Large Language Models zur Bewertung der semantischen Ähnlichkeit von Textbeiträgen.

Eine Untersuchung ihres Einsatzes zur Identifikation von Duplikaten im Kontext einer Co-Creation Hub Applikation.

Studentin

Neva Nann

Aufgabenstellung: Das Start-up BrainE4 hat eine Co-Creation Hub Applikation entwickelt, um Lösungs-Ideen, Meinungen und Vorschläge auf offene Fragestellungen zu sammeln. In einem Mitmach-Dialog werden zu einer Frage die Beiträge durch Nutzer erfasst und in Paarvergleichen zur Bewertung gegenübergestellt.

In der Praxis hat sich gezeigt, dass der Moderator eines Dialogs erhebliche Arbeit mit der Identifikation von ähnlichen Ideen (Duplikate) hat, die im laufenden Dialog zu einer Kernidee zusammengefasst werden.

In dieser Projektarbeit wird der Einsatz von Large Language Models (LLMs) bei der automatisierten Erkennung von Duplikaten evaluiert. Es wird deren Fähigkeit zur Quantifizierung der semantischen Ähnlichkeit zwischen Texten untersucht und mit Textbeiträgen aus Dialogdaten der BrainE4 verifiziert.

Ergebnis: Drei Modelle zeigen eine besonders gute Performance: paraphrase-multilingual-mpnet-base-v2, multilingual-e5-base und all-MiniLM-L6-v2. Die Modelle paraphrase-multilingual-mpnet-base-v2 und all-MiniLM-L6-v2 erfahren eine Performancesteigerung durch die vorgängige Übersetzung der Texte ins Englische. Das Modell multilingual-e5-base klassifiziert die deutschen Originaldaten besser als übersetzte englische Daten.

Ein Ensemble-Modell mit Mehrheitsentscheid zeigt leichte Performancesteigerungen gegenüber den einzelnen Modellen und liefert konsistente und robuste Ergebnisse.

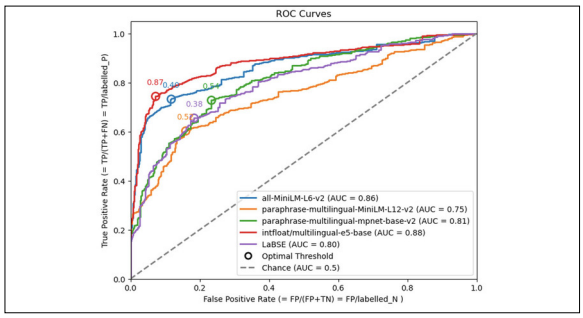
Ein Hybrid-Modell mit Kombination von semantischer Ähnlichkeit und Keyword-Suche ist erfolgreich beim Auffinden von Duplikaten. Es zeigt sich, dass die semantische Suche sinnvoll von der Keyword-Suche ergänzt wird.

Fazit: Das Modell multilingual-e5-base zeigt die beste Leistung gemessen am F1-Score und der Accuracy und hat zudem den Vorteil, dass die Performance auf den deutschen Originaltexten erreicht wird, ohne dass eine Übersetzung ins Englische notwendig ist.

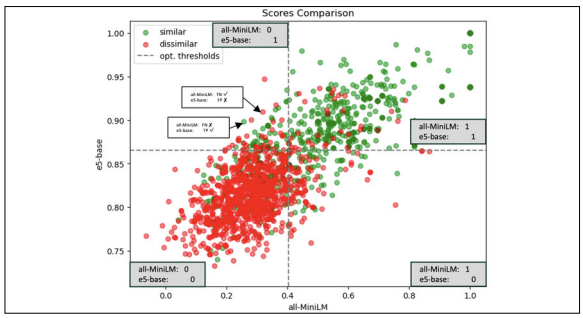
Die Performanceverbesserung durch Ensemble-Modelle liefert einen relativ geringen Mehrwert auf Kosten eines wesentlich höheren Rechenaufwands. Da im vorliegenden Anwendungsfall ein gewisser Interpretationsspielraum gegeben ist, ist die Verwendung des Einzelmodells multilingual-e5-base ausreichend.

Die Hybrid-Suche mit Kombination von semantischer Ähnlichkeit und Keyword-Suche ist zu empfehlen, um neue Textbeiträge auf Duplikate in der bestehenden Datenbasis abzugleichen.

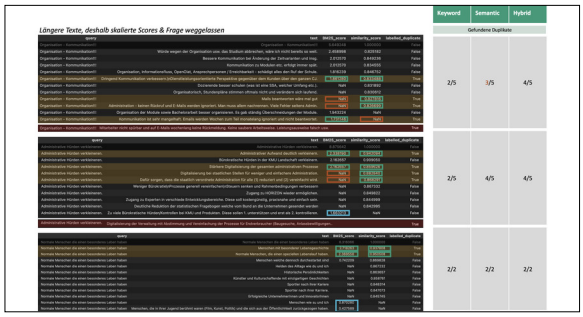
ROC Kurven geprüfter LLMs
Eigene Darstellung



Gegenüberstellung der Ähnlichkeitswerte zweier Modelle
Eigene Darstellung



Untersuchung der Ergebnisse aus der Hybrid-Suche
Eigene Darstellung



Referent
Prof. Dr. Lin Himmelmann

Themengebiet
Data Science

Projektpartner
BrainE4, 6330 Cham, Zug