NLP Project Proposal

Group

Ruiqi Li, Yixi Liang, He Huang

Writing is one of the most important skills that everyone must learn. It is a significant task to help pre-high school students to learn how to write. Therefore, we will focus on analyzing the structures and elements of the writing samples from students in grades 6-12. The main data source is from Kaggle dataset of "Evaluating Student Writing" which is presented by Georgia State University.

The initial plan of model choices involves pretrained NLP and rule-based models to perform the text classification. Customization will be done based on the pipeline construction. Besides the packages like numpy, scikit, and scipy for general data handling, nltk will be expected to use for text processing. Transformers (from Tensorflow and PyTorch) are considered for building models. NLP tasks that may be performed includ tokenization, normalization, stopwords handling, sentiment analysis, etc.

At this point, we consider using the confusion matrix, roc-auc, Cohen's Kappa, Matthews correlation, and embedded evaluation function for performance evaluation. Moreover, we can submit our pipeline onto Kaggle to receive its score.

Here is some schedule to follow:

- By Nov 16 – Data Wrangling

- By Nov 30 – General Pipeline of training and testing

- By Dec 7 - Evaluation

- Dec 12 – Presentation

Dataset URL: https://www.kaggle.com/competitions/feedback-prize-2021/data