Universität Regensburg
DIDAKTIK DER CHEMIE

BENJAMIN MÜNCH & OLIVER TEPNER

Sprachmodellgestützte Klassifikation schriftlicher Unterrichtsreflexionen

| Abstract

Schriftliche Reflexionen von Unterrichtserfahrungen stellen in der universitären Lehrkräftebildung ein etabliertes Mittel zur professionellen Entwicklung dar. Die Herausforderung besteht darin, einen Kompromiss aus einer ausführlichen Rückmeldung und dem dafür erforderlichen zeitlichen Aufwand zu finden. Methoden des Machine Learning (ML) und Natural Language Processing (NLP) bieten eine Möglichkeit, schriftliche Reflexionen ressourcenschonend auf Basis etablierter Kriterien zu analysieren. In dieser Studie wird ein Tool entwickelt, das angehenden Chemielehrkräften datengestütztes, automatisiertes Feedback zur Verfügung stellt und eine Alternative zu den klassischen Kodierverfahren bietet. Dazu wird ein sogenannter Klassifikator mit den Ergebnissen bisheriger klassischer Kodiermanuale trainiert. Im Rahmen des Trainings des Modells werden Herausforderungen aufgrund unausgeglichener Datensätze mit Methoden des Up- bzw. Downsamplings adressiert, mit dem Ziel, die beste Trainingsmethode zu identifizieren. Im Beitrag werden erste Ergebnisse, sowie Herausforderungen und Chancen des Vorgehens berichtet.

| Theoretischer Hintergrund

Large Language Models

- Vortrainierte Sprachmodelle als Grundlage für die Entwicklung von State of the Art (SOTA)-Modellen in einem breiten Anwendungsspektrum (Devlin, 2018)
- Natural Language Processing, um systematischen Zugriff auf Wissen- und Sprachnutzung von Probanden zu erhalten (Wulff, 2020)

Reflexionskompetenz

- Reflexion als gezieltes Nachdenken über bestimmte Handlungen oder Geschehnisse im Berufsalltag und Ableiten begründeter Konsequenzen für das weitere Handeln (Wyss, 2013)
- Zuwachs im prozeduralen Reflexionswissen sowohl durch Selbst- als auch Fremdreflexion möglich (Kobl, 2021)

| Ziele & Forschungsfragen

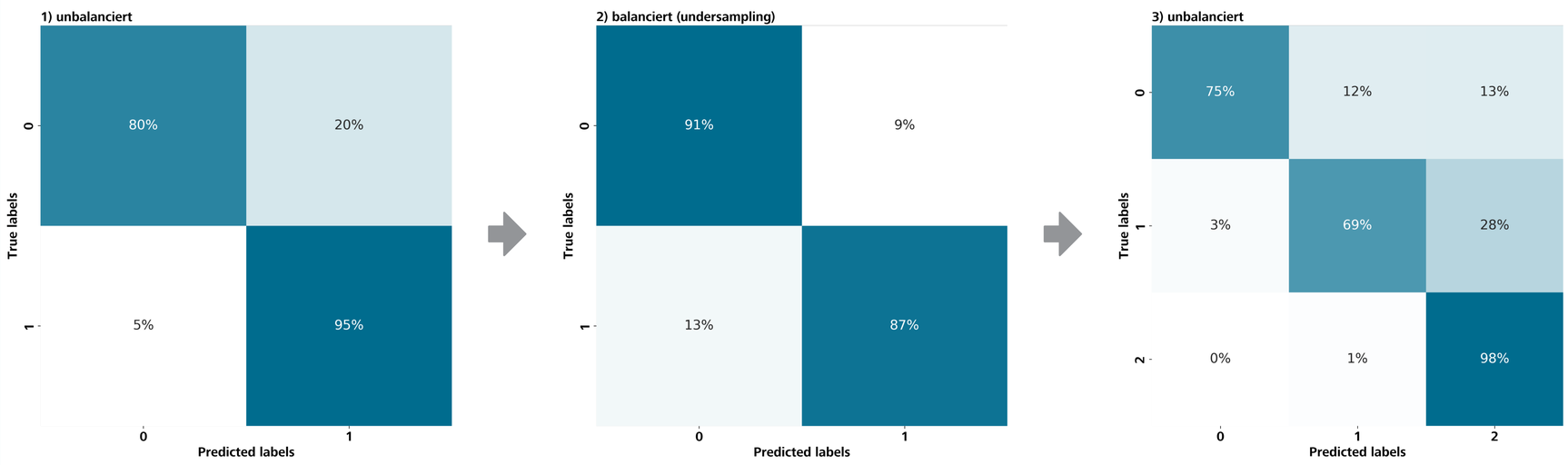
Ziele

- Training eines Klassifikators (Modell) zur Evaluation spezifischer Aspekte der Reflexionskompetenz von Studierenden des Chemielehramts
- Weiterentwicklung des Klassifikators als Feedback-Tool für Dozierende

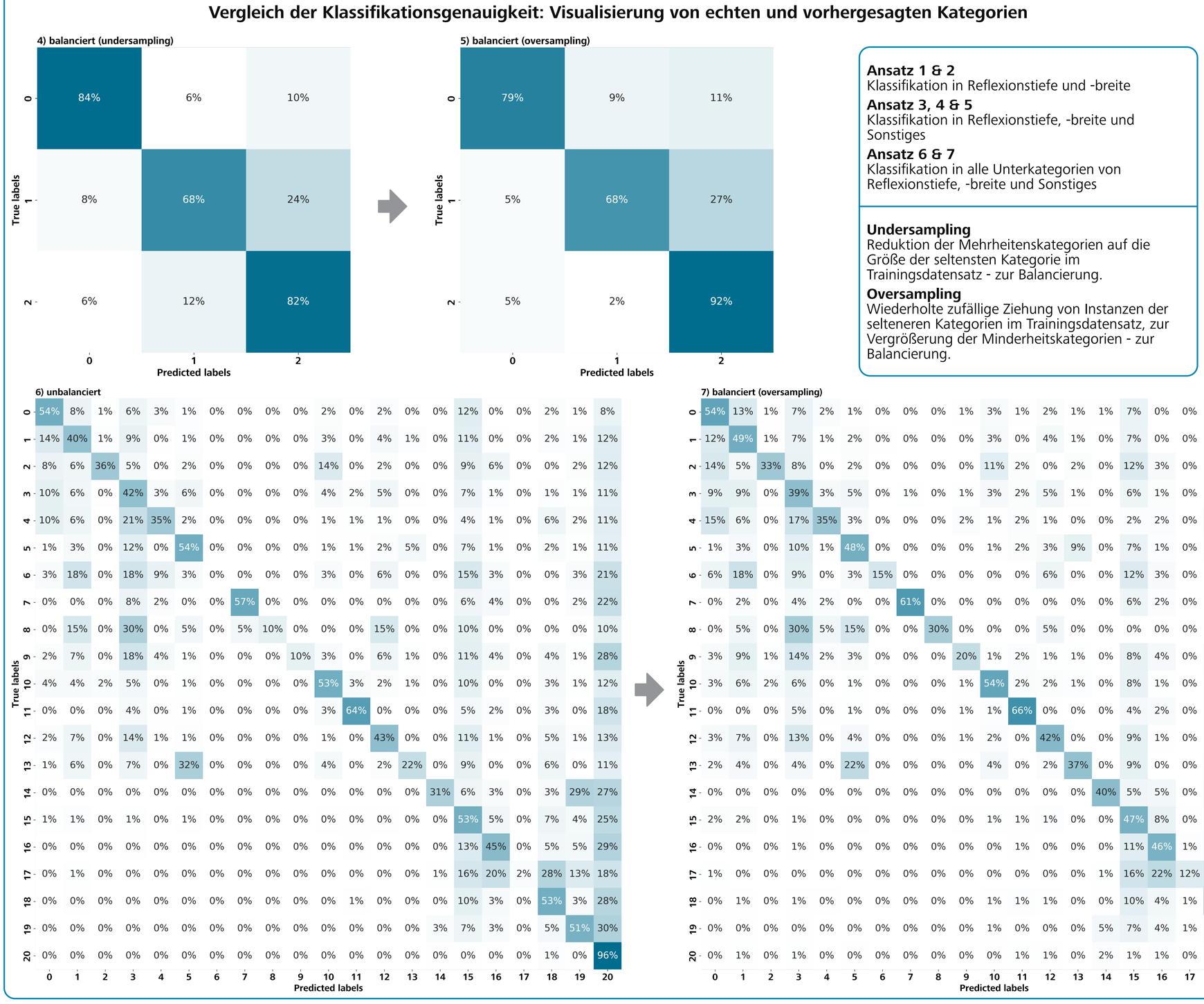
Forschungsfragen

- Kann ein Machine-Learning-Modell zur Erfassung von Reflexionskompetenz von Chemielehramts-studierenden 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



Ansatz 1 & 2
Klassifikation in Reflexionstiefe und -breite

Ansatz 3, 4 & 5
Klassifikation in Reflexionstiefe, -breite und Sonstiges

Ansatz 6 & 7
Klassifikation in alle Unterkategorien von Reflexionstiefe, -breite und Sonstiges

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

*Arbeitsansatz mit reduzierter Labelanzahl zur Optimierung der Testergebnisse.

Metrik	1)	2)	3)	4)	5)	6)	7)	8)*
Accuracy	0.92	0.89	0.81	0.76	0.79	0.68	0.64	0.63
Precision	0.95	0.90	0.84	0.77	0.82	0.67	0.65	0.64
Recall	0.95	0.86	0.81	0.75	0.79	0.68	0.64	0.63
F1	0.95	0.88	0.81	0.76	0.79	0.66	0.63	0.63
Micro F1	0.92	0.89	0.81	0.76	0.79	0.68	0.64	0.63
Macro F1	0.88	0.89	0.81	0.75	0.78	0.43	0.43	0.52
Cohen's κ	0.77	0.78	0.69	0.60	0.66	0.57	0.52	0.52
Loss	0.21	0.35	0.47	0.58	0.58	1.05	1.32	1.26

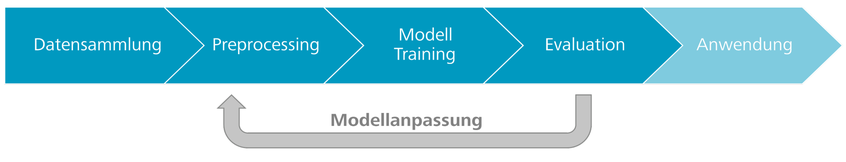
Accuracy: Der Anteil der korrekt klassifizierten Beispiele an der Gesamtanzahl der Beispiele.
Precision: Der Anteil der richtig positiven Vorhersagen an allen als positiv klassifizierten Beispielen.
Recall: Der Anteil der richtig positiven Vorhersagen an allen tatsächlich positiven Beispielen.
F1: Das 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.

| Kontakt

Benjamin.Muench@ur.de
+ 49 941 943 5576

| Untersuchungsdesign

- 168 schriftliche Reflexionen / 29.367 kodierte Segmente
- Validierung des Einsatzes in der Lehramtsausbildung anhand eines bereits entwickelten Kodiermanuals (Kobl, 2021; Reimer & Tepner, eingereicht)



| Ausblick

- Überprüfung und Anpassung der Kategorienanzahl
- Vergleich mit erklärbaren ML-Ansätzen zur Identifikation von Merkmalen der Kategorien:
  - Naive Bayes
  - logistische Regression
  - Random Forest
- Erweiterung und Balancierung des Datensatzes durch synthetische Daten
- Validierung der synthetischen Daten durch das trainierte Modell und bestehende Kodiermanual

| Literatur