CODING CHALLENGE – NLP / Sentiment Classification

Aufgabenteil A: NLP

Zusammen mit dieser Challenge wurde Ihnen ein Datensatz bestehend aus 2 .csv Dateien (Trainings- und Validation-Datensatz) zur Verfügung gestellt. Der Datensatz enthält Tweets zu verschiedenen Produkten sowie eine zugehörige Bewertung (Sentiment) des Tweets. Ihre Aufgabe besteht in der Erstellung eines Modells zu Klassifizierung der Sentiments.

Die Ergebnisse ihrer Auswertungen stellen Sie bitte im Rahmen des etwa 60-minütigen technischen Interviews vor. Dabei haben Sie zunächst 15 Minuten Zeit, um die aus ihrer Sicht wesentlichen Ergebnisse vorzustellen. Die Form der Präsentation (frei, Folien, Jupyter-Notebook,…) wählen Sie dabei selbst. Anschließend findet eine Diskussion statt, in deren Rahmen Ihnen Fragen zu ihrer Vorgehensweise sowie zu Ihrem Programmcode gestellt werden. Es sollte anhand des Codes gut nachvollziehbar sein, wie sie dabei vorgegangen sind.

Gehen Sie dabei in folgenden Teilschritten vor:

- **Explorative Datenanalyse:** Machen Sie sich mit dem Datensatz und den für die Klassifikation relevanten Eigenschaften vertraut.

- **Modellbuilding:** Erstellen Sie mindestens ein Ihnen geeignet erscheinendes Modell für dieses Problem. Das Modell sollte „state-of-the-art“ Technologien in NLP verwenden, die Privatnutzern frei zur Verfügung stehen und die im Rahmen dieser Coding Challenge sinnvoll (insbesondere ohne die Notwendigkeit GPU-Zeit einzukaufen) einsetzbar sind.
- **Evaluation:** Evaluieren Sie die Performance des/der Modelle anhand geeigneter Kriterien / Metriken.

Aufgabenteil B: ML Engineering

Neben der Hauptaufgabe können Sie zusätzlich ihre Fähigkeiten und Kenntnisse im Bereich des Machine Learning Engineerings demonstrieren:

- **Aufbau REST-API:** Hierfür soll der Inferenz-Schritt Ihres trainierten Modells per „einfachem“ REST-Service verfügbar gemacht werden.
- **Containerisierung:** Zusätzlich soll der erstellte REST-Service mittels „Dockerfile“ in einem Container laufen.

Diese Coding Challenge ist in wenigen Stunden gut zu erledigen. Wir erwarten nicht, dass Sie mehr Zeit investieren. Bitte teilen Sie sich ihre Zeit entsprechend ein und konzentrieren Sie sich auf das – aus ihrer Sicht - Wesentliche. Eine erschöpfende Tiefenanalyse der Daten und oder eine aufwändige Optimierung vieler verschiedener Modelle werden dementsprechend nicht von Ihnen erwartet.

Wichtig:

Bitte stellen Sie uns den von Ihnen im Rahmen der Coding Challenge erstellten Quellcode spätestens 24h vor dem Technischen Interview in geeigneter Form zur Verfügung. Da der Spam-Filter der HUK-COBURG recht restriktiv ist würden wir Sie bitten, uns den Code idealerweise als Download-Link (Dropbox, We-Transfer,…) oder GitHub Repository zur Verfügung zu stellen und nicht als Email-Anhang zu versenden.