudierenden und Lehrkräfte in Hinblick auf die Umsetzung der Reflexionselemente auszuwerten sowie Qualitätsabschätzun-

gen automatisiert vorzunehmen (Mientus et al., 2023). Die Rückmeldung erfolgt dann beispielsweise in der Form, dass den Studierenden mitgeteilt wird, welche Elemente sie in welchem Umfang in ihren Texten umgesetzt haben. Erste Ergebnisse zur Akzeptanz der Rückmeldungen von Seiten der Studierenden deuten darauf hin, dass Studierende die Rückmeldung durchaus als fachlich korrekt beurteilen. Allerdings schätzen sie die Rückmeldungen als zu unpersönlich ein (Mientus et al., 2021; Wulff et al., 2021). Das trainierte Reflexionsmodell konnten wir an einem weiteren Standort zur Verfügung stellen[2]. Auch an diesem Standort funktionierten die ML-Modelle akzeptabel (Genauigkeit 67 %). Nach weiterem sog. Feintuning (engl.: *fine-tuning*) der Modelle anhand der Unterrichtsreflexionen vom neuen Standort konnte die Klassifikationsgenauigkeit hin zu substantieller Übereinstimmung (Genauigkeit 76 %) verbessert werden.

4.4 Exploration von Themen in Unterrichtsreflexionen mit unsupervised ML

Unsupervised ML-Ansätze können unstrukturierte Datensätze wie etwa unkodierte Unterrichtsreflexionen auf Satz- oder Textebene gruppieren. In Hinblick auf Unterrichtsreflexionen ist das Gruppieren eine wichtige Methode, um automatisiert zu extrahieren, über welche Themen die Studierenden in ihren Texten schreiben. Bisher ist unklar, welche Themen (beispielsweise bestimmte Reflexionsauslöser) von den Studierenden beispielsweise besonders häufig oder selten aufgegriffen werden. Auch bei diesen ML-Algorithmen gibt es ein Spektrum an Komplexität. Einfache Gruppierungsverfahren arbeiten auf Basis der Abstandsbestimmung verschiedener Datenpunkte, während fortgeschrittene Algorithmen Optimierungsverfahren verwenden, um beispielsweise angemessene Repräsentationen der Daten zu erhalten, die dann wiederum gruppiert werden können. Vortrainierte Sprachmodelle (sog. *foundation models*) können ebenfalls insbesondere bei der Repräsentation der Daten helfen, die dann gruppiert werden können. In unserer Teilstudie haben wir das frei verfügbare Sprachmodell BERT verwendet (Devlin et al., 2018), um die Sätze in den Reflexionen in einem hochdimensionalen Vektorraum darzustellen, der Facetten der Bedeutung der Sätze berücksichtigt. Für die Unterrichtsreflexionen angehender Physiklehrkräfte zu einer standardisierten Unterrichtssituation (Vignette) konnten wir feststellen, dass durch die Verwendung der Sprachmodelle und entsprechender Gruppierungsalgorithmen (HDBSCAN, siehe McInnes et al., 2017) inhaltlich ab-

2 Siehe: https://www.ipn.uni-kiel.de/de/forschung/projekte/automatisiert-reflexionstexte-evaluieren-mit-ki/arete.ki [Letzter Aufruf: 18. 05. 2023].

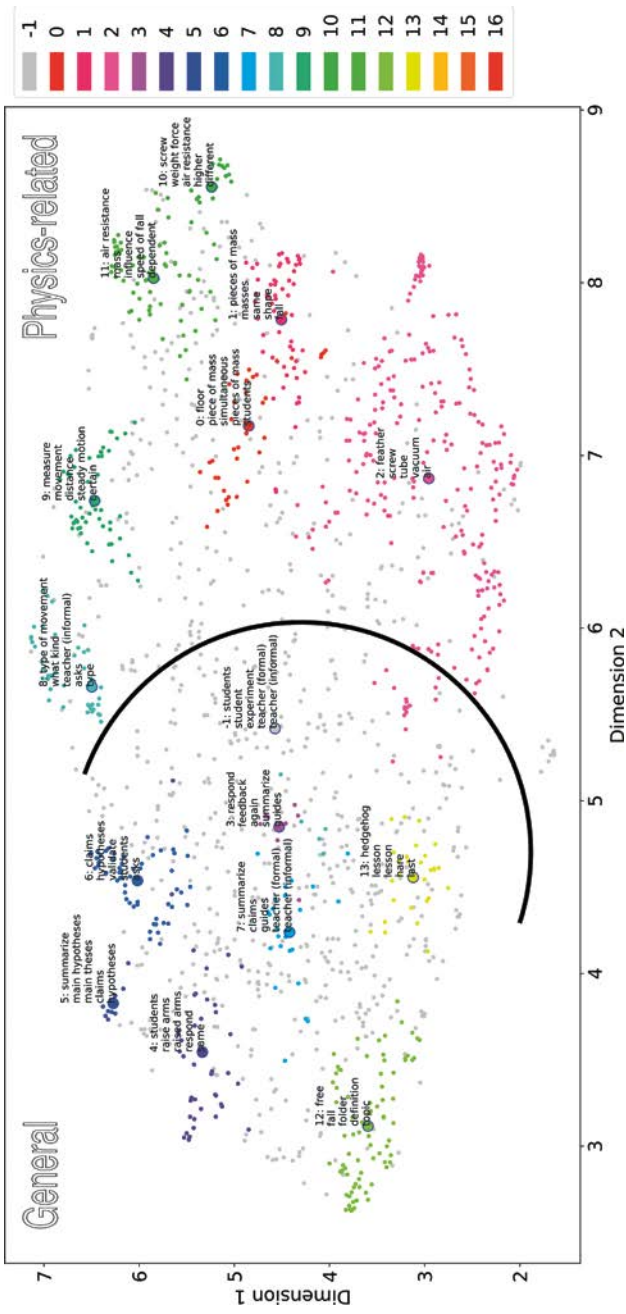

Abbildung 1 Zweidimensionale Darstellung des semantischen Raumes, in dem die Sätze aus den Reflexionen Themen (farblich markiert) zugeordnet sind.

grenzbare Themen identifizierbar waren (Wulff, Buschhüter et al., 2022). Die identifizierten Themen ließen sich in eher allgemeine Themen und fachspezifischere Themen unterteilen (siehe Abb. 1). Themen wie *Hypothesen formulieren oder zusammenfassen* (Nr. 5 und 6) oder das *Abschreiben der Definition des Freien Falls* (Nr. 12) waren eher allgemeinerer Natur. Fachspezifische Themen waren etwa die *Diskussion der Abhängigkeit der Fallbewegung vom Luftwiderstand* (Nr. 11) oder die *Durchführung des Experiments mit Feder und Schraube in der Vakuum-Fallröhre* (Nr. 2).

In einer Folgestudie zeigte sich, dass fachspezifische Themen in Bewertungen eher von Physik-Lehramtsstudierenden (Expert:innen) identifiziert wurden und weniger von Studierenden anderer Fächer, die über weniger physikspezifisches Vorwissen verfügen (Wulff et al., 2023). Auch für diese Themenanalyse wäre es denkbar, den Studierenden für eine standardisierte Situation wie in der Vignette automatisiert eine Rückmeldung zu geben, beispielsweise welche Themen angesprochen und welche ausgelassen wurden.

5 DISKUSSION UND AUSBLICK

KI-basierte Methoden wie ML und NLP bieten neue, evidenzbasierte Forschungsansätze, um die Professionalisierung angehender Lehrkräfte im Allgemeinen und schriftliche Unterrichtsreflexionen als Professionalisierungstool im Besonderen zu untersuchen und in der Praxis der Lehrkräfteausbildung umzusetzen (Ullmann, 2019). Die Ergebnisse zu den Forschungsfragen 1 und 2 zeigen, dass sowohl einfache Kodierungen als auch explorative Auswertungen möglich sind. Mittlerweile konnten die trainierten ML-Modelle bereits in die Praxis, hier in Form einer Webanwendung, implementiert werden. Erste Evidenz deutet darauf hin, dass bei wiederholter Anwendung von Reflexionen nach dem Reflexionsmodell die beschreibenden Anteile in den Unterrichtsreflexionen sinken und die Anteile zu Alternativen sowie Konsequenzen ansteigen, von initial sehr geringen Werten. Inwieweit Reflexion aber tatsächlich zur Entwicklung professioneller Handlungskompetenzen beiträgt, ist bislang – zumindest im Fach Physik – ungeklärt.

Ebenso offen ist die Frage wie eine lernwirksame Rückmeldung zu Reflexionen gestaltet werden sollte. Zwar können wir auf Basis unserer Modelle Anteile an Elementen zurückmelden, allerdings kann diese Form der Rückmeldung lediglich oberflächen-strukturell helfen. Interessante Möglichkeiten bieten hier generative KI-Modelle wie etwa GSM. Auch dort sind allerdings noch basale Fragen nach der Transparenz von Entscheidungen oder der Verzerrung durch Trainingsdaten ungeklärt, ebenso wie generelle datenschutzrechtliche Fragen, etwa inwie-

weit die Weitergabe von Forschungsdaten an private Unternehmen, die die GSM betreiben, ungünstige Abhängigkeiten erzeugt, inwieweit Gruppen, die nicht angemessen in den Trainingsdaten repräsentiert sind, Nachteile bei GSM-basierten Entscheidungen zu befürchten haben, oder inwieweit probabilistische Lernansätze überhaupt formal nachprüfbare Erkenntnisse erzeugen (Caliskan et al., 2017). Eine Nutzung in der Praxis muss deshalb kritisch gesehen werden. Im Gegensatz zu LLM-basierten Anwendungen wie ChatGPT hat der vorgeschlagene Ansatz in unseren Studien den Vorteil, dass die ML-Modelle lokal vorliegen (also keine Daten an kommerziell arbeitende Konzerne geliefert werden) und die Modelle systematisch und beliebig getestet werden können, sodass insgesamt eine größere Kontrolle über den Forschungsprozess möglich ist. Des Weiteren können unsere Modelle einfach über Forschungskontexte geteilt werden, da alle Analysen auf frei zugänglicher Software (Python sowie entsprechender Bibliotheken) basieren und auf vergleichsweise einfachen Heimcomputern gerechnet werden können.

Für die Zukunft der Lehrkräfteausbildung sowohl in der ersten (hochschulischen) Phase als auch in den weiteren Phasen wären ML-Modelle zur automatisierten Auswertung und Rückmeldung zu Unterrichtsreflexionen wünschenswert. Einerseits könnten diese einen umfassenderen und intensiveren Einsatz von Reflexion in Praxisphasen (und ebenso Theoriephasen) gewährleisten und andererseits dazu beitragen, evidenzbasiert Erkenntnisse über den Ablauf und die Qualität reflexionsbezogener Denkprozesse (von Aufschnaiter et al., 2019) zu erlangen. Die vorgestellten Analysen stellen einen ersten Schritt in diese Richtung dar. Sie zeigen, dass ML-Algorithmen verwendet werden können, um modellbasierte Auswertungen von Unterrichtsreflexionen vorzunehmen. Ein weiterer wichtiger Forschungsbereich wird deshalb sein aufzuzeigen, auf welche Weise Lehrkräfte solche KI-basierten Analysen und Rückmeldungen gewinnbringend für die Analyse von Lehr- und Lernprozessen im Allgemeinen nutzen können.

Literaturverzeichnis

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R. & McWilliams, M. (2017). Towards Reflective Writing Analytics: Rationale, Methodology and Preliminary Results. *Journal of Learning Analytics, 4*(1), 58–84. https://doi.org/10.18608/jla.2017.41.5

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K. H. K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., … Wilson, C. D. (2019). The Refined Consensus Model of Pedagogical Content Knowledge. In A. Hume, R. Cooper & A. Borowski (Hrsg.), *Repositioning Pedagogical Content Knowledge in Teachers' Professional Knowledge* (S. 77–94). Springer. https://doi.org/10.1007/978-981-13-5898-2_2

Darling-Hammond, L. (2012). *Powerful Teacher Education: Lessons from Exemplary Programs*. Jossey-Bass.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Donnelly, D. F., Vitale, J. M. & Linn, M. C. (2015). Automated Guidance for Thermodynamics Essays: Critiquing Versus Revisiting. *Journal of Science Education and Technology, 24*(6), 861–874. https://doi.org/10.1007/s10956-015-9569-1

Eldan, R. & Shamir, O. (2016). The Power of Depth for Feedforward Neural Networks. *arXiv:1512.03965v4*. https://doi.org/10.48550/arXiv.1512.03965

Hatton, N. & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U

Korthagen, F. A. (2005). Levels in reflection: core reflection as a means to enhance professional growth. *Teachers and Teaching, 11*(1), 47–71. https://doi.org/10.1080/1354060042000337093

Korthagen, F. A. & Kessels, J. (1999). Linking Theory and Practice: Changing the Pedagogy of Teacher Education. *Educational Researcher, 28*(4), 4–17. https://doi.org/10.3102/0013189X028004004

Krüger, D. & Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 26*(1), 157–172. https://doi.org/10.1007/s40573-020-00118-7

Lai, G. & Calandra, B. (2007). Using Online Scaffolds to Enhance Preservice Teachers' Reflective Journal Writing: A Qualitative Analysis. *International Journal of Technology in Teaching and Learning, 3*(3), 66–81.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G. & Misra, V. (2022). Solving Quantitative Reasoning Problems with Language Models. *arXiv:2206.14858*. https://doi.org/10.48550/arXiv.2206.14858

Marsland, S. (2015). *Machine Learning. An Algorithmic Perspective* (2. Aufl.). Chapman & Hall/CRC Press. https://doi.org/10.1201/b17476

McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software, 11*(2), 205. https://doi.org/10.21105/joss.00205

Mena-Marcos, J., García-Rodríguez, M.-L. & Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education, 36*(2), 147–163. https://doi.org/10.1080/02619768.2012.713933

Mientus, L., Hume, A. C., Wulff, P., Meiners, A. & Borowski, A. (2022). Modelling STEM Teachers' Pedagogical Content Knowledge in the Framework of the Refined Consensus Model: A Systematic Literature Review. *Education Sciences, 12*(6), 385. https://doi.org/10.3390/educsci12060385

Mientus, L., Wulff, P., Nowak, A. & Borowski, A. (2021). ReFeed: computerunterstütztes Feedback zu Reflexionstexten: Ein Lehrkonzept zur Förderung der Reflexionskompetenz angehender Physiklehrkräfte an der Universität Potsdam. In M. Kubsch, S. Sorge, J. Arnold & N. Graulich (Hrsg.), *Lehrkräftebildung neu gedacht. Ein Praxishandbuch für die Lehre in den Naturwissenschaften und deren Didaktiken* (S. 160–165). Waxmann. https://doi.org/10.31244/9783830993490

Mientus, L., Wulff, P., Nowak, A. & Borowski, A. (2023). Fast-and-frugal means to assess reflection-related reasoning processes in teacher training – Development and evaluation of a scalable machine learning-based metric. *Zeitschrift für Erziehungswissenschaft, 26*(3), 677–702. https://doi.org/10.1007/s11618-023-01166-8

Nehm, R. H. & Härtig, H. (2012). Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software. *Journal of Science Education and Technology, 21*(1), 56–73. https://doi.org/10.1007/s10956-011-9282-7

Nowak, A., Kempin, M., Kulgemeyer, C. & Borowski, A. (2019). Reflexion von Physikunterricht. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Kiel 2018* (S. 838). Universität Regensburg. https://doi.org/10.25656/01:16753

Odden, T. O. B., Marin, A. & Caballero, M. D. (2020). Thematic analysis of 18 years of physics education research conference proceedings using natural language processing. *Physical Review Physics Education Research, 16*(1), 010142. https://doi.org/10.1103/PhysRevPhysEducRes.16.010142

Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J. & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education, 37*(3), 348–373. https://doi.org/10.1080/02619768.2014.892479

Sherin, B. (2013). A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data. *Journal of the Learning Sciences, 22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sorge, S., Neumann, I., Neumann, K., Parchmann, I. & Schwanewedel, J. (2018). Was ist denn da passiert? Ein Protokollbogen zur Reflexion von Praxisphasen im Lehr-Lern-Labor. *MNU Journal, 6,* 420–426.

Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257. https://doi.org/10.1007/s40593-019-00174-2

von Aufschnaiter, C., Fraij, A. & Kost, D. (2019). *Reflexion und Reflexivität in der Lehrerbildung, 2*(1), 144–159. https://doi.org/10.4119/UNIBI/HLZ-144

Wang, H.-C., Chang, C.-Y. & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education, 51*(4), 1450–1466. https://doi.org/10.1016/j.compedu.2008.01.006

Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H., Stede, M. & Borowski, A. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology, 30*(1), 1–15. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A. & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning – A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology, 31*(4), 490–513. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Mientus, L., Nowak, A. & Borowski, A. (2021). Stärkung praxisorientierter Hochschullehre durch computerbasierte Rückmeldung zu Reflexionstexten. *die hochschullehre, 7*(11), 93–99. https://doi.org/10.3278/HSL2111W

Wulff, P., Mientus, L., Nowak, A. & Borowski, A. (2022). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education, 33*(3), 439–466. https://doi.org/10.1007/s40593-022-00290-6

Wulff, P., Westphal, A., Mientus, L., Nowak, A. & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing – Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education, 7,* 1061461. https://doi.org/10.3389/feduc.2022.1061461

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), 35027. https://doi.org/10.1088/1361-6552/acc5cf

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C. & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757