Automatic short answer grading with SBERT on out-of-sample questions

Aubrey Condor
University of California, Berkeley
aubrey_condor@berkeley.edu

Max Litster
University of California, Berkeley
maxlitster@berkeley.edu

Zachary Pardos
University of California, Berkeley
pardos@berkeley.edu

ABSTRACT

We explore how different components of an Automatic Short Answer Grading (ASAG) model affect the model's ability to generalize to questions outside of those used for training. For supervised automatic grading models, human ratings are primarily used as ground truth labels. Producing such ratings can be resource heavy, as subject matter experts spend vast amounts of time carefully rating a sample of responses. Further, it is often the case that multiple raters must come to a census before a final ground-truth rating is established. If ASAG models were developed that could generalize to out-of-sample questions, educators may be able to quickly add new questions to an auto-graded assessment without a continued manual rating process. For this project we explore various methods for producing vector representations of student responses including state-of-the-art representation methods such as Sentence-BERT as well as more traditional approaches including Word2Vec and Bag-of-words. We experiment with including previously untapped question-related information within the model input, such as the question text, question context text, scoring rubric information and a question-bundle identifier. The out-of-sample generalizability of the model is examined with both a leave-one-question-out and leave-one-bundle-out evaluation method and compared against a typical student-level cross validation.

Keywords

ASAG, Assessment, SBERT, Generalizability

1. INTRODUCTION

Automatic Short Answer Grading (ASAG) is an emerging field of research, as the education community has started to embrace the use of technology to assist students and education professionals. It has been shown that the use of open-ended (OE) questions helps facilitate learning [7], but educators are often deterred from their use because grading requires much more time than that for multiple choice [12]. In addition, human ratings may contain bias and vary in consistency, as rating choices are often subjective [28]. ASAG systems may be an important tool for educators, allowing more frequent use of OE questions, and more objective ratings for both formative and summative assessments.

A key challenge with supervised automatic grading models is gathering a large enough sample of labelled data for training. While some labeling tasks for supervised learning may be straightforward such as identifying an image as either a dog or a cat, others such as rating student responses require careful consideration. In high-stakes assessment scenarios, two or more ratings by different experts are often necessary to form a reliable consensus rating. Thus, obtaining labelled data to train an ASAG system can be arduous. It follows that quickly introducing new questions to an existing system may not be feasible if a data collection as well as the meticulous rating of new responses is necessary. Further, a new model would have to be trained and tuned with the newly collected responses. If we create more generalizable ASAG models, educators may have the flexibility to add new questions to an existing assessment with very little effort, thus increasing the practical use of the ASAG system.

We hypothesize that the inclusion of extra question related information within the model input may improve both the classification performance, and the generalizability of the model. For the purposes of this project, we formally define the generalizability of an ASAG model as the capacity to classify responses from out-of-training-sample questions.

This research contributes to the field of automatic grading in three related ways. We focus on classification performance and generalizability of the supervised grading model in terms of 1) the textual representation type, 2) the content of the input and 3) the classification model. We compare three different representation types, including those of state-of-the-art models: Sentence-BERT, Word2Vec, and Bag of Words. In terms of input content, we experiment with including previously untapped resources relating to the questions in the model input. Such resources include a question-bundle identifier, the question stem text, question context text, and rubric information. Extra input content is vectorized (if the source is textual) and concatenated to the response vectors

to be used as input to the classification model. Finally, we compare a non-neural model (a multinomial logistic regression) and a simple neural (a three layer feed forward network) model.

In order to examine the generalizability of the model for each experiment, we use a leave-one-question-out evaluation procedure where we train the model on N-1 questions, and use the one left-out question data as our test set. Thus, during training, the model has not yet seen responses from the question for which we use solely to evaluate the model. We go one step further in testing the generalizability of the model with a leave-one-bundle-out evaluation procedure where we train the model on M-1 question bundles (groupings of questions that are related in context), and use the questions for the left-out bundle as the test set. We conceptualize the leave-one-bundle-out method as a more extreme test of the model's ability to classify out of sample questions because even questions that are related in context have not been seen by the model during training. Additionally, we compare results of our experiments against and a typical student level cross validation. Results from a majority class classifier are included as well for a baseline comparison.

2. RELATED WORK

This section outlines notable work relating to ASAG and the more general use of NLP for education. For this project, we build on the previous literature by considering lessons learned in preceding research, and employing novel approaches that, to our knowledge, have not yet been explored.

2.1 ASAG work

A systematic review of trends in ASAG [3] illustrates an increasing interest in the field of automatic grading for education. Unsupervised methods have been explored such as concept mapping, semantic similarity, and clustering to assign ratings. For example, Mohler and Mihalcea [19] compared knowledge based and corpus based semantic similarity measures for automatic grading, Klein et al. [14] implemented a latent semantic analysis approach, and Basu et al. [2] used clustering to provide rich feedback to groups of similar responses. In addition, many types of supervised classification methods have been utilized for ASAG. Noteable examples include Hou and Tsao [13] who incorporated POS tags and term frequency with a Support Vector Machine classifier, and Madnani et al. [16] who made use of simple features such as a count of commonly used words and length of response with a logistic regression classifier.

More recent ASAG research exploits deep learning methods. Noteable work includes Zhang et al. [31] who used a combination of feature engineering and deep belief networks, Liu et al. [15] who employed multi-way attention networks, and Yang et al. [30] who considered a deep autoencoder model specific to Chinese responses. Additionally, Qi et al. [21] created a hierarchical word-sentence model with a CNN and Bi-LSTM model and Tan et al. [25] explored the use of a graph convolutional network (GCN) to encode a graph of all student responses.

Further, much of the newest ASAG work makes use of state-of-the-art transformer based models, including Gaddipati et al. [11] who evaluated four different types of response embeddings, ELMo, GPT, BERT, and GPT-2 for their performance on an ASAG task, Camus and Filighera [4] who compared the performance of transformer models for ASAG in terms of the size of the transformer and the ability to generalize to other languages, and Sung et al. [24][23] who examined the effectiveness of pre-training BERT, including further pre-training the model on relevant domain texts.

2.2 NLP for Education

Literature addressing the general application of natural language processing (NLP) for various uses in field of education has grown quickly in recent years as well. For example, Fonseca et al. [10] used NLP to automatically classify the programming assignments for students within given academic context, Thaker et al. [26] incorporated textual similarity techniques to recommend remedial readings to students, and Arthurs and Alvero [1] examined bias in word representations for college admissions essays. Additionally, Xiao et al. [29] employed NLP and transfer learning methods for problem detection in peer assessments, Venant and d'Aquin [27] utilized a concept graph to predict semantic complexity of short essays by written by English language learners, and Chen et al. [6] leveraged a variety of textual analysis methods to predict student satisfaction in the context of online tutorial dialogues.

We build on the previous literature by incorporating state-of-the art representation methods such as Sentence-BERT and a neural classification model. The novel contribution of this project includes both our focus on the generalizability of the model to out-of-training-sample questions, as well as the leveraging of previously untapped, question related information as input to the model.

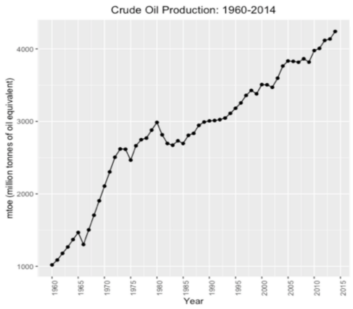
3. DATA SET

The data we will use for this project was sourced from a 2019 field test of a Critical Reasoning for College Readiness (CR4CR) assessment [17] created at the Berkeley Evaluation and Assessment Research (BEAR) center. The data consists of 5,550 student responses from 558 distinct students to 33 different items. The field test included other items that were multiple choice, but these questions were filtered out of the data for our use. The mean number of responses per question is 179 with the minimum being 128 and the maximum being 313. Most of the items belong to an item bundle - a grouping of items that are related in context and/or share a common question context. Additionally, the items were administered in four different test forms, where some items were included in multiple forms. The items all relate to four constructs about student understanding of algebra.

An example of one of the items, labelled 'Crude Oil 4ab' is included in Figure 1. For this item, students are presented with two images relating to oil production - one being a line graph and the other, a table. With the given context, students are presented with a choice between the graph or the table for which would be better to represent the historical patterns, or change over time of the oil production. The correct answer for this question is the graph, and students are expected to provide reasons why this is the right choice.

An example of a student response to the Crude Oil 4ab question (shown in Figure 1) rated at the highest (most correct)

The following graph and table show annual crude oil production in million tonnes of oil equivalent (mtoe) from 1960 to 2014.



Year	Oil Produced (mtoe)	Year	Oil Produced (mtoe)	Year	Oil Produced (mtoe)
1960	1,020	1980	2,986	2000	3,507
1961	1,087	1981	2,814	2001	3,504
1962	1,179	1982	2,695	2002	3,470
1963	1,267	1983	2,671	2003	3,595
1964	1,369	1984	2,732	2004	3,763
1965	1,467	1985	2,695	2005	3,833
1966	1,301	1986	2,807	2006	3,828
1967	1,503	1987	2,835	2007	3,816
1968	1,705	1988	2,945	2008	3,864
1969	1,904	1989	2,991	2009	3,817
1970	2,106	1990	3,026	2010	3,976
1971	2,303	1991	3,011	2011	4,006
1972	2,505	1992	3,023	2012	4,117
1973	2,619	1993	3,046	2013	4,136
1974	2,615	1994	3,111	2014	4,240
1975	2,465	1995	3,181		
1976	2,663	1996	3,254		
1977	2,747	1997	3,357		
1978	2,769	1998	3,426		
1979	2,880	1999	3,380		

[a] For a group project, you and your classmates have to present the overall historical patterns of annual oil production as a poster. Due to limited space, only one of the following representations can be included in the poster: a table, a graph, or a set of equations. Which representation should you use?

[b] Briefly explain your answer choice in [a].

Figure 1: An example of an item from the data set.

score category is shown: *"the graph easily displays patterns over time whereas the table and equations require more analyzing."* In contrast, a student response to the same question rated at the lowest (most incorrect) category is shown: *"the table is more clear, the information is seen in the table."*

Responses were rated from 0 (fully incorrect) to 4 (fully correct) by multiple researchers and subject-matter experts at the BEAR center. The quality and consistency of ratings were evaluated by an inter-rater reliability score, and when a high percentage of rating mismatches between raters existed, incongruous ratings were discussed until a consensus was reached by the raters.

4. METHODS

In this section, We briefly introduce the representation methods and model classes that we include in our experiments. Additionally, we describe the question related information, beyond that of the responses, that are used as inputs to the model. Further, we outline our experiments in detail including methods for evaluation and comparison.

4.1 Input Representations

We chose to include three distinct, yet commonly used, representation types in our experiments: A count-based method that elicits the distinct vocabulary of our data (Bag of Words), a simple neural method that utilizes pre-trained word vectors (Word2Vec), and a state-of-the-art, contextual neural method (Sentence-BERT). A short description of each representation type is included.

4.1.1 Sentence-BERT

Sentence-BERT (SBERT) is a modification of the BERT network that utilizes siamese and triplet network structures to create semantically meaningful sentence embeddings [22]. SBERT fine-tunes the BERT network on a combination of the SNLI dataset and the Multi-Genre NLI datasets, totaling about 1 million sentence pairs. Although sentence embeddings can be derived from the original BERT model using methods such as averaging the BERT output layers or using the [CLS] token embedding, it has been shown that such methods yield poor sentence embeddings [20]. In comparison, SBERT sentence embeddings outperformed other state-of-the-art methods such as InferSent [8] and Universal Sentence Encoder [5] on the SentEval [8] benchmark, which gives an idea of the quality of sentence embeddings for various tasks such.

4.1.2 Word2Vec

Word2Vec (W2V) ma13, is a neural model that creates vector representations of words that have been shown to be semantically meaningful and useful in different NLP tasks [22]. We use an extension of the previously introduced Skip-gram model [18] that incorporates sub-sampling of frequent words during training in order to speed up training, and improves accuracy of representations of less frequent words. For this project, we use the Google News corpus of pre-trained word embeddings. Vectors of size 300 are created for each word, and in order to construct response embeddings from the individual word vectors, we employ a simple but popular method: averaging the vectors of all words in the response.

4.1.3 Bag-of-words

The bag-of-words (BOW) model represents a document as a vector, or "bag," of length equal to the number of unique words in the entire corpus and values of the vector equal to the frequency with witch its corresponding word occurred in the document. In our application, the bags are student short answers and additional question information.

4.2 Input Content

In an item design context, there are various untapped sources of information relating to a particular question that may be useful to include as input to a classification model. We explore the use of four different sources of information, outside that of the response itself. A brief description of the such sources are included below.

4.2.1 Question Text

The question text consists of the direct question stem. As in the example item provided in Figure X, the question text would be: *"Briefly explain your answer choice in [a]."*

4.2.2 Question context

The question context includes any textual information beyond that of the question stem that is related to the question, and might be useful for the respondent to produce a response. In the question example in Figure 2, the question context would include: *"The following graph and table show annual crude oil production in million tonnes of oil equivalent (mtoe) from 1960 to 2014. For a group project, you and your classmates have to present the overall historical patterns of annual oil production as a poster. Due to limited space, only one of the following representations can be included in the poster: a table, a graph, or a set of equations. Which representation should you use?"* We note that not all of the included items have question context text beyond that of the question text itself. For such items, it was not possible to include question context as part of the input.

4.2.3 Rubric Text

Level	Description	Example Response
4	Student provides a fully correct positive and negative justification for ...	"The graph best illustrates the visual trend of oil production. The table or the equation won't provide an overall ..."
3	Student provides a fully correct positive justification for selected representation ...	"The table cannot show a trend in the oil production effectively. "
2	Student provides a partial/general positive justification for selected ...	"The graph helped me with more questions."
1	Student provides incorrect justification for selected representation.	"Because the graph gives us better projections."
0	Student makes no attempt to provide a justification for selected representation.	"I am not sure why."

Figure 2: An example of a scoring rubric corresponding to the item in Figure 1.

As part of the assessment cycle as previously mentioned, an important step in a measurement process is defining the outcome space for each item through a scoring guide, which is essentially a rubric. The scoring guide includes a detailed description of the reasons for which a response would be rated at a certain score, and is used as a guide for human raters. In addition, the scoring guide often includes example responses for each rating level. An example of a scoring guide for the 'Crude Oil 4ab' item in Figure 1 is included shown in Figure 2. When the rubric text is included in the input text, we include both the level descriptions as well as the example response(s).

4.2.4 *Bundle Identifier*

As described above in the data section, most of the questions belong to a bundle of questions - those that are linked based on a similar image, or context text. As a question bundle identifier, we concatenate a one-hot vector to the input vectors. Although items within the same bundle will often share the same context text, we include the one-hot bundle identifier within our extra input text experiments so that we can infer whether the model makes use of semantics within the context text, or rather just a general indication of similar questions.

4.3 Classification Models

We compare a multinomial logistic regression model with a simple neural network classification model. We chose these classification methods representing a linear transformation of the feature space to a label (regression) and a non-linear transformation (neural network). A brief overview of each model is included.

Multinomial logistic regression (MLR) is a classification model that predicts probabilities of different outcomes for a categorical dependent variable. In order to generalize to a K-class setting, the model runs K-1 independent binary logistic regression models where one outcome is chosen as a "pivot" and other K-1 outcomes are separately regressed against the pivot outcome. We use the Limited-memory Broyden-Fletcher-Goldfarb-Shannon (LBfGS) algorithm for optimization [9], and incorporate L2 regularization.

Additionally, we use a simple feed forward neural network on a categorical cross entropy loss function with 2 hidden layers of size 100, using rectified linear unit (ReLU) activation functions for both hidden layers. We include dropout of 0.4 and utilize Adam optimization. We train the model for 16 epochs and use a batch size of 36.

4.4 Evaluation and Model Comparison

In this section, we enumerate our experiments and the evaluation methods chosen for comparison.

4.4.1 *Experiments*

As input to our model, we experiment with 8 different combinations of content to vectorize and concatenate to the response vectors before training our classification models:

1) response
2) question + response
3) question context + response
4) scoring rubric + response
5) bundle one-hot + response
6) question + scoring rubric + response
7) bundle one-hot + question + scoring rubric + response
8) bundle one-hot + question + question context + scoring rubric + response

For each of the 8 combinations listed above, we create three different vector representations with the aforementioned methods: 1) Sentence-BERT, 2) Word2Vec, and 3) Bag-of-words, resulting in 24 distinct input types. We fit a classification model for each of the input types, for both of our classification models. Thus, we compare 48 separate versions of an ASAG model with three types of evaluation.

4.4.2 *Leave-one-question/bundle-out Evaluation*

In order to assess the generalizability of the ASAG model to out-of-training-sample questions for each of the 48 experiments, we average the results of N (where N is the number of questions) independent models. For each of the N models, we train the classifier on data from N-1 questions, and test on data exclusive to the left-out-of-training question. In the case of the leave-one-question-out results, it is important to note that although the model has not seen data specific to the left-out question, it has seen questions that are part of the same question bundle and are therefore related.

To expand our evaluation of generalizability further, we include a leave-one-bundle-out metric for each experiment. For such, we average the results from M (where M is the number of bundles) independent models where we train the classifier on data from M-1 bundles, and test on data exclusive to the left-out-of-training questions which belong to a single bundle. So, these results give us an idea of whether the model can successfully rate responses from questions that have not been used for training, and when the model has not seen questions related by context during training.

4.4.3 *Evaluation Metrics*

We report our results in both multilabel accuracy, and weighted F1 score because multilabel accuracy is both widely used and easy to interpret, and the weighted F1 score captures both the precision and recall and accounts for class imbalance.

Multilabel accuracy represents the degree to which our model classifications agree with the ground truth labels (for this

Table 1: Experiment Results: Multilabel Accuracy

	Response Text	Bundle ID	Question Text	Context Text	Rubric Text	Random Holdout			Question Holdout			Bundle Holdout				
						SBERT	W2V	BOW	SBERT	W2V	BOW	SBERT	W2V	BOW	Average (ACC)	Average (Wf1)
Majority Class							0.34715			0.37044			0.30521		0.34093	0.25429
LogReg	x					0.58034	0.49765	0.54665	0.35870	0.35338	0.36517	0.31806	0.25316	0.26995	0.39367	0.38323
LogReg	x	x				0.59603	0.53154	0.55602	0.37225	0.37595	0.36097	0.32855	0.25784	0.28364	0.40698	0.39513
LogReg	x		x			0.63062	0.55748	0.58702	0.40378	0.41290	0.32506	0.36276	0.30162	0.25982	0.42678	0.40317
LogReg	x			x		0.62972	0.56036	0.58486	0.40352	0.41815	0.33053	0.34539	0.31713	0.27005	0.42886	0.40411
LogReg	x				x	0.61657	0.55982	0.58846	0.38488	0.38379	0.42198	0.28175	0.23273	0.31443	0.42049	0.39944
LogReg	x		x		x	0.61818	0.56432	0.59152	0.40967	0.38968	0.37571	0.34737	0.26432	0.30380	0.42940	0.40213
LogReg	x	x	x		x	0.62484	0.56144	0.59261	0.41534	0.40174	0.36048	0.32196	0.25499	0.28627	0.42441	0.40195
LogReg	x	x	x	x	x	0.61406	0.55963	0.59009	0.41494	0.39999	0.36104	0.31797	0.26117	0.29534	0.42380	0.39920
NN	x					0.60249	0.56450	0.60973	0.35736	0.34029	0.36930	0.31540	0.25760	0.26633	0.40922	0.40709
NN	x	x				0.60540	0.59802	0.61963	0.39011	0.37083	0.35526	0.31598	0.23819	0.28716	0.42006	0.41686
NN	x		x			0.65800	0.61009	0.65532	0.40633	0.37576	0.34075	0.35574	0.32317	0.27179	0.44411	**0.42600**
NN	x			x		0.66070	0.61388	0.66198	0.39309	0.38079	0.33176	0.36227	0.31851	0.28206	**0.44500**	0.42470
NN	x				x	0.64017	0.61226	0.62234	0.36721	0.35595	0.40597	0.33015	0.23390	0.33769	0.43395	0.41403
NN	x		x		x	0.62503	0.61009	0.62198	0.41137	0.39257	0.33095	0.34568	0.25618	0.31157	0.43394	0.41456
NN	x	x	x		x	0.63206	0.59981	0.62938	0.40885	0.36204	0.39455	0.36535	0.26301	0.32980	0.44276	0.42075
NN	x	x	x	x	x	0.60557	0.59405	0.61478	0.42270	0.39130	0.35288	0.30132	0.27360	0.31397	0.43002	0.40366
Average (ACC)						**0.62124**	0.57468	0.60452	0.39500	0.38157	0.36140	0.33223	0.26919	0.29273		
Average (WF1)						**0.60830**	0.55063	0.59135	0.37680	0.35320	0.33350	0.31472	0.24976	0.28697		

project, human ratings). It is calculated simply as the number of correct predictions divided by the number of total number of examples. The F1 score for a certain class is the harmonic mean of its precision and recall, where precision is calculated as true positives divided by false positives and true positives, and recall is calculated as true positives divided by false negatives and true positives. In order to account for class imbalance, we specifically use the weighted F1 score. This metric calculates the F1 score for each class independently, and the overall score for all the classes is the average weighted by class size.

5. RESULTS

Results of our experiments are detailed in Table 1, reported in multilabel accuracy. For column and row averages, the weighted F1 score is presented as well. In the left-most half of the table, an x is present for a given row if the information type, indicated by the column header, is included in the model input. For example, results in the first row represent an input of only the response text and results in the second row represent an input of both the Bundle ID and the response. Additionally, the top half of the table results are those from the multinomial logisitic regression classifier, and the bottom half of the table results are those for the neural network classifier (as indicated by the leftmost column). For each of our evaluation methods, random holdout, question holdout, and bundle holdout, we present results for the three textual representation methods: SBERT, W2V, and BOW.

In terms of the general performance of our classification models, we consider the random holdout evaluation method. Overall, SBERT representations performed best when averaging across the classification methods and input combinations, followed by the BOW representations (accuracy of 0.621 for SBERT compared to 0.575 and 0.605 for W2V and BOW, respectively). Additionally, the neural network achieves higher accuracy than the logistic regression in general. Both SBERT and BOW perform notably well when the input includes the question text, or the question content.

To assess the generalizability to grading answers to questions unseen in the training set, we focus on the question holdout and bundle holdout results. Across the board, we see much lower accuracy for the question and bundle holdout experiments than that of the random holdout, with the bundle holdout being the lowest. This is in line with what one might expect because for the question holdout, the model has not yet seen responses for the particular question in the test set and for the bundle holdout, the model has not seen questions even related to the test set question.

Similar to the random holdout experiments, we see the same overall pattern for the question and bundle holdout experiments: SBERT is generally superior, followed by BOW and W2V, respectively. One notable difference for the question holdout experiments compared to those of the random holdout is that we see increased performance when we include multiple extra sources of information. For example, with SBERT and bundle holdout, we achieve 0.365 accuracy with the neural network classifier when we include the rubric text, question text, and bundle ID. We might explain this result as, when the model is lacking previous information about the test question from training, extra input information might provide guidance for the model.

For the question and bundle hold out experiments, the addition of the rubric text improves performances particularly well with the use of BOW representations, for both the logistic regression and neural network classifiers. With SBERT, the addition of the question text seemed to help the generalizability of the model as well. Interestingly, we do not see the same pattern between the classification models for the question and bundle holdout methods: where the neural net was clearly superior in the random holdout experiments, results are more similar between the logistic regression and neural network for the bundle and question holdout experiments.

We see from the row averages in the right most columns of the table that across all experiments and text representation types, the response and question text, as well as the response and context text achieve the highest evaluation scores. Additionally, the column averages further confirm that the SBERT representations perform best.

Further, we include results from a majority class classifier on the top row for a baseline comparison. We emphasize that, across all random holdout experiments, the classification models outperform the majority class classifier significantly. However, this is not the case for the question holdout and bundle holdout experiments. For the question holdout experiments, many of the SBERT experiments out-

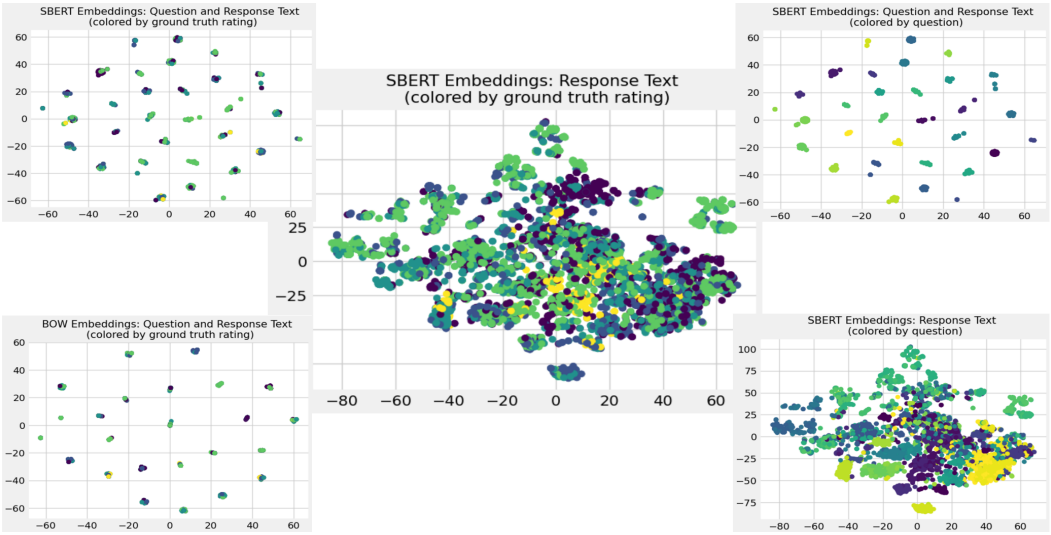


Figure 3: 2D visuals of input vector representations. Plots vary by representation type, input content, and color labeling.

performed the majority class classifier, but on average, the W2V and BOW experiments performed a bit worse than majority class. For the bundle holdout experiments, on average, SBERT performs slightly better than majority class, but the W2V and BOW experiments do not.

Further interpretations can be made from Figure 4, which includes two dimensional, t-distributed Stochastic Neighbor Embedding (TSNE) reduced vector representations. The center image includes SBERT embeddings of only the response text and the colors represent the ground truth ratings. From the top left and right images, as well as the bottom left image we can see very distinct question clusters with the inclusion of question text. However, from the bottom right image, we can visualize question specific clusters, but they are not as distinct as the representations that include the question text.

Thus, we conclude that in terms of question holdout, the model can generalize to out-of-sample questions with only slight improvement over majority class, with a state-of-the-art representation method like SBERT. Certain extra pieces of input information aid our models more than others like question text and the best performing models use the neural network classifier.

6. DISCUSSION

Although our results are not promising for the generalizability of autograding models to unseen questions, we emphasize the importance of finding more generalizable models to decrease time spent on the laborous task of creating ground-truth human ratings. Our intention is that this work will influence researchers to consider further innovative methods to increase the generalizability of ASAG models. Further, because we did find that including certain question-related text may improve model performance, it may be of use to the ASAG research community to continue to explore how extra sources of information about a question may be incorporated into an ASAG system.

As is evident in our literature review, there has been increased adoption of state-of-the-art textual representation methods such as SBERT, and transformer-based models such as BERT and XLNet, within the field of NLP in Education. Our results support that such models may achieve superior performance for certain tasks.

To build on this work further, we could consider other methods, beyond that of concatenation to the input text, to include the extra question information in our model. We could further pre-train a transformer-based model such as BERT or XLNet with the extra textual information by either tuning the existing weights or altering the existing architecture with an extra encoder layer of weights trained on our text alone. Moreover, we may focus more closely on how the classification model itself might be altered such that it might better generalize to out-of-training-sample questions, instead of only focusing on the input content.

We believe that beyond model performance, the practical utility of an ASAG system must be considered in order for educators to continue to adopt new technologies that employ advanced methods in artificial intelligence. Recent years have seen vast improvements in the field of machine learning and language processing. Embracing such technologies for applications in education may be pivotal to provide the assistance that both educators and learners need. However, we do not suggest that machine learning systems such as ASAG should be used to replace human judgements in education, especially in high stakes testing scenarios. We emphasize the ASAG systems should be used to support educators, not replace them. This project represents a continued effort to explore the ways in which we can make use of new technologies to improve learning.

7. REFERENCES

[1] N. Arthurs and A. J. Alvero. *Whose Truth is the "Ground Truth"?* College Admissions Essays and Bias in Word Vector Evaluation Methods.

[2] 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*, 1:391–402, 2013.

[3] S. Burrows, I. Gurevych, and B. Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117, 2015.

[4] L. Camus and A. Filighera. Investigating transformers for automatic short answer grading. In *International Conference on Artificial Intelligence in Education*, pages 43–48. Springer, Cham, 2020.

[5] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, and Y. Sung. Universal sentence encoder., 2018.

[6] G. Chen, R. Ferreira, D. Lang, and D. Gasevic. *Predictors of Student Satisfaction: A Large-Scale Study of Human-Human Online Tutorial Dialogues*. International Educational Data Mining Society, 2019.

[7] M. T. Chi, N. De Leeuw, M. H. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477, 1994.

[8] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. preprint, 2017.

[9] R. Fletcher. *Practical Methods of Optimization(2nd ed.)*. John Wiley Sons, -0-471-91547-8, New York, 1987.

[10] S. C. Fonseca, F. D. Pereira, E. H. Oliveira, D. B. Oliveira, L. S. Carvalho, and A. I. Cristea. *Automatic subject-based contextualisation of programming assignment lists*. International Educational Data Mining Society, 2020.

[11] S. K. Gaddipati, D. Nair, and P. G. Pl"oger. Comparative evaluation of pretrained transfer learning models on automatic short answer grading. arxiv. preprint, 2020.

[12] C. L. Hancock. Enhancing mathematics learning with open-ended questions. *The Mathematics Teacher*, 88(6):496, 1995.

[13] W. J. Hou and J. H. Tsao. Automatic assessment of students' free-text answers with different levels. *International Journal on Artificial Intelligence Tools*, 20(2):327–347, 2011.

[14] R. Klein, A. Kyrilov, and M. Tokman. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *G. RoBling, T. Naps, C. Spannagel (Eds.), Proceedings of the 16th annual joint conference on innovation and technology in computer science education . Darmstadt: ACM*, pages 158–162. Darmstadt: ACM, 2011.

[15] T. Liu, W. Ding, Z. Wang, J. Tang, G. Y. Huang, and Z. Liu. June). automatic short answer grading via multiway attention networks. In *International conference on artificial intelligence in education.*, pages 169–173. Springer, Cham., 2019.

[16] N. Madnani, J. Burstein, J. Sabatini, and T. O. Reilly. Automated scoring of a summary writing task designed to measure reading comprehension. In J. B. Tetreault and C. Leacock, editors, *Proceedings of the 8th workshop on innovative use of nlp for building educational applications . Atlanta*, pages 163–168. Association for Computational Linguistics, 2013.

[17] J. Mason, M. Wilson, A. E. Arneson, and D. Wihardini. *A framework for the college ready algebraic thinking assessment (CRATA)*. University of California, Berkeley Evaluation and Assessment Research Center, 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality., 2013.

[19] M. Mohler and R. . Mihalcea. March). *Text-to-text semantic similarity for automatic short answer grading*, 12:567–575, 2009.

[20] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. pages 1532–1543. In Proceedings of the conference on empirical methods in natural language processing, 2014.

[21] H. Qi, Y. Wang, J. Dai, J. Li, and X. Di. Attention-based hybrid model for automatic short answer scoring. pages 385–394. SIMUtools 2019. LNICST, vol. 295, . Springer, Cham, 2019.

[22] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. preprint, 2019.

[23] C. Sung, T. Dhamecha, and N. Mukhi. Improving short answer grading using transformer-based pre-training. In *International Conference on Artificial Intelligence in Education*, pages 469–481. Springer, Cham, 2019.

[24] C. Sung, T. Dhamecha, S. Saha, M. Tengfei, V. Reddy, and A. Rishi. Pre-training bert on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6071–6075, Hong Kong, CH., 2019. Association for Computational Linguistics.

[25] H. Tan, C. Wang, Q. Duan, Y. Lu, H. Zhang, and R. Li. Automatic short answer grading by encoding student responses via a graph convolutional network. *Interactive Learning Environments*, pages 1–15, 2020.

[26] K. Thaker, L. Zhang, D. He, and P. Brusilovsky. Recommending remedial readings using student knowledge state. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pages 233–244, 2020.

[27] R. Venant and M. d'Aquin. Towards the prediction of semantic complexity based on concept graphs. In *12th International Conference on Educational Data Mining*, pages 188–197, 2019.

[28] S. A. Wind and M. E. Peterson. Sa systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 2(35):161–192, 2018.

[29] Y. Xiao, G. Zingle, Q. Jia, S. Akbar, Y. Song, M. Dong, and E. Gehringer. *Problem detection in peer assessments between subjects by effective transfer learning and active learning*. 2020.

[30] X. Yang, Y. Huang, F. Zhuang, L. Zhang, and S. Yu. Automatic chinese short answer grading with deep autoencoder. In *International Conference on Artificial*

Intelligence in Education., pages 399–404. Springer, Cham, 2018.

[31] Y. Zhang, R. Shah, and M. Chi. *Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading*. International Educational Data Mining Society, 2016.