FG Software Engineering
Leibniz Universität Hannover
Welfengarten 1
3. Stock, Flur G

Bachelorarbeit

- zu vergeben -

Erstellung eines deutschen Datensatzes zur Stimmungsanalyse von Entwickleraussagen

Hintergrund

Um Stimmung innerhalb von Entwicklern messen zu können, wurden unter anderem Sentimentanalysetools wie Senti4SD oder RoBERTa entwickelt. Diese verwenden Methoden des machinellen Lernens und brauchen daher Daten fürs Training. Es existierend einige englische Datensätze, die aus Quellen wie GitHub oder Stack Overflow stammen. Diese wurden durch mehrere Personen mit Polaritäten wie Positiv, Neutral und Negativ oder mit Emotionen anhand eines Emotionsmodells gelabelt. Jedoch fehlt ein solcher Datensatz für den deutschsprachigen Raum, um Tools in diesem Raum entwickeln und einsetzen zu können.

Aufgabe

Im Rahmen der Arbeit soll ein deutscher Datensatz (mit mindestens 3000 Aussagen) erstellt werden. Dafür können bestehende Datensätze geeignet übernommen und übersetzt werden, sofern dies möglich ist. Zudem soll in einer ersten Recherche untersucht werden, welche öffentlich zugängliche deutsche Plattformen mit Kommunikationsdaten von Entwicklern vorhanden sind, um diese zu crawlen und zu labeln. Es sollen Teilnehmer gesucht werden, die dann eine Menge an deutschen Aussagen von Entwicklern anhand einer Guideline in Emotionen labeln sollen, basierend auf ein Emotionsmodell. Anschließend sollen Tools mit diesen Datensätzen trainiert und ihre Performanz ausgewertet werden.

Diese Arbeit gliedert sich in die folgenden Schritte:

1. Einlesen in die Grundlagen der Stimmungsanalyse
    a. Emotionsmodelle
    b. Sentiment Analysis Tools
2. Suchen/Crawlen deutscher Kommunikationsdaten von Entwicklern
3. Labeln des deutschen Datensatzes anhand eines Emotionsmodells
    a. Suchen von Teilnehmern
    b. Planen und Durchführung eines kleinen Workshops mit Erstellung einer Guideline fürs Labeln
4. Evaluation des Datensatzes
    a. Training vorhandener Tools mit dem Datensatz

Organisatorisches

Betreuer: Martin Obaidi
Beginn: ab sofort möglich