Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

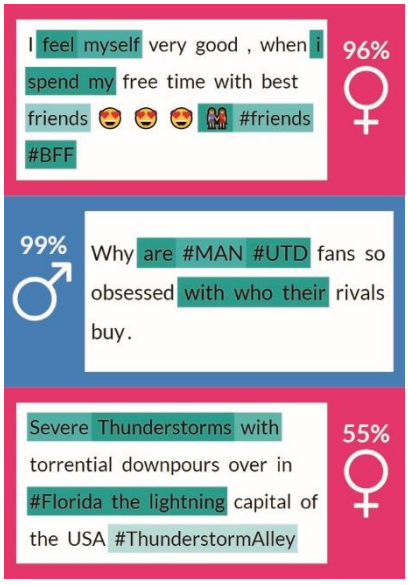
InIT Institut für angewandte
Informationstechnologie

Age und Gender Detection mit Word Embeddings und Deep Learning

Diplomierende
Florin Hardegger
Don Anson Kodiyan

Dozierende
Mark Cieliebak
Stephan Neuhaus

Diese Bachelorarbeit beschäftigt sich mit der Fragestellung, wie sich demografische Merkmale wie Alter oder Geschlecht von Twitter-Benutzern anhand ihrer Tweets möglichst korrekt bestimmen lassen (das sog. Author Profiling-Problem). Author Profiling ist ein wichtiges Instrument der Sprachanalyse und kommt heute bereits zur Anwendung, zum Beispiel um im Marketing die geeignete Zielgruppe für eine bestimmte Werbung zu ermitteln.
Diese Forschungsarbeit ermöglicht es, Alter und Geschlecht von Autoren exakter als in einer vorausgegangenen Projektarbeit vorherzusagen. Ausserdem wurde die Problemstellung erweitert, sodass nun auch die Erkennung von Sprachvariation (Dialekte) möglich ist. Für die Entwicklung der Modelle wurden Daten verwendet, die aus einem wissenschaftlichen Wettbewerb stammten.
Der Vergleich verschiedener neuronaler Netzwerke und die Verwendung ergänzender Datensätze führt zu einer wesentlichen Verbesserung der Performanz in Author Profiling. Dabei wurde der Fokus auf die Verwendung von Convolutional Neural Networks (CNNs) und Reccurent Neural Networks (RNNs) gerichtet.
Mit dem System wird eine Genauigkeit (Anteil korrekter Vorhersagen im Verhältnis zu allen Vorhersagen) von ca. 50 % bei der Vorhersage des Alters, von fast 80 % bei der Vorhersage des Geschlechts und von fast 80 % bei der Erkennung des Dialekts englischsprachiger Twitter-Benutzer erzielt. Das stellt in allen Bereichen eine deutliche Verbesserung gegenüber den besten Ergebnissen von PAN-2016 dar. Besonders schwierig war dabei die Erkennung des Alters, denn der PAN-Datensatz enthielt nur sehr wenige Benutzer im Alter von über 65 Jahren, was das Training und die Leistung der Modelle stark beeinträchtigt hat. Die Lösung für dieses Problem war die Erweiterung der Datenbasis für diesen Datensatz. Dabei wurden weitere Benutzer hinzugenommen, für die im Twitter-Profil ein Geburtsdatum angegeben war, und bei denen das geschätzte Alter gemäss Profilbild zum angegebenen Alter passte.
Die oben genannten Ergebnisse wurden mit einem Modell aus drei CNNs und einem RNN erzielt. Weitere vielversprechende Resultate liefert der Ansatz mit Bidirectional Gated Recurrent Units mit einem Attention-Mechanismus, der seit einiger Zeit in der maschinellen Übersetzung Bestmarken erreicht. Die Resultate dieses Ansatzes liegen leicht über denen, die mit CNNs und RNNs erzielt wurden.



In dieser Abbildung ist die Geschlechtsklassifizierung englischer Tweets ersichtlich. Dazu wurden Tweets dreier Autoren ausgesucht: Ein Tweet einer Frau, ein Tweet eines Mannes und ein Tweet einer Wetterstation. Je dunkler das Grün desto stärker wird ein Wort für die Klassifizierung gewichtet. Die Tweets der natürlichen Personen wurden mit einer geschätzten Sicherheit von über 95% richtig vorhergesagt. Der letzte Tweet enthält keine geschlechtsrelevante Informationen und wird dementsprechend klassifiziert.