ZUSAMMENFASSUNG

Medizinische Daten können für Analysen und Forschung im medizinischen Bereich verwendet werden, jedoch bleibt ein großer Teil der Daten unstrukturiert und somit ungenutzt. Hierfür sind medizinische Berichte wie Arztbriefe ein Beispiel: Sie dienen der Übermittlung von medizinischen Informationen, die in natürlicher Sprache für ärztliches Fachpersonal formuliert wurden. Doch damit die in der natürlichen Sprache enthaltenen Informationen auch von automatisierten Systemen genutzt werden können, müssen diese Informationen zunächst extrahiert werden.

Die Eigennamenerkennung ist ein Teilgebiet der Informationsextraktion im Bereich der Computerlinguistik und befasst sich damit, jedes Wort eines Dokumentes einer zuvor definierten Kategorie zuzuordnen. Die Kategorien, die innerhalb dieser Arbeit zugewiesen werden, sind Diagnose, Behandlung und Medikation. Unser Ziel ist es, die Forschung auf dem Gebiet der natürlichen Sprachverarbeitung in deutscher Sprache im medizinischen Bereich voranzutreiben. Dafür untersuchen wir geeignete Ansätze, um die Vorhersagequalität auf einem Vergleichsdatensatz zu verbessern.

Hierzu analysieren wir zunächst den Datensatz namens Berlin-Tübingen-Oncology Corpus (BRONCO). Mit Hilfe dessen trainieren, validieren und evaluieren wir sechs Transformer-Modelle: die deutschsprachigen Modelle GBERT, MedBERT und GELECTRA sowie die mehrsprachigen Modelle mBERT, XLM-RobERTa und XLM-RoBERTa GER. Wir trainieren die Modelle sowohl mit einem Merkmalsextraktions- als auch mit einem Feinabstimmungsansatz und führen eine Hyperparameter-Optimierung durch. Die Ergebnisse zeigen auf, dass der Feinabstimmungsansatz mit Hyperparameter-Optimierung zu den besten Ergebnissen führt.

Außerdem untersuchen wir die Umsetzbarkeit von domänenadaptivem Vortrainieren auf einem der Modelle. Aufgrund der geringen Menge an Daten zum Vortrainieren konnten wir bei diesem Ansatz keine signifikanten Verbesserungen feststellen. Schließlich führen wir eine detaillierte Untersuchung der Ergebnisse des besten deutschsprachigen Modells GELECTRA und des besten mehrsprachigen Modells XLM-RoBERTa durch.

Unseres Wissens nach erzielt GELECTRA den bisher höchsten F1-Wert auf dem zurückgehaltenen **BRONCO!** (BRONCO!)50-Testdatensatz mit einer Gesamtpunktzahl von 82.2 Prozent. Wir übertreffen damit die Ergebnisse, die von Kittner et al. für diesen Datensatz veröffentlicht wurden, je nach Kategorie um 2.3-7.7 Prozentpunkte.

Die Ergebnisse zeigen, dass die Adaption von Modellen auf dem aktuellen Stand der Forschung an die deutsche medizinische Eigennamenerkennung ein offenes Forschungsfeld mit vielversprechender Zukunft ist.