Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

Entwicklung eines Frameworks für visuelle Korpusexploration und Vergleich von Machine Learning- und Deep Learning-Algorithmen

Einen Textkorpus zu verstehen ist überaus zeitaufwändig, jedoch ein wesentlicher Qualitätsbestandteil jeglicher darauf aufbauender Analysen oder Klassifizierungen.

In einer vorhergehenden Arbeit wurde ein System entwickelt, das eine intuitive Benutzeroberfläche zur Textklassifizierung anbietet. Darauf aufbauend werden weitere Module hinzugefügt und das bestehende System erweitert.

Im Rahmen des Projektes werden zwei neue Module geplant und entwickelt. Das erste analysiert einen Textkorpus mittels einer robusten Methode, Dateien einzulesen und die Analyse durchzuführen, welche Statistiken über Themen, Textlängen und relevante Wörter für sowohl den ganzen Korpus wie auch aufgeschlüsselt nach Klassen, liefert. Weiter werden verschiedene Sentence Splitters und Word Tokenizers durch das visuelle Hervorheben der Unterschiede verglichen.

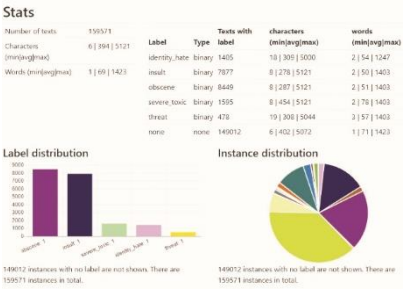
Das zweite Modul fügt einen grafischen Vergleich zum Klassifizierungsframework aus der vorherigen Arbeit hinzu. Die Klassifizierungsresultate der Modelle werden verglichen und gefiltert: Zum Beispiel nach kompletter Übereinstimmung oder nach Fällen, bei denen alle Modelle die falsche Klasse vorhergesagt haben. Von jedem Filter werden einige Beispiele in der Benutzeroberfläche präsentiert, wobei die Vorhersagen der Algorithmen farblich kodiert dargestellt werden.

Die primären Herausforderungen waren das Aufbereiten und Präsentieren der Ausgabe von verschiedenen Algorithmen und der Korpusanalyse. Dabei wurde darauf geachtet, dass der Benutzerfluss natürlich und die Darstellung intuitiv ist. Rückmeldungen aus einem Benutzertest bestätigen, dass diese Herausforderungen gemeistert wurden.

Es gibt viele Funktionen, die hinzugefügt werden können, wie zum Beispiel, dass die besten Parameter für die Klassifizierungsalgorithmen automatisch gefunden werden. Ferner könnte das Framework um weitere Module erweitert werden, wie beispielsweise die Möglichkeit, einzelne Texte zu untersuchen und analysieren sowie nach Begriffen im Korpus zu suchen.

Diplomierende
Linus Janis Metzler
Nadina Siddiqui

Dozent
Mark Cieliebak



Der obere Teil der Korpusanalyse zeigt Statistiken zu Wort- und Textlängen an. Ebenfalls wird die Klassen- und Instanzenverteilung als Diagramm dargestellt, um einen Einblick in den Korpus zu erhalten.

Der visuelle Vergleich von verschiedenen Klassifizierungen zeigt, welche Klassen von allen Algorithmen falsch vorhergesagt wurden. Die vorhergesagten Klassen werden mit der korrekten Klasse verglichen und das Resultat graphisch dargestellt.