NLP state of the art Modelle – Entdecken von Hasskommentaren in Tweets

Studiengang: MAS | Vertiefung: MAS Data Science

Kann BERT - ein ‘state of the art language model for NLP’ - helfen Hass auf Twitter zu erkennen? Sind solche Modelle praxistauglich und erkennen sie zuverlässig Hasskommentare? Können sie ohne grossen Aufwand in Streaming- und Analyseumgebungen eingebaut werden? Auch wenn eine Firma nicht über die Ressourcen wie Google oder Facebook verfügt? Diese und andere Fragen bezüglich der technischen Machbarkeit beantwortet diese Masterarbeit.

Hasskommentare, Trolling und Belästigung im Internet nehmen immer grössere Ausmasse an. Auch bekannte Personen des öffentlichen Lebens äussern vermehrt diffamierende Aussagen in sozialen Medien, niemand scheint dabei Konsequenzen befürchten zu müssen. Dadurch sinkt die allgemeine Hemmschwelle und der Umgangston wird harscher. Es entsteht der Eindruck, dass es in den sozialen Medien kaum mehr Raum für Nachdenklichkeit, Reflexion und Empathie gibt.

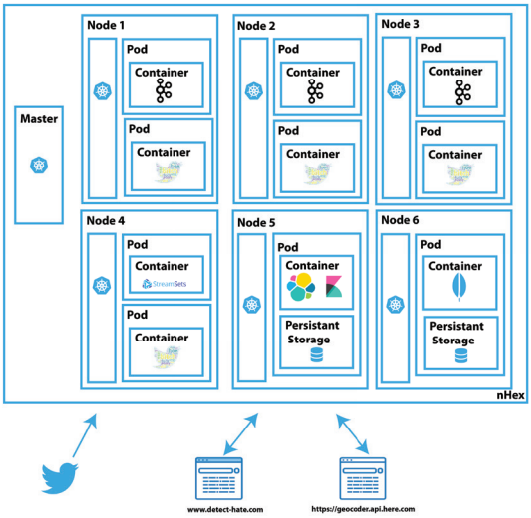
Meines Erachtens ist es zwingend, dass sich diesbezüglich etwas ändern muss: die Moderation von Inhalten auf Plattformen wie Twitter, Facebook oder auch Internetforen allgemein muss ausgebaut und verbessert werden. Darum untersucht diese Masterarbeit, ob eine automatisierte Erkennung von Hasskommentaren mittels ‘state of the art’ Algorithmen in der Praxis unter vertretbarem Aufwand möglich ist.

Zwar existieren zum Thema ‘hate speech detection’ bereits einige Forschungsarbeiten und Veröffentlichungen, diese sind jedoch meist rein wissenschaftlicher Natur. Zudem werden dabei grösstenteils klassische ‘machine learning’ Verfahren verwendet. In diesen Arbeiten fehlen aber Aussagen zur Praxistauglichkeit der Modelle. Zusätzlich zur Betrachtung gängiger Metriken - wie zum Beispiel F1 score, accuracy oder recall - müssen auch die Modellgrösse (benötigter Speicherplatz), Trainings- und Inferenzzeiten in einer praxistauglichen Anwendung betrachtet werden. Die Auswirkungen solcher Modelle auf die Hardware und der Aufwand für eine Implantation müssen ebenfalls analysiert werden.

Diese Masterarbeit prüft die technische Machbarkeit von BERT (Bidirectional Encoder Representations from Transformers) Modellen in einer Streaming- und Analyseumgebung, welche auf gängiger ‘on premise’ Hardware aufgebaut worden ist.

Verena Mai
verenamai@gmail.com

Word Cloud der häufigsten Wörter aus negativen Tweets von Trainingsdaten



Architektur der Streaming- und Analyseumgebung