

**Improving Automated Assessment for Student Open-responses in  
Mathematics**

by

Sami Baral

A Thesis Proposal  
Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the  
Degree of Master of Science

in

Computer Science  
April, 2022

APPROVED:

---

Professor Neil T. Heffernan, Master's Thesis Advisor

---

Professor George Heineman, Thesis Reader

---

Professor Craig E. Wills, Head of Department

## **Abstract**

Open-ended questions in mathematics are commonly used by teachers to monitor and assess students' deeper conceptual understanding of content. Student answers to these types of questions often exhibit a combination of language, drawn diagrams and tables, and mathematical formulas and expressions that supply teachers with insight into the processes and strategies adopted by students in formulating their responses. While these student responses help to inform teachers about their students' progress and understanding, the amount of variation in these responses can make it difficult and time-consuming for teachers to manually read, assess, and provide feedback on student work. For this reason, there has been a growing body of research in developing AI-powered tools to support teachers in this task. This work seeks to build upon the prior work that presents a model designed to help automate the assessment of student responses to open-ended questions in mathematics through sentence-level semantic representations. We conduct an error analysis of this model, to examine characteristics of student responses that may be considered to further improve the method. We find that this model performs poorly in presence of mathematical terms and images in student responses. We then introduce a model as a step toward the improvement of this method in presence of mathematical terms and we find that this new model outperforms the previously published benchmarks across three different metrics.

## **Acknowledgements**

I would like to express my gratitude to my advisor, Prof. Neil T. Heffernan for his advice and encouragement. I would like to thank my mentor Prof. Anthony F. Botelho for his continuous support throughout this work. I also want to thank the contributors to this research (John, Priya, Karthik and Anzhuo). I would like to thank my reader Prof. George Heineman for his feedback. Finally, I would like to thank my family and friends for the love and support throughout the journey.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>6</b>
2.1	Related Works . . . . .	6
2.2	A Benchmark Comparison . . . . .	7
<b>3</b>	<b>Methodology</b>	<b>10</b>
3.1	Dataset . . . . .	11
3.2	SBERT-Canberra Model . . . . .	13
3.3	Evaluation of SBERT . . . . .	14
3.4	Approach to Error Analysis of the SBERT-Canberra Method . . . . .	17
3.4.1	Uni-level Linear Model . . . . .	17
3.4.2	Multi-level Linear Model . . . . .	18
<b>4</b>	<b>Results</b>	<b>20</b>
4.1	SBERT Model . . . . .	20
4.2	Error Analysis of SBERT . . . . .	21
<b>5</b>	<b>SBERT-MTF Model</b>	<b>24</b>
5.1	Model Design . . . . .	26
5.2	Model Performance . . . . .	27

<b>6</b>	<b>Limitations and Future Works</b>	<b>28</b>
<b>7</b>	<b>Conclusion</b>	<b>31</b>
<b>8</b>	<b>Acknowledgement</b>	<b>33</b>

# List of Figures

1.1	The number of open response problems assigned over the course of a school year with the ASSISTments learning platform, aggregated from 2010-2020. . . . .	4
1.2	The percent of student open-response answers that were scored and given written feedback by a teacher before and during remote learning in response to COVID-19. . . . .	5
2.1	Example of an open ended question taken from openupresources.org . . . .	8
3.1	The design of the SBERT-Canberra method, that suggests scores based on similarity between the answers. . . . .	15
5.1	The design of the ensembled SBERT-MTF method, that suggests scores for student open responses. . . . .	25

# List of Tables

2.1	Sample student responses (selected from across multiple problems for illustrative purposes) and the teacher provided scores on a scale of 0 to 4 to the open ended questions in mathematics. . . . .	9
3.1	Features for the Linear Model of Error analysis of SBERT-Canberra model	13
3.2	Rasch Model Performance compared to the models developed in Erickson et al.[30] . . . . .	16
4.1	The resulting model coefficients for the uni-level linear regression model and random and fixed effects of the multi-level linear model of absolute error. . . . .	22
5.1	Rasch Model Performance comparing the performance of the proposed SBERT-MTF model with the observed SBERT-Canberra model . . . . .	27

# Chapter 1

## Introduction

In the past decade, development in artificial intelligence and machine learning methods have led to advancements in online learning platforms, transforming learning experiences and teaching practices. From personalized learning to augmenting teaching processes through automated assessment methods [4], [8], [12], [13], [27], [33], the benefits of these platforms have been significant. With this, several prior works have leveraged machine learning methods and natural language processing-based techniques to automate the assessment of students' work, both in mathematical [20], [30] and non-mathematical domains [23], [24]. With these methods and models of student learning becoming deeply integrated into normal instructional and educational practices, it is important to understand the strengths and weaknesses in their application. Within this, it is important to not only identify areas where existing methods under-perform, but it is also important to develop methods to improve such models to alleviate risks to fairness.

In many K-12 mathematics classrooms, teachers have come to rely on the use of open-ended questions to assess their students' knowledge and understanding of assigned content. These questions are designed such that teachers can learn about their students' understandings and misunderstandings. Unlike close-ended problems, where there is a single



or finite-number of accepted answers (e.g. a multiple-choice question), open-ended questions allow students to justify and express their thinking processes through language; it is common that students may combine language, images, drawn diagrams, tables, or other mathematical expressions, equations, and terminologies to illustrate their knowledge and understanding of the material.

The use of open-ended questions is not found only in mathematical contexts, but the aspects of this domain make it particularly difficult to develop teacher supports for these types of question. Within computer-based learning platforms, research across fields of study have led to the development of a multitude of teacher-augmentation tools [29] and methodologies that leverage machine learning techniques. Among these supports, automated methods have been developed and deployed to help teachers assess student essays and short answers in several domains [6], [14], [17], [24]. As was highlighted in [30], the arduous task of manually assessing and providing feedback to student open-ended work may explain the decline of open-ended questions assigned over the course of a school year (e.g. Figure 1.1 which shows the number of open response questions assigned within the ASSISTments learning platform, aggregated over the last 10 years). In addition to this decline, as was also reported in [30], very few student responses to open-ended questions are ever scored by the teacher, with even fewer ever receiving feedback. Figure 1.2 illustrates this, as well as the subsequent plot of these values from February through October of 2020, during COVID-19 induced remote learning.

There are several notable challenges in developing automated supports to help teachers assess student open-ended work. In the case of close-ended problems, where there is a finite number of accepted correct answers, auto-scoring methods can apply simple matching techniques to compare the student answer with the list of correct answers and consistently achieve near-perfect accuracy. In regard to open-ended problems, however, the correctness of student responses is more subjective, where teachers commonly assess

students based on an explicit or implicit rubric that identifies key points that must be included in a student response to sufficiently demonstrate comprehension. It is also the case that student responses to open-ended questions differ in the context of mathematical and non-mathematical domains. One such difference, for example, is that many non-mathematical domains such as history or language arts, student “open-ended” essays and short answers are often comprised of multiple sentences and paragraphs [9], [15], [23], [24], whereas in mathematics, responses are generally shorter (maybe one or two, often incomplete sentences) [20], [30] that combine language with mathematical symbols, expressions, or other visuals. Aside from these response-level characteristics, however, several other student-, problem-, and even teacher-level factors can make the development of these automated supports more challenging; consider, for example, the variation in how teachers approach the assessment of student answers, using different inherent rubrics and pedagogical philosophies [3], [6], [7], [10].

While the examination of student answers to open-ended poses challenges in developing automated assessment supports for teachers, prior work has shown promise in this context [30]. In that work, the authors explore several machine-learning and natural language processing (NLP) methods to predict teacher-provided scores to open-ended problems, offering an evaluation method and benchmark of comparison for similar methods.

In this research <sup>1</sup>, we build upon prior research presented in [30] to develop and evaluate an automated assessment model of student open-responses in mathematics. We observe a model to automatically score student open-responses in mathematics that is based on sentence-level semantic representation of the student responses (Sentence BERT or SBERT;[28]), which we call “SBERT-Canberra” method. We evaluate this method based on the dataset and evaluation methods presented in [30], and find that this data method

---

<sup>1</sup>It is important to acknowledge that this work was published at the International Educational Data Mining Society(EDM) 2021[31] and the International Conference on Artificial Intelligence in Education(AIED) 2022[32]

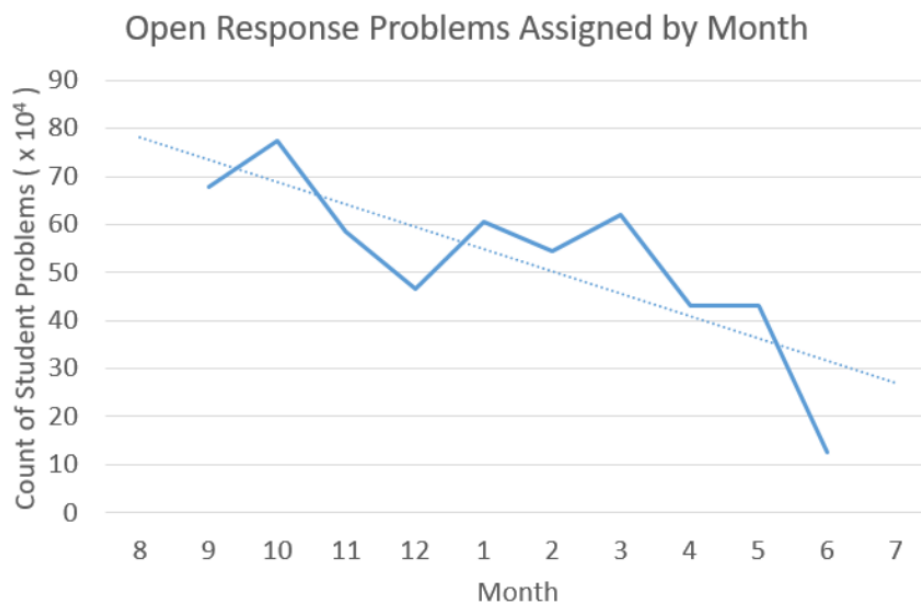


Figure 1.1: The number of open response problems assigned over the course of a school year with the ASSISTments learning platform, aggregated from 2010-2020.

outperforms the prior works on automated scoring from Erickson et al. [30]. Further, with the goal of exploring the limitations and further improving the SBERT-Canberra method for assessment, we perform an exploratory error analysis to identify the areas of improvement that may be addressed by future iterations of these methods. We identify that the existence of non-linguistic terms is positively correlated with model prediction error. Following these findings, we then propose a simple targeted method to resolve this problem, and demonstrate how this method can be combined with the SBERT-Canberra model to improve the performance. Towards this, we specifically seek to address the following research questions:

1. How does a model utilizing Sentence-BERT compare to previously developed approaches in predicting teacher given assessment scores for student response to open-ended problems?
2. What are the characteristics of student answers that correlate with errors observed

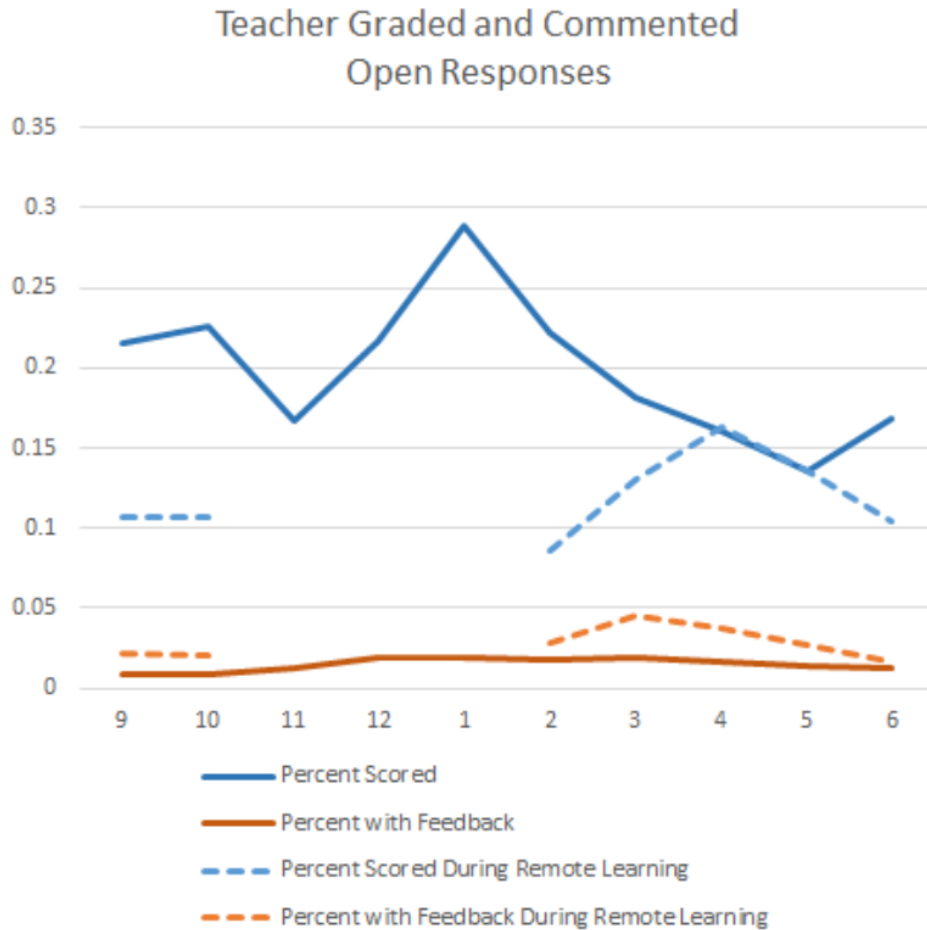


Figure 1.2: The percent of student open-response answers that were scored and given written feedback by a teacher before and during remote learning in response to COVID-19.

in our Sentence-BERT model?

3. Which of student-, problem-, or teacher-level characteristics most explain the variance of error observed when the model is applied in real learning environments?
4. How does accounting for non-linguistic terms in the prior Sentence-BERT based method through the proposed model affect the performance of the auto-assessment methods on existing benchmarks?

# Chapter 2

## Background

### 2.1 Related Works

There have been several works related to the automated scoring of open-ended responses in the past. Most of such works utilize a combination of Natural language Processing (NLP) and machine learning techniques of ranging complexity to process open-ended responses. Much of the existing work in this area has been applied in the context of non-mathematical content. Developments such as C-rater[6] is a well-cited approach that uses such methodologies to estimate the assessed correctness of answers to short answer questions. This method uses grading rubrics and breaks down scores into multiple knowledge components to evaluate each student response. Other works [14], [17] have implemented clustering techniques to grade short textual answers to questions. More recently, studies have based their approach around deep learning methods, which have led to promising improvements over previous benchmarked results [23], [24]. While most of these works have been on non-mathematical domains, studies like [20] explore mathematical language processing using clustering techniques and the bag-of-words approaches for automated assessment of open-ended response in mathematics. However, this study only considers

the mathematical content, discarding the non mathematical texts.

Many of these more-recent studies have utilized publicly-released embedding methods trained on large corpuses of data, including those of Word2Vec [16] and GloVe [21], to model the semantic meaning of words. However, word embeddings capture limited information about the semantics of a sentence, where the sequence of words may have large impacts on interpreted meaning. To capture the contextual information within sentences and further increase the generalization capabilities of NLP embedding methods, techniques such as Universal Sentence Encoders [25] and Sentence-BERT [28] generate a single embedding that is designed to be representative of the entire sentence while preserving the semantic and contextual information of the words within such sentences.

One of the most commonly-used NLP embedding methods in recent years has been that of Bidirectional Encoder Representations from Transformers (BERT, [26]). Building upon and distinguishing itself from other methods such as GloVe, the BERT method is designed to incorporate contextual information into generated embeddings to distinguish words that may have the same spelling but different meanings depending on usage (e.g. the word “bank” referring either a financial institution or perhaps a slope of land near a river); BERT has been shown to outperform many other approaches in a number of NLP tasks including, as is important for this work, semantic textual similarity (STS) [26]. Sentence-BERT, or SBERT [28], modifies the pre-trained BERT network to reduce the computational overhead of BERT in order to also generate a sentence-level embedding of a given series of words.

## **2.2 A Benchmark Comparison**

In this work, we are exploring the use of this SBERT method to build upon the prior benchmark set in Erickson et al., 2020 ([30]) in assessing student answers to open-ended

Elena, Lin, and Noah all found the area of Triangle Q to be 14 square units but reasoned about it differently, as shown in the diagrams. Explain *at least one* student's way of thinking and why his or her answer is correct.

copied for free from [openupresources.org](https://openupresources.org)

Type your answer below:

Figure 2.1: Example of an open ended question taken from [openupresources.org](https://openupresources.org)

problems in mathematics. In that work, the authors discuss the challenges in developing models to predict teacher assigned grades for student open responses in mathematics, using a dataset of authentic student responses within the ASSISTments [18] learning platform. Erickson et al. compares 6 models utilizing machine learning (e.g. random forest and XGBoost [22]) and more complex deep learning (e.g. LSTMs [2]) techniques, combined with natural language processing algorithms to assess responses that are combinations of mathematical expressions and non-mathematical text. For the feature extraction process from the open response data, the study uses the Stanford Tokenizer [19] combined with Global Vectors for Word Representation (GloVe) [21].

Table 2.1: Sample student responses (selected from across multiple problems for illustrative purposes) and the teacher provided scores on a scale of 0 to 4 to the open ended questions in mathematics.

Sample Responses	Score
$y=4x-2$	4
I counted	4
I multiply -3 and 2x	2
diagram is on paper	3
Yes Because $Y=mx+b$	0
I got 2/9 by dividing by 4	3
I was not in class for this so I don't know.	1
I went multiplication first then division then multiplication	3
I got this by doing 45/75. I knew that $75 + 75 = 150$ and 150 goes into 450 3 times and $3 \times 2 = 6$ . So the answer is 6.	4
You would need an example and then you would need to draw a line and find out far away your shape is from the line and mark it and then do that on the rest of your lines on the shape	4
The distributive property means that a number outside a set of parentheses can be multiplied by each of the numbers within the parentheses and the answer will be the same. It works because it would be the same as multiplying each number by the number outside the parentheses and then adding them together.	1



# Chapter 3

## Methodology

In this study, we build upon the work of [30] and present an automated scoring method based on the SBERT methodology; as will be detailed further, we refer to this model as the SBERT-Canberra model throughout the remainder of this work. We evaluate this method and compare it to the prior methods in automated scoring from [30]. Then, in a secondary analysis, we utilize real data collected from a pilot study of our model running within a computer-based tool that provides teachers with suggested scores to explore the limitations of our approach through an exploratory error analysis. Finally based on the findings from the analysis of the SBERT-Canberra model, this work attempts to improve the method by developing a simple, targeted method, drawing inspiration from closed-ended assessment methods. We call this proposed method the “Math Term Frequency” (MTF) model and demonstrate how it can be combined with the SBERT-Canberra model to improve its performance. Our data and approach to these analyses are described in this section.

## 3.1 Dataset

In this work, we utilize two datasets of student answers to open-ended questions, paired with teacher-provided assessment scores. An example of one of these open-ended mathematics questions is shown in Figure 2.1. In this example, students are not asked to find the area of the triangles, but rather explain in their own words what one of the figures is illustrating an approach to solving the problem.

For the development of our SBERT-Canberra model, we use the dataset (and evaluation code) from the Erickson et al. study [30]. This dataset is comprised of student answers to open response questions within the ASSISTments[18] online learning platform; the dataset consists of 150,477 total student responses from 27,199 unique students to 2,076 unique problems graded by 970 unique teachers. As was performed in [30], we omit any case where a student response contained no characters (e.g. an empty response or one containing only whitespace characters), or contained nothing but an image (cases where there was an image accompanied by other text or non-whitespace characters is not omitted). The removal of such empty responses resulted in the dataset dropping to 141,612 graded student responses, 25,069 unique students, 2,042 unique problems, and 891 unique teachers. Within this data, each response is accompanied by a teacher-provided assessment score that follows an integer ordinal 5-point scale from 0-4; a “4” here is synonymous with a student receiving a 100% for the response.

Table 2.1 lists several student answers contained within the dataset, chosen from across multiple problems for illustrative purposes. As was noted in the introduction, these responses highlight some of the challenges of this modeling task. First, the length of responses varies greatly between students as well as across problems. In addition to this, the interleaving of mathematics and linguistic text likely makes it difficult for pre-trained embedding models to interpret. Similarly, the variation in mathematical representation

(i.e. the use of the term “dividing” rather than the “/” operator), may lead to confusion in a machine learning model trained over such data. As the mathematical variables are also represented by recognized english characters (e.g. “y”), it may be difficult to derive semantic meaning for such tokens. It is for this reason that we hypothesize that a contextual-based embedding approach, such as BERT and SBERT, may be superior to traditional embedding methods that do not account for context within the sentence. Finally, the noise in ground truth labels become evident from the table. The student who answered “I counted” but still received full credit, for example, exemplifies that some teachers may score students based on completion or other factors unrelated to their demonstration of understanding or mastery. This is not to say that any one scoring method is more correct or valid than another, but rather that there is likely large variation in these labels, making it difficult for machine learning models to effectively learn associations between student answers and these scores in some cases.

The second dataset used in this work is comprised of student responses collected during the pilot testing of a teacher-augmentation tool designed to aid in the assessment of student open response answers within ASSISTments [18]. This tool, called QUICK-Comments, used our developed model to predict the scores of student answers to open response questions in mathematics. Models were trained over the same open educational resource (OER) curricula from which the problems used in the first dataset were collected and produce estimates using the same grading scale as the first study. During the pilot study, 12 middle school mathematics teachers were given access to the tool and compensated for their time to assign, assess, and provide feedback to student open ended work during the Spring and Fall of 2020. This dataset consists of 30,371 graded student open responses to 915 unique open response problems solved by 1,628 unique students.

Table 3.1: Features for the Linear Model of Error analysis of SBERT-Canberra model

Title	Description	Mean
Answer Length	Length of the answer	10.39
Average character per word	the average number of characters per words	3.54
Numbers count	total number of digits	3.54
Operators count	total mathematical symbols in the response	1.47
Equation percent	percentage of mathematical equations in answer	0.27
Presence of Images	Indicator of presence of images in the answer	0.15

## 3.2 SBERT-Canberra Model

The model developed for this work follows a 2-stage process to generate estimates of teacher-assigned scores for a set of given student answers. In approaching this model, we propose a reframing of the initial problem. In [30], the problem was posed as a traditional supervised learning problem; in other words, given a set of student answers  $A$ , train a model  $f(\cdot)$  such that  $Y = f(A)$ . Instead, we propose a more unsupervised approach as depicted in Figure 3.1. If we have a set of historic answers  $A_{0\dots n-1}$ , and want to predict the score of a new answer  $A_n$ , a logical choice of score may be that corresponding with the historic answer that is most similar to the new answer  $A_n$ . In this way, the problem is posed as a similarity ranking problem rather than a supervised learning problem.

There are several potential advantages to this approach. First, when utilizing a pre-trained model of SBERT, described in Section 2, no actual model training is necessary (so long as a reasonable distance metric is identified). Second, as SBERT is optimized for contextual similarity tasks, the problem is better suited to utilize the embedding method’s strengths. Finally, in a practical sense, as no model training is necessary (beyond utilizing the pre-trained embedding model), such a model can be more easily applied at scale, requiring just a pool of historic answers to compare against. We hypothesize that this method may also require fewer example answers than traditional machine learning methods as well, but this claim is not deeply explored in this current work.

In applying this method, the set of historic answers  $A_{0\dots n-1}$  are fed through the pre-trained SBERT model to produce a 768-valued feature vector for each answer; these vectors are then stored for later access.. Given a new answer,  $A_n$ , a feature vector is similarly produced. In stage two of our method, all pairwise comparisons are then made between  $A_n$  and  $A_{0\dots n-1}$ , calculating Canberra distance [11] for each pair. Canberra distance, as opposed to other common distance metrics such as Euclidean or Cosine similarity, is a distance metric calculated over ranked lists. With this metric calculated for all pairs, the  $A_{0\dots n-1}$  historic answers are then min-sorted to identify the most similar historic answer,  $A_s$ , to our new answer  $A_n$ . The score associated with  $A_s$  is then used as the prediction for the given answer  $A_n$ . The design to this approach is outlined in Figure 3.1.

As an additional component of this model, a “fallback” condition is implemented to be able to produce scoring estimates for problems where there are no historic answers on which to compare. In this case, we train a single multinomial regression model over all known answers, utilizing 1) the number of words in the answer and 2) the average length of each word in the answer; this model produces a probability distribution over 5 categorical labels (observing the 0-4 grading scale as a multinomial regression formulation). This one model is trained over all known answers and used then only in the case that no historic answers are available for the SBERT-Canberra model. This component is viewed as being part of our SBERT-Canberra approach.

### **3.3 Evaluation of SBERT**

To evaluate our SBERT-Canberra scoring method, we utilize the same data and code presented in [30]. In that paper, the authors present the usage of a 2-parameter rasch model [1] (equivalent to a traditional item response theory, or IRT, model). The purpose of this model is to learn a separate parameter for each student and problem presented, represent-

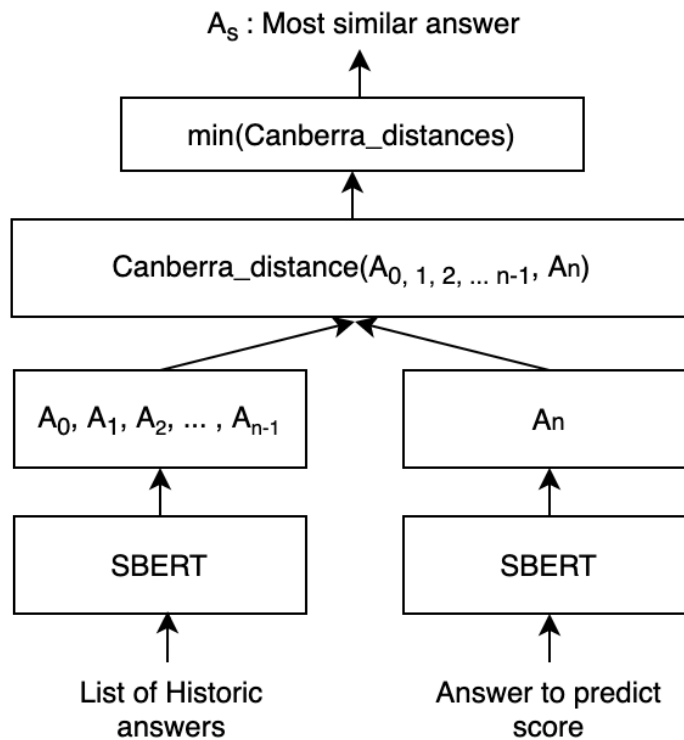


Figure 3.1: The design of the SBERT-Canberra method, that suggests scores based on similarity between the answers.

ing student ability and problem difficulty, respectively. The intuition behind the use of this model is to evaluate an NLP automated scoring model based solely on its ability to interpret the words in each student answer. As the score of each answer is likely correlated with student ability (or knowledge) and problem difficulty (e.g. easy problems are likely to exhibit higher scores), such a model provides a reasonable minimum baseline of comparison. By adding a model’s scoring estimates as covariates to the rasch model and then comparing the performance of such a model to the rasch model without covariates, we are able to observe the true value-added performance of the NLP scoring model.

Following the same procedure as conducted in [30], we are able to directly compare

Table 3.2: Rasch Model Performance compared to the models developed in Erickson et al.[30]

Model	AUC	RMSE	Kappa
Current Work			
<b>Rasch* + SBERT-Canberra</b>	<b>0.856</b>	<b>0.577</b>	<b>0.476</b>
Erickson et al. 2020			
Baseline Rasch	0.827	0.709	0.370
Rasch + Number of Words	0.829	0.696	0.382
Rasch* + Random Forest	0.850	0.615	0.430
Rasch* + XGBoost	0.832	0.679	0.390
Rasch* + LSTM	0.841	0.637	0.415

\*These rasch models also included the number of words.

our Sentence-BERT method to those presented in that prior work. The models are trained and evaluated using a 10-fold student-level cross validation, and model performance is compared based on 3 performance metrics. First, treating the label as multinomial, rather than ordinal, AUC is calculated using the method described in [5]. Second, the root mean squared error (RMSE), is calculated over the ordinal prediction and label. Finally, a multi-class kappa is calculated, again using the multinomial label representation. The multinomial representations were argued to be appropriate due to the likely non-linear distribution of scores, while then RMSE provides insight into a more linear assumption of the data. Arguably an additional rank-based metric such as Spearman’s Rho would also be a suitable metric of comparison, but is not included for more direct comparisons to the previous work.

## 3.4 Approach to Error Analysis of the SBERT-Canberra Method

In evaluating the SBERT-Canberra method, it is important to explore limitations of the approach in order to identify where the model does well and where it may yet improve through future iteration. As such, we also conduct an exploratory error analysis of the method using the data collected from the QUICK-Comments pilot study. Toward this, we observe two regression models that observe absolute model error as a dependent variable. By exploring characteristics of student answers in the context of this modeling error, we can observe which aspects correlate most with higher prediction error. Similarly, we apply then a multi-level model to observe which of student-, problem-, and teacher-level identifiers most explains any observed modeling error.

### 3.4.1 Uni-level Linear Model

The uni-level linear model is based on student answer level characteristics. The student answer level characteristics are comprised of a set of six answer-level features extracted from the student open response data. These features are listed in Table 3.1. In calculating these features, the answer is first tokenized using the Stanford NLP tokenizer[19], dividing each textual answer into smaller tokens. For example, if the response to a particular problem is “I got 2/9 by dividing by 4”, a simple tokenizer splits this response text by spaces which would give the list of tokens as: (“I”, “got”, “2/9”, “by”, “dividing”, “by”, “4”). Then from the tokenized data, we separate the tokens consisting of either digits or mathematical symbols. The number of such tokens is divided by the total number of tokens to calculate the equation percentage<sup>1</sup>. The average equation percentage calculated

---

<sup>1</sup>We acknowledge that this feature is a misnomer as it includes numeric terms, operators, and expressions as well as equations, but chose this feature name for sake of brevity.



by the procedure mentioned above is 27% across the entire dataset. For calculating the length of the answer text, we count the total words in the text simply by splitting them by space. The average length of answers across the dataset is 10.39. Similarly, within each response, the number of numeric digits (i.e. Numbers count) and number of operator characters (i.e. Operators count) are counted independent of the tokens.

ASSISTments[18] allows students to upload images as part of the response to open-ended questions; this is most commonly a picture taken of work done on paper. The response text in such cases includes the URL of the uploaded image to the system. About 15% of the total responses in the dataset contains images. Some of such responses are entirely images, whereas in others, some text is provided as context. Since these scoring models are not yet designed to support images, we hypothesize that the images' presence contributes significantly to the modeling error.

A simple linear regression model is fit to the pilot study student answers, observing absolute model error as the dependent variable. This value is calculated by simply subtracting the predicted score from the teacher-provided label (as a linear label), and taking the absolute value. In this case, a value of 0 would indicate a correct estimate, while higher values (up to 4) represent greater prediction error; we do not differentiate between under- and over-predicting in this analysis.

### **3.4.2 Multi-level Linear Model**

The uni-level linear model observes features that describe characteristics of the student responses, but as described in Section 3.1, modeling error may not be confined to just characteristics of the responses themselves. It is very likely that modeling error can be attributable to other external factors at the student-, problem-, and teacher-levels.

To explore this possibility, we apply a multi-level linear model observing the student answer characteristics as fixed effects, and student, problem, and teacher identifiers as

three separate level-2 random effect variables. As it is the case that the same student may write multiple answers within our data, this structure is similar to that of a repeated-measure analysis.

$$\begin{aligned} abs(\text{model error}) = & \text{Answer Covariates} + (1|\text{student identifier}) \\ & + (1|\text{problem identifier}) \\ & + (1|\text{teacher identifier}) \end{aligned} \tag{3.1}$$

Again observing absolute prediction error as the dependent variable, this analysis will be able to answer 1) whether the majority of explainable variance exists at the student-answer level or at a higher level, and 2) which of student-, problem-, and teacher-level identifiers most explains variance in our modeling error (e.g. which of these identifiers is most correlated with the error). The equation, expressed as its R code formulation, is reported as Equation 3.1.

# Chapter 4

## Results

### 4.1 SBERT Model

The results of the SBERT model is compared directly to the results from Erickson et al.[30] as shown in Table 3.2. As can be seen in that table, the SBERT-Canberra method outperformed the baseline as well as all previous models across all three metrics. While the difference in AUC values between our method and the previous best approach is notably small, the difference in both RMSE and Kappa appears to be comparatively larger. To interpret these two metrics, these results suggest that we should expect teachers to agree with our method's estimates 47% of the time accounting for random chance, and is likely to be wrong by just over half a grade-point on average. This also does suggest, however, that there is still plenty of room for improvement of these models.

What is also worth noting from the results of Erickson et al. [30], is the high performance of the baseline rasch model. This emphasizes the difficulty of this NLP modeling task in that the baseline model is using nothing other than the student and problem identifiers; it is able to seemingly predict teacher-provided scores with an AUC above 0.8 without using any part of the student response; there is only a 0.03 AUC difference between

that baseline model and our current proposed method. This suggests that these external factors may be explaining a large portion of the student scores, and may subsequently explain a large portion of our prediction error.

## 4.2 Error Analysis of SBERT

In exploring this further, the results of the error analysis of the SBERT-Canberra method are presented in Table 4.1. It is found that the uni-level linear model explains 38.6% of the variance of the outcome as given by r-squared. Out of the six student answer-level features, nearly all were found to be statistically reliable predictors of model error; in verifying these results, it was found that all included covariates exhibited inter-correlations less than 0.3 (suggesting a moderately low impact of multicollinearity potentially skewing the interpretation of these results). In close examination of the coefficients of these features, however, despite being statistically reliable, many are found to be close to 0, suggesting a very little meaningful correlation with the modeling error. This is not the case, however, for two of these variables, Equation Percent and Presence of Images, we see a more meaningful coefficient. This suggests, due to the direction of this value, that the presence of mathematical elements as well as the presence of images (unsurprisingly) both correlate with higher prediction error. It further follows, then, that further improvements to the SBERT-Canberra method should explore methods of better representing and accounting for these mathematical terms in student responses; similarly, though likely much more difficult, incorporating an aspect of image recognition could be another area worth exploring.

In regard to the multi-level linear model, accounting for student, problem, and teacher identifiers each as random effects, we see that the inclusion of these level-2 factors explains some of the impact of the fixed effects (also in Table 4.1). Here it is found that

Table 4.1: The resulting model coefficients for the uni-level linear regression model and random and fixed effects of the multi-level linear model of absolute error.

	Uni-level Linear		Multi-level Linear	
	Variance	Std. Dev.	Variance	Std. Dev.
<i>Random Effects</i>				
Student	—	—	0.034	0.185
Problem	—	—	0.313	0.559
Teacher	—	—	0.048	0.851
	B	Std. Error	B	Std. Error
<i>Fixed Effects</i>				
Intercept	0.581***	0.017	0.772***	0.070
Answer Length	-0.008***	0.001	-0.009***	0.001
Avg. Word Length	-0.014***	0.003	-0.013**	0.003
Numbers Count	<0.001	<0.001	<0.001	<0.001
Operators Count	-0.006***	0.001	0.002	0.001
Equation Percent	0.443***	0.018	0.080***	0.022
Presence of Images	2.248***	0.021	1.858***	0.028

\*p < 0.05 \*\*p < 0.01 \*\*\*p < 0.001

all but two of the fixed effects are statistically reliable. It is also found that the magnitude of the coefficients for the Equation Percent and Presence of Images is also reduced. This suggests that, perhaps, the student and/or problem identifiers partially explain these correlations (some problems may be more likely to have responses with images or mathematical terms in them, or some students may be more inclined to use images or such terms more than others). What is worth noting, however, is that it was found that the level-2 variables account for 55.5% of the variance of the outcome. This suggests that a majority of the modeling error can be explained by these factors that are external to the student answers.

Looking at the variance of the random effects, it can be seen that the problem level identifiers contribute most in terms of explaining the variance of the outcome. It is certainly the case that the SBERT-Canberra method is accounting for each individual prob-

lem in producing its estimates (e.g. it only observes historic answers within each unique problem), but it would seem that there are other problem-level factors that are not being accounted for within this approach.

# Chapter 5

## SBERT-MTF Model

The methods presented in this work target the specific problem of non-linguistic terms contributing to prediction error. The previously-developed SBERT-Canberra model outperformed previous decision-tree- and deep-learning-based approaches [30] by leveraging pre-trained Sentence-BERT embeddings. The use of pre-trained embeddings provides several advantages in that they are commonly built using very large corpuses of data; by training such embedding methods on such sources as Wikipedia or online news sources, the models can learn the semantic meaning of words and sentences based on their proximity to other words and sentences within observed documents. In short, pre-trained models can learn language representation from large datasets that can then be used to increase the predictive power in smaller datasets. The challenge, however, is that only a finite number of words (and sentences, by extension) can be recognized by these methods; traditionally, unrecognized words and phrases may be given a default embedding. When observing non-linguistic terms such as numbers and expressions, many such terms may not be represented within the embeddings (e.g. representing “the answer is 4.3333” with the same embedding as, for example, “the answer is 2.987” if neither of the numbers are recognized). Particularly in mathematics contexts, such non-numeric terms are likely

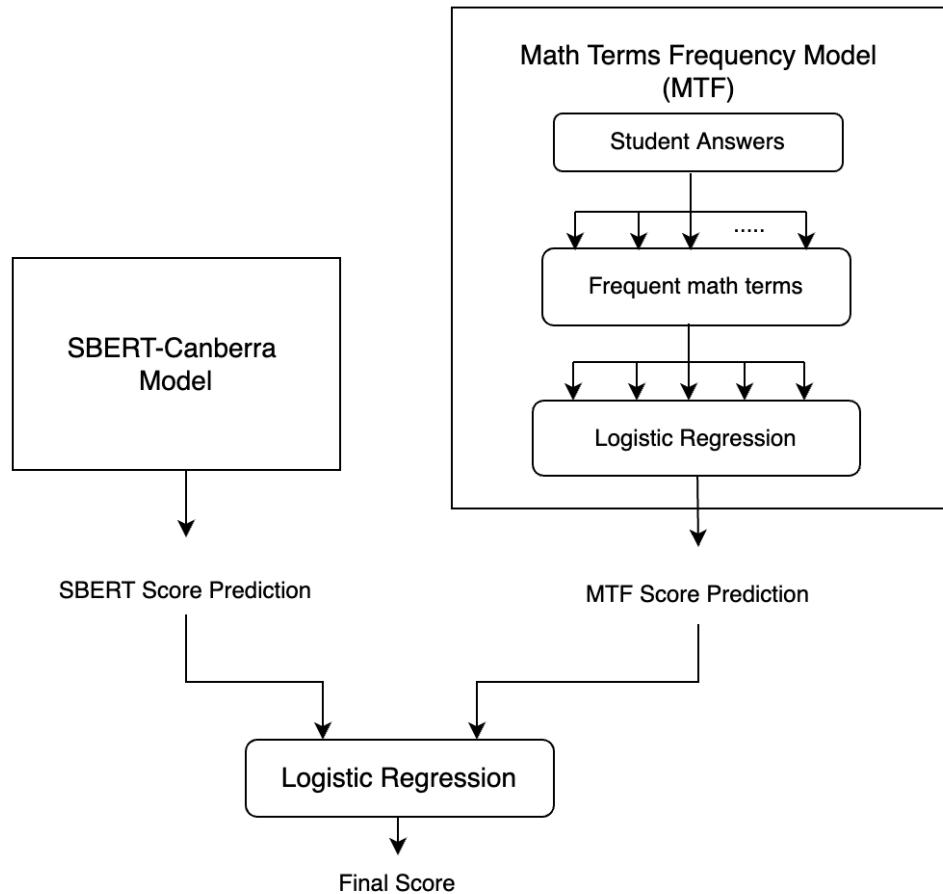


Figure 5.1: The design of the ensembled SBERT-MTF method, that suggests scores for student open responses.

to greatly inform the correctness of the student response. As such, one possible solution is to expand the embedding space to include such terms; while plausible, this would likely require large datasets of mathematics responses, but even then would not be able to represent every possible number or expression (given that these are infinite). Instead, we propose the “Math Term Frequency” (MTF) method which takes a much simpler approach, drawing inspiration from assessment methods applied for close-ended problems. The goal of the MTF method is to supplement the previously-developed SBERT-Canberra model through ensembling, resulting in what we are calling the “SBERT-MTF” model, as illustrated in Figure 5.1.



## 5.1 Model Design

The MTF method works by first parsing student answers to identify non-linguistic terms. The function works through a sequence of steps, which includes splitting a student answer by spaces, removing alphabet-only terms (accounting for commas, uncommon punctuation, and contractions), combining equations separated by spaces, removing extraneous parentheses, and rounding off decimals, among other optimizations. Much of this pruning is done with regular expressions. Once the non-linguistic terms have been identified, the MTF method involves identifying the most frequently-occurring terms for each possible integer score as a means of learning a kind of rubric. It is hypothesized that the correct student answers are likely to exhibit a smaller number of certain terms, with lower scores exhibiting a larger variety of terms. There will likely be some terms that are common throughout all scored answers (e.g. if the students reference a number from the problem text), but there are likely to be some terms that demonstrate high comprehension; similarly, students exhibiting common misconceptions may arrive at a similar set of incorrect answers. With this in mind, we select the five most-frequent terms,  $a_1, a_2, a_3, a_4, a_5$ , from the list of parsed non-linguistic terms. For each student response  $s$  in the training data for problem  $p$ , let  $S$  denote the set of mathematical terms in  $s$ . The input associated with  $s$  is the 5-vector  $\langle \mathbb{1}_S(a_1), \mathbb{1}_S(a_2), \mathbb{1}_S(a_3), \mathbb{1}_S(a_4), \mathbb{1}_S(a_5) \rangle$ , where  $\mathbb{1}_S$  denotes the indicator function for a term in  $s$ . In other words, features used in this method indicate whether a newly-observed student response contains any of the most frequent terms most commonly associated with each given score. These features are used in a multinomial logistic regression (treating each score as an independent category, following previous works) that is trained separately for each problem.

The score predictions from the MTF model are then ensembled with the SBERT-Canberra predictions using another logistic regression model, referred to as the SBERT-

Table 5.1: Rasch Model Performance comparing the performance of the proposed SBERT-MTF model with the observed SBERT-Canberra model

Model	AUC	RMSE	Kappa
<b>Rasch* + SBERT-MTF</b>	<b>0.871</b>	<b>0.524</b>	<b>0.508</b>
Rasch* + SBERT-Canberra	0.856	0.577	0.476

MTF model; to clarify, this ensemble regression model observes ten features corresponding to the probability estimates produced for each of the five possible scores for each of the two observed models. The goal of this is to combine the semantic representation captured by the SBERT method, while taking advantage of the non-linguistic term matching from the MTF method.

## 5.2 Model Performance

As to directly compare this new method to the SBERT-Canberra model, we use similar evaluation method and dataset used in the previous section 3.3 for the evaluation of SBERT-Canberra model utilizing a 2-parameter Rasch model, and three evaluation metrics. The Rasch model performance of the Math terms frequency model as compared to the performance of the prior models for scoring open-ended responses is presented in Table 5.1. The results suggests that the proposed SBERT-MTF model outperforms the previous highest-performing SBERT-Canberra model across all three evaluation metrics.

# Chapter 6

## Limitations and Future Works

In regard to our approach as well as in light of our findings, there are several limitations and opportunities for future directions. While the SBERT-Canberra approach, utilizing sentence-level embeddings, outperforms the previously-developed models in predicting scores for open responses, the difference in AUC is rather small; the fact that the method produces a classification (as opposed to a probability as is often the case with such models) likely impacts its AUC performance. The manner in which the method makes its prediction can be considered a greedy approach in that only the closest historic answer is used to predict the score. Instead, a weighted vote approach using all historic scores (or a subset of similar scores above an identified threshold) may improve the model by allowing for some degree of uncertainty. Similarly, the use of the word count model as a fallback may further be improved; while it was the case that there were very few instances of problems not having enough data within the cross validation, improving this fallback method may help to improve the model when applied in practical settings where the “cold start” problem is more prevalent; as the method currently relies heavily on having a sufficiently-sized pool of human-scored historic answers, future research can focus on utilizing unlabeled student answers or exploring other unsupervised methods that may

additionally support these methods in cases where labeled data is scarce.

While the SBERT-Canberra model performed arguably well, the error analysis revealed several areas where this approach, as well as others, may focus in future works. A limitation of this, however, is that both models left variance unexplained in the outcome. We chose to look at these factors based on hypotheses and anecdotal observations, but there may be other large factors that can explain more of the error that we are seeing. Subsequent works could conduct more thorough surveys of both answer-level and higher-level factors. Future works can also explore additional model structures and language features that may lead to improvements to performance. The analyses presented in this work, however, can act as a baseline to further evaluate if future iterations of our approach truly improve upon these identified areas.

The MTF method represents an intentionally-simple approach to address a targeted weakness in the presence of non-linguistic terms in the SBERT based method and seemingly led to positive impacts on the performance. With that, there are still several areas in which these models could be improved, in addition to improving the accuracy of the parsing function. While this proposed solution is an initial step towards addressing non-linguistic terms in these NLP based methods, more advanced methods based on Mathematical language processing, MathBERT could be explored for the better performance of these models. While the current works are based on textual open-ended responses, there are other forms of open-ended responses in mathematics including drawn diagrams and graphs, hand written formulas and expression uploaded as images, and other forms of audio and video responses. This work could be further expanded to the other forms of student open-ended responses (images, diagrams, etc.) through advanced methods of image processing.

It is also the case that this work focuses only on models that predict numeric assessment scores, while we strongly believe that it will be equally, if not more important to

additionally develop methods to suggest or generate directed feedback for for these student answers; teachers use textual feedback messages to offer constructive guidance to students, but it is often a very time-consuming task to write these messages for each students' answer. We believe that the SBERT-Canberra approach can be extended to support this task as well, where such a model may be able to recommend feedback to new student answers that has been previously given to an identified similar historic answer. Future work is intended to explore these methods further for such feedback-suggestion tasks.

# Chapter 7

## Conclusion

In this thesis, we have presented an approach in addressing and formulation of the problem of automating the assessment of student open-ended work. We have illustrated that our SBERT-Canberra method outperformed a previously-established benchmark, but still exhibits areas where it may be able to improve. Through the conducted error analysis, we have identified areas where more advanced methods of image processing and natural language processing (or math language processing), may lead to further improvements. With all of this, however, it was also identified that problem-level features appear to be most impactful in explaining the variance of modeling error; this is particularly surprising as variations in teacher grading were previously hypothesized to be a larger factor in this context.

With the findings from the study, we then propose a method to address the limitation of the SBERT-Canberra method, specifically in presence of non-linguistic terms. We demonstrate that this new method, based on the occurrence of frequent math terminologies in student responses in predicting scores, when combined with the SBERT-Canberra model improves the model performance across all the observed metrics. This approach is intentionally a simple. It is the goal of this work to act as a step toward building better

teacher supports for these types of open-ended problems, as well as provide others with guidance toward the same or similar goals.

# Chapter 8

## Acknowledgement

We thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306 ), and the EIR. We also thank the Office of Naval Research (N00014-18-1-2768) and finally Schmidt Futures as well as a second anonymous philanthropy.



# Bibliography

- [1] B. D. Wright, “Solving measurement problems with the rasch model,” *Journal of educational measurement*, pp. 97–116, 1977.
- [2] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [3] D. R. Thompson and S. L. Senk, “Implementing the assessment standards for school mathematics: Using rubrics in high school mathematics courses,” *The Mathematics Teacher*, vol. 91, no. 9, pp. 786–793, 1998.
- [4] J. Burstein, C. Leacock, and R. Swartz, “Automated evaluation of essays and short answers,” 2001.
- [5] D. J. Hand and R. J. Till, “A simple generalisation of the area under the roc curve for multiple class classification problems,” *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.
- [6] C. Leacock and M. Chodorow, “C-rater: Automated scoring of short-answer questions,” *Computers and the Humanities*, vol. 37, no. 4, pp. 389–405, 2003.
- [7] J. Z. Sukkarieh, S. G. Pulman, and N. Raikes, “Automarking: Using computational linguistics to score short free- text responses,” 2003.
- [8] E. Chen, M. Heritage, and J. Lee, “Identifying and monitoring students’ learning needs with technology,” *Journal of Education for Students Placed at Risk*, vol. 10, no. 3, pp. 309–332, 2005.
- [9] S. Dikli, “An overview of automated scoring of essays,” *The Journal of Technology, Learning and Assessment*, vol. 5, no. 1, 2006.
- [10] S. L. Meier, B. S. Rich, and J. Cady, “Teachers’ use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring,” *Assessment in Education*, vol. 13, no. 01, pp. 69–95, 2006.
- [11] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello, “Canberra distance on ranked lists,” in *Proceedings of advances in ranking NIPS 09 workshop*, Citeseer, 2009, pp. 22–27.
- [12] M. Mohler and R. Mihalcea, “Text-to-text semantic similarity for automatic short answer grading,” in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 2009, pp. 567–575.

- [13] L. Cutrone, M. Chang, *et al.*, “Auto-assessor: Computerized assessment system for marking student’s short-answers automatically,” in *2011 IEEE International Conference on Technology for Education*, IEEE, 2011, pp. 81–88.
- [14] S. Basu, C. Jacobs, and L. Vanderwende, “Powergrading: A clustering approach to amplify human effort for short answer grading,” *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 391–402, 2013.
- [15] H. Chen and B. He, “Automated essay scoring by maximizing human-machine agreement,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1741–1752.
- [16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [17] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende, “Divide and correct: Using clusters to grade short answers at scale,” in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 89–98.
- [18] N. T. Heffernan and C. L. Heffernan, “The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching,” *International Journal of Artificial Intelligence in Education*, vol. 24, no. 4, pp. 470–497, 2014.
- [19] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit,” in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
- [20] A. S. Lan, D. Vats, A. E. Waters, and R. G. Baraniuk, “Mathematical language processing: Automatic grading and feedback for open response mathematical questions,” in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 2015, pp. 167–176.
- [21] J. Pennington, R. Socher, and C. Manning, “Global vectors for word representation,” 2015.
- [22] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [23] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee, “Investigating neural architectures for short answer scoring,” in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 159–168.
- [24] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan, “A memory-augmented neural model for automated grading,” in *Proceedings of the fourth (2017) ACM conference on learning@ scale*, 2017, pp. 189–192.
- [25] D. Cer, Y. Yang, S.-y. Kong, *et al.*, “Universal sentence encoder,” *arXiv preprint arXiv:1803.11175*, 2018.

- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [27] K. Holstein, G. Hong, M. Tegene, B. M. McLaren, and V. Aleven, “The classroom as a dashboard: Co-designing wearable cognitive augmentation for k-12 teachers,” in *Proceedings of the 8th international conference on learning Analytics and knowledge*, 2018, pp. 79–88.
- [28] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [29] P. An, K. Holstein, B. d’Anjou, B. Eggen, and S. Bakker, “The ta framework: Designing real-time teaching augmentation for k-12 classrooms,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–17.
- [30] J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan, “The automated grading of student open responses in mathematics,” in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 615–624.
- [31] S. Baral, A. F. Botelho, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, “Improving automated scoring of student open responses in mathematics.,” *International Educational Data Mining Society*, 2021.
- [32] K. Seetharaman, S. Baral, A. Botelho, A. Wang, G. Heineman, and N. T. Heffernan, “Enhancing auto-scoring of student open responses in the presence of mathematical terms and expressions,” *International Conference on Artificial Intelligence in Education*, 2022.
- [33] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, “An automatic short-answer grading model for semi-open-ended questions,” *Interactive learning environments*, vol. 30, no. 1, pp. 177–190, 2022.