Leveraging Physical and Semantic Features of text item for Difficulty and Response Time Prediction of USMLE Questions

Gummuluri Venkata Ravi Ram[1], Kesanam Ashinee[2], Anand Kumar M[3]
Department of Information Technology, National Institute of Technology Karnataka
Surathkal, Karnataka, India
[1]toraviram2003@gmail.com, [2]ashineekesanam@gmail.com, [3]m_anandkumar@nitk.edu.in

Abstract

This paper presents our system developed for the Shared Task on Automated Prediction of Item Difficulty and Item Response Time for USMLE questions, organized by the Association for Computational Linguistics (ACL) Special Interest Group for building Educational Applications (BEA SIGEDU). The Shared Task, held as a workshop at the North American Chapter of the Association for Computational Linguistics (NAACL) 2024 conference, aimed to advance the state-of-the-art in predicting item characteristics directly from item text, with implications for the fairness and validity of standardized exams. We compared various methods ranging from BERT for regression to Random forest, Gradient Boosting(GB), Linear Regression, Support Vector Regressor (SVR), k-nearest neighbours (KNN) Regressor, Multi-Layer Perceptron(MLP) to custom-ANN using BioBERT and Word2Vec embeddings and provided inferences on which performed better. This paper also explains the importance of data augmentation to balance the data in order to get better results. We also proposed five hypotheses regarding factors impacting difficulty and response time for a question and also verified it thereby helping researchers to derive meaningful numerical attributes for accurate prediction. We achieved a RSME score of 0.315 for Difficulty prediction and 26.945 for Response Time.

1 Introduction

The automated prediction of item difficulty and item response time is a critical task in the field of educational assessment, with implications for the fairness and validity of standardized exams. Traditionally, item characteristics such as difficulty and response time have been obtained through labor-intensive pretesting processes, posing challenges related to time, cost, and security. To address these challenges, there is a growing interest in leveraging natural language processing (NLP) techniques to predict item characteristics directly from the item text. (Baldwin et al., 2021)

In this paper, we present our system developed for the Shared Task on Automated Prediction of Item Difficulty and Item Response Time, organized by the Association for Computational Linguistics (ACL) Special Interest Group for Building Educational Applications (BEA SIGEDU). The Shared Task was held as a workshop at the North American Chapter of the Association for Computational Linguistics (NAACL) 2024 conference. Our participation in this Shared Task aimed to advance the state-of-the-art in predicting item characteristics and contribute to the ongoing efforts to improve the efficiency and fairness of standardized testing.

In this paper, we provide an overview of our system architecture, including methodologies employed for predicting item difficulty and item response time. We describe the features utilized, the model architectures, and the training procedures. Furthermore, we present the experimental setup, including the dataset used for training and evaluation, data augmentation, as well as the evaluation metrics employed to assess the performance of our system as prescribed by shared task.

Through our participation in the Shared Task, we aim to demonstrate the effectiveness of our approach in predicting item characteristics and contribute to the collective efforts in developing more accurate and efficient models for automated assessment in educational contexts. Additionally, we discuss the implications of our findings and potential future directions for research in this area. We believe that our system holds promise for enhancing the fairness and effectiveness of standardized testing, ultimately benefiting both test developers and test takers alike.

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 534–541

2 Related Work

The paper **"Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory"** by Li-Huai Lin, Tao-Hsing Chang, Fu-Yuan Hsu focuses on utilizing Long Short-Term Memory (LSTM) to predict the difficulty of test items in reading comprehension. Traditional methods of estimating item difficulty relied on expert validation or pretests, which were labor-intensive and costly. By automating the prediction process using LSTM, the study aims to overcome these challenges. Experimental results indicate that the proposed method shows a good prediction agreement rate. The use of LSTM in predicting item difficulty offers a more efficient and accurate approach compared to manual methods, showcasing the potential of machine learning in educational assessment (Štěpánek et al., 2023)

The paper **Question Difficulty Prediction for READING Problems in Standard Tests** by Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y. and Hu, G. initially involves converting the input into embeddings, followed by passing it through a Bidirectional Long Short-Term Memory (BI-LSTM) network to capture semantic relationships. Subsequently, an Attention Layer is employed to identify words within the document or option that hold significant relevance to a given question. This process aids in selecting pertinent information. Finally, the Prediction Layer displays the predicted difficulty scores. (Huang et al., 2017)

Jump-Starting Item Parameters for Adaptive Language Tests by the authors McCarthy, A.D., Yancey, K.P., LaFlair, G.T., Egbert, J., Liao, M. and Settles, B address the challenge of calibrating test item difficulties in high-stakes language assessments, either with limited pilot test data or without any prior information. They propose a multi-task generalized linear model utilizing BERT features to jump-start the estimation of item difficulties. With only 500 test-takers and a small sample of item exposures from a large item bank, their approach rapidly improves the quality of difficulty estimates. This joint model facilitates the comparison of test-taker proficiency, item difficulty, and language proficiency frameworks such as the Common European Framework of Reference (CEFR). Moreover, it allows for the generation of new item difficulty estimates without the need for piloting, reducing item exposure and enhancing test security. The authors validate their method using operational data from the Duolingo English Test, demonstrating strong correlations between the derived difficulty estimates and lexico-grammatical features associated with reading complexity. (McCarthy et al., 2021)

3 Task Description

In recent times, Efforts have been made to enhance the prediction of item parameters for high-stakes medical exams such as the USMLE, have been hampered by challenges in sharing exam data. To address this gap, A Shared Task is proposed focusing on predicting item parameters using practice item content and characteristics from the USMLE Examination. Refer (Yaneva et al., 2024)

The objective of this Shared Task is to advance the state-of-the-art in item parameter prediction, specifically focusing on two tracks:

- **Track 1 - Predicting Item Difficulty:** Given the item text and associated metadata, participants are tasked with predicting the item difficulty variable. Item difficulty represents the proportion of examinees who answer the item/question correctly, providing insights into the relative complexity of the item.
- **Track 2 - Predicting Item Response Time:** Given the item/question text and metadata, participants are challenged to predict the time intensity variable, reflecting the time required by examinees to respond to the item. Understanding item response time aids in optimizing exam administration and identifying potential issues with overly time-consuming items.

4 Dataset Description

The dataset for this Shared Task comprises 466 previously utilized and retired Multiple Choice Questions (MCQs) from the United States Medical Licensing Examination (USMLE) Steps 1, 2 CK, and 3. The USMLE is a series of examinations handled and developed by the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB) to support medical licensure decisions in the United States.

The dataset is structured with the following attributes:

ItemNum: Consecutive number assigned to the item in the dataset.

ItemStem_Text: Textual description of the clinical case or scenario presented in the MCQ stem.

Answer_A to Answer_J: Text for response options A to J. Unused columns remain blank for items with fewer than J response options.
Answer_Key: Letter denoting the correct answer for the item.
Answer_Text: Text corresponding to the correct response for the item.
ItemType: Denotes whether the item contained an image (PIX) or not (Text). Images are not part of the dataset.
EXAM: Indicates the Step of the USMLE exam to which the item belongs (Step 1, Step 2, or Step 3).
Difficulty: Measure of item difficulty where higher values indicate more difficult items.
Response_Time: Mean response time for the item measured in seconds, including initial response and revisits by examinees.

The guidelines for MCQ construction emphasize adherence to a standard structure, avoiding extraneous material, misleading information, and grammatical cues. The items were authored by experienced subject matter experts to assess medical knowledge.

The training data consists of 466 samples. Additonally, to augment the sample dataset, we employed paraphrasing on the provided textual questions (ItemStem_Text) and expanded the training dataset size.

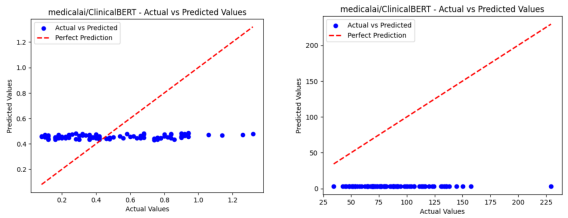
5 Methodology

5.1 Baseline Model

We tried BERT for regression as baseline model. We fine-tuned BERT specifically for regression tasks, utilizing BERT embeddings of the questions. Leveraged the CamembertTokenizer to process the textual descriptions from our dataset.

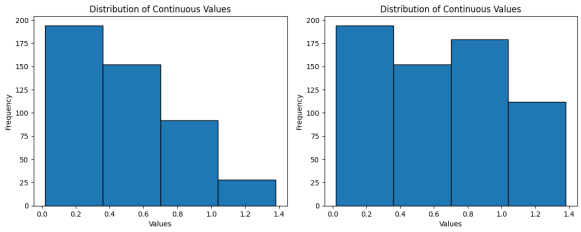
To ensure with BERT's maximum input sequence length of 512 tokens, we set a maximum input sequence length of 300 tokens. Any descriptions exceeding this length were filtered out to avoid truncation, ensuring the integrity of the input data.

The BERT architecture consists of an embedding layer and 12 stacked transformers. Each input sequence yields a sequence of vectors as output, with each vector representing a token in the input. However, for regression tasks, only the final hidden state of the first token, denoted by the "[CLS]" token, is utilized. In line with BERT's architecture, we appended a dense linear layer with dropout after the "[CLS]" token to serve as the final regression layer. This layer facilitates the regression task by mapping the BERT embeddings to the corresponding output labels.



(a) Difficulty (b) Response Time

Figure 1: Predicted v/s True Value plot on validation set on finetuning BERT as regressor



(a) Original Data (b) After augmntation

Figure 2: True Value Distribution in 4 bins before and after data augmentation

5.2 Data Augmentation

The training dataset provided comprises 466 samples. Upon analyzing the distribution of difficulty values, we observed a scarcity of samples with difficulty greater than 0.7. Consequently, we utilized GPT-3.5 LLM to generate additional instances through paraphrasing techniques. Passed on the question samples into the LLM and gave a prompt to paraphrase the given samples. Refer Fig.2a for imbalanced data and Fig.2b for balanced data

5.3 Data Engineering

We propose the following hypotheses based on literature review and reviews from students, based on experience:

- "The readability of a question influences its difficulty and response time" : The tougher the question is to read, the more the student gets confused and hence difficulty and average response time increases.
- "The average length of a question affects response time and subsequently, difficulty" : longer questions take long time to read.

- "The number of options may lead to confusion, potentially increasing difficulty"
- "The average length of options impacts response time and difficulty"
- "The similarity among options influences decision-making, thus affecting difficulty and response time"

Consequently, we extracted these features from the provided dataset. For readability assessment, we utilized the SMOG index (Lin et al., 2019), which is used in educational and medical settings to calculate readability of a document.

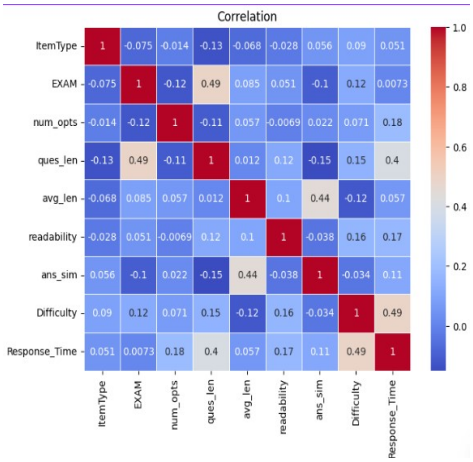


Figure 3: Correlation of extracted features with target variables

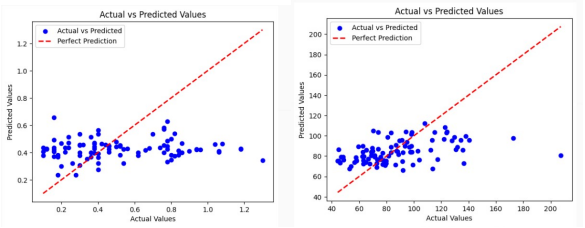
We can see the correlation heat-map as in Fig.3. Clearly the extracted features seem to have good correlation with difficulty and response time, thence justifying the hypotheses.

5.4 Bio-BERT Embeddings

The dataset, being from the medical domain, necessitated the utilization of BioBERT to extract embeddings. We fine-tuned BioBERT specifically based on question-difficulty pairs. The embeddings encapsulate contextual information aligned with the respective difficulty levels. (Yaneva et al., 2019) (Yaneva et al., 2020). In our exploration, we experimented with various methodologies and approaches

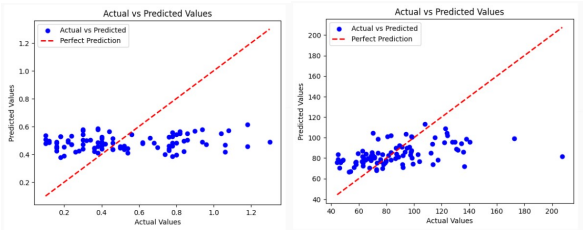
5.5 Approach I - BERT + ANN

We designed 2 distinct Artificial Neural Networks (ANNs) to explore the relationship between the features extracted from the dataset, particularly in the context of question difficulty.



(a) Difficulty (b) Response Time

Figure 4: Predicted v/s True Value plot on Val set for ANN 1 trained on Embeddings + Num Features



(a) Difficulty (b) Response Time

Figure 5: Predicted v/s True Value plot on Val Set for ANN 2 trained on embeddings + Num Features

For the first ANN architecture, we leveraged BioBERT embeddings, which are representations derived from a pre-trained language model specifically tailored for the biomedical domain. These embeddings, comprising vectors of size 768, served as one input stream. Concurrently, we processed seven numerical features independently. These features likely included various attributes such as question length, readability scores, and other relevant metrics. Each stream of inputs traversed through its respective hidden layers before being concatenated at a later stage, in order to capture intricate relationships between textual and numerical features.

The second ANN configuration adopted a different strategy. Here, we fused both the text embeddings obtained from BioBERT and the numerical feature vector derived from the dataset. By concatenating these representations, we aimed to create a unified feature set that encapsulates both textual and numerical attributes of the questions. This combined input was then fed through the hidden layers of the neural network, potentially enabling the model to discern intricate patterns and correlations between the textual content and numerical characteristics of the questions.

By employing these two distinct architectures, we aim to explore and compare the effectiveness of different approaches in utilizing BioBERT embeddings and numerical features to predict question

difficulty levels within the medical domain.

Table 1: Results Of the 2 ANN Models

Target labels	ANN1	ANN2
Difficulty	0.32	0.29
Response Time	26.65	26.11

5.6 Approach 2 - Word2Vec + ML Models

The Deep learning models such as ANNs rely more on larger databases for optimal performance, we've opted for an alternative strategy. Hence we've transitioned to utilizing Word2Vec embeddings, a widely-used technique for generating word embeddings based on the distributional semantics of words within a corpus. Unlike BERT, which thrives on large datasets to capture contextual nuances, Word2Vec offers a computationally efficient means to represent words in a continuous vector space, thereby capturing semantic relationships.

For this, we trained regression models on the Word2Vec embeddings and specifically, we employed the following regression models:

1. Random Forest: An ensemble learning method capable of handling non-linear relationships and high-dimensional data, Random Forest constructs a multitude of decision trees during training and outputs the mean prediction of individual trees.

2. Linear Regression: A regression technique that models the relationship between the dependent variable and one or more independent variables by assuming a linear relationship between them.

3. Support Vector Regression (SVR): A regression algorithm based on the Support Vector Machine (SVM) framework, SVR is adept at handling non-linear relationships by mapping data into a higher-dimensional feature space.

By leveraging Word2Vec embeddings and training on these regression models, we aim to capture the intricate relationships between the textual representations of medical questions and their corresponding difficulty levels. (Yaneva et al., 2021)

Table 2: Word2Vec + ML Model (Linear Regression, SVR, Random Forest Regressor

Target Values	LR	SVR	RFR
Difficulty	0.37	0.356	0.324
Response Time	79.59	86.227	27.24

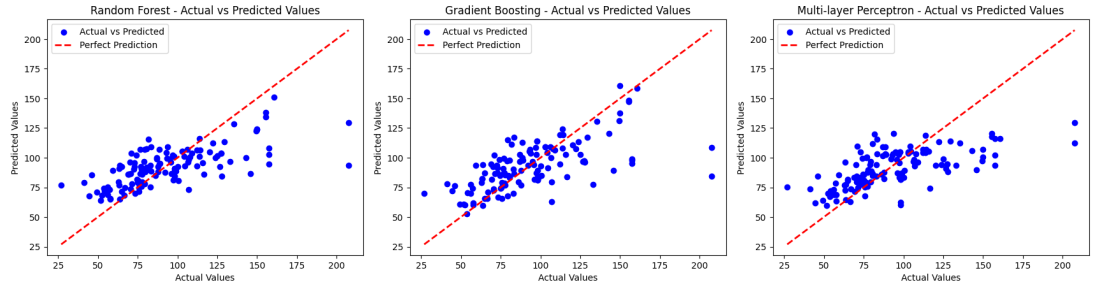
5.7 Approach 3 - BERT + ML Models

We performed experimentation utilizing BioBERT embeddings in three distinct configurations: only with text embeddings, only with numerical features, and with a concatenated dataset combining text embeddings and numerical features. The numerical features encompassed attributes such as average length, readability scores, number of options, average length of options, and similarity scores derived from the dataset. The concatenated dataset combines the text embeddings from BioBERT with the numerical features, aiming to leverage both the textual and quantitative aspects of the data for improved regression performance. Each of these datasets underwent training on a range of regression models, including Random Forest, Gradient Boosting, Linear Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). By systematically exploring various feature combinations and regression algorithms, we aimed to discern the most effective methodologies for predicting the desired output labels.This comprehensive approach enables us to evaluate the performance of various models and feature combinations, thereby gaining insights into the most suitable methodologies for our regression task. (Settles et al., 2020)

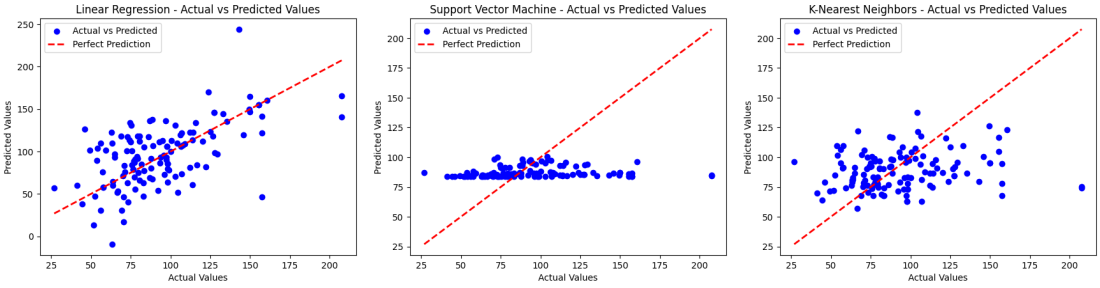
6 Experimental Results and Discussion

Our baseline model, BERT Regressor, achieved RMSE scores of 0.307 for predicting Difficulty and 88.502 for predicting Response Time. These scores demonstrate the model's performance in predicting both the difficulty level of exam items and the time intensity required for examinees to respond to them. Fig.1a and Fig.1b shows that most of predicted values are in a specified range and hence we assumed that the imbalance in data as shown in Fig.2a. Hence we balanced the data. We also extracted few numerical features as discussed in feature engineering section and experimented with them.

Instead of simply finetuning BERT, we trained Bio-BERT embeddings with ANN and results are as shown in table 1. We tried two ANNs whose architecture is as mentioned in methodology section, former concatenating numerical features and text embeddings in a hidden layer and the latter initially concatenating both. As shown in Fig.4a, Fig.4b and Fig.5a, Fig.5b the dispersion of predicted values increased but still not upto the mark. ANN1

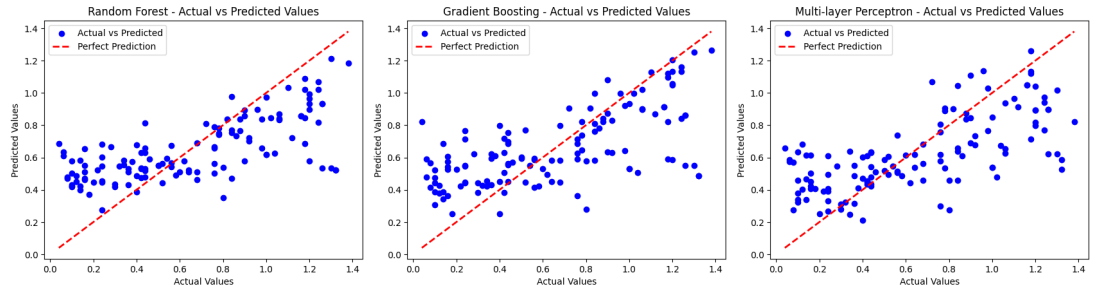


(a) Random Forest Regressor (b) Gradient Boosting (c) Multilayer Perceptron

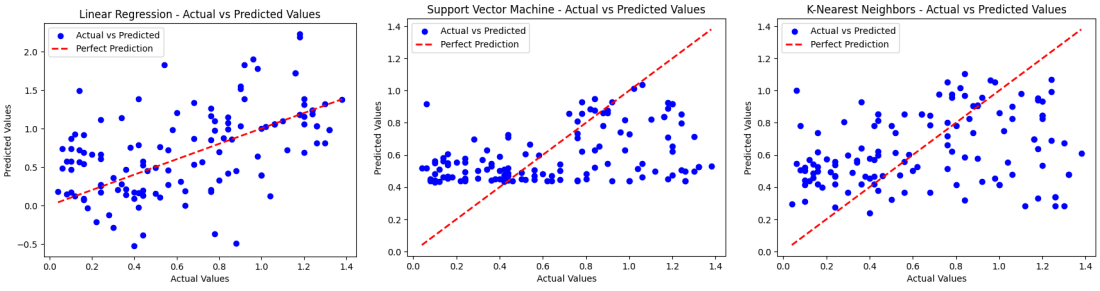


(d) Linear Regression (e) Support Vector Regressor (f) KNN Regressor

Figure 6: Plot of Predicted V/S true labels for validation dataset for **Difficulty** variable upon training ML models on concatenated Input (Text Embedding + numerical features)



(a) Random Forest Regressor (b) Gradient Boosting (c) Multilayer Perceptron



(d) Linear Regression (e) Support Vector Regressor (f) KNN Regressor

Figure 7: Plot of Predicted V/S true labels for validation dataset for **Response Time** variable upon training ML models on concatenated Input (Text Embedding + numerical features)

Target labels	RFR	GB	LR	SVM	MLP	KNN
Difficulty	0.294	0.283	0.480	0.296	0.329	0.293
Response Time	24.029	24.508	28301.258	31.959	25.363	26.918

Table 3: RMSE Scores for 6 models on two target labels: Difficulty and Response Time using only question embeddings

Target labels	RFR	GB	LR	SVM	MLP	KNN
Difficulty	0.346	0.331	0.370	0.363	0.390	0.417
Response Time	28.86	29.37	32.28	32.24	33.21	34.47

Table 4: RMSE Values across 6 models on two target labels: Difficulty and Response Time using only numerical data

Target labels	RFR	GB	LR	SVM	MLP	KNN
Difficulty	0.292	0.297	0.489	0.362	0.39	0.288
Response Time	24.05	24.47	31.48	32.27	33.38	25.05

Table 5: RMSE Values accross 6 models on two target labels: Difficulty and Response Time using concatenated data

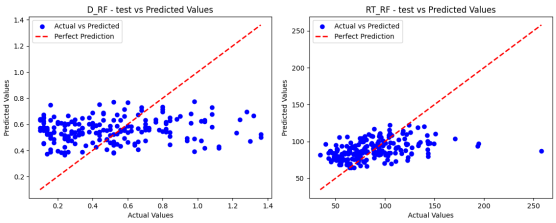
performed better than ANN2 in increasing range of prediction. Hence we assumed it might be due to BERT being a Large Language Model is unable to capture the essence or overall context with such small dataset, and hence shifted to more general model Word2Vec with ML Models as we presume DL models need more data.

Moving on to our 2nd approach consisting of training ML models with Word2Vec embeddings, the results are as in Table 2 . Clearly results are worser when compared to that of training BERT embeddings with ANN.

Hence we considered the issue is in ANN. Since ANN being Deep Learning Model, with such limited data it is unable to capture patterns essentially and hence we tried training ML models with Bio-BERT embeddings. They outperformed ANN, hence we came to conclusion of using ML models for prediction.

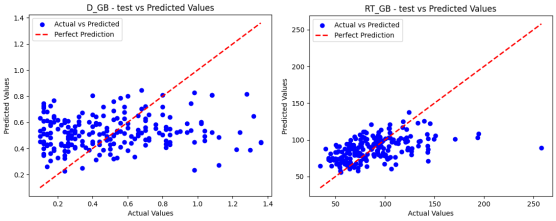
In order to understand importance of extracted numerical features, we used the same ML models to perform regression on only question embeddings and only numerical data and results for each are shown in Table 3 and Table 4 respectively. This clearly states that both text-embeddings and numerical features engineered by our hypotheses are crucial for predicting values.

Hence we concatenated both and trained the ML models to get results as shown in Table 5. Clearly Gradient Boosting, Random Forest Regressor and Multi Layer Perceptron have performed best and hence we considered them to be best models for submission. Fig. 6a - Fig. 6f shows the plots for actual v/s predicted Diffculty values. Fig. 7a - Fig. 7f shows the plots for actual v/s predicted Response Time values



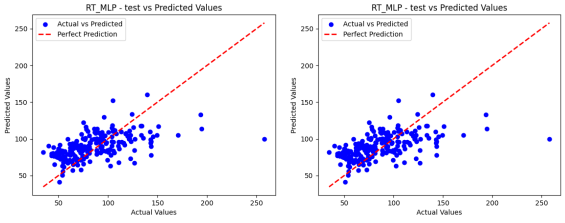
(a) Difficulty (b) Response Time

Figure 8: Actual v/s Predicted value plots for Random Forest Regressor on gold_label test data



(a) Difficulty (b) Response Time

Figure 9: Actual v/s Predicted value plots for Gradient Boosting on gold_label test data



(a) Difficulty (b) Response Time

Figure 10: Actual v/s Predicted value plots for Multi-Layer Perceptron on gold_label test data

Table 6: Test Data Results

Target labels	RFR	GB	MLP
Difficulty	0.315	0.322	0.336
Response Time	28.768	27.481	26.945

7 Error Analysis

The final test data for shared task had 201 data points and Team ScalarLab had made three submissions/prediction files obtained by best three models of which we trained viz. Random Forest Regressor, Gradient Boosting and Multilayer Perceptron trained on concatenated Bio-BERT embeddings and extracted numerical features. The RMSE scores are as reported in Table6. Fig.8, Fig.9 and Fig.10, we clearly can see they outperformed ANN and BERT for regressor.

8 Conclusion and Future Work

We have achieved 0.315 RMSE for difficulty prediction and 26.945 RMSE for response time prediction. We successfully compared and explained why Deep Learning model ANN failed in making better predictions, we discussed the importance of data augmentation and how results improved, and also proposed five hypotheses that seem to impact difficulty, response time of MCQs. As future work, we would like to explore how Deep Learning Models can learn better with limited data and which embeddings are better fir such tasks where limited data is available. We would also explore what are ther factors that impact difficulty and response time of questions (MCQs) and incorporate that info in models to be trained to achieve better RMSE scores.

References

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Li-Huai Lin, Tao-Hsing Chang, and Fu-Yuan Hsu. 2019. Automated prediction of item difficulty in reading comprehension using long short-term memory. In *2019 international conference on asian language processing (ialp)*, pages 132–135. IEEE.

Arya D McCarthy, Kevin P Yancey, Geoffrey T LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 883–899.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.

Lubomír Štěpánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19):4104.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 11–20.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2020. Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6812–6818.

Victoria Yaneva, Daniel Jurich, Le An Ha, and Peter Baldwin. 2021. Using linguistic features to predict the response process complexity associated with answering clinical mcqs. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, An Ha Le, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.