ML-basierte Bewertung von Medizin-Studierenden anhand von klinischen Fallbeispielen

Studiengang: BSc in Informatik
Vertiefung: Data Engineering
Betreuer: Prof. Dr. Jürgen Vogel
Experte: Reto Trinkler (StepFour)

Ein Prototyp wurde entwickelt, der mithilfe von Natural Language Processing (NLP) die Qualität studentischer Fallzusammenfassungen im Kontext virtueller Patient*innen bewertet. Eine kombinierte regel- und LLM-gestützte Pipeline erzeugt Teilbewertungen und visualisiert diese nachvollziehbar in einer Webapplikation.

Motivation und Ziel

An der Medizinischen Fakultät der Universität Bern werden virtuelle Patient*innen (VP) eingesetzt, um klinische Entscheidungsfindung zu trainieren. Studierende verfassen dazu narrative Fallzusammenfassungen. Die manuelle Bewertung ist jedoch zeitaufwendig und liefert keine automatisierten, inhaltsspezifischen Rückmeldungen. Ziel ist die Entwicklung eines Prototyps, der studentische Zusammenfassungen automatisch mit einer Musterlösung vergleicht, deren inhaltliche Qualität bewertet (0–2 Punkte je Kriterium) und die Ergebnisse nachvollziehbar visualisiert.

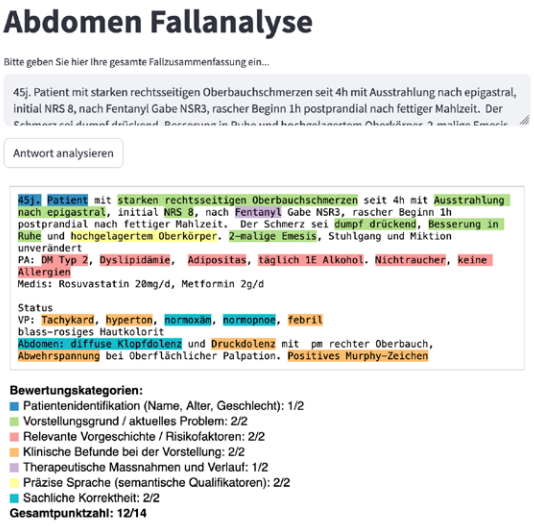
Methodik

Als Datengrundlage diente ein annotierter Korpus mit rund 100 Fallzusammenfassungen (60 abdominale, 40 neurologische Fälle), die von Fachpersonen anhand definierter Subkriterien (z. B. Patientendaten oder Vorgeschichte) bewertet wurden. Für strukturierte Informationen wie Alter, Name oder Medikamente kamen regelbasierte Verfahren (Regex, Jaro-Winkler) zum Einsatz. Komplexere Kategorien wurden mithilfe von GPT-4o-mini, einem Large Language Model (LLM), unter Anwendung von Zero-/Few-Shot-Prompting bewertet. Die Ergebnisse werden in einer interaktiven Webanwendung (Streamlit) visualisiert. Dazu gehören farbige Markierungen relevanter Textstellen, Einzelbewertungen pro Subkriterium, konkrete Verbesserungshinweise sowie ein Vergleich mit Durchschnittswerten.

Resultate

Die Genauigkeit variiert je nach Subthema. Strukturierte Informationen wie Patientendaten (Name, Alter, Geschlecht) oder Medikation lassen sich mit regelbasierten Verfahren sehr zuverlässig erkennen (Genauigkeit: ~0.95). LLM-basierte Bewertungen zeigten bei klaren Kriterien mittlere Genauigkeiten (~0.8), bei schwach definierten deutlich geringere (~0.6). Insgesamt zeigte sich, dass eine hybride Kombination aus regelbasierten und modellgestützten Verfahren vielversprechend ist, jedoch bei schwach formalisierten Aspekten an Grenzen stösst. Ein grösserer Datensatz würde die Trainingsmöglichkeiten für Machine-Learning-Modelle deutlich verbessern. In der Evaluation mit sechs Medizinstudierenden wurde insbesondere die farbliche Markierung der Textstellen positiv bewertet.

Angelo Lienhard
077 422 29 87
lienhard.angelo@gmail.com



Screenshot der Webapplikation