Enhancing Transfer Learning of LLMs through Fine-Tuning on Task-Related Corpora for Automated Short-Answer Grading

Nazmul Kazi
School of Computing
University of North Florida
Jacksonville, FL, USA
nazmulkazi@oxiago.com
0000-0003-3610-455X

Indika Kahanda
School of Computing
University of North Florida
Jacksonville, FL, USA
indika.kahanda@unf.edu
0000-0002-4536-6917

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00255

***Abstract*—Automated short-answer grading (ASAG) is a crucial element of any intelligent tutoring platform. Machine Learning (ML) has shown great promise for ASAG. However, this task remains challenging even for Deep Learning (DL) approaches and Large Language Models (LLMs), requiring semantic inference and textual entailment recognition. The SemEval-2013 Task 7, The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge, is a benchmark widely used for research on ASAG. The SciEntsBank data included in this collection contains nearly 11,000 answers to 197 assessment questions in 15 different science domains. Despite the popularity, only a few researchers have explored the potential of DL or LLMs for this task. In this project, we explore the effectiveness of the RoBERTa Large model, an LLM trained on an extensive text corpus for language comprehension. By fine-tuning the model on the Multi-Genre Natural Language Inference (MNLI) corpus for semantic inference and subsequently on the SciEntsBank dataset, with a focus on the 3-way labels of correct, incorrect, and contradictory, we achieved a weighted F1-score of 0.77, 0.72, and 0.72 on unseen answers, questions, and domains, respectively. Notably, our model significantly benefits from fine-tuning on the MNLI corpus, particularly in enhancing its performance on the *contradictory* class (which constitutes only 10% of the dataset) through transfer learning leading to significant improvements on the more challenging test sets: unseen questions and unseen domains.**

***Index Terms*—Automated Short Answer Grading, ASAG, Large Language Models, Transfer Learning, Semantic Inference, Recognizing Textual Entailment, MNLI, SemEval, SciEntsBank**

I. Introduction

Adaptive testing and assessment capture the cognitive level of students which is essential for formulating personalized learning routes [1]. Typically, standardized multiple-choice (i.e., closed-ended) questions are used in adaptive assessments due to their simplicity of implementation and assessment. However, multiple-choice questions can only test a limited range of knowledge, mainly focusing on factual recall rather than a deeper understanding or application of concepts [2]. In some cases, students might randomly guess correct answers without a full grasp of the material, leading to an inaccurate reflection of their cognitive proficiency in the assessment [1].

Besides, multiple-choice questions commonly provide binary feedback (correct or incorrect), which might limit the student's ability to learn from their mistakes and improve their understanding of the material. On the contrary, open-ended questions often require students to think critically and creatively. This promotes meta-cognitive skills and allows for a more comprehensive evaluation of a student's understanding and application of concepts [3]. In open-ended questions, students are asked to generate their own answers and explain their thought processes and reasoning. This can help instructors gain valuable insight into their student's learning progress and identify areas for improvement.

However, grading open-ended questions manually can be time-consuming and tedious [4]. This process can detract instructors from their primary objectives of teaching objectives and providing personalized feedback to help students improve upon their mistakes. Therefore, automated short-answer grading (ASAG) is a crucial component of any intelligent tutoring system [1]. ASAG automatically assesses the correctness of short answers, providing real-time feedback to students and helping instructors evaluate students more efficiently. This would allow the instructors to devote more time to teaching and supporting students.

Therefore, there has been a growing interest in developing automated systems that can evaluate student responses to open-ended questions across various domains and topics [1]. ASAG is a challenging task, as it requires both semantic inference and recognizing textual entailment (RTE). Besides, ASAG presents an additional layer of complexity by requiring transfer learning to ensure cross-domain compatibility, making it inherently a data-driven problem. The SemEval-2013 Task 7 and the SciEntsBank dataset included in this challenge are widely used benchmarks for ASAG research.

Various approaches have been proposed but only a handful of researchers have explored the potential of LLMs for this task and the SciEntsBank dataset. LLMs have shown great promise in various natural language processing tasks, including language modeling, machine translation, and sentiment analysis due to their ability to capture complex contextual

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00255

information and relationships between words and phrases [5], [6]. Compared to classical ML approaches, LLMs have the potential to improve the accuracy and efficiency of ASAG. Dzikovska et al. [7] reported 0.63 as the best weighted-average F1 for 3-way labeling and all the reported models use some form of classical ML. Recent studies demonstrate the superior performance of LLMs in ASAG. Sung et al. [8] achieved a weighted-average F1 of 0.68 using BERT-base [5], while Zhu et al. [1] reported a weighted-average F1 of 0.67 using BERT-base and 0.69 using a BERT-based DNN network. Camus and Filighera [9] fine-tune and compare the model performance of various LLMs including BERT, RoBERTa, AlBERT, XLM, and XLMRoBERTa; RoBERTa Large fine-tuned over the MNLI corpus performed the best with a weighted-average F1 of 0.72.

In this project, we investigate two main research questions related to automated short-answer grading using LLMs. Firstly, we explore whether classical ML approaches could establish a better baseline than the existing lexical baseline from SemEval-2013 Task 7. We test several models, including Decision Tree, Random Forest, Support Vector Machines (SVM), and Multi-layer Perceptron (MLP) with various feature extraction techniques, such as bag-of-words and TF-IDF. Secondly, we investigate whether fine-tuning RoBERTa-Large on a more extensive and diverse corpus, such as the Multi-Genre Natural Language Inference (MNLI) corpus [10], could help the model with semantic inference and transfer learning to boost the model performance. By addressing these research questions, we aim to contribute to the ongoing efforts to improve the accuracy and efficiency of ASAG using LLMs.

II. Dataset

We utilized a portion of the Student Response Analysis (SRA) corpus [11], known as the SciEntsBank dataset, for this experiment. The dataset has been annotated with SRA labels by human annotators [7]. The dataset was released with three distinct labeling versions: 5-way, 3-way, and 2-way. In our experiment, we employed the 3-way labeling, where each sample is labeled as either *correct*, *contradictory*, or *incorrect*.

The SciEntsBank dataset comprises four distinct sets: a training set and three test sets. The test sets are tailored to assess the adaptability of a model across various problems and domains. The details regarding the creation and purpose of these test sets are outlined below:

1) **Unseen domains (UD):** The authors set aside the complete set of questions and answers of three science domains from training to create this set. The purpose of this set is to evaluate a model's versatility and adaptability across diverse knowledge domains.
2) **Unseen questions (UQ):** Within the 12 domains selected for training, the authors randomly selected a subset of questions and held out all responses to these selected questions to create this set. This set evaluates a model's capacity to accommodate novel questions within familiar domains.
3) **Unseen answer (UA):** From the questions selected for the training set, the authors withheld a subset of randomly selected responses from the training set to create this set. This set of unseen answers is the most typical approach to model evaluation. Its purpose is to assess the model's proficiency in grading responses it has not previously encountered.

The training set consists of samples that are not present in any of the three test sets. Table I displays the distribution of the 3-way labels in the dataset. Please note that the dataset is imbalanced. The *Contradictory* class has a significantly low number of samples (only 10% of the dataset) compared to the other two classes.

TABLE I
THE 3-WAY LABEL DISTRIBUTION OF SCIENTSBANK DATASET.

Labels	Train	Test		
		Unseen Answers	Unseen Questions	Unseen Domains
Correct	2,008	233	301	1,917
Contradictory	499	58	64	417
Incorrect	2,462	249	368	2,228
Total	4,969	540	733	4,562
		5,835		
	10,804			

III. Models

A. Classical ML Models (Baseline)

In this study, we establish a baseline for our deep learning models using four popular classical ML models. These models require feature engineering, selection, and extraction. We use the TfidfVectorizer with the word analyzer to generate features. To reduce the dimensionality of the feature space, we remove all English stop words and lemmatize the remaining words. We also include uni-grams and bi-grams in our feature space but limit it to the top 10,000 features for training.

We experiment with two tree-based models: Decision Tree (DT) and Random Forest (RF). DT is a non-parametric model that generates parameters from the provided features and builds tree structures for the decision process. On the other hand, RF fits multiple decision trees on different subsets of the dataset and decides on a label by averaging over all decision trees. We use an RF model consisting of one hundred decision trees to estimate the labels.

Another model we explore is the Support Vector Machines (SVM), a supervised learning model that tries to define a hyperplane on the feature space that distinctly separates the data points. SVM has demonstrated promising results in various NLP tasks. We use the linear SVM classifier from Sci-Kit Learn, and we experiment with three C values, finding the best performance for $C = 1$ (see Table II).

Additionally, we investigate the performance of an artificial neural network (ANN) model, specifically, the Multilayer Perceptron (MLP) model, a fully connected feedforward neural network. We use ReLU as the activation function and the *Adam*

TABLE II
F1-SCORES OF LINEAR SVM FOR DIFFERENT C VALUES.

	Macro			Weighted		
Value of C	0.1	0.5	1.0	0.1	0.5	1.0
Unseen Answers (UA)	0.43	0.48	**0.51**	0.54	0.59	**0.61**
Unseen Questions (UQ)	0.32	0.34	**0.35**	0.42	0.45	**0.46**
Unseen Domains (UD)	0.34	0.36	**0.39**	0.37	0.45	**0.47**

optimizer with two hidden layers with one thousand and one hundred neurons, respectively, considering the complexity of the problem.

B. Large Language Models (LLMs)

Large Language Models (LLMs) such as RoBERTa Large have shown remarkable success in natural language processing tasks. RoBERTa Large is a pre-trained LLM that has been trained on a massive amount of unlabelled text data. Fine-tuning a pre-trained RoBERTa Large model on a specific task can lead to significant improvements in performance [6].

For this experiment, we employed the pre-trained RoBERTa Large model without any additional pre-training. The fine-tuning process for the RoBERTa Large model involved using the Adam optimizer with the following hyperparameters: a learning rate of 2e-5, an Adam epsilon of 1e-08, a batch size of 5, a warm-up step of 500, and a weight decay of 0.01. The model was fine-tuned for 20 epochs, and we recorded its performance at the end of each epoch. We fine-tuned a RoBERTa Large model solely on the SciEntsBank dataset which we will refer to as "RoBERTa Large". Additionally, we fine-tune another RoBERTa Large model on the Multi-Genre Natural Language Inference (MNLI) corpus, a widely used benchmark dataset for natural language inference, and subsequently on the SciEntsBank dataset. We will refer to this model as "RoBERTa Large MNLI".

C. Experimental Setup

To evaluate the performance of our models, we use the same three metrics as used by [7]: a) accuracy, b) macro-average F1, and c) weighted-average F1. Macro-average F1 calculates the average F1 score for all classes without considering the size of each class. Please note that the dataset is imbalanced as shown in Table I, with the *Contradictory* class having a significantly smaller number of samples compared to the other two classes. Weighted-average F1 considers the size of each class in the calculation and returns a balanced score over imbalanced datasets.

For classical ML models and evaluation metrics, we used the Sci-Kit Learn[1] library. We implemented the LLMs using PyTorch[2] and the Hugging Face Transformers[3] library.

[1]https://scikit-learn.org/
[2]https://pytorch.org/
[3]https://huggingface.co/

IV. RESULTS AND DISCUSSION

Table III presents the performances of four classical ML models: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) on three test sets. The table also includes the lexical baseline from SemEval-2013 Task 7 [7]. MLP outperforms all other classifiers on the Unseen Answers test set in all metrics. On the Unseen Questions test set, RF achieves the highest scores in all metrics while achieving the same macro-averaged F1 as the lexical baseline. On the Unseen Domains test set, RF yields the highest accuracy and weighted-average F1 while MLP returns the best macro-averaged F1 among our models but none of the models outperform the lexical baseline. Overall, MLP shows superior performance over other ML models for intra-domain classification whereas RF shows a high potential for transfer learning. We establish a new baseline for the LLMs by taking the maximum score per test set and metric—a method chosen to highlight the best potential performance of classical ML across different aspects—which is denoted as "Baseline" in Table IV.

Table IV presents the performance of LLMs on all test sets. The baseline indicates that the LLMs have performed significantly better. The SemEval 2013 Task 7 results are also presented in the table, which shows the maximum scores achieved by any model/algorithm in that competition. These scores serve as another point of comparison for the LLMs. The BERT-base fine-tuned by Sung et al. [8] outperforms the baseline in all test sets and metrics except UQ and UD on the accuracy, whereas the BERT-base fine-tuned by Zhu et al. [1] outperforms the baseline in each test set and metric. The BERT-base fine-tuned by Sung et al. [8] performs slightly better than the BERT-base fine-tuned by Zhu et al. [1] based on weighted-average F1 (0.68 vs 0.67). BERT-based DNN performs equally or better than both BERT-base models except for shows under-performance for UA and UD on macro-averaged F1 compared to the BERT-base fine-tuned by Sung et al. [8] (0.71 vs 0.72 and 0.56 vs 0.58, respectively). With a weighted average F1 of 0.69, BERT-based DNN outperforms both BERT-base models. RoBERTa Large underperforms compared to BERT-based DNN on UA and UQ but outperforms the model on UD in all metrics. RoBERTa Large MNLI outperforms all models on all test sets and all metrics. RoBERTa Large MNLI significantly improves upon the UQ and UD sets compared to any other models on any metrics. Notably, RoBERTa Large MNLI shows closely similar performance on the UQ and UD sets portraying little to no difference between unseen questions and domains while narrowing down the performance gap with UA.

The RoBERTa Large model fine-tuned by Camus and Filighera [9] on the MNLI corpus demonstrates similar performance to our own. While their model achieved higher performance for the UA set on all metrics, our model excels on the UQ set. Notably, both models exhibit almost identical performance on the UD set. We strongly attribute the performance differences to our choice of hyperparameters, emphasizing that batch size can significantly impact the scores as well. Importantly, our results not only validate their model

TABLE III
PERFORMANCE OF CLASSICAL MACHINE LEARNING MODELS. ACC: ACCURACY, M-F1: MACRO-AVERAGE F1, W-F1: WEIGHTED-AVERAGE F1.

Classifier	Unseen Answers			Unseen Questions			Unseen Domains		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
Lexical Baseline [7]	0.56	0.41	0.52	0.54	**0.39**	0.52	**0.58**	**0.42**	**0.55**
Decision Tree (DT)	0.65	0.57	0.64	0.48	0.33	0.44	0.48	0.34	0.43
Random Forest (RF)	0.63	0.54	0.62	**0.56**	**0.39**	**0.53**	0.55	0.37	0.52
Support Vector Machines (SVM)	0.63	0.52	0.61	0.47	0.34	0.46	0.50	0.38	0.47
Multilayer Perceptron (MLP)	**0.67**	**0.63**	**0.67**	0.38	0.33	0.38	0.47	0.39	0.47

TABLE IV
PERFORMANCE OF LARGE LANGUAGE MODELS (LLMS). ACC: ACCURACY, M-F1: MACRO-AVERAGE F1, W-F1: WEIGHTED-AVERAGE F1.

Classifier	Unseen Answers			Unseen Questions			Unseen Domains		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
Baseline	0.67	0.63	67	0.56	0.39	0.53	0.58	0.42	0.55
SemEval 2013 Task 7 (the best score per set and metric) [7]	0.72	0.65	0.71	0.66	0.47	0.63	0.64	0.49	0.62
BERT Base by Sung et al., 2019 [8]	0.76	0.72	0.76	0.65	0.58	0.65	0.64	0.58	0.63
BERT Base by Zhu et al., 2022 [1]	0.74	0.69	0.73	0.66	0.55	0.65	0.66	0.56	0.64
BERT-based DNN by Zhu et al., 2022 [1]	0.77	0.71	0.76	0.69	0.58	0.67	0.66	0.56	0.65
RoBERTa (Large + MNLI) by Camus and Filighera, 2020 [9]	0.79	0.78	0.79	0.66	0.66	0.66	0.72	0.71	0.72
RoBERTa Large	0.76	0.71	0.75	0.65	0.54	0.66	0.67	0.60	0.67
RoBERTa Large MNLI	0.77	0.74	0.77	0.72	0.67	0.72	0.72	0.70	0.72

TABLE V
THE IMPACT OF MNLI CORPUS AND TRANSFER LEARNING IN MODEL PERFORMANCE. CR: CORRECT, CN: CONTRADICTORY, AND IN: INCORRECT.

Classifier	Unseen Answers			Unseen Questions			Unseen Domains		
	CR	CN	IN	CR	CN	IN	CR	CN	IN
RoBERTa Large	0.77	0.58	0.78	0.68	0.24	0.70	0.67	0.40	0.72
RoBERTa Large MNLI	0.78	0.65	0.78	0.70	0.55	0.76	0.71	0.64	0.75
Improvement	0.01	**0.07**	0.00	0.02	**0.31**	0.06	0.04	**0.24**	0.03

performance but also reciprocally affirm our findings. It is crucial to emphasize that the objective of this experiment is not solely to validate their model's performance but rather to investigate the effect of fine-tuning on the MNLI corpus, shedding light on its influence on model performance and its utility in transfer learning—an aspect unexplored by Camus and Filighera [9].

Table V presents a performance comparison between RoBERTa Large and RoBERTa Large MNLI based on F1 scores. It is evident that fine-tuning the model on the MNLI corpus has significantly enhanced its capacity for semantic inference, resulting in a notable boost in performance, especially within the contradictory class. This improvement is most pronounced in the UQ and UD sets, with increases of 0.31 and 0.24, respectively. Additionally, the UA set demonstrates a respectable increase of 0.07.

Notably, the SciEntsBank dataset comprises a considerably smaller number of training samples within the contradictory class, accounting for only 10% of the dataset (see Table I). As a consequence, RoBERTa Large initially exhibits comparatively lower performance for *contradictory* when contrasted with both *correct* and *incorrect* across all test sets. However, the fine-tuning of the model on the MNLI dataset has played a pivotal role in enhancing its performance within this minority class through effective transfer learning.

V. CONCLUSION AND FUTURE WORK

The ability to comprehend and accurately categorize student answers is a pivotal aspect of automated scoring systems, and it holds great potential to assist students in their educational pursuits. However, this task presents challenges, given the diverse and intricate nature of student responses. Traditional rule-based approaches often struggle to capture the subtleties and nuances of language usage. In contrast, Large Language Models (LLMs) have exhibited exceptional performance in this domain compared to classical machine learning methods.

Through our experimentation, we have uncovered a significant enhancement in model performance by fine-tuning LLMs on the Multi-Genre Natural Language Inference corpus. This fine-tuning equips the model with a profound understanding of

semantic inference, thereby contributing to its improved performance. Our findings underscore the effectiveness of transfer learning, a critical component for tasks such as Automated Short Answer Grading, particularly for ensuring cross-domain adaptability and compliance.

In future research, this experiment can be expanded by including the SRA corpus, which incorporates the Beetle dataset alongside the SciEntsBank dataset. Notably, the SciEntsBank dataset, a part of the SRA corpus, also provides 5-way labeling. An intriguing avenue for investigation would be to discern which specific classes within the 5-way labeling scheme benefit the most from fine-tuning the model on the MNLI corpus. Furthermore, exploring the potential impact of fine-tuning the model on other similar corpora, such as Stanford Natural Language Inference (SNLI) and SciTail, in conjunction with MNLI, could lead to further improvements in model performance. In addition, considering the continuous advancements in language models, assessing the performance of more recent models like GPT-3 or GPT-4 on the SRA corpus presents another promising avenue for future exploration.

VI. Acknowledgement

We are grateful to the National Research Platform for providing access to the Nautilus HyperCluster, which was used for the computations presented in this paper.

References

[1] X. Zhu, H. Wu, and L. Zhang, "Automatic short-answer grading via bert-based deep neural networks," *IEEE Transactions on Learning Technologies*, vol. 15, no. 3, pp. 364–375, 2022.

[2] O. L. Liu, H.-S. Lee, C. Hofstetter, and M. C. Linn, "Assessing knowledge integration in science: Construct, measures, and evidence," *Educational Assessment*, vol. 13, no. 1, pp. 33–55, 2008.

[3] S. Bonthu, S. Rama Sree, and M. Krishna Prasad, "Automated short answer grading using deep learning: A survey," in *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*. Springer, 2021, pp. 61–78.

[4] S. Kumar, S. Chakrabarti, and S. Roy, "Earth mover's distance pooling over siamese lstms for automatic short answer grading." in *IJCAI*, 2017, pp. 2046–2052.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv*, 2019.

[7] M. O. Dzikovska, R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," University of North Texas, Tech. Rep., 2013.

[8] C. Sung, T. I. Dhamecha, and N. Mukhi, "Improving short answer grading using Transformer-based pre-training," in *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*. Springer, 2019, pp. 469–481.

[9] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*. Springer, 2020, pp. 43–48.

[10] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101

[11] M. O. Dzikovska, R. Nielsen, and C. Brew, "Towards effective tutorial feedback for explanation questions: A dataset and baselines," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 200–210.