Klinische Textanalyse und ICD-Code Zuweisung mit vortrainierten NLP-Modellen

Studiengang: BSc in Medizininformatik
Betreuer: Prof. Dr. Murat Sariyar

Die steigende Komplexität der ICD-10 Codes in Spitälern, begleitet von einer massiven Zunahme der Diagnosecodes, stellt medizinische Kodierer vor Herausforderungen. Vortrainierte Natural Language Processing (NLP)-Modelle können dazu beitragen, den Kodierungsprozess zu unterstützen und zu vereinfachen.

Einleitung

Die kontinuierliche Weiterentwicklung der ICD-10 Kodierung in Spitäler und die Erweiterung der Diagnosecodes von etwa 18‘000 auf über 155‘000 stellen eine immer grösser werdende Aufgabe für medizinische Kodierer dar. Die Übersetzung komplexer klinischer Informationen in standardisierte Codes ist ein zeitaufwendiger und fehleranfälliger Prozess. Die Anwendung von vortrainierten NLP-Modellen kann hierbei unterstützen, indem sie die Genauigkeit in der Diagnosekodierung verbessert und den Kodierungsprozess vereinfacht.

Methodik

Ein Algorithmus zur automatisierten ICD-10 Kodierung wird mithilfe der vorhandenen MIMIC-IV Daten entwickelt. Dieser Ansatz beruht auf einer gründlichen Literaturrecherche und nutzt eine breite Palette effizienter und bewährter NLP-Modelle aus der Hugging-Face-Bibliothek. Die Datenanalyse erfolgt durch den Einsatz von Power Query, während MongoDB als Speicherlösung für die MIMIC-IV Daten dient.

Ergebnisse

Die vortrainierten Modelle BioBERT, ClinicalBERT, ClinicalLongformer, ClinicalBigBird und Llama2 wurden für das Projekt adaptiert. Spezifische Preprocessing-Methoden wurden eingesetzt, um eine effiziente Nutzung der limitierten Kontextlänge zu ermöglichen. Dies umfasste das Entfernen von Zahlen und unnötigen Zeichen sowie die Auswahl der Tokens ab dem Stichwort «discharge». 512 Tokens - BioBERT, ClinicalBERT, ClinicalLongformer und ClinicalBigBird zeigen eine ähnliche Accuracy von 92% bei den Top 50 ICD-10 Codes. 1‘024 Tokens - ClinicalLongformer und ClinicalBigBird erreichen eine Accuracy von 93.4% bzw. 93.6%.

Diese Ergebnisse zeigen, dass eine erhöhte Anzahl an Tokens einen - wenn auch kleinen - positiven Effekt auf die Accuracy hat. Bei 4‘096 - Tokens zeigt ClinicalLongformer eine Accuracy von 94.6%. Dies bestätigt den positiven Einfluss einer erhöhten Token-Anzahl auf die Accuracy. Die maximale Token-Kapazität von ClinicalBigBird und Llama2 konnte aufgrund der begrenzten Serverkapazität nicht vollständig ausgeschöpft werden. Die Nutzung bereitgestellter Informationen auf einer zentralen Plattform wie GitHub, insbesondere in Form detaillierter technischer Dokumentationen, soll neuen Studierenden den Einstieg in die NLP-Analyse mithilfe von Large-Language-Modellen erleichtern.

Diskussion

Die Token-Kapazität spielt eine entscheidende Rolle, da Modelle wie BioBERT und ClinicalBERT bei längeren Texten an ihre Grenzen stossen. Die Leistung beider Modelle entspricht weitgehend den Ergebnissen aus wissenschaftlichen Studien. Hingegen erzielt ClinicalLongformer bessere Ergebnisse bei seiner maximalen Tokenlänge. Obwohl unsere Ergebnisse im Vergleich zu Fachstudien geringer sind, beruhen sie auf einer kleineren Datenmenge. Dies unterstreicht die Wichtigkeit eines effektiven Preprocessings und des Umgangs mit Stoppwörtern für NLP-Modelle. In diesem Projekt wurden Stoppwörter beibehalten, da die Modelle auf den gesamten Kontext angewiesen zu sein scheinen.

Mariem Mansour
Advanced Data Processing

Fatma Yilmaz
Advanced Data Processing