Karlsruher Institut für Technologie

aifb

# Bonusabgabe
# Anwendungen der Künstlichen Intelligenz

Dr. Michael Färber, Dr. Tobias Käfer, M.Sc. Shuzhou Yuan

16.12.2023

## Aufgabenstellung

Gerade in den letzten Jahren wurde immer mehr deutlich, was für eine Reichweite Kommentare und Beiträge im Internet haben können. Der Großteil der Meinungsäußerung findet online statt. Dadurch steigt jedoch auch die Anzahl an unangemessen und schädlichen Inhalten, die von Mobbing bis hin zu Shit-Storms ausarten können.

Ihre Aufgabe bei dieser Herausforderung wird es sein, die Rolle einer Content-Polizei zu übernehmen. Um ein zivilisiertes und ordnungsgemäßes Verhalten auf einer imaginären Social Media-Plattform zu gewährleisten, müssen Sie alle unerwünschten Inhalte aussortieren. Zu diesem Zweck wird Ihnen ein Trainingsset mit anonymisierten Benutzerbeiträgen zur Verfügung gestellt, die jeweils als hate speech, offensive language oder neither gekennzeichnet sind. Verwenden Sie diese Trainingsmenge, um ein Modell zu entwickeln, das Beiträge in diese drei Klassen möglichst gut klassifizieren kann.

### Daten

Die Datei train.csv umfasst Ihre Trainingsdaten. Jede Zeile enthält den Text eines Beitrags und ein Label, getrennt durch ein Komma. Die Klassenlabels sind „0“ für hate speech, „1“ für offensive language und „2“, für neither.

Die Datei test_no_label.csv enthält die Testdaten, die am Ende zur Bewertung Ihres Modells verwendet werden. Sie enthält keine Klassenlabels.

### Formalitäten

(1) Verwenden Sie bestehende Methoden des maschinellen Lernens für die Klassifizierung.

(2) Sie können existierende Programmiersprachen (vorrangig Python[1], R[2] auch möglich, falls bereits erlernt), Libraries (e.g. Pandas[3], Sklearn[4], NLTK[5], Keras[6]) und verwenden.

(3) Die Aufgabe darf alleine oder zu zweit bearbeitet werden. Falls Sie zu zweit daran arbeiten, notieren Sie bitte mit wem Sie die Gruppe gebildet haben und geben Sie jeweils einzeln ab.

[1] https://www.python.org
[2] https://www.r-project.or
[3] https://pandas.pydata.org
[4] https://scikit-learn.org/stable/
[5] https://www.nltk.org/
[6] https://keras.io

### Abgabe

Die Abgabe ist bis zum 30. Januar 2024, 23:59 Uhr über Ilias in einem dann entsprechend angelegten Abgabe-Ordner möglich. Einreichungen, die nach der Frist eingehen, werden von der Bewertung ausgeschlossen. Die folgenden Unterlagen müssen eingereicht werden:

1. Eine Kopie der Datei test_no_label.csv mit den von Ihnen vorhergesagten Klassenlabels (die Reihenfolge der Zeilen muss der ursprünglichen Datei entsprechen und verwenden Sie die im Training benutzten Klassenlabels. Bitte benennen Sie Ihre Datei test_with_label.csv)
2. Programmcode, der für das Training verwendet wurde.
3. PDF-Datei mit:
    - Vollständigem Namen, u-Kürzel und Matrikelnummer
    - Ggf. Teampartner und dessen u-Kürzel
    - Kurze Beschreibung des Ansatzes für die Klassifizierung in 4-5 Sätzen (z.B. Preprocessing, Trainingssplit, Modell, wichtige Parameter, ..)

Bitte laden Sie alle oben genannten Informationen/Daten in einem ZIP-Archiv von angemessener Größe (maximal einige MB) hoch und benennen Sie dieses wie folgt: u-Kürzel.zip. Abgaben, die formal nicht korrekt sind (z.B. inkorrekte Dateinamensgebung, falsches Dateiformat oder fehlende Dateien in der Abgabe), werden nicht bewertet.

## Ergebnisse

### Auswertung

Die von Ihnen vorhergesagten Labels für die Einträge in Testdatei werden mit Hilfe eines gewichteten F1 Scores bewertet.

### Bonuspunkte

Wenn Ihre Vorhersagen einen Weighted-F1-Score von 0,80 erreichen, erhalten Sie einen Bonuspunkt, bei 0,85 zwei Bonuspunkte und bei 0,90 drei Bonuspunkte auf Ihre bestandene Prüfung. Die beste Abgabe von allen erhält zusätzlich noch einmal drei Bonuspunkte.

Happy policing.