14.10.2024

Das haben wir gemacht:

* Git Repository erstellt, dieses enthält eine Main und 2 weitere Branches namens micha\_dev und luca\_dev, um unabhängig voneinander im selber Programm arbeiten zu können. Fertige Inkremente werden dann in die Main geladen.

Diese Probleme sind aufgetreten:

* Keine Probleme

16.10.2024

Das haben wir gemacht:

* Daten aus Kaggle als csv Datensatz heruntergeladen
* Daten in Python unter Verwendung von Pandas als DataFrame hereingeladen
* Spaltenüberschriften werden anschließend festgelegt
* Tokenisierung der Daten mit Hilfe von transformers BertTokenizer

Diese Probleme sind Aufgetreten:

* Probleme beim Encoding, der Datensatz lag in Kaggle nicht in UTF8 vor; Datensatz musste in Python encoded werden in ISO8859-1
* Wie bereinigen wir die Daten? Stopwords entfernen? Dinge wie @ am Anfang entfernen, um die Daten strukturierter darzustellen, es kann aber nichtmehr erkannt werden, ob es ein Kommentar oder eine Antwort auf ein Kommentar ist.

Ein Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung

**Token:**

Token IDs: Dies ist eine Liste von Token-IDs, die den Text repräsentieren. BERT verwendet eine spezielle Tokenisierungsmethode, bei der jedes Wort oder Subwort in eine eindeutige ID umgewandelt wird.

Spezielle Token: Die IDs 101 und 102 sind spezielle Token für den Anfang ([CLS]) und das Ende ([SEP]) eines Satzes.

Padding: Die 0-Werte am Ende sind Padding-Token, die verwendet werden, um alle Sequenzen auf eine einheitliche Länge zu bringen (in diesem Fall 128 Token).

Attention Maske:

Attention-Maske: Diese Maske zeigt an, welche Token tatsächlich Teil des Textes sind (mit 1 markiert) und welche Token Padding sind (mit 0 markiert). Dies hilft dem Modell, die Padding-Token zu ignorieren.

Label: Label: Dies ist das Ziel-Label für den Text. In diesem Fall ist das Label 0, was wahrscheinlich eine negative Stimmung repräsentiert (abhängig von der Kodierung in Ihrem Datensatz).

Zusammenfassung

Token IDs: Repräsentieren den Text in einer Form, die das BERT-Modell versteht.

Attention-Maske: Hilft dem Modell, relevante Teile des Textes von Padding zu unterscheiden.

Label: Das Ziel-Label für die Sentiment-Analyse.

Diese vorbereiteten Daten können nun verwendet werden, um ein vortrainiertes BERT-Modell zu trainieren oder zu evaluieren.