Übersicht des Quellcodes und der Datensätze

Der in der Arbeit verwendete Code ist in fünf JupyterNotebooks enthalten.

Die Ausführung der Notebooks ist auf googleColab ausgelegt. Eine lokale Ausführung über JupyterLab oder andere Anwendungen sollte auch möglich sein. Dabei könnte es jedoch mit Imports oder anderen Sachen zu Problemen führen.

Das Ausführen der Notebooks in Colab ist einfach. Die Code-Zellen sind so angeordnet, sodass man sie von Anfang bis Ende, der Reihe nach ausführen kann.

Die Notebooks und die verwendeten Daten sind in den beiligenden Ordnern vorliegend.

Eine kurze Beschreibung des Inhalts jedes Notebooks:

- JsonFormatting - Es werden die Shared Task Daten in das gewünschte Format gebracht.
- Datenerweiterung - Politische Texte werden mit durch lexikalem Abgleich mit dem Opinion Role Lexicon schwach klassifiziert.
- Modell_aufbauen - Es wird ein BERT-Modell auf den Shared Task Daten trainiert (Modell v1)
- Modell_Datenerweiterung - Es wird ein BERT-Modell auf den schwach klassifizierten Daten trainiert, und dann auf den Shared Task Daten (Modell v2)
- scoring - Es wird ein trainiertes Modell geladen, mit welchem der Testdatensatz klassifiziert wird und mit den Gold-Standart Labels verglichen wird.

Eine Übersicht über die Dateien in ihren Ordnern:

- Shared Task Daten - Die jsonl-Dateien sind die originalen Shared Task Daten die in JsonFormatting zu den drei Text-Dateien umformatiert werden.
- Datenerweiterung Daten - Beinhaltet das Opinion Role Lexicon, die Plenarprotokolle in den originalen XML-Dateien und der schwach-klassifizierten Variante
- Jupyter Notebooks - JupyterNotebooks

Zum vollständigen Ausführen von Modell_aufbauen und Modell_Datenerweiterung wird ein Huggingface

Account benötigt, um das Trainierte Modell imHugginface Modell-Hub abzuspeichern.

Links zu den Repositories:

https://github.com/xBosse/datasets

https://github.com/xBosse/datenerweiterung

https://github.com/xBosse/notebooks