

Automatisiertes Assessment schriftlicher Reflexionen im Chemielehramtsstudium

Abstract

Schriftliche Reflexionen fördern die Entwicklung professionellen Wissens im Lehramtsstudium, indem sie theoretische Konzepte mit Praxiserfahrungen verknüpfen [3]. Der Aufwand für individuelles, kriteriengeleitetes Feedback stellt jedoch eine Herausforderung dar. Im vorliegenden Projekt wurde ein Machine-Learning-Modell trainiert, das automatisiertes Feedback zu Reflexionen von Chemielehramtsstudierenden ermöglicht. Die Reflexionen entstammen 20-minütigen Unterrichtssequenzen von Studierenden in der Studienabschlussphase. Grundlage ist ein validiertes Kodiermanual [2][5], das verschiedene Reflexionskategorien differenziert. Trainiert wurde ein BERT-Transformer-Modell [1][4] auf ca. 18.000 manuell codierten Satzsegmenten. Die Evaluation erfolgte über 5-fache Kreuzvalidierung mit Metriken wie Accuracy, Precision, Recall, F1 (micro/macro/weighted), Kappa und Loss. Das Modell ist in eine anonyme Webanwendung eingebettet, über die Studierende direkt nach dem Verfassen ihrer Reflexion Rückmeldung erhalten. Diese erfolgt zweistufig: durch ein Diagramm zur kategorialen Verteilung und ein automatisch generiertes schriftliches Feedback. Die OpenAI-API wertet dafür die zugrunde liegenden Daten aus und liefert individuell formulierte Hinweise zur Weiterentwicklung der Reflexion. Das Poster zeigt den Entwicklungsprozess, die Modellgüte und die Gestaltung des Feedbacks. Es diskutiert Perspektiven für den Einsatz in der Lehrkräftebildung, darunter eine geplante Studierenden-Evaluation mit Interviews sowie eine mögliche Erweiterung auf das Fach Physik.

Theoretischer Hintergrund

Large Language Models (LLM)

- Kombination von Machine Learning und Prinzipien des Natural Language Processing, zur systematischen Analyse der Wissen- und Sprachnutzung von Studierenden [7]

Reflexionskompetenz

- Reflexion als bewusstes und kriteriengeleitetes Nachdenken über Handlungen, woraus begründete Konsequenzen für das weitere Handeln ableitet und umgesetzt werden [8]
- Wesentlich im Professionalisierungsprozess einer Lehrkraft [6]
- Weiterentwicklung des prozeduralen Reflexionswissens sowohl durch Selbst- als auch Fremdreflexion möglich [2]

Ziele & Forschungsfragen

Ziele

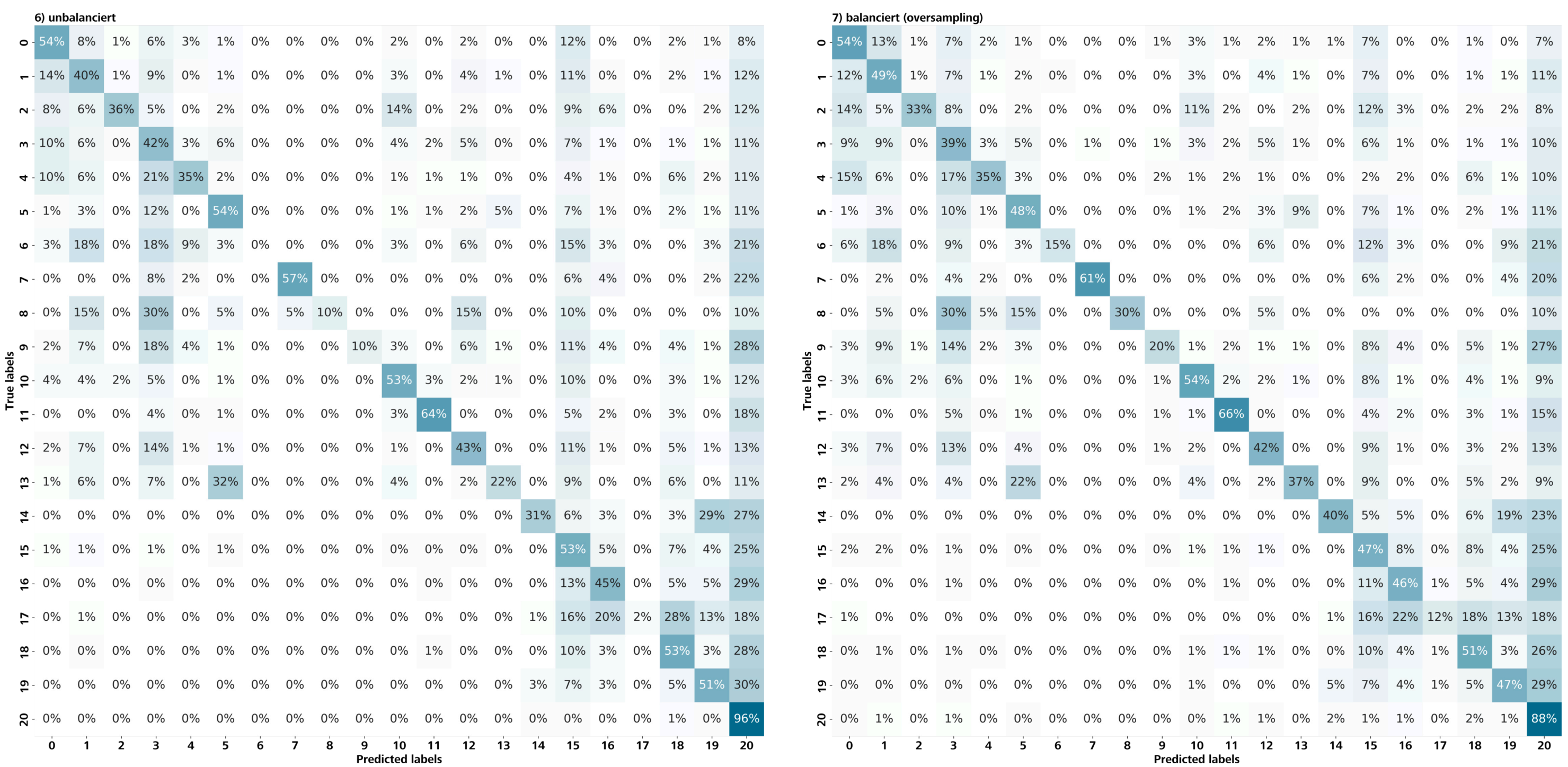
- Finetuning eines Large Language Modells (LLM) zur Evaluation spezifischer Aspekte der Reflexionskompetenz von Studierenden des Chemielehramts
- Implementierung des finalen Modells in ein Webtool zur Unterstützung der Dozierenden

Forschungsfragen

- Kann ein Machine-Learning-Modell zur Erfassung von Reflexionskompetenz von Chemielehramtsstudierenden entwickelt werden, das eine Alternative zu klassischen Kodiermanualen bietet?
- Welche Aspekte von Reflexionskompetenz kann ein Machine Learning basiertes Tool valide erfassen?

Bisherige Ergebnisse der 5-Fold Crossvalidation (Mittelwerte)

Vergleich der Klassifikationsgenauigkeit: Visualisierung von echten und vorhergesagten Kategorien



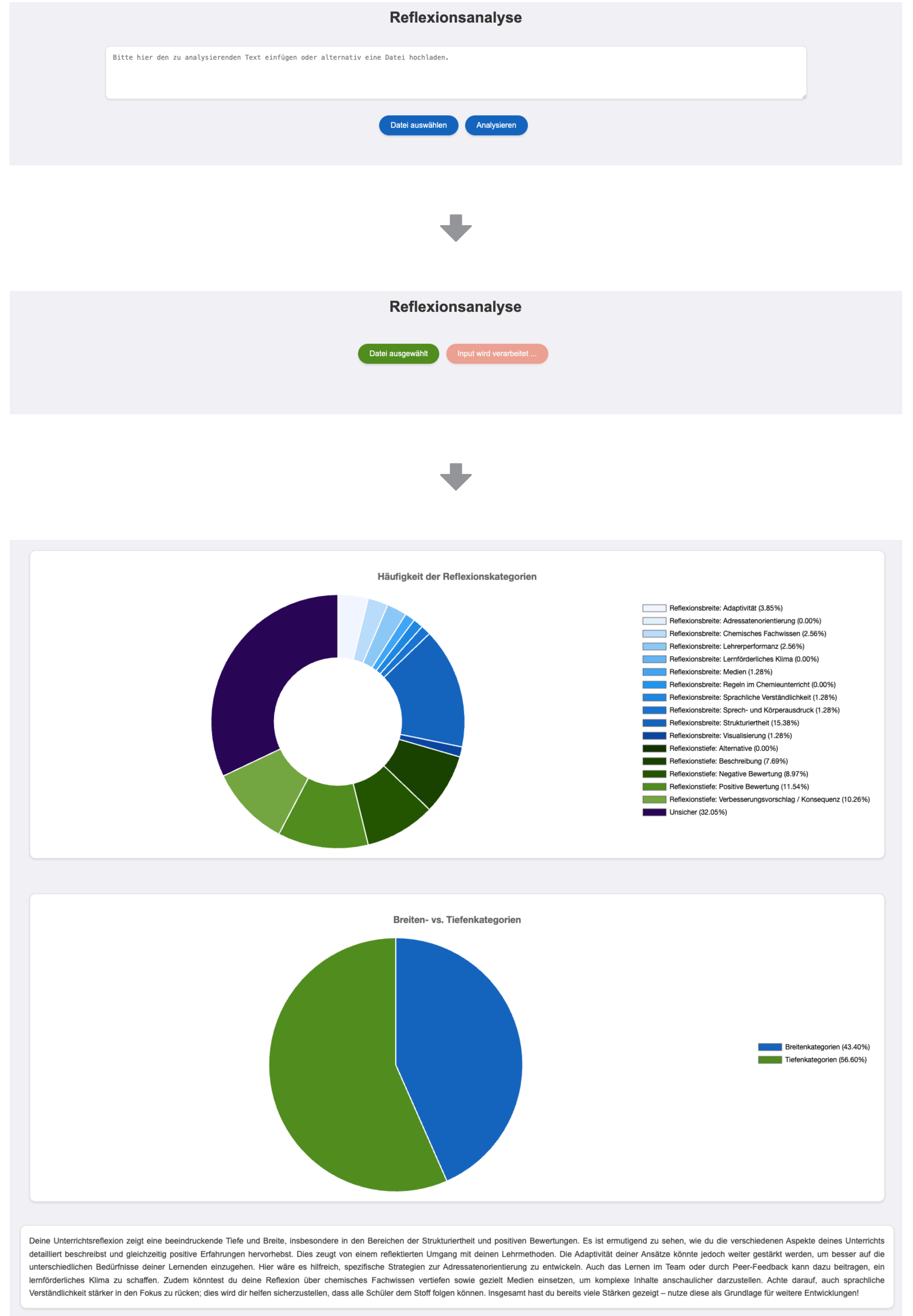
Metrik \ Modell	6)	7)	8)*
Accuracy	0.68	0.64	0.63
Precision	0.67	0.65	0.64
Recall	0.68	0.64	0.63
F1	0.66	0.63	0.63
Micro F1	0.68	0.64	0.63
Macro F1	0.43	0.43	0.52
Cohen's κ	0.57	0.52	0.52
Loss	1.05	1.32	1.26

Accuracy: Der Anteil der korrekt klassifizierten Beispiele an der Gesamtanzahl der Beispiele.
Precision: Der Anteil der richtig positiven Vorhersagen an allen als positiv klassifizierten Beispielen.
Recall: Der Anteil der richtig positiven Vorhersagen an allen tatsächlich positiven Beispielen.
F1: Das harmonische Mittel von Precision und Recall, welches ein ausgewogenes Maß für beides darstellt.
Micro F1: Der F1-Score berechnet über alle Klassen hinweg, wobei jede Instanz gleich gewichtet wird.
Macro F1: Der durchschnittliche F1-Score über alle Klassen, wobei jede Klasse gleich gewichtet wird.
Cohen's κ: Maß für die Übereinstimmung zwischen zwei Beurteilern (hier: Mensch / Maschine), das die Zufallskorrelation berücksichtigt.
Loss: Eine Funktion, die die Differenz zwischen den vorhergesagten Werten und den tatsächlichen Werten misst und den Fehler des Modells angibt.
Undersampling Reduktion der Mehrheitskategorien auf die Größe der seltensten Kategorie im Trainingsdatensatz - zur Balancierung.
Oversampling Wiederholte zufällige Ziehung von Instanzen der selteneren Kategorien im Trainingsdatensatz, zur Vergrößerung der Minderheitskategorien - zur Balancierung.

Kategorie (Label) / Labelkodierung	Häufigkeit	6,7	8*
Reflexionsbreite Adaptivität	386	0	2
Reflexionsbreite Adressatenorientierung	615	1	2
Reflexionsbreite Chemisches Fachwissen	66	2	4
Reflexionsbreite Lehrerperformanz	606	3	5
Reflexionsbreite Lernförderliches Klima	197	4	3
Reflexionsbreite Medien	459	5	0
Reflexionsbreite Organisationsform	34	6	3
Reflexionsbreite Regeln im Chemieunterricht	51	7	3
Reflexionsbreite Schülerexperiment	20	8	11
Reflexionsbreite Sonstiges	113	9	11
Reflexionsbreite Sprachliche Verständlichkeit	378	10	4
Reflexionsbreite Sprech- und Körperausdruck	452	11	1
Reflexionsbreite Strukturiertheit	325	12	3
Reflexionsbreite Visualisierung	113	13	0
Reflexionstiefe Alternative	513	14	6
Reflexionstiefe Beschreibung	4420	15	7
Reflexionstiefe Negative Bewertung	2021	16	8
Reflexionstiefe Perspektive	120	17	11
Reflexionstiefe Positive Bewertung	2989	18	9
Reflexionstiefe Verbesserungsvorschlag/Konsequenz	3761	19	10
Unkategorisiert (Sonstiges)	11737	20	11

*Arbeitsansatz mit reduzierter Labelanzahl zur Optimierung der Testergebnisse.

Webtool



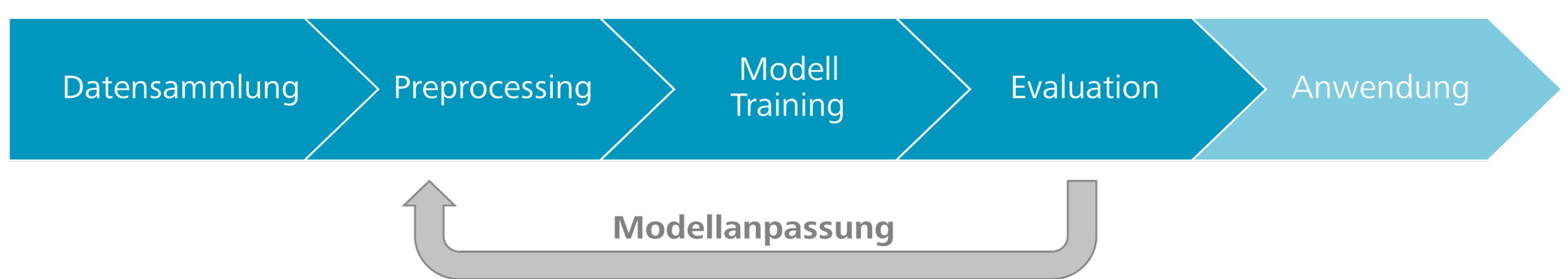
Kontakt



Benjamin.Muench@ur.de
+ 49 941 943 5789

Untersuchungsdesign

- 168 schriftliche Reflexionen / 29.367 kodierte Segmente
- Validierung des Einsatzes in der Lehramtsausbildung anhand eines bereits entwickelten Kodiermanuals [2][5]



Limitationen & Ausblick

Limitationen

- Hohe Modellperformanz kaum erklärbar
- Nicht alle Kategorien des Ausgangsmanuals sind performant

Ausblick

- “Go-Live” des Webtools
- Erhebung zur Usability / Akzeptanz / Effektivität
- Data Retrieval durch Websitenutzung zur ergänzenden Validierung des Modells

Literatur

- [1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- [2] Kohl, C. (2021) Förderung und Erfassung der Reflexionskompetenz im Fach Chemie, Logos Verlag Berlin.
- [3] Kulgeneyes, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Reinhold, P., Riese, J., Schecker, H., Schröder, J., and Vogelsgang, C. (2021) Exploring the impact of pre-service science teachers' reflection skills on the development of professional knowledge during a field experience. International Journal of Science Education, 43 (18), 3035–3057.
- [4] Latif, E., Lee, G.-G., Neuman, K., Kastorff, T., and Zhai, X. (2024) G-SciBERT: A Contextualized LLM for Science Assessment Tasks in German.
- [5] Reimer, S., and Tepner, O. (submitted) Development of preservice science teachers' reflective skills.
- [6] von Aufschneider, C., Fraji, A., and Kost, D. (2019) Reflexion und Reflexivität in der Lehrerbildung. Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion, 144-159 Seiten.
- [7] Wulff, P., Buschhüter, D., Westphal, A., and Borowski, A. (2020) Potentiale automatischer Sprachverarbeitung für die Fachdidaktik. Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Wien 2019, 49–54.
- [8] Wyss, C. (2013) Unterricht und Reflexion: eine mehrperspektivische Untersuchung der Unterrichts- und Reflexionskompetenz von Lehrkräften, Waxmann.