Die Masterarbeit "An Efficient Algorithm for Computing Space-Time-Linguistics Similarities and Labelling Social Media Posts" von Stefan Zimmer befasst sich mit der Entwicklung einer Methode zur Extraktion von Emotionsinformationen aus Social Media Posts. Genauer gesagt wurde ein bestehendes Verfahren, das Ähnlichkeiten im geographischen, zeitlichen und linguistischen Raum zwischen Posts berechnet, hinsichtlich der Klassifikationsleistung sowie Laufzeit verbessert. Im Anschluss wird ein semi-überwachter, Graphen-basierter Maschine Learning Algorithmus ausgeführt um jeden Social Media Post mit seiner primär assoziierten Emotionskategorie – z. B. Fröhlichkeit, Wut oder Traurigkeit – zu versehen.
Die Arbeit fußt auf den schlechten Skalierungseigenschaften bei der Berechnung des Ähnlichkeitsgraphen durch die existierende Verarbeitungspipeline TwEmLab (quadratische Komplexität). Zur Steigerung der Geschwindigkeit wurde die Software zunächst CPU-parallelisiert. Dies brachte jedoch nicht den gewünschten Performancesprung. Zur weiteren Verbesserung kam deshalb die intelligentere Methode der Set Similarity Joins zum Einsatz. Dies erforderte zudem die Entwicklung von Verfahren zur Entschlüsselung der räumlichen, zeitlichen und linguistischen Eigenschaften von Posts hinsichtlich ihrer emotionalen Botschaft sowie deren Kodierung in eine geeignete Token-Repräsentation. Es wurden linguistische Features verwendet, die einzelne Wörter und Satzelemente sowie die schriftliche Ausdrucksweise eines Posts analysieren. So kamen neben der Emoji-, Akronym- und Onomatopöie-Stimmungsvalenz, der Art und Häufigkeit der Interpunktion, unter anderem auch N-Gramme zum Einsatz. Diese Features ermöglichen die semantische Charakterisierung von Einträgen in sozialen Medien in Bezug auf ihre Emotionslage und werden abseits der räumlichen und zeitlichen Dimension bei der Ähnlichkeitsberechnung verwendet.
In der Arbeit konnte gezeigt werden, dass der top-k Set Similarity Join unter Verwendung des Jaccard-Koeffizienten der Ausgangsmethode bei der Auswertung großer Datenmengen rechentechnisch überlegen ist. Insbesondere die Variante des Algorithmus, die auf tausenden Grafikkartenkernen parallel ausgeführt wird und eine quasi-lineare Komplexität aufweist, zeigt eine fundamentale Performancesteigerung im Rahmen von zwei Größenordnungen bei der Ähnlichkeitsberechnung im Vergleich zur existierenden Graphenerzeugung. Des Weiteren verhält sich der Algorithmus auch bei schiefen Verteilungen der Token ausreichend robust und ist demnach geeignet für ein größeres Spektrum an Eingangsdatensätzen. Das Klassifikationsergebnis konnte insbesondere bei weniger dominanten Emotionsklassen optimiert werden.