Word Complexity Prediction

https://sites.google.com/view/lcpsharedtask2021

Task Description:

Lexical complexity plays a crucial role in reading comprehension. Predicting lexical complexity accurately can enable a system to better guide a user to an appropriate text, or tailor a text to their needs. NLP systems have been developed to simplify texts for second language learners, native speakers with low literacy levels, and people with reading disabilities. Thus, the main objective of LCP is to predict the difficulty of the word in the passage.
This problem is sub-divided into two tasks:

- Task-1: predicting the complexity score of **single words**
- Task-2: predicting the complexity score of **multi-word expressions**.

Dataset Description:

Dataset consists of trial data, train data and validation data. Each dataset consists of id, corpus, sentence, token (single word or multiword) and complexity of the token. Annotators have assigned complexity from 1-5 and then normalized between 0-1 complexity score. Dataset consists of text from three domains bible, biomedical and Europarl.

Approach:

We can approach this problem as regression task as the complexity score is real value ranging between 0-1. I have explored various features such as length of the token, pos tag of the token, no of synonym/antonym, no of vowels, no of syllables, occurrence of target token in training sentences. In addition, I have used grove 300D embedding of the target token. Furthermore, I have also used extracted contextual features with the help of BERT model.

Experiments and Results:

I have experimented with various models as shown below. Lexical complexity prediction task uses two evaluation metrics: Pearson score and R2 score.

- **Sub-task 1 Results:**

Model	Pearson score	R2 score
Lasso Regression	0.73	0.53
Adaboost Regressor	0.72	0.52
Gradientboosting Regressor	0.74	0.55
Bagging Regressor	0.73	0.54

3 layered Neural Network	0.75	0.55
Stacking (linear regression)	0.76	0.55
Stacking (KNN regression)	0.75	0.56

- **Sub-task 2 Results:**

Model	**Pearson score**	**R2 score**
Lasso Regression	0.79	0.62
Adaboost Regressor	0.78	0.61
Gradientboosting Regressor	0.80	0.63
Bagging Regressor	0.80	0.62
3 layered Neural Network	0.78	0.61
Stacking (linear regression)	0.81	0.61
Stacking (KNN regression)	0.81	0.63