Name: Suguru Naresh

Email address: sugurunaresh111@gmail.com

Contact number: +91 8096569188

Anydesk address: **485 090 771**

Years of Work Experience:

Date: **25th June 2020**

Self Case Study -2: ***Kaggle Competition- CommonLit Readability Prize***

Overview:

CommonLit, Inc., is a nonprofit education technology organization serving over 20 million teachers and students with free digital reading and writing lessons for grades 3-12. Together with Georgia State University, an R1 public research university in Atlanta, they are challenging Kagglers to improve readability rating methods.

Currently, most educational texts are matched to readers using traditional readability methods or commercially available formulas. However, each has its issues. Tools like Flesch-Kincaid Grade Level are based on weak proxies of text decoding (i.e., characters or syllables per word) and syntactic complexity (i.e., number or words per sentence). As a result, they lack construct and theoretical validity. At the same time, commercially available formulas, such as Lexile, can be cost-prohibitive, lack suitable validation studies, and suffer from transparency issues when the formula's features aren't publicly available.

So the CommonLit organization conducted this challenge to use the Machine learning skills to rate the educational texts and engage the students with texts of the right level of challenge which can help them to develop good reading skills.

In this competition, we shall build the ML models to rate the complexity of reading passages for grade 3-12 classroom use with the help of the vast dataset provided that includes readers from a wide variety of age groups and a large collection of texts taken from various domains.

Performance metric:

Submissions are scored on the root mean squared error. RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $\hat{y}$ is the predicted value, y is the original value, and n is the number of rows in the test data.

Data Description:

We have been provided with the excerpts from several time periods and with a wide range of reading ease scores.Here we need to predict the readability score for these excerpt texts

The train/test dataset contains below columns:

- `ID` - unique ID for excerpt
- `url_legal` - URL of source
- `license` - license of source material
- `excerpt` - text to predict reading ease of
- `target` - reading ease
- `standard_error` - measure of spread of scores among multiple raters for each excerpt

Research-Papers/Solutions/Architectures/Kernels:

1. https://www.kaggle.com/ragnar123/commonlit-readability-roberta-tf

The above solution is contributed by Ragnar in kaggle which helps the challengers with a good start.He tokenized the text data with RoBERTa Tokenizer and applied RoBERTa base models with 4 KFolds to achieve a good score.

2. https://arxiv.org/pdf/1907.11692.pdf

It is the research paper presented by Yinhan Liu along with few others at Facebook AI.It is the finely tuned BERT model which is trained on a vast dataset(nearly 160GB) and can be used for transfer learning for NLP applications effectively.

3.https://huggingface.co/roberta-base:

It is the official website which provides all the documentation regarding RoBERTa and repositories for RoBERTa basemodels.

First Cut Approach:

I would like to experiment with the data with traditional methods and apply TFIDF Vectorizer after preprocessing the data and compare the results with Decision Trees,XGBT ,BERT before using the RoBERTa Model.