Utilizing Machine Learning to Predict Question Difficulty and Response Time for Enhanced Test Construction

Rishikesh Fulari
Purdue University, Fort Wayne
fularp01@pfw.edu

Jonathan Rusert
Purdue University, Fort Wayne
jrusert@pfw.edu

Abstract

In this paper, we present the details of our contribution to the BEA Shared Task on Automated Prediction of Item Difficulty and Response Time. Participants in this collaborative effort are tasked with developing models to predict the difficulty and response time of multiple-choice items within the medical domain. These items are sourced from the United States Medical Licensing Examination® (USMLE®), a significant medical assessment. In order to achieve this, we experimented with two featurization techniques, one using lingusitic features and the other using embeddings generated by BERT fine-tuned over MS-MARCO dataset. Further, we tried several different machine learning models such as Linear Regression, Decision Trees, KNN and Boosting models such as XGBoost and GBDT. We found that out of all the models we experimented with Random Forest Regressor trained on Linguistic features gave the least root mean squared error, securing fourteenth rank out of 43 for Item Difficulty Prediction and ninth rank out of 34 for Response Time Prediction. We made our code publicly available on GitHub.[1].

1 Introduction

To conduct fair standardized tests for evaluating the learning outcomes of students, it is necessary to design tests that cover variety of questions of all difficulty levels such as ‘easy’, ‘moderate’ and ‘difficult’ ones. Allowed exam time is another component that impacts the difficulty of exam. Allowing ample amount of time to solve the questions can considerably reduce the difficulty whereas providing very little time to solve the exam questions can on other hand, make the exam unreasonably difficult. Thus, the difficulty level of questions and the time taken to solve the questions(response time) are two critical factors to determine the overall difficulty of exam.

Determining the difficulty of items as well as the response time for this task, is a challenge in itself. Conventionally, item difficulty and the response time are gathered through pretesting, where new items are incorporated into live exams alongside scored items. However, this process is labor-intensive and costly, often limiting the number of items that can be created. Furthermore, the reliance on pretesting poses security risks, as items may be copied or leaked due to their repeated usage.

To tackle these challenges, there’s a growing interest in predicting item characteristics such as difficulty and response time directly from the item text. This approach, known as the "cold-start parameter estimation problem" (McCarthy et al., 2021) aims to streamline the process and enhance fairness by reducing the reliance on pretesting. By utilizing predictive models, estimates of item difficulty and response time can be generated, enabling a more efficient parameter estimation process with a smaller sample of test-takers.

In this paper, we examine several approaches which build on predictive machine learning models (for example, linear regression, decision trees) and deep learning models(such as BERT). Our best model for the task of item difficulty achieved RMSE of 0.31 and the best model for the task of predicting response time achieved RMSE of 31.68. We hope that the exploration of models in this paper is able to help future researchers in the evaluation of exams.

2 Related Work

One of the earliest applications of predicting item difficulty emerged in the realm of language testing. Here, a framework was introduced to assess learners’ language proficiency in English, German, or French (Beinborn et al., 2015). Controlling the

[1] https://github.com/rishikeshF/sig-edu-bea-2024-predicting-response-time-and-question-difficulty

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 528–533
June 20, 2024 ©2024 Association for Computational Linguistics

difficulty of tests has also been important for automated generation of MCQ format tests(Alsubait et al., 2013). Another application can be found in context of automated grading where question difficulty estimates guide test creation(Padó, 2017). Thus, predicting item difficulty has been a subject of growing research and with passage of time has extended to high-stakes applications such as medical or clinical exam(Yaneva et al., 2020).

In order to automate difficulty prediction, machine learning and NLP based approaches using word lengths, sentence lengths and tf-idf featurization were proposed (Settles et al., 2020). A further improvement to it can be seen in the form of introduction of linguistic features (Yaneva et al., 2021) which drastically improved the performance of approaches based on using machine learning models.

On similar lines, machine learning and deep learning approaches have been researched upon for predicting the response time (Baldwin et al., 2021). Other techniques employed in this regard include using transfer learning (Xue et al., 2020) and language models such as BERT (Devlin et al., 2019).

3 Experiments

For assessing performance, Root Mean Squared Error (RMSE) serves as the metric for predicting response times and item difficulty in regression tasks. Its suitability stems from the intuitive nature of RMSE, making it well-aligned with the nature of these regression tasks.

3.1 Dataset

The data for both tasks, response time prediction and difficulty prediction, consists of 667 previously used and now retired Multiple Choice Questions (MCQs) from USMLE Steps 1, 2 CK, and 3. The USMLE is a series of examinations (called Steps) to support medical licensure decisions in the United States that is developed by the National Board of Medical Examiners (NBME) (Yaneva et al., 2024) and Federation of State Medical Boards (FSMB). Here is a sample question from USMLE Step 1.

> Q. A 65-year-old woman comes to the physician for a follow-up examination after blood pressure measurements were 175/105 mm Hg and 185/110 mm Hg 1 and 3 weeks ago, respectively. She has well-controlled type 2 diabetes mellitus. Her blood pressure now is 175/110 mm Hg. Physical examination shows no other abnormalities. Antihypertensive therapy is started, but her blood pressure remains elevated at her next visit 3 weeks later. Laboratory studies show increased plasma renin activity; the erythrocyte sedimentation rate and serum electrolytes are within the reference ranges. Angiography shows a high-grade stenosis of the proximal right renal artery; the left renal artery appears normal. Which of the following is the most likely diagnosis?
> (A) Atherosclerosis
> (B) Congenital renal artery hypoplasia
> (C) Fibromuscular dysplasia
> (D) Takayasu arteritis
> (E) Temporal arteritis

The part describing the case is referred to as stem, the correct answer is referred to as key, and the incorrect answer options are known as distractors. All items are MCQs that test medical knowledge and were written by experienced subject matter experts following a set of guidelines, stipulating adherence to a standard structure. These guidelines require avoidance of “window dressing” (extraneous material not needed to answer the item), “red herrings” (information designed to mislead the test-taker), and grammatical cues (e.g., correct answers that are longer or more specific than the other options). The goal of standardizing items in this manner is to produce items that vary in their difficulty and discriminating power due only to differences in the medical content they assess. The items were administered within a standard nine-hour exam. For this shared task, the item characteristic data was derived from first-time examinees from accredited US and Canadian medical schools.

Each item is tagged with the following item characteristics:

- **Item difficulty** A measure of item difficulty where higher values indicate more difficult items.

- **Time intensity** Arithmetic mean response time, measured in seconds, across all examinees who attempted a given item in a live exam. This includes all time spent on the item from the moment it is presented on the screen until the examinee moves to the next item, as well as any revisits.3. Feature engineering

- **ItemNum** denotes the consecutive number of the item in the dataset (e.g., 1,2,3,4,5, etc).
- **ItemStem_Text** contains the text data for the item stem (the part of the item describing the clinical case).
- **Answer_A** contains the text for response option A
- **Answer_B** contains the text for response option B
- **Answer_C** contains the text for response option C.

  (...)
- **Answer_J** contains the text for response option J. For items that have fewer than J response options, the remaining columns are left blank. For example, if an item contains response options A to E, the fields for columns F to J are left blank for that item.
- **Answer_Key** contains the letter of the correct answer for that item.
- **Answer_Text** contains the text of the correct response for the item.
- **ItemType** denotes whether the item contained an image (e.g., an x-ray image, picture of a skin lesion, etc.) or not. The value "Text" denotes text-only items that do not contain images and the value "PIX" denotes items that contain an image. Note that the images are not part of the dataset.
- **EXAM** denotes the Step of the USMLE exam the item belongs to (Step 1, Step 2, or Step 3). For more information on the Steps of the USMLE see https://www.usmle.org/step-exams.
- **Difficulty** contains the item difficulty measure. Higher values indicate more difficult items.
- **Response_Time** contains the mean response time for the item measured in seconds.

The training set comprised of 466 examples and the test set contained 201 items. The combined length of question, multiple choices and the answer was mostly less than 200 words and maxed out at 379. Figure 1 shows a distribution on the number of words in the examples.

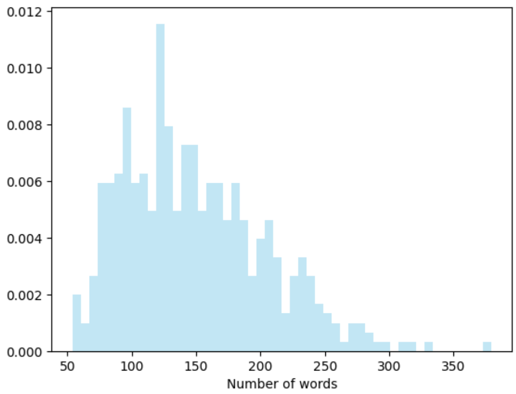

Figure 1: Number of words histogram

4 Methods

4.1 Feature Engineering

For both the tasks, two featuring engineering approaches were tried. First, using embeddings the entire text (comprised of question, multiple choices and corresponding answer) was converted into a 768 dimensional vector. The second approach used linguistic features. The details of both the approaches are as follows:

4.1.1 Embeddings

In order to represent the textual features, sentence Transformer based embeddings were used. The sentence transformer model used is pritamdeka/S-PubMedBert-MS-MARCO (Deka et al., 2022) (from HuggingFace). It maps sentences and paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search. This sentence transformer model has been developed by fine-tuning microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext (Gu et al., 2020) model on MS-MARCO (Nguyen et al., 2016) dataset. It can be used for the information retrieval task in the medical or health domain.

4.1.2 Text based/Linguistic Features

Another method explored for text representation involved leveraging specific linguistic features: word count, number of unique words, number of additives, number of unique additives, number of normalized additives, as well as counts of numbers and letters. Additionally, two additional features, namely 'ItemType' (indicating the presence of a picture) and 'EXAM' (exam level), were incorporated. Each question and word were consequently encoded into a nine-dimensional vector, encompassing these precise linguistic characteristics for subsequent analysis.

4.2 Machine Learning Models

Various machine learning models were explored, encompassing classical approaches like linear regression, Decision Tree Regressor, and K-Nearest Neighbours regressor, as well as advanced techniques such as fine-tuned language models like BERT, simple one-neuron networks, and ensemble methods including random forest regressor. Additionally, boosted models like gradient boosted decision trees regressor and XGBoost regressor were investigated.

4.2.1 Hyperparameters for difficulty prediction

- **Decision Trees** Max-depth: 3
- **KNN** Number of neighbors: 7
- **XGBRegressor** Max-depth: 5
  Number of estimators: 700
- **GBDT**: Max-depth: 3
  Number of estimators: 600
- **Random Forest Regressor**: Max-depth: 3
  Number of estimators: 700

4.2.2 Hyperparameters for response time prediction

- **Decision Trees** Max-depth: 3
- **KNN** Number of neighbors: 7
- **XGBRegressor** Max-depth: 5
  Number of estimators: 600
- **GBDT**: Max-depth: 3
  Number of estimators: 600
- **Random Forest Regressor** Max-depth: 6
  Number of estimators: 800

5 Observations and Results

Table 1 depicts the Root Mean Squared Error obtained for different machine learning models and neural networks along with the corresponding featurization method used.

In our investigation employing various machine learning models and featurization techniques, we found Fine-Tuned BERT to yield consistently stable results, with the lowest RMSE of 0.31 in the task of difficulty prediction. Conversely, our analysis revealed that the Random Forest Regressor, particularly when paired with Linguistic Features, exhibited superior performance in predicting response time with RMSE of 31.68. These results were based on the models trained on a subset of training dataset instead of entire training set, as a smaller subset was used for validation and testing prior to the release of test set.

After training the models on entire training data, the results obtained differed from the previous results. This time, Linguistic features used with Linear Regression gave the lowest RMSE of 0.302 on Difficulty prediction and RMSE of 26.181 on Response Time prediction. These were closely matched by Random Forest Regressor with scores of 0.303 and 26.234 for Difficulty prediction and Response Time prediction respectively. Table 1 displays the RMSE obtained for different models.

A noteworthy observation here is that Linear Regression performed the worst(RMSE of 0.614) with embeddings as features for Difficulty prediction task but performed the best(RMSE of 0.302) when used with Linguistic features, surpassing all other models. This substantial improvement in the performance can be attributed to the fact that total number of input vector size was reduced from 768 dimensions(when used with embeddings) to 9 dimensions(when used with linguistic features), thus eliminating the 'curse of dimensionality'.

Notably, linguistic features, encompassing syntactic aspects such as word count and the presence of additives, emerged as pivotal predictors for response time estimation (Baldwin et al., 2021). Models utilizing embeddings exhibited an average RMSE for the response time task exceeding that of models leveraging linguistic features by 12 seconds. This observation aligns with the intuitive notion that a greater word count in a question correlates with increased time required for student comprehension and analysis, consequently resulting in extended response times. The rationale lies in the fact that candidates typically need more time to read a question with a higher word count, thereby automatically increasing the response time.

6 Conclusion

In conclusion, our research demonstrates the efficacy of machine learning models and feature engineering in addressing key challenges of standardized testing. Linear Regression coupled with lin-

Serial number	Model	Featurization	RMSE for Task 1: Predicting Difficulty	RMSE for Task 2: Predicting Response Time
1.	One neuron network	Embeddings	0.368	32.708
2.	Fine-Tuned BERT	Embeddings	0.321	78.837
3.	Linear Regression	Embeddings	0.614	49.583
4.	Decision Trees	Embeddings	0.320	29.927
5.	KNN	Embeddings	0.332	29.727
6.	XGBoost	Embeddings	0.319	29.657
7.	GBDT	Embeddings	0.32	29.927
8.	Random Forest	Linguistic Features	0.303	26.234
9.	**Linear Regression**	**Linguistic Features**	**0.302**	**26.181**
10.	Decision Trees	Linguistic Features	0.348	28.862
11.	KNN	Linguistic Features	0.324	29.574
12.	XGBoost	Linguistic Features	0.353	28.644
13.	GBDT	Linguistic Features	0.348	28.862

Table 1: RMSE for different models and the corresponding featurization method

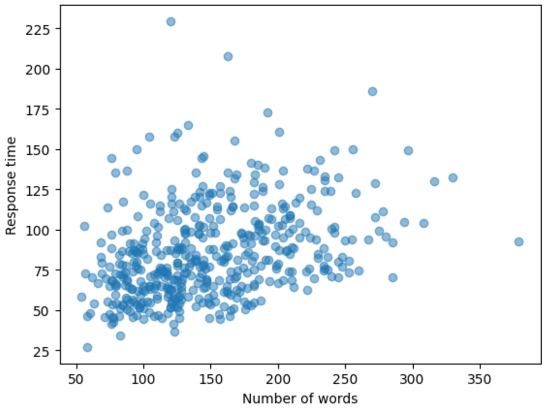

Figure 2: Number of words versus Response time

guistic features gave the lowest RMSE scores of 0.302 and 26.181 for the Difficulty prediction and response time prediction respectively. These findings highlight the potential of predictive models to streamline assessment processes and improve fairness. By reducing reliance on labor-intensive pretesting, our approach offers a scalable alternative while ensuring the integrity of assessment materials. Future research should explore additional techniques and validate findings across diverse educational contexts. Overall, our work advances educational assessment by offering innovative solutions to test design challenges.

References

Tahani Alsubait, Bijan Parsia, and Uli Sattler. 2013. A similarity-based theory of controlling mcq difficulty. pages 283–288.

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.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Pritam Deka, Anna Jurek-Loughrey, and P Deepak. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Ulrike Padó. 2017. Question difficulty – how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online. Association for Computational Linguistics.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 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, Marseille, France. European Language Resources Association.

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. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–232, Online. 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.