

Automatic Short Answer Grading on High School's E-Learning Using Semantic Similarity Methods

Daniel Wilianto¹, Abba Suganda Girsang¹

¹ Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Abstract– Grading students' answers has always been a daunting task which takes a lot of teachers' time. The aim of this study is to grade students' answers automatically in a high school's e-learning system. The grading process must be fast, and the result must be as close as possible to the teacher assigned grades. We collected a total of 840 answers from 40 students for this study, each already graded by their teachers. We used Python library sentence-transformers and three of its latest pre-trained machine learning models (all-mpnet-base-v2, all-distilroberta-v1, all-MiniLM-L6-v2) for sentence embeddings. Computer grades were calculated using Cosine Similarity. These grades were then compared with teacher assigned grades using both Mean Absolute Error and Root Mean Square Error. Our results showed that all-MiniLM-L6-v2 gave the most similar grades to teacher assigned grades and had the fastest processing time. Further study may include testing these models on more answers from more students, also fine tune these models using more school materials.

Keywords– automated grading, sentence-transformers, machine learning, cosine similarity, mean absolute error, root mean square error

DOI: 10.18421/TEM121-37

<https://doi.org/10.18421/TEM121-37>

Corresponding author: Daniel Wilianto,
Computer Science Department, Binus Graduate Program –
Master of Computer Science, Bina Nusantara University,
Jakarta, Indonesia.


Email: daniel.wilianto@binus.ac.id

Received: 17 September 2022.

Revised: 16 November 2022.

Accepted: 24 November 2022.

Published: 27 February 2023.

 © 2023 Daniel Wilianto & Abba SugandaGirsang; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

1. Introduction

The development of information technology over the past decade has brought major changes to all fields of work and aspects of human life. All fields of work have utilized information technology to improve the quality of their services, including education. The implementation of the teaching and learning process, giving out and collecting school assignments, administering exams, all can now be done remotely with the help of technology.

One of the most important parts of the teaching and learning process is the assessment of student works to find out how much knowledge has been successfully absorbed by the students and how far they understand the topic being taught. This assessment process is often a tedious, monotonous, and time-consuming process for the teacher. In the private school where the corresponding author works, a teacher can teach 5 classes in one level, each class has 40 students on average. So, a teacher must check the work of 200 students on average every time he gives an exercise.

According to [1], the use of computer technology can reduce the burden on teachers and allow teachers focus on the human side. Currently, teachers still spend a lot of time to grade exams and school assignments. This repetitive task reduces teachers' time for teaching, research, self-development, and interaction with students.

Grading answers of objective questions, such as multiple choice, true or false questions, and filling in the blanks by choosing from the list of words that have been provided; is very easy for software to do, because there is only one correct/absolute answer. The problem is that objective questions cannot accurately assess students' knowledge and understanding, because there is a luck factor when students are only required to choose a correct answer among several choices. Subjective questions, such as questions that require short answers, are the best choice to test students' understanding and knowledge on the subject. To write short answers, students have

to put their thoughts into their own sentences. However, most school exam questions have more multiple-choice questions and fewer question-answers, because assessing students' short answers takes up much more time, also it has consistency problems due to the possibility of a large variety of answers. Automatic assessment by software algorithms, especially with artificial intelligence can be a solution for this short answer assessment process [2].

People can distinguish objective questions and subjective questions easily; but it may be more difficult to distinguish between short answer questions and essay questions. To be able to distinguish them, there are five minimum criteria for a question to be considered a short answer question. First, the question must be answered by requiring external knowledge; that is the answer must not be contained in the question. Second, the question must be answered with natural language (language that is created naturally because of the evolution of communication between humans). Third, the answer to the question must be at least one phrase long and a maximum of one paragraph long. Fourth, the assessment of the answers to these questions must be based on the content, not the writing style. Fifth, the answer to the question may not be a simple yes or no, and it may not ask students' personal opinions [3].

Automatic Short Answers Grading as described above is the focus of research in this thesis. In making an ASAG system, there are 4 components that must be considered, namely algorithms, technology, datasets, and evaluation techniques used. The four components are equally important and determine the effectiveness of the system, as shown in Figure 1.

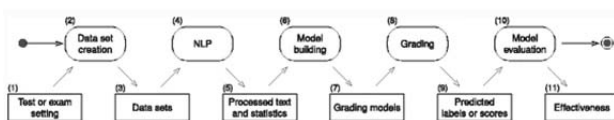


Figure 1. Processes in an ASAG system

The widespread use of this system will greatly facilitate teachers and students, especially during the pandemic where assignments and exams are carried out online using the school's e-Learning system. On the student side, the use of an automated grading system will allow students to get feedback and assessment results immediately after the exam ends, without waiting for the teacher. On the teacher's side, the use of an automated grading system would free them from the time-consuming and error-prone assessment process. According to [4], students often have to wait a long time to wait for the results of their exam assessments, due to manual assessments by teachers; and after a long wait, they sometimes

got different scores when to other classmates who had almost the same answer.

The aim of our study is to find the best machine learning models to be used on real high school students' answers. The grading process has to have both good processing speed and grade accuracy. Speed is important, so that our students can see the grades of their answers as soon as possible after submitting their works to the school's e-learning system. Accuracy is also important. The usage of software assures consistency, since computer codes are static unlike human mind, but we need to know if software can really replace our teachers in grading our students' answers.

2. Literature Review

We are reviewing some papers related to the application of artificial intelligence in education and study related to automatic short answer grading in this section.

2.1. Application of Artificial Intelligence in Education

According to [5], the use of Artificial Intelligence in everyday life has increased exponentially. We use AI services when searching on Google. We use AI services when communicating with the iPhone assistant, Siri. We also trust AI to handle our personal data, our medical data, and our financial data. Why don't we entrust our children's education to AI?

A few companies such as Carnegie Learning and Content Technology, have used AI to help with learning, administer exams, and receive feedback from students, from preschool to college level. There are also companies that use AI to study student textbooks and pinpoint the key areas that need attention the most. This AI also generates practice questions for students automatically [6].

Meanwhile, according to [7], a system that can automatically assess student answers (non-multiple choice) has been explored by humans for more than a decade. The proposed approach includes grouping students' answers into groups of similar answers and assigning scores to the entire group instead of individual answers, scoring based on manually constructed rules or ideal answer models and automatically assigning scores based on the semantic similarity of answers to answers, given reference.

According to [8], the semantic similarity measurement algorithm consists of 2 main groups, namely corpus-based and knowledge-based. The corpus-based algorithm processes a collection of text documents, extracts information from them and uses the information obtained to determine the similarity

between text elements (words, phrases). Knowledge-based algorithms obtain semantic similarity by using information from a semantic repository (ontology, semantic network).

Meanwhile, according to [9], the answer assessment technique has 5 main groups, namely Natural Language Processing (NLP), Information Extraction and Pattern Matching, Machine Learning, Document Similarity, and Clustering.

2.2. Study Related to Automatic Short Answer Grading

[10] used machine learning models BERT and XLNET in their study on creating an automated grading system for short answers. For the dataset, they used SemEval 2013 which was deliberately designed by the Association for Computational Linguistics as processed data for use by natural language processing researchers. They compared the results of their research with other papers and found that using machine learning models BERT and XLNET for short answer assessment had results that were equivalent to or slightly better than other popular methods with the same dataset.

[11] in their study tries to propose an automatic short answer grading system that is claimed to be fast and simple. They added Question Demoting and Term Weighting techniques to the Python's scikit-learn model to increase the accuracy of the assessment. For the dataset, they also used SemEval 2013 where they only took questions and answers in 15 science areas. The result is an automatic short answer grading system with a claimed pretty good performance, where their system manages to grade an average of 33 answers per minute on a computer with a CPU speed of 2.25 GHz. Unfortunately, the term weighting technique is admittedly less influential because most of the answers from the dataset used in this study only consist of a few words so that there is no difference between keywords and words that are less important. Further research is needed using datasets that have longer answers.

[12] used three different frameworks for the embedding process in their study: SBERT, Word2Vec, and Bag of Words. For the dataset, they used the Berkeley Evaluation and Assessment Research (BEAR) Center. Datasets are processed in the program using various combinations, some process only used answers, while some process used questions and answers at the same time. The result is SBERT consistently beats Word2Vec and Bag of Words. The weakness of this study that the authors themselves admitted is that they haven't tested it on real world scenario to see if it produces the same result.

[13] developed their own algorithm in their study on Automatic Short Answer Grading, where they combined the sentence-to-sentence similarity method (embeddings using InferSent) with a token processing algorithm which was coded by themselves. There are three datasets used, namely the Large-Scale Industry Dataset, SemEval 2013, and Mohler. As the result, the combination of these two methods was claimed to have accuracy that rivals and even exceeds state of the art methods, especially in the answers that use too many paraphrases / strays far from the question domain. They only showed accuracies in this study, though. They didn't display how fast their program execution time is compared to state-of-the-art methods.

3. Materials and Methods

We are explaining where and how we get the materials for this study in this section. Then we explain how we perform automatic grading on the collected students' answers.

3.1. Materials

The questions and answers for this study were obtained from Immanuel Christian Junior High School located in Pontianak City, Indonesia. There are a total of 21 questions, each answered by 40 students, resulting in a collection of 840 answers. Each answer was graded by the teacher already, ranging from zero (for answers that were totally wrong) to one (perfect answers). The teacher had also provided the correct answer for each question. The subject of these questions is English, where the students were required to answer questions in short sentences after reading some short stories in English.

Table 1 shows some examples of these answers and their grades. The question is "Why did she knock at the door?" and the correct answer which provided by the teacher is "She knocked at the door because she reasoned that someone could require assistance.". Be informed that we left the grammatical errors which were made by the students the way they were originally written on purpose, because that's how they are treated in real world scenario.

Table 1. Examples of Students' Answers and Their Grades

Student's Answer	Teacher's Grade
She knocked at the door because she needed assistance.	0.7
She knock the door because she thought there's someone that need help.	1
She knocked at the door because she heard a loud crash followed by tears.	0.6
She knocked at the door because she heard yelling coming from the last office, which	1

was located at the end of the corridor and there was a loud crash, followed by tears. She reasoned that someone could require assistance.	
She knocked on the door because he heard screaming from the office at the end of the corridor.	0.6
because she reasoned that someone could require assistance.	1
She knocked at the door because she hear yelling coming from the last office, Then there was a loud crash, followed by tears.	0.6
She knock at the door because she hear there was a loud crash and followed by tears.	0.6
She knocked at the door because she hear a loud crash, followed by tears and she reasoned that someone could require assistance.	1
She knock at the door because there was a loud crash, followed by tears and she think someone coildreyquireassitance.	0.9

For the grading task, we used a standard PC with 6 Gigabytes of RAM and Intel Core i3 processor clocked at 1.70 Gigahertz. It used an SSD for data storage, with a maximum read and write speed of 500 Megabytes per second. The operating system used was Microsoft Windows 10 version 21H1. The school materials were stored in MySQL server version 5.7.39. We had also installed Python version 3.9 which was required to write codes and run the methods which are described in the next section.

3.2. Methods

Prior to the year 2017, methods for measuring semantic similarity of texts were mostly based on Recurrent Neural Network (RNN) algorithms. RNN is a form of Artificial Neural Networks (ANN) architecture which is specially designed to process sequential data. RNN has vast applications, including natural language processing (NLP), speech recognition, machine translation, character-level language modeling, image classification, image captioning, stock predictions, and financial engineering. Long Short-Term Memory, or LSTM was perhaps the most successful and famous RNN because it overcomes the problems of training a recurrent network.

This situation changed when [14] presented their Transformer Model during the 31st Conference on Neural Information Processing Systems, which was held on Long Beach Convention Center, California in December 2017. This first Transformer Model was explained in their study paper “Attention is All You Need”. Ever since then, natural language processing had mostly moved to variants of Transformer Model.

The most notable Transformer Model is BERT. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based machine learning technique for natural language processing pre-training which was developed by Google. BERT was created and published by [15].

The interesting thing is, for many tasks, the latter parts of these transformer models are the same as those in RNNs — often a couple of feedforward NNs that output model predictions. It’s the input to these layers that changed. The dense embeddings created by transformer models are so much richer in information that we get massive performance benefits despite using the same final outward layers. These increasingly rich sentence embeddings can be used to quickly compare sentence similarity for various use cases.

As for our study, which is performing Automatic Short Answer Grading on high school students’ answers, we used sentence-transformer models. The first sentence-transformer model which was introduced by [16] in their paper was called SBERT. SBERT was a modification of the pretrained BERT that use Siamese and triplet network structures to derive semantically meaningful sentence embeddings, that can be then compared using cosine similarity. This reduces the effort for finding the most similar pair of 10,000 sentences from 65 hours with BERT to about 5 seconds while maintaining the accuracy of BERT. This sentence-transformer model has been archived by the time we conducted our study, and had been replaced by newer, state-of-the-art sentence-transformer models.

We picked three pre-trained models to be used on our study, which were recommended by the official website of sentence-transformer itself: all-mpnet-base-v2, all-distilroberta-v1, and all-MiniLM-L6-v2. These are all-purpose models which are suitable for performing sentence embeddings on our school materials.

The stages of our study are:

- i. Teacher-provided answers, students’ answers and teacher’s grades for each student answers were stored in database.
- ii. Sentence-transformer models performed sentence embeddings for each teacher-provided answers and students’ answers.
- iii. Computer grades were calculated using cosine similarity formula. Cosine similarity formula result ranges from 0 to 1, just like teacher’s grades. They were calculated by pairing sentence embedding values from each student’s answer and teacher-provided answer which referred to the same question.
- iv. Computer grades were stored on different columns on database depending on which

sentence-transformer models performed the sentence embeddings.

- v. The time which was needed by each model was also recorded.
- vi. Mean Absolute Error and Root Mean Square Error were calculated using each set of computer grades against the grades that teacher assigned to each student’s answers, to find out how close they were.

Cosine similarity is calculated using equation (1).

$$\text{Sim}(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

Where A_i are the values of the first sentence’s embeddings and B_i are the values of the second sentence’s embeddings, which were all produced by sentence-transformer models.

Mean Absolute Error is calculated using equation (2).

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (2)$$

While Root Mean Square Error is calculated using equation (3).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \quad (3)$$

In both equation (2) and equation (3), y_i are the grades given by computer software, x_i are grades given by a human teacher, and n are the number of data being compared.

4. Analysis Results

Table 2 shows the results of our study. The first column shows the name of the pre-trained sentence-transformer model which performed the sentence embeddings. The second column and the third column show the Mean Absolute Error and Root Mean Square Error respectively, which were obtained by calculating the computer grades against teacher-assigned grades. The fourth column shows the time needed in seconds to grade all 840 student answers by going through all stages described above using each different model.

Table 2. The Study Result of Each Sentence-Transformer Models

Pre-Trained Model Name	MAE Value	RMSE Value	Time Needed in Seconds
all-mpnet-base-v2	0.17420	0.21907	144.42955
all-distilroberta-v1	0.17263	0.22004	78.84022
all-MiniLM-L6-v2	0.16625	0.21295	31.25612

We are able to see from Table 2 that all-MiniLM-L6-v2 has the best performance in our study. It allowed us to grade 840 answers automatically in 31 seconds, and it also has the lowest MAE values compared to the other models. That means the grades it produced are the closest to our teacher-assigned grades. Compared to MAE, its’ RMSE value may have smaller differences to the other two models, but it still has the lowest error value.

We looked further into the individual grades which were given by each processes using different models, to find out why can’t the error values be lower. Table 3 shows some of the most notable cases, when we compare the grades that our software gave and the grades that the teacher gave. Be informed that the correct answer is supposed to be “She worked six hours a week.”; as provided by the teacher.

Table 3. Computer-Produced Grades Compared to Teacher-Assigned Grades

Student’s Answer	Teacher	all-mpnet-base-v2	all-distilroberta-v1	all-MiniLM-L6-v2
She works a couple of hours a week.	0.9	0.731122	0.774580	0.894837
She worked 6 hours in a week.	1	0.964758	0.980591	0.982214
In a week she works for 6 hours.	1	0.804523	0.874066	0.886175
She work 6 hours a week.	1	0.851350	0.888775	0.926250

As observed on Table 3, our human teacher seemed to be quite lenient to his students, giving very good grade (0,9) to an answer which was clearly wrong (calling it a couple of hours instead of six hours) and perfect grades (1) on answers which were wrong grammatically. Coincidentally, the grades which were obtained using all-MiniLM-L6-v2 model for the sentence embedding’s part were all higher than the others’ grades. Which in the end, made all-MiniLM-L6-v2 the best sentence-transformer model to be used on high school student answers, in our study.

5. Conclusions

Our paper showed how computer software can be used to perform automatic grading on high school students’ short answers. We have also discussed the state-of-the-art methods for measuring sentence to sentence similarity from time to time; and applied the best current method in our study. Our results show that current technology is already reliable enough to

grade students' answers in place of human teachers. We also need to make a statement that computer assigned grades will never have zero MAE or RMSE values when measured against teacher-assigned grades anyway, in any future study. As shown on our results discussion, the problem lies with the humans. As normal humans, our teachers are imperfect beings; they gave good scores to students out of kindness sometimes; and they might accidentally made mistakes while grading students' answers.

For our future work, more students' answers and teacher-assigned grades may be collected and processed on the software we have prepared, to assure which model is the best for grading short answers automatically. Also we can move to newer, better sentence to sentence similarity measuring methods when they come out. Computer technology keeps improving after all.

Acknowledgement

We are thankful for IMMANUEL Christian Junior High School for providing us with the materials for this study, especially for the cooperation of the headmaster and the English teacher. We are also thankful for BINUS University Graduate Program for supporting this study, especially for all the lecturers who have given their critics and suggestions during the making of this paper.

References

- [1]. Huang, J., Saleh, S., & Liu, Y. (2021). A review on artificial intelligence in education. *Academic Journal of Interdisciplinary Studies*, 10(3), 206-206.
- [2]. Lubis, F. F., Putri, A., Waskita, D., & Sulistyanningtyas, T. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. *International Journal of Technology*, 12(3), 571-581.
- [3]. Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60-117.
- [4]. Galhardi, L. B., & Brancher, J. D. (2018, November). Machine learning approach for automatic short answer grading: A systematic review. In *Ibero-american conference on artificial intelligence* (pp. 380-391). Springer, Cham.
- [5]. Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour*, 1(3), 1-3.
- [6]. Kengam, J. (2020). Artificial Intelligence in Education. *Science and Technology Department Bournemouth University*. United Kingdom.
- [7]. Filighera, A., Steuer, T., & Rensing, C. (2020, July). Fooling automatic short answer grading systems. In *International conference on artificial intelligence in education* (pp. 177-190). Springer, Cham.
- [8]. Rozeva, A., & Zerkova, S. (2017, December). Assessing semantic similarity of texts—methods and algorithms. In *AIP Conference Proceedings* (Vol. 1910, No. 1, p. 060012). AIP Publishing LLC.
- [9]. Hasanah, U., Permanasari, A. E., Kusumawardani, S. S., & Pribadi, F. S. (2019). A scoring rubric for automatic short answer grading system. *Telkonnika (Telecommunication Computing Electronics and Control)*, 17(2), 763-770.
- [10]. Ghavidel, H. A., Zouaq, A., & Desmarais, M. C. (2020). Using BERT and XLNET for the Automatic Short Answer Grading Task. In *CSEDU (1)* (pp. 58-67).
- [11]. Sultan, M. A., Salazar, C., & Sumner, T. (2016, June). Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1070-1075).
- [12]. Condor, A., Litster, M., & Pardos, Z. (2021). Automatic Short Answer Grading with SBERT on Out-of-Sample Questions. *International Educational Data Mining Society*, 345-352.
- [13]. Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018, June). Sentence level or token level features for automatic short answer grading?: Use both. In *International conference on artificial intelligence in education* (pp. 503-517). Springer, Cham.
- [14]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All you Need 31st Conference on Neural Information Processing Systems (NIPS 2017). *Long Beach, CA, USA*, 1-11.
- [15]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).
- [16]. Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992).