

Die Auswertung von Schweizerdeutschen Texten erfolgt durch Polizisten und Polizistinnen hauptsächlich manuell oder mit der Hilfe einer Stichwortsuche. Es fehlen intelligente Lösungen, welche die automatisierte Verarbeitung der Texte zulässt. Daher soll anhand der Master Thesis aufgezeigt werden, ob eine auf Machine Learning basierte Klassifikation von Schweizerdeutschen Textnachrichten möglich ist.

Ausgangslage

Polizisten und Polizistinnen nutzen für die Auswertung von sichergestellten Textdaten Computerprogramme, welche die Texte mit Machine Learning Algorithmen in verschiedene Kategorien einteilen. Die Schweizerdeutsche Sprache wird jedoch bei kommerziellen Produkten nicht unterstützt, weshalb oft nur eine Stichwortsuche angewendet werden kann. In der vorliegenden Masterthesis soll anhand von Schweizerdeutschen Grooming-Nachrichtenverläufen aufgezeigt werden, ob eine maschinelle Klassifizierung von Schweizerdeutschen Texten möglich ist.

Vorgehen und Technologien

Da in der Schweizerdeutschen Sprache eine einheitliche Orthographie fehlt, ist eine auf lexikalischen Features basierende Klassifizierung nicht zielführend. Aufgrund dessen, wurde das sprachliche Verhalten des Täters oder der Täterin anhand bestehender Literatur und Interviews mit Fachpersonen, analysiert. Mit den Erkenntnissen konnten so aus den lexikalisch unabhängigen Metainformationen der Daten diskrete Features generiert werden, welche dialektübergreifend geltend sind.

Für die Analyse wurde mit den Technologien scikit-learn, NLTK und Python ein Jupyter-Notebook-Dokument erstellt. Diese Kombination der Technologien bewährt sich für die Analyse von Machine Learning Aufgaben und ist weit verbreitet.

Durch überwachtes Lernen wurde die binäre Klassifikation mit geeigneten Machine Learning Modellen getestet. Das Aufspalten der Nachrichtenverläufe in Blöcke erlaubte die optimale Kombination aus dem Informationsgehalt und der Anzahl der Samples. Anhand einzelner und kombinierten Featuresets wurden mit Hilfe der Kreuzvalidierung die Hyperparameter der Klassifikatoren fortlaufend optimiert.

Ergebnisse

Im Zuge dieser Master Thesis konnte aufgezeigt werden, dass eine Unterscheidung von Schweizerdeutschen Grooming- und Alltagsnachrichten mit Machine Learning Algorithmen möglich ist.

Trotz der geringen Menge an zur Verfügung stehenden Daten war es durch das Aufteilen der Nachrichtenverläufe in mehrere Samples möglich, aussagekräftige Ergebnisse zu generieren. Für die Problemstellung eignen sich die Klassifikatoren Naive Bayes und Gradient Boost. Mit den aus den Metadaten generierten Features resultierte ein F_1 -Score von 0.68 bei einem Anteil der positiven Klasse von 30%.

Wird die bag-of-words Präsentation der Daten verwendet erreicht der F_1 -Score sogar Werte bis zu 0.98. Für eine Schweizweit abgestützte Verlässlichkeit sind jedoch mehr Trainingsdaten erforderlich.

Ausblick

Das weitere Ziel ist die Erstellung eines Produktes, welches den Vollzugsbehörden zur Verfügung gestellt werden kann. Dies beinhaltet die Erweiterung des Delikte-Katalogs. Interessant sind dabei häufig auftretende Delikte wie Online-Betrug oder Betäubungsmittel-Widerhandlung. Dazu gilt es entsprechende Trainingsdaten zu sammeln und diese zu verarbeiten. In einer letzten Phase soll nach den Tests mit einem Prototyp eine marktreife Applikation erstellt und verbreitet werden.



Tobias Saladin