# Student Essay Score Predicting with Neural Networks

Yu Tao

yt560@georgetown.edu

## 1. Abstract

In this project, the author constructed an essay rating system that utilized the student essays to predict their scores. The author focused on deep learning models, building LSTM model from scratch, and fine-tuning pre-trained DistilBERT and DeBERTa models. These models are trained on student essays available from Kaggle.com. The outcomes were evaluated on quadratic weighted kappa. The final results indicated that DeBERTa model overperformed the others, yields kappa = 0.80, as the optimal choice for student essay scoring tasks.

## 2. Motivation

Essay writing is a method to evaluate student learning performance. Thousands of students take essays as the test on the same topics. However, it is time-consuming for educators to grade manually. Developing a reliable Automated Essay Scoring (AES) system can not only reduce educators' workload, but also provide students with in-time feedback. In addition to saving time and effort, AES is aimed to eliminating human rater errors and to create a fairer essay scoring system than human grading (Kusuma et al., 2022).

The goal of this project is to build a neural network model to predict the student essay scores that assists essay grading system.

## 3. Introduction

This project is inspired by a competition held on Kaggle.com, named Automated Essay Scoring (AES) 2.0. The first AES competition was held in 2012 sponsored by the William and Flora Hewlett Foundation, and the second competition was held in 2024.

Based on previous studies, essay scoring can be modeled via linear regression, support vector machine (SVM), multiclass linear discriminate analysis (LDA), and regression trees. These models require do feature engineering manually to get extra information from the text data, including number of words, number of sentences, number of vocabularies, number of grammar errors, etc. The drawback of these models is their lack of the ability of semantic understandings (Kusuma et al., 2022).

Essay scoring can also be modeled using neural network models such as CNN, LSTM, and BERT. The strength of these models is that the underlying semantics can be understood through word embeddings, and there is usually no need for manual feature extraction. However, pre-trained models like BERT have a limitation on the input size, accepting only a maximum length of 512 tokens, the remaining words will be ignored. Same as LSTM model, which requires a fixed number of input lengths in the embedding layer, making it impossible to consider all words. Thus, deep learning models cannot consider features like word count and sentence count.

## 4. Data Collection and Data Preprocessing

The dataset can be downloaded from Kaggle.com, it is divided into a training set and testing set. The whole dataset comprises about 24,000 student-written essays. Approximately 17,000 of them are in the training set that can be utilized for training and validating. The whole test set is invisible for the public, but Kaggle.com provided the first three records as a sample of submission.

There are three variables included in the training file, they are "essay_id" (the unique ID of the essay), "full_text" (the full essay response), and "score" (holistic score of the essay on a 1-6 scale). The score measures the quality of essays, 1 is the lowest score and 6 is the highest score. According to the grading criteria, an essay scored 1 indicates a lack of critical thinking, lack of evidence to support its position, disorganized on the essay structure, and errors in vocabulary ang grammar. An essay scored 6 indicates critical thinking, examples provide a solid support to its position, the structure is well-organized, and minimal errors in vocabulary and grammar.

The training essays is divided in a ratio of 80-20, split into a training set with 13,846 essays and a validation set with 3,461 essays before modeling. And the scores are rescaled from 1-6 to 0-5.

## 5. Exploratory Data Analysis

New features including word count, sentence count, paragraph count, and vocabulary size are extracted from the text data for exploratory data analysis.

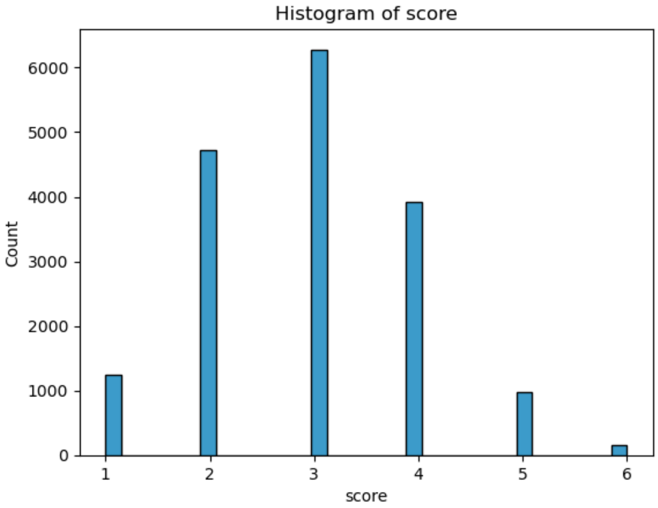


Figure 1. The histogram of score

As shown in figure 1, the distribution of essay scores is imbalanced and follows a normal distribution. The proportion of essays with a score of 3 was the largest, with the majority of essay scores falling within the range of 2 to 4. There are few essays obtained extremely high or extremely low scores. About 1200 essays are scored 1 and 5 separately. Essays with the highest score 6 have a count less than 200.

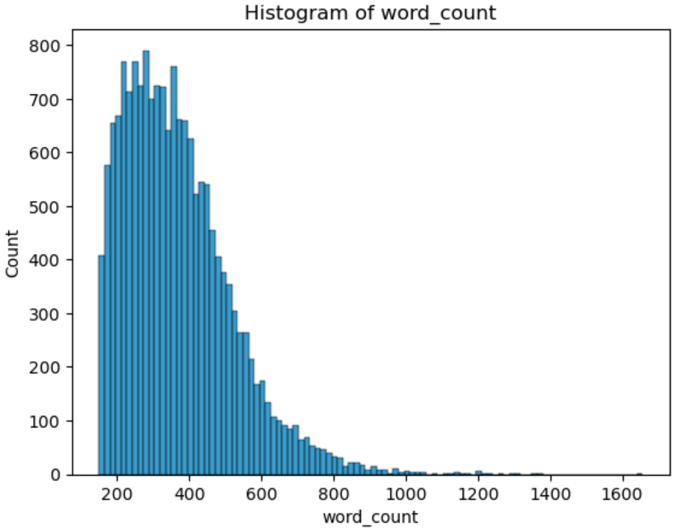


Figure 2. The histogram of word count

Figure 2 reveals the distribution of word count over all training text. There are no essays fewer than 100 words, most concentrated around 400 words, and very few essays exceeded 600 words. Through the visualization of word count, the author believes that cutting of one essay after 512 words is a reasonable number when using deep learning models.

| Topic | Keywords |
|---|---|
| 1 | venus, planet, earth, author, would, like, could, also, surface, humans |
| 2 | cars, car, driverless, would, people, driving, could, driver, many, drive |
| 3 | seagoing, people, get, luke, animals, cowboys, help, program, also, cowboy |
| 4 | electoral, face, vote, college, people, mars, president, would, states, electors |
| 5 | students, could, technology, emotions, would, help, facial, computer, student, system |

Table 1. Potential topics listed over the essays.

Latent Dirichlet Allocation (LDA) technique is used for topic modeling to extract potential topics from given essays. As we can see from Table 1, the essays cover a wide range of topic, such as the universe (topic 1), society (topic 2), nature (topic 3), politics (topic 4), and technology (topic 5).

## 6. Modeling

The project focused on developing deep learning models for essay scoring. In the modeling part, both RNN-based and transformer-based models will be considered and compared in the very end.

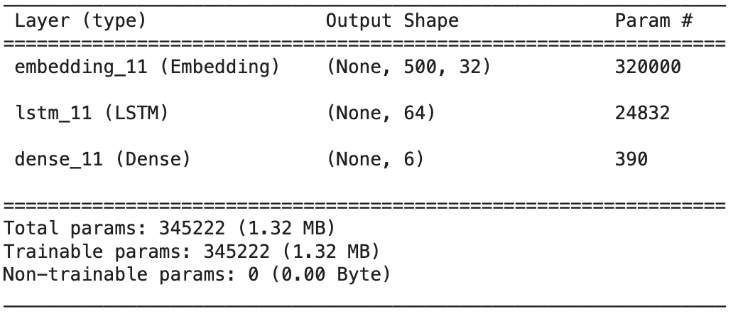
Recurrent Neural Network (RNN) is a neural network that has an internal memory. Unlike other neural networks that the outputs only depend on the current inputs, the output of RNN depends not only on the current input, but also all the past inputs. The fact that RNN can accept sequential data makes it suitable for tasks like natural language processing, speech recognition, and time-series analysis. RNN could learn long-term dependencies in theoretical, but it does not in practice. Another problem that RNN-type models have is that they process data in sequential and could not do parallelized computations.

Transformer is another architecture of neural networks. There are no recurrent units in the architecture, the model learn context in parallel, thus requires less computing time compared RNN models. The attention mechanism is one of the key components inside transformer, which allows information to flow directly from any position to any other position without flowing through the positions between them.

Bi-directional Encoder Representations and Transformers (BERT) is a Transformer structured deep learning model that used as end-to-end model in AES. Several researchers compared the model performance on BERT and ensemble-based approach, figured out that the BERT has a better performance (Beseiso & Alzahrani, 2020).

## 6.1. Long Short Term Memory (LSTM)

LSTM is a variation of a Recurrent Neural Network model. Compared to RNN model, LSTM model introduced forget gate, input gate, and output gate in one cell so that making it remember long-term data and prevent vanishing gradient possible (Mittal, 2021). The model has been used for multiple NLP tasks and has showed great performance.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_11 (Embedding)    (None, 500, 32)           320000

 lstm_11 (LSTM)              (None, 64)                24832

 dense_11 (Dense)            (None, 6)                 390

=================================================================
Total params: 345222 (1.32 MB)
Trainable params: 345222 (1.32 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 3. LSTM model structure

Figure 3 is the LSTM model structure used in this project. The simple LSTM model is used as the benchmark and it contains three layers: the embedding layer, the LSTM layer, and the output dense layer. The model takes the most frequent 10,000 vocabulary from the corpus, truncates all the essays to 500 words, and embeds each token to 32 dimensions. The output layer used softmax activation for multi-class classification problems. The model is compiled with “adam” optimizer, loss is calculated on categorical cross entropy, and it also returns the quadratic weighted kappa value.

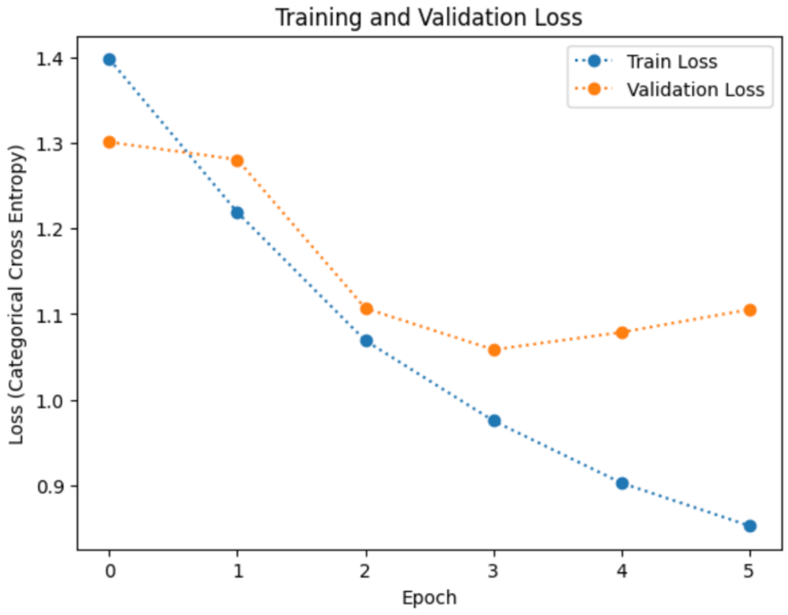


Figure 4. Training and validation loss for LSTM model

Figure 4 shows the training and validation loss during training procedure. The training loss keeps decreasing, but the validation loss does not always go down and there is some evidence of overfitting after 3 epochs. The training procedure stopped after 5 epochs because of the introduce of early stopping, and the best model obtained is the model at epoch 3.

## 6.2. DistilBERT

DistilBERT is a scaled-down version of BERT model. In terms of model structure, DistilBERT has only half of the number of layers compared to the BERT base model with 12 layers. In terms of the number of parameters, BERT base model has 110m parameters, while DistilBERT has only 66m parameters, which is 40% smaller than BERT. Although DistilBERT has a smaller size and faster inference speed, it still retains 97% of BERT performance (Efimov, 2023).

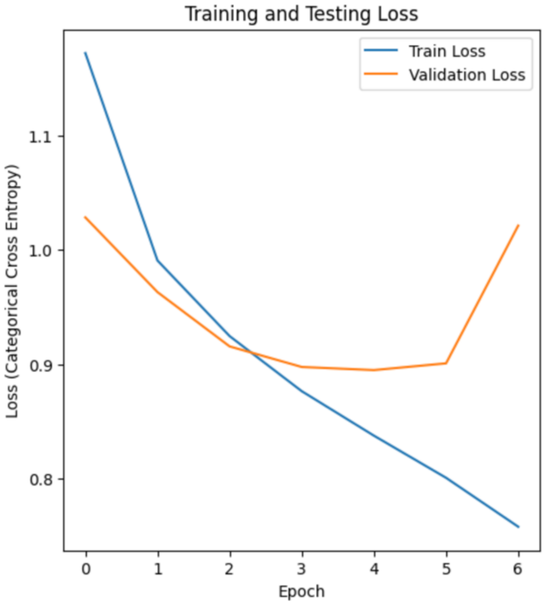


Figure 5. Training and validation loss for DistilBERT model

The author used a customized DsitilBERT model by adding one additional dense layer with dropout and ReLU activation onside the DistilBERT output layer. Figure 5 illustrates the training procedure. The training stopped after epoch 6 because of the increase on validation loss, and the optimal model was obtained at epoch 4. The loss curves are similar compared to the LSTM curves: training loss continuous to decrease, and the model appears to be overfitting after epoch 6. However, the DistilBERT model obtained a smaller loss on both training and validation data sets.

## 6.3. Decoding-Enhanced BERT with Disentangled Attention (DeBERTa)

DeBERTa is an upgraded version of BERT model invented by Microsoft in 2021, it is also the most popular model applied in NLP field. There are two updates on the DeBERTa model compared to the BERT model. First is disentangled attention. The DeBERTa model separates the word embedding into two vectors that contains word information and position information separately in the transformer block, which enables the model to figure out the importance in both words and positions. The second upgrade is the enhanced mask decoder. In the decoding layer, absolute position encoding is added into the model, which enables the model to distinguish two different instances although they are similar in the vector spaces (Efimov, 2023). Consider the sentence “a new store opened beside the new mall”, the DeBERTa model can figure out that the “store” is the subject and is different with the “mall” (He et al., 2020).

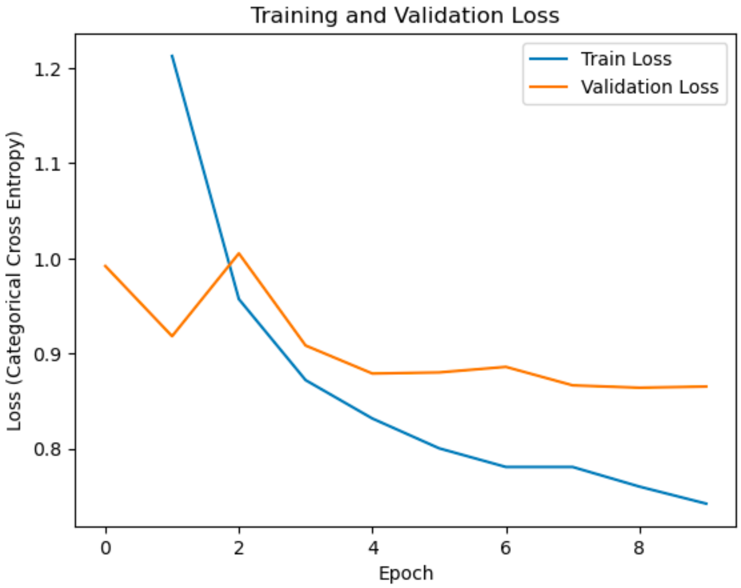


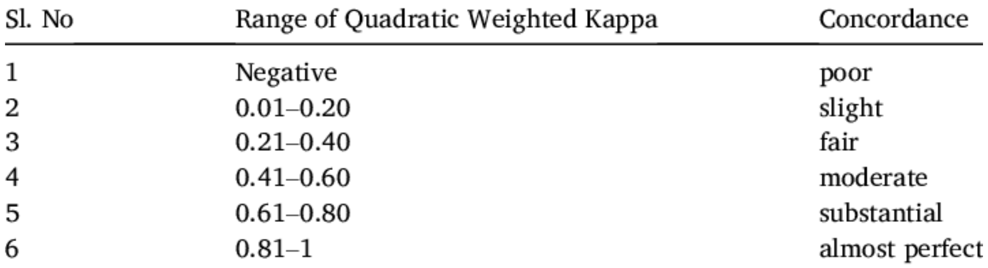
Figure 6. Training and validation loss for DeBERTa model

The model structure used in this project is the basic DeBERTa model without customization. A Trainer from transformer package is used for training. Figure 6 illustrates the training procedure of DeBERTa model. The author trained the model for 10 epochs with early stopping. Training stopped after 10 epochs and the best model was retrieved at epoch 7. The validation loss after epoch 7 did not increase significantly as the DistilBERT model, but the minimal validation loss value is similar at around 0.9.

# 7. Results

## 7.1. Evaluation Metric

The model performance is evaluated on quadratic weighted Kappa (QWK), which is a metric for evaluating ordered classes. QWK measures the agreement between the model prediction and the ground truth. The introduction of quadratic will punish more on a larger disagreement and punish less on a smaller disagreement. The value typically ranges from 0 (random agreement between raters) to 1(complete agreement between two raters), and the value may go below 0 if the disagreement is larger than the agreement (Aroraaman, 2018).



| Sl. No | Range of Quadratic Weighted Kappa | Concordance |
|---|---|---|
| 1 | Negative | poor |
| 2 | 0.01–0.20 | slight |
| 3 | 0.21–0.40 | fair |
| 4 | 0.41–0.60 | moderate |
| 5 | 0.61–0.80 | substantial |
| 6 | 0.81–1 | almost perfect |

Table 2. Interpretation of quadratic weighted kappa.

Table 2 displays the range of kappa values and the corresponding meanings. If the prediction of a model received a value between 0.61 and 0.80 indicating that the outcome is acceptable, and a value greater than 0.81 can be considered as almost perfect predictions (Abraham & Nair, 2019).

## 7.2. Summary Table of Modeling Results

| Model | Train/Validation Loss | Validation QWK |
|---|---|---|
| LSTM | 0.90/1.08 | 0.70 |
| DsitilBERT | 0.76/0.89 | 0.73 |
| DeBERTa | 0.76/0.86 | 0.80 |

Table 3. Modeling loss and quadratic weighted kappa summary table.

The table above is the summary of the modeling results. The BERT models have a smaller training and validation loss compared to RNN model. DsitilBERT and DeBERTa model received the same training loss, but DeBERTa model got a slightly smaller validation loss. The quadratic weighted kappa of LSTM model is 0.70, the DistilBERT model got a value of 0.73, and the DeBERTa model got a value of 0.80. Based on the loss and kappa values, the DeBERTa model is the optimal model for solving the essay scoring problems.

## 7.3. Confusion Matrix

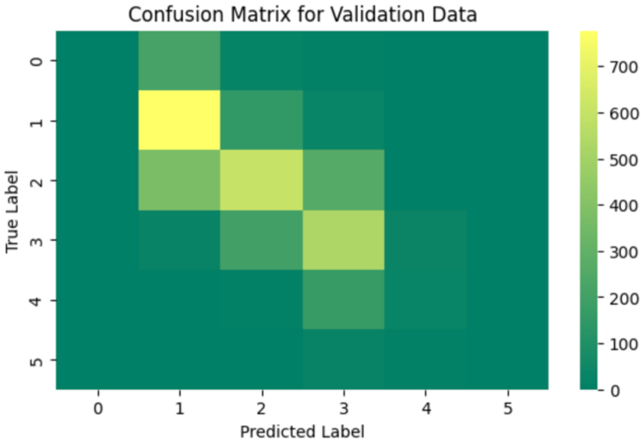


Figure 7. Confusion Matrix for LSTM model

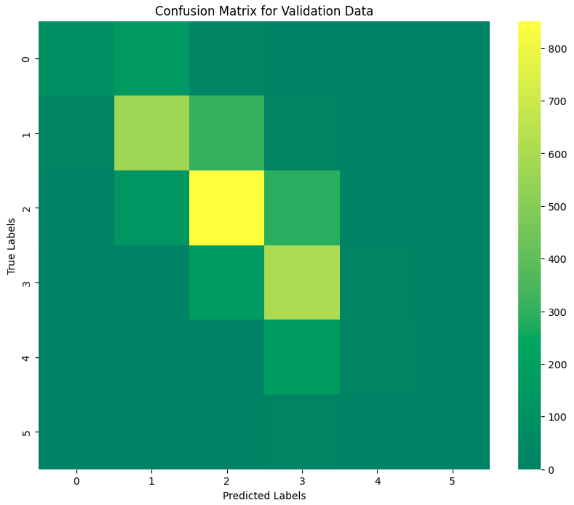


Figure 8. Confusion matrix for DistilBERT model

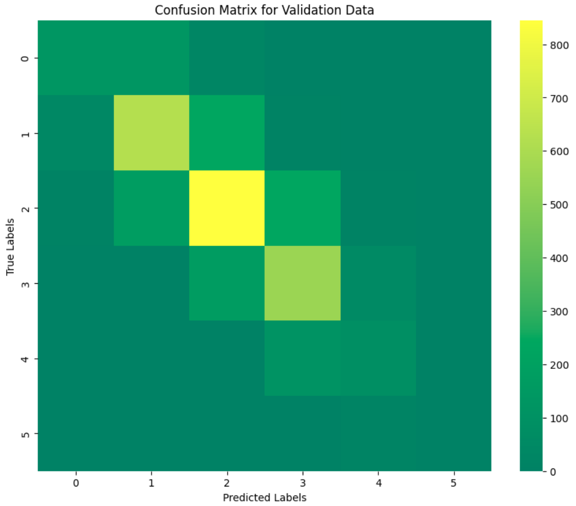


Figure 9. Confusion matrix for DeBERTa model

Figure 7 – 9 displays the confusion matrix result on validation dataset of all the model listed in table 3. The bright square on diagonals indicates that all the models tried in this project are capable of classifying essay scores. It can be found that the models are more accurate in classifying essays with score 2 – 4. That is because the labels are imbalanced, so the more data provided, the more features the model can learn from it. Both LSTM and DistilBERT models are more conservative in identifying essays with score 1, 5, and 6, tending to assign the essays to a value between 2 - 4. The LSTM and DistilBERT models can identify the essays with middled scores.

It is also worth noting that the DeBERTa model performs better on the essays with little data. From the confusion matrix we can see that given an essay scored 1, most of time the DistilBERT model will predict it as score 2. But there is improvement on DeBERTa model, essays labeled as 1 are rated as 1 or 2 with the probability of 50%-50%. Same as those essays with score 6: The DistilBERT model always assign a value of 4 to a scored 6 essay, but the DeBERTa model will assign them with value 5. Since the agreement between 5 and 6 is larger than the agreement between 4 and 6, the quadratic weighted kappa will be larger on DeBERTa model, that is why the DeBERTa model is considered as a better model.

## 8. Conclusions and Future Works

This study indicates that without the need for feature engineering, using deep learning models for essay scoring is possible. Although neural network model's characteristic is more like a black box, deep learning models are able to detect underlying patterns, such as essay structure, grammar, vocabulary, and critical thinking, which play vital roles in determining essay scores, based on the validation results.

In this project, the author explored various models for essay scoring, including one RNN-based LSTM model and two pretrained Transformer-based large language models (DistilBERT and DeBERTa). The validation quadratic weighted kappa showed that Transformer-based models overperform RNN-based models. In addition, by introducing enhanced decoding and disentangled attention to the BERT model can significantly increase the model performance in essay scoring tasks. Finally, the DeBERTa model got the highest quadratic weighted kappa compared to other models and was considered as the optimal one.

However, due to the imbalance in the labels, these models still have difficulty on predicting high-scoring and low-scoring essays. To address this issue, one approach is to do downsampling, and another approach is to acquire extra data for training. Exploring additional datasets from previous AES competitions on Kaggle.com and utilizing them as part of training data should be considered as a further step.

On the other hand, although the pre-trained DeBERTa model is proved to be the optimal choice at the current stage, it is obvious that single large language models do not take manually extracted features into considerations. Noticing the fact that traditional machine learning models from the past have also been successful in essay scoring tasks, manual feature engineering remains to be a valuable part and should not be ignored. By integrating both text data and additional features into the model and applying techniques such as feature union might yield a more ideal result on essay scoring tasks.

## 9. References

Abraham, B., & Nair, M. S. (2019). Automated grading of prostate cancer using convolutional neural network and ordinal class classifier. *Informatics in Medicine Unlocked*, *17*, 100256. https://doi.org/10.1016/j.imu.2019.100256

Aroraaman. (2018, December 30). *Quadratic kappa metric explained in 5 simple steps*. Kaggle. https://www.kaggle.com/code/aroraaman/quadratic-kappa-metric-explained-in-5-simple-steps

Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, *11*(10).

Efimov, V. (2023, October 8). *Large language models: Distilbert-smaller, faster, cheaper and lighter*. Medium. https://towardsdatascience.com/distilbert-11c8810d29fc

Efimov, V. (2023b, November 30). *Large language models: Deberta-decoding-enhanced Bert with disentangled attention*. Medium. https://towardsdatascience.com/large-language-models-deberta-decoding-enhanced-bert-with-disentangled-attention-90016668db4b

He, P., Liu, X., Gao, J., & Chen, W. (2020). *Deberta: Decoding-enhanced bert with disentangled attention*. arXiv preprint arXiv:2006.03654.

Kusuma, J. S., Halim, K., Pranoto, E. J., Kanigoro, B., & Irwansyah, E. (2022). *Automated Essay Scoring Using Machine Learning. 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*. https://doi.org/10.1109/icoris56080.2022.10031338

*Learning agency lab - automated essay scoring 2.0*. Kaggle. (n.d.). https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

Mittal, A. (2021, August 26). *Understanding RNN and LSTM*. Medium. https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e