Benjamin Münch[1] [1]Universität Regensburg
Oliver Tepner[1]

Sprachmodellgestützte Klassifikation schriftlicher Unterrichtsreflexionen

Theoretischer Hintergrund

Die Reflexionskompetenz von angehenden Lehrkräften spielt eine zentrale Rolle in der Professionalisierung von Lehrkräften (Baumert & Kunter, 2006) und Qualitätssicherung des Unterrichts (Abels, 2011; KMK, 2004). Sie umfasst die Fähigkeit, über vergangene Unterrichtserfahrungen nachzudenken, diese zu bewerten und Konsequenzen für zukünftige Handlungen abzuleiten (Wyss, 2013). Die Analyse der Reflexionskompetenz erfolgt in der Regel manuell mit validierten Kodiermanualen (Kobl, 2021; Reimer & Tepner, eingereicht). In der Lehrkräfteausbildung sind (schriftliche) Reflexionen daher ein etabliertes Mittel, um diesen Kompetenzerwerb zu fördern (von Aufschnaiter et al., 2019). Die Herausforderung liegt jedoch in der aufwendigen Analyse der Reflexionen durch klassische Kodiermanuale. Mit dem Aufkommen von Methoden des Machine Learning (ML) und Natural Language Processing (NLP) eröffnen sich neue Wege, diese Reflexionen automatisiert und ressourcenschonend auszuwerten.

Im vorliegenden Promotionsprojekt wird der Einsatz von Machine-Learning-Modellen, insbesondere von großen Sprachmodellen (Large Language Models, LLMs) zur Evaluation von Reflexionskompetenz untersucht. Ziel ist es, zu prüfen, inwieweit diese Modelle eine valide Alternative zu klassischen Kodiermanualen darstellen und welche spezifischen Aspekte der Reflexionskompetenz sie identifizieren können.

Sprachbasierte Modelle, insbesondere LLMs wie BERT (Devlin et al., 2018), bieten die Möglichkeit, natürliche Sprache auf systematische Weise zu analysieren. Diese Modelle sind in der Lage, Zusammenhänge in Textdaten zu verstehen und anhand festgelegter Kategorien zu klassifizieren. Im Bildungsbereich können solche Modelle Muster im Text von Lehramtsstudierenden identifizieren (Wulff et al., 2022), die Rückschlüsse auf die Qualität der Reflexionen erlauben.

Ziele (Z) und Fragestellungen (F)

Z1: Training eines Klassifikators (Modell) zur Evaluation spezifischer Aspekte der Reflexionskompetenz von Studierenden des Chemielehramts

Z2: Weiterentwicklung des Klassifikators als Feedback-Tool für Dozierende

F1: Kann ein Machine-Learning-Modell zur Erfassung von Reflexionskompetenz von Chemielehramtsstudierenden entwickelt werden, das eine Alternative zu klassischen Kodiermanualen bietet?

F2: Welche Aspekte von Reflexionskompetenz kann ein Machine Learning basiertes Tool valide erfassen?

Methodik

Das Modelltraining basiert auf 168 schriftlichen Reflexionen von Studierenden des Chemielehramts. Diese Reflexionen wurden mithilfe eines bereits etablierten Kodiermanuals manuell in insgesamt 19 Kategorien (siehe Tab. 1) der Reflexionsbreite und -tiefe eingeordnet (Kobl, 2021; Reimer & Tepner, eingereicht). Eine weitere Kategorie wurde ergänzt, um Teile eines Reflexionstextes abzubilden, die sich weder der Reflexionstiefe noch der

Reflexionsbreite zuordnen lassen. Insgesamt wurden 29.367 Textsegmente kodiert, ein Segment stellt hierbei einen Satz oder einen Teilsatz dar.

Tabelle 1: Kategorien des Datensatzes für das Modelltraining

	Kategorie (Label)	Häufigkeit
Reflexionsbreite	Adaptivität	386
	Adressatenorientierung	615
	Chemisches Fachwissen	66
	Lehrerperformanz	606
	Lernförderliches Klima	197
	Medien	459
	Organisationsform	34
	Regeln im Chemieunterricht	51
	Schülerexperiment	20
	Sonstiges	113
	Sprachliche Verständlichkeit	378
	Sprech- & Körperausdruck	452
	Strukturiertheit	325
	Visualisierung	113
Reflexionstiefe	Alternative	513
	Beschreibung	4420
	Negative Bewertung	2021
	Perspektive	120
	Positive Bewertung	2989
	Verbesserungsvorschlag/Konsequenz	3761
	Nicht kategorisiert	11737

Ein zentraler Aspekt der Untersuchung war der Umgang mit den zunächst unbalancierten Daten. Da einige Kategorien häufiger vertreten waren als andere, wurden Techniken wie Oversampling angewendet (Mohammed et al., 2020; Ta et al., 2022), um eine ausgeglichene Verteilung der Kategorien im Trainingsdatensatz zu gewährleisten. Bei Oversampling wird aus den vorhandenen Beispielen wiederholt zufällig gezogen, bis die Kategorie(n) eine bestimmte Häufigkeit aufweisen.

Ergebnisse

Die Ergebnisse der Analyse zeigen, dass das Modell (Latif et al., 2024) eine hohe Genauigkeit bei der Klassifikation der Reflexionensegmente in 20 Kategorien erreicht.

Die besten Ergebnisse wurden bei der 5-Fold Crossvalidation mit unbalanciertem Datensatz erzielt, der einen F1-Score von .66 erreichte (siehe Tab. 2). Der F1-Score stellt den harmonischen Mittelwert aus Recall und Precision dar (Sokolova et al., 2006) und bewegt sich auf einer Skala von 0 bis 1. Je höher der F1-Score, desto stärker stimmt die Vorhersage des Modells mit der tatsächlichen kategoriellen Zuordnung des Satzsegments überein. Dies zeigt,

dass das Modell in der Lage ist, valide Rückschlüsse auf die Reflexionsbreite und -tiefe zu ziehen.

Tabelle 2: Ergebnisse der 5-Fold Crossvalidation

	unbalanciert	**balanciert (oversampling)**
Accuracy	0.68	0.64
Precision	0.67	0.65
Recall	0.68	0.64
F1	0.66	0.63
Micro F1	0.68	0.64
Macro F1	0.43	0.43
Cohens κ	0.57	0.52
Loss	1.05	1.32

Diskussion

Die Ergebnisse deuten darauf hin, dass LLMs eine vielversprechende Alternative zu traditionellen Kodiermanualen darstellen können. Die automatisierte Klassifikation ermöglicht es, große Mengen an Reflexionen in kurzer Zeit auszuwerten, ohne die Validität der Ergebnisse zu beeinträchtigen.

Ein zentrales Problem bleibt die Verarbeitung unausgeglichener Datensätze. Trotz des Einsatzes von Oversampling-Techniken war die Klassifikation seltener Kategorien weniger präzise als bei Verwendung des unbalancierten Datensatzes. Eine mögliche Erklärung für dieses Ergebnis ist, dass sich die Textvariation innerhalb einer Kategorie nicht erhöht und das Modell deshalb vom Training mit dem vergrößerten, aber redundanten Datensatz nicht profitiert. Zudem bleibt unklar, auf welchen Merkmalen die Zuordnung des Modells beruht.

Fazit und Ausblick

Zusammenfassend lässt sich festhalten, dass Machine Learning, insbesondere Large Language Models (LLMs), das Potenzial hat, die Analyse von Unterrichtsreflexionen in der Lehrkräfteausbildung grundlegend zu erweitern. Die Ergebnisse dieser Studie zeigen, dass solche Modelle valide Rückschlüsse auf Reflexionskompetenz ziehen können und eine ressourcenschonende Alternative zu traditionellen Kodiermethoden bieten. Ein zentraler Vorteil ist die Skalierbarkeit: LLMs ermöglichen es, große Mengen an Reflexionen automatisiert auszuwerten, wodurch der zeitliche und personelle Aufwand erheblich reduziert wird. Damit wird auch sehr zeitnahes Feedback an die Lernenden möglich, welches für die Reflexion und weitere Planungen und Handlungen im Rahmen der Lehrkräfteaus- und fortbildung genutzt werden kann. Zudem fördern LLMs eine konsistente Bewertung, indem sie subjektive Interpretationen minimieren, die in manuellen Kodierungen durch unterschiedliche Kodierende auftreten können.

Zukünftige Bestrebungen dieses Forschungsprojektes werden sich darauf konzentrieren, die Modelle im Sinne von *explainable machine learning* weiter zu optimieren, um die Kategorienzuordnung transparenter zu machen. Auch der Einsatz synthetischer Daten wird näher erforscht, um mit balancierten Trainingsdatensätzen die Klassifizierung der Studierendenreflexionen weiter zu verbessern. Das übergeordnete Ziel dieser Forschungsvorhaben ist die Entwicklung eines praktischen Tools, das Dozierende in der Lehre unterstützt und den Studierenden automatisiertes Feedback zu ihren Reflexionen bietet.

Literatur

Abels, S. (2011). *LehrerInnen als „Reflective Practitioner“: Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht*. VS Verlag für Sozialwissenschaften GmbH.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), Article 4. https://doi.org/10.1007/s11618-006-0165-2

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, 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 Science Education. In A. Hume, R. Cooper, & A. Borowski (Hrsg.), *Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science* (S. 77–94). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_2

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

KMK (Hrsg.). (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. D. F. vom 07.10.2022)*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Kobl, C. (2021). *Förderung und Erfassung der Reflexionskompetenz im Fach Chemie*. Logos Verlag Berlin. https://doi.org/10.30819/5259

Latif, E., Lee, G.-G., Neuman, K., Kastorff, T., & Zhai, X. (2024). *G-SciEdBERT: A Contextualized LLM for Science Assessment Tasks in German*. https://doi.org/10.48550/ARXIV.2402.06584

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Reimer, S., & Tepner, O. (eingereicht). Förderung der adaptiven Erklärkompetenz angehender Chemielehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, 1–46.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In A. Sattar & B. Kang (Hrsg.), *AI 2006: Advances in Artificial Intelligence* (Bd. 4304, S. 1015–1021). Springer Berlin Heidelberg. https://doi.org/10.1007/11941439_114

Ta, T., Butt, S., Angel, J., Sidorov, G., & Gelbukh, A. (2022). *The Combination of BERT and Data Oversampling for Relation Set Prediction*.

von Aufschnaiter, C., Fraij, A., & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. *Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption*, *Gestaltung und Diskussion*, 144-159 Seiten. https://doi.org/10.4119/UNIBI/HLZ-144

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*. https://doi.org/10.1007/s40593-022-00290-6

Wyss, C. (2013). *Unterricht und Reflexion: Eine mehrperspektivische Untersuchung der Unterrichts-und Reflexionskompetenz von Lehrkräften*. Waxmann.