

ULRR

Performance evaluation of the SBERT model for automatic short answer grading

Item Type	Meetings and Proceedings
Authors	Ahmed, Abbirah;Joorabchi, Arash;Hayes, Martin
Citation	Human-Centred AI Education & Practice Conference (HCAI-ep '23)
Rights	Attribution-NonCommercial-ShareAlike 4.0 International
Download date	2026-06-08 17:36:23
Item License	http://creativecommons.org/licenses/by-nc-sa/4.0/
Link to Item	https://hdl.handle.net/10344/31508

Performance Evaluation of the SBERT model for Automatic Short Answer Grading

Towards Enhancing Educational Assessment (Work in Progress)

Abbirah Ahmed
Department of Electronic and
Computer Engineering
University of Limerick
Ireland
abbirah.ahmed@ul.ie

Arash Joorabchi
Department of Electronic and
Computer Engineering
University of Limerick
Ireland
arash.joorabchi@ul.ie

Martin J. Hayes
Department of Electronic and
Computer Engineering
University of Limerick
Ireland
martin.j.hayes@ul.ie

ABSTRACT

Automated Short Answer Grading (ASAG) represents an actively researched domain within the field of automated assessment and blended learning. This study is based on the utilization of a Sentence-Transformer Model for ASAG and the assessment of the model's generalizability across a spectrum of publicly available datasets. These datasets encompass a wide range of subjects, including university-level computer science, Natural Language Processing, and science subjects spanning grades 3 to 8. Remarkably, the SBERT model demonstrated exemplary performance across all the datasets examined. However, it is important to fine tune the model on domain-specific data in order to make it efficient for a particular domain, enabling the model to acquire and comprehend the specialized vocabulary relevant to that domain. Additionally, we delve into key considerations essential for designing a domain-specific, yet generalized model for ASAG.

CCS CONCEPTS

• Computing Methodologies • Artificial intelligence • Natural Language Processing • Lexical Semantics

KEYWORDS

Natural Language Processing, Automated Assessment, Automatic Short Answer Grading

1 Introduction

Within the educational framework, written assessments play a critical role in assessing students' knowledge levels. Among the various assessment techniques available, written exams stand out as a foundational method for evaluating students' understanding. These exams incorporate a variety of question types, including

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Multiple Choice Questions (MCQs), short-answer questions, and essays, all serving as instruments to appraise student learning. According to the study [1], short answer questions help to demonstrate a better level of understanding as compared to the MCQs. While grading MCQs is a straightforward process, assessing short answers and essays is time-consuming and requires a contextual understanding and in-depth analysis of the content. The primary distinction between these evaluation methods lies in the grading criteria applied. Short-answer responses are graded based on their relevance to a predetermined reference or correct answer, with less emphasis on the quality of the written student response. Conversely, the grading of essays places a premium on the overall quality of the response, taking into consideration aspects such as spelling, grammar, and coherence, in addition to the content itself.

The grading of written assessments is a crucial task, especially in the large class sizes. With the evolution and integration of digital technologies into the education systems automatic grading has become one of the popular fields of active research among Artificial Intelligence and Natural Language Processing researchers.

An Automatic Short Answer Grading (ASAG) system compares student's responses with a reference/model answer for a given question using a machine learning model and that model assigns a mark automatically based on the result of the comparison [2, 3]. In the past, various ASAG systems have been proposed using various text similarity methods from classical text similarity techniques [2, 4-7] to state-of-the-art deep learning and large language models [8-17]. However, the development of a generalized short answer grading system, capable of efficiently assessing responses across diverse questions and fields of study, while simultaneously accommodating the heterogeneity in student responses—characterized by disparities in vocabulary, sentence structure, and grammatical errors—presents a formidable challenge.

In recent years, there have been significant advancements in the field of Natural Language Processing (NLP) and deep learning, leading to the introduction of several state-of-the-art language models, including BERT, RoBERTa, GPT-3, and GPT-3.5 (also

known as ChatGPT) [18]. A key feature of these models is their utilization of a transfer learning approach. Rather than training a model from scratch, these models can be fine-tuned on task-specific or domain-specific, relatively smaller datasets to effectively execute a variety of downstream tasks. Among these models, BERT has exhibited competitive performance in a range of text-based tasks, including tasks related to text similarity and text search.

This study represents a continuation of the work [18], in which authors leveraged various pre-trained SBERT models to address the ASAG task. The primary objective of this previous research was to assess the efficacy of pre-training data within these models when applied to domain-specific data, specifically for tasks related to text semantic similarity. Notably, the study identified an SBERT-based model as one of the highest-performing models, attributed to the relevance of its pre-training data. In our study, we adopted the same SBERT-based model to evaluate its performance across a range of datasets encompassing diverse domains and varying dataset lengths. This model is fine-tuned using these datasets and model performance is evaluated across those domains.

2. Methodology

2.1 Model

Sentence BERT framework and the Sentence-transformers library was introduced by Reimers and Gurevych [19], in 2019. This framework is a modification of the pre-trained BERT network, incorporating triplet and Siamese network architectures, with the specific aim of producing semantically meaningful sentence embeddings. These sentence embeddings are then compared using the cosine similarity measure. A wide variety of pretrained models are available in the Sentence-transformers library. Authors [18] assessed various pre-trained models based on SBERT and identified “All-distilroberta-v1”, as the best performing model for ASAG. This model is trained on a huge corpus of 1 billion sentence pairs collected from various sources such as wiki answers, yahoo search, and stack exchange. The same model is utilized in this study using a transfer learning approach. We finetuned the SBERT model on various datasets discussed in the next section and analyzed the performance of the model with respect to various factors such as the impact of the length of responses, size of dataset, and subject domain.

2.2 Datasets

In this research, five publicly available datasets were utilized, each comprising questions, reference answers, and student-provided responses in English language. Table 1 provides a list of these datasets along with a concise description for each.

The Mohler’s dataset is considered as the benchmark dataset, released in 2009 [3] and is based on the assignments of an

undergraduate course on data structures at University of Texas. It contains responses from 30 students for three assignments which consist of seven questions each. These answers are graded by two instructors independently and the average of their score is considered as the final score. In 2011, authors published extended dataset [2], containing student answers to 10 assignments and two exam papers. There were four to seven questions in each assignment, and 10 questions per exam paper. All the student answers are graded by two instructors.

Table 1. Datasets for ASAG

Dataset	No. of Responses	Score Range	Domain
ASAP ¹	17,043	0-3	Biology, Physics, Science, etc.
Mohler’s dataset ²	2558	0-5	Computer Science
Stita ³	333	0-1	Statistics
CU-NLP ⁴	171	0-100	Computer Science
Scientbank ⁵	139	0-1	Science

The ASAP dataset was released by the Hawlett Foundation on Kaggle for Automated Student Assessment Prize competition. The dataset encompasses responses for science subjects from students ranging from 8th grade to 10th grade, with response lengths not exceeding 50 words. Comprising 10 prompts, one for each question, the dataset comprises a total of 17,043 responses, which are assessed using grading scale ranging from 0 to 3. Similar to Mohler’s dataset, two independent instructors marked the responses and the average of the score provided by two independent instructors is taken as final score.

The CU-NLP dataset is generated from the final examination of the Natural Language Processing course at Cukurova University in 2019. This examination comprised two open-ended questions, to which 86 students provided responses. The reference answers for each question were curated by the course instructor, and subsequent grades of each student response were normalized to a scale ranging from 0 to 1.

The SciEntBank dataset is based on the science subjects from 3rd to 6th grade. In this study, a two-way dataset is used in which answers are labeled as either correct or incorrect. However, in their study, Gobbo et al. [20] undertook the process of standardizing the dataset to align it with other datasets. Two experts were asked to assess the responses, and the final score was determined by averaging the scores provided by these experts.

The Stita Dataset contains responses to 6 questions from statistics subject in higher education. Original dataset is in Italian language,

¹ <https://www.kaggle.com/c/asap-sas>

² https://github.com/dbbrandt/short_answer_grading_capstone_project

³ GitHub - edgresearch/dataset-automaticgrading-2022

⁴ <https://bmb.cu.edu.tr/uorhan/CuNLP.htm>

⁵ https://github.com/dbbrandt/short_answer_grading_capstone_project

however only the English translated version of the dataset is publicly available.

2.3 Evaluation metrics

Root Mean Square Error: It is a measure to calculate the error value between predicted and observed values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_p - x_o)^2} \quad (1)$$

Mean Absolute Error: It is the arithmetic average of the absolute difference between predicted and observed values:

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_p - x_o| \quad (2)$$

Pearson's correlation: In ASAG task, is used to measure the correlation between the marks assigned by the instructors and the marks predicted by the models:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (3)$$

where X and Y for two distributions, x_i and y_i are the i th value of distributions and \bar{x} and \bar{y} are the mean values for both distributions respectively.

3. Experimental setup

The SBERT models simultaneously process two sentences, thus we supplied the model with pairs of reference and student answers. While various studies employ different pre-processing techniques including tokenization [16], stopwords removal [21], and lemmatization or stemming, in our approach, we only used case normalization, as the tokenization process is inherently integrated into the SBERT framework. Moreover, based on the recommendation [19], original scores are normalized to the range 0-1. We finetuned the same model on all the datasets separately with batch size 16, 5 epochs and 500 evaluation steps. Once the model is trained on the respective data, sentence pairs were encoded into the embeddings which were then compared using the cosine similarity score. In the end a Linear Regression model is trained on the cosine similarity features to compare the results with the observed scores. For the model's performance evaluation, we used RMSE, MAE and Pearson's correlation for previously mentioned datasets.

4. Results and Discussions

The performance scores achieved by the SBERT model, finetuned on different datasets, are presented in Table 2.

In the table below, disparities in performance scores are evident, and these variations may be attributed to various factors. The model exhibited strong performance on the Stita dataset, while producing nearly equivalent results for the Mohler, CU-NLP, and SciEntBank datasets. It is interesting to note that model did not perform well on the largest dataset, namely ASAP, in comparison to the others.

Table 2. Performance Scores Achieved by SBERT Model

Study	Dataset	RMSE	MAE	Pearson's Correlation	Accuracy
[22]	Mohler	0.42	0.77	0.61	-
[23]	SciEntBank	-	-	-	0.80
[24]	Mohler	0.80	-	0.57	-
[15]	Mohler	0.88	-	0.66	-
[10]	Mohler	0.91	-	0.63	-
[25]	SciEntBank	-	-	-	0.79
[26]	SciEntBank	-	-	-	0.82
[17]	SciEntBank	-	-	-	0.73
[20]	CUNLP	0.81	0.64	-	-
[20]	ASAP	0.64	0.27	-	-
[20]	Stita	0.42	0.28	-	-
[18]	Mohler	0.69	-	0.81	-
This Study	Mohler	0.34	0.35	0.80	-
	ASAP	0.29	0.23	0.63	-
	SciEntBank	0.33	0.27	0.50	-
	CUNLP	0.21	0.15	0.54	-
	Stita	0.17	0.12	0.81	-

However, this performance of the model can be attributed to several factors:

1. Training Data Quality: It is important to have sufficient data points for each type of sentence pairs to ensure robust performance, an adequate representation of sentence pairs, ranging from incorrect to fully correct answers, is essential. Unfortunately, many datasets encounter a significant challenge in the form of class imbalance, a factor that can exert a substantial influence on the model's learning process.

2. Data Pre-Processing: It plays a crucial role in the context of Automated Short Answer Grading (ASAG). In this domain, it is imperative that student responses are meticulously prepared to ensure optimal alignment with reference answers. If the student's answer includes text that is not a part of the reference answer, such as segments of the question text, it can lead to reduced similarity scores. This, in turn, has the potential to impact on the overall performance of the ASAG model.

3. Data Size: For the model training the size of the dataset plays a crucial role to achieve the model's efficiency. The larger data size will provide a better and diverse range for the model to better learn the underlying patterns. Moreover, it can also help models to adapt the complexities of the task more effectively.

Apart from above, a more important observation pertains to the model's consistently impressive performance. This can be attributed to the inherent alignment between the pre-training data utilized by the model and the specific task for which the model was purposefully designed.

From the foregoing observations, it becomes evident that for a generalizable Automated Short Answer Grading (ASAG) model, the curation of a training dataset encompassing a wide array of study materials from diverse domains or subjects is imperative. However, constructing such an extensive system presents great challenges, including the substantial data requirements, privacy, biasness, and associated costs.

A potential solution to address the challenge of training data involves a domain-specific approach. For instance, in the case of an ASAG model tailored for higher education in the Computer Science domain, the training dataset should be thoughtfully curated from resources such as relevant textbooks, reading materials, research articles, or even pertinent public forums. This focused approach allows for a more efficient and contextually accurate training process, ultimately yielding a specialized model that excels in its designated field of study.

5. Conclusion

In this study, we evaluated the performance of the SBERT model on various datasets publicly available for ASAG task. These datasets are different from each other in terms of subjects, sizes, and response lengths. We observed variations in the model's performance for each dataset based on the previously mentioned factors. It can be concluded that model's performance can be improved if larger and domain-specific datasets could be used, that will eventually help instructors to ease the burden. Additionally, to ensure fairness and equity in ASAG, it's crucial to address biases by carefully selecting training data and features by continuously monitor and correct biases throughout the system's lifespan.

REFERENCES

- [1] Zhang, L., Huang, Y., Yang, X., Yu, S. and Zhuang, F. An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 30, 1 (2022/01/02 2022), 177-190.
- [2] Mohler, M., Bunescu, R. and Mihalcea, R. *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments*. City, 2011.
- [3] Mohler, M. and Mihalcea, R. *Text-to-text semantic similarity for automatic short answer grading*. City, 2009.
- [4] Gomaa, W. H. and Fahmy, A. A. Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3, 11 (2012).
- [5] Sultan, M. A., Salazar, C. and Sumner, T. *Fast and easy short answer grading with high accuracy*. City, 2016.
- [6] Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R. and Sengupta, B. *Sentence level or token level features for automatic short answer grading?: Use both*. Springer, City, 2018.
- [7] Roy, S., Dandapat, S., Nagesh, A. and Narahari, Y. *Wisdom of students: A consistent automatic short answer grading technique*. City, 2016.
- [8] Magooda, A. E., Zahran, M., Rashwan, M., Raafat, H. and Fayek, M. *Vector based techniques for short answer grading*. City, 2016.
- [9] Surya, K., Gayakwad, E. and Nallakaruppan, M. Deep learning for short answer scoring. *Int. J. Recent. Technol. Eng. (IJRTE)*, 7, 6 (2019).
- [10] Gomaa, W. H. and Fahmy, A. A. *Ans2vec: A Scoring System for Short Answers*. Springer International Publishing, City, 2020.
- [11] Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V. and Arora, R. *Pre-training BERT on domain resources for short answer grading*. City, 2019.
- [12] Sung, C., Dhamecha, T. I. and Mukhi, N. *Improving short answer grading using transformer-based pre-training*. Springer, City, 2019.
- [13] Gong, T. and Yao, X. An attention-based deep model for automatic short answer score. *International Journal of Computer Science and Software Engineering*, 8, 6 (2019), 127-132.
- [14] Xia, L., Guan, M., Liu, J., Cao, X. and Luo, D. *Attention-Based Bidirectional Long Short-Term Memory Neural Network for Short Answer Scoring*. Springer, City, 2020.
- [15] Prabhudesai, A. and Duong, T. N. B. *Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression*. City, 2019.
- [16] Sasi, Nair, D. and Paul Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading. *arXiv pre-print server* (2020-09-02 2020).
- [17] Ghavidel, H. A., Zouaq, A. and Desmarais, M. C. *Using BERT and XLNET for the Automatic Short Answer Grading Task*. City, 2020.
- [18] Ahmed, A., Joorabchi, A. and Hayes, M. J. *On the Application of Sentence Transformers to Automatic Short Answer Grading in Blended Assessment*. IEEE, City, 2022.
- [19] Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [20] Del Gobbo, E., Guarino, A., Cafarelli, B. and Grilli, L. GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. *Knowledge and Information Systems* (2023), 1-40.
- [21] Tulu, C. N., Ozkaya, O. and Orhan, U. Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. *IEEE Access*, 9 (2021), 19270-19280.
- [22] Kumar, S., Chakrabarti, S. and Roy, S. *Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading*. City, 2017.
- [23] Saha, S., Dhamecha, T. I., Marvaniya, S., Foltz, P., Sindhgatta, R. and Sengupta, B. Joint multi-domain learning for automatic short answer grading. *arXiv preprint arXiv:1902.09183* (2019).
- [24] Hassan, S., A. A. and El-Ramly, M. Automatic Short Answer Scoring based on Paragraph Embeddings. *International Journal of Advanced Computer Science and Applications*, 9, 10 (2018).
- [25] Camus, L. and Filighera, A. *Investigating transformers for automatic short answer grading*. Springer, City, 2020.
- [26] Lun, J., Zhu, J., Tang, Y. and Yang, M. *Multiple data augmentation strategies for improving performance on automatic short answer scoring*. City, 2020.