CIT. Journal of Computing and Information Technology, Vol. 30, No. 2, June 2022, 85–99
doi: 10.20532/cit.2022.1005478

Development of an Automated Scoring Model Using SentenceTransformers for Discussion Forums in Online Learning Environments

Bachriah Fatwa Dhini and Abba Suganda Girsang

Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Due to the limitations of public datasets, research on automatic essay scoring in Indonesian has been restrained and resulted in suboptimal accuracy. In general, the main goal of the essay scoring system is to improve execution time, which is usually done manually with human judgment. This study uses a discussion forum in online learning to generate an assessment between the responses and the lecturer's rubric in the automated essay scoring. A SentenceTransformers pre-trained model that can construct the highest vector embedding was proposed to identify the semantic meaning between the responses and the lecturer's rubric. The effectiveness of monolingual and multilingual models was compared. This research aims to determine the model's effectiveness and the appropriate model for Automated Essay Scoring (AES) used in paired sentence Natural Language Processing tasks. The *distiluse-base-multilingual-cased-v1* model, which employed the Pearson correlation method, obtained the highest performance. Specifically, it obtained a correlation value of 0.63 and a mean absolute error (MAE) score of 0.70. It indicates that the overall prediction result is enhanced when compared to the earlier regression task research.

ACM CCS (2012) Classification: Computing methodologies → Modeling and simulation → Model development and analysis → Model verification and validation

Applied computing → Education → Distance learning

Keywords: Automatic Essay Scoring, Discussion Forum, SentenceTransformers, Monolingual Model, Multilingual Model

1. Introduction

Online education is designed as a service for lecturers in the distance learning system. Principally, distance learning is online learning [1]. It is designed to facilitate the asynchronous interaction between lecturers and students through the Learning Management System (LMS) platform, which is used to display the distance learning system [2]. The features offered by the LMS include discussion forums, inbound messages, coursework, and learning administration tools such as scheduling and learning histories. These systems also provide teaching resources.

Aside from particular assignments, one of the assessment metrics in online learning modes is discussion scoring [3]. During online learning, assessments include 30% of the final score and discussion forums 70% of the final examination. In general, responses to questions in discussion topics are written in short fragments, shorter answers, or even paragraphs in forum discussions. Thus, scoring of each of these items is still done manually. The influencing factors in evaluating posts accurately and precisely on discussion forums include the complexity and difficulty of the category [4].

Rubrics are used as a reference in discussion forums; the handcrafted system is closely related to the scoring rubric [5], which has been designed as a criterion for assessing essays. Manual assessments use rubrics to obtain appro-

priate essay scores and consistency in assessment. Since human evaluations are expensive, time-consuming, and prone to subjective bias, automated assessments have sparked the interest of researchers. Rubric scores are significant for automated essay scoring [5]. Taghipour's doctoral dissertation utilizes iterative neural networks to improve accuracy in predicting holistic scores and applying rubric assessments, including organization and strength of argument [6].

Automated essay scoring has been presented as a quick, efficient, and cost-effective solution to the issue of students' essay scoring because text-based grading exams often require a large workforce, substantial training, and skills in evaluating reliable marks in general. Automated Essay Scoring (AES) is a system that achieves the same level of agreement with human graders as with one another [7]. The concept of essay scoring in a discussion forum uses pairwise sentences to figure out semantic similarity, which compares responses to a discussion and the reference answers from the lecturer.

Research on AES has been widely conducted in optimizing AES performance by implementing deep learning that has revolutionized Natural Language Processing (NLP). Modernization of language models pre-trained on a vast scale was done utilizing unlabeled text datasets. Transfer learning has achieved good performance even when labeled data is scarce. A sentence transformer based on BERT is a state-of-art text embedding that constructs text vector representations of the highest quality through the Python framework. It has produced significant breakthroughs in NLP tasks [8].

An organization named the Ubiquitous Knowledge Processing (UKP) Lab led many pre-trained SentenceTransformers models. However, most of the efforts have been focused on English and rarely on Indonesian [9]. Reimers Expanding the SentenceTransformers model includes two fine-tuning stages, following the teacher and student model approach. The student model is pre-trained on the multilingual model known as the distillation of multilingual knowledge. Authors in [9] state that monolingual as well as multilingual models create aligned vector spaces wherein closely related inputs from several languages are mapped. The models are trained on 50+ languages, including Indonesian (id).

Although AES has been extensively explored, only a few research efforts have recently been done studying Indonesian. Research progress on this language in NLP is slow due to the lack of available resources. Most of the methods related to AES in English are supported by many publicly available datasets such as the Automated Student Assessment Prize (ASAP), SemEval STS, Quora-QP, *etc.* Therefore, most of them are pre-trained in English, in contrast with the Indonesian language dataset.

This study presents monolingual and multilingual models. It aims to determine the effectiveness of the models and find out which model can be appropriately used in paired sentence NLP tasks for AES through dataset sources on response discussion forums in Indonesian having in mind the overall processing time. Consequently, comparing monolingual and multilingual models for single-language tasks can help researchers decide whether training a model for a target language is worthwhile or whether a multilingual model is sufficient. This study is organized as follows. The next section shows a brief literature review of AES in Indonesian and a pre-trained model from SentenceTransformers. Section 3 depicts the methodology of the study. Finally, results and discussion are provided in Section 4.

2. Literature Review

Replacing human graders with automated scoring is often challenging. Some suggested automated essay-scoring techniques have been well-studied to overcome difficult circumstances. Automated identification of potential human grader faults is one of 'AES's essential functions. The study of AES related to semantics has attracted some researchers in this last decade and has grown fast. The first AES originated from Allis Batten Page with Project Essay Grade (PEG) in 1968. Subsequently, [10] proposed an Intelligent Essay Assessor that used Latent Semantic Analysis (LSA) to calculate the semantic similarity value between texts. Some performed semantic correlation studies for essay scoring to obtain good performance [11]–[13]. Other studies have utilized a much

more sophisticated approach utlizing LSTM neural networks. Such approaches were tested on essay scoring, obtaining 78% accuracy [14], and [15] on short answers, obtaining almost a perfect accuracy of 93%. Most AES research uses the Automated ASAP [16] as experimental data. ASAP has been tested with modern language models like LSTM, BERT, and derivation BERT, such as XLNet [17], DistillBERT, Roberta, Mobile BERT [18], *etc.*

Although it is still limited and the assessment results are not good enough, several AES studies in Indonesia have followed the development of the NLP method, starting from the previous methods using more traditional algorithms to the state-of-the-art ones. Traditional algorithms like the Jaccard Coefficient are also utilized for these purposes. With the parameters of sentence similarity and keywords, [19] proposed automatic assessment using the Longest Common Subsequence (LCS), Cosine Coefficient (CC), Jaccard Coefficient (JC), and Dice Coefficient (DC) methods to assess the essay model for short answers. Combining two sentence similarity values and keywords can increase the correlation. Using the same data, [20] explored automated essays scoring by removing the semantic meaning of short answers. The apprach generates word embedding using continuous bag-of-words (CBOW). In [20] the authors proposed a solution to increase the Pearson correlation score by 0.5 and reduce the MAE error value by 0.24 (from 0.94 to 0.70).

Latent Semantic Analysis (LSA) is used to evaluate the proposed short answer scoring, classify it using Support Vector Machine (SVM) based on the subject, and remove unrelated answers [21]. The average accuracy of this approach is 72.01%. The same algorithm implementing essay assessment on e-learning platforms with short answer types proposes LSA to find meaning or document concepts by comparing semantic similarities [22]. Paired sentences between students' answers and the teacher's answer key are then applied to an additional dictionary for synonyms. The result of the accuracy of this system is 83.3%.

Through the most popular RNN LSTM, the proposed method increased Quadratic Weighted Kappa (QWK) and accuracy by 2.38% and 2.05%, respectively. The order distribution does not capture the sentence semantics for the Indonesian short essay tasks. Indonesia's revolutionary short essay scoring technique employs the transfer learning dependency tree LSTM. Additionally, it proves that comparisons of LSTM using no learning transfer get 48.26% QWK and 64.58% QWK with LSTM using learning transfer.

In 2019, the Gadjah Mada University Research Group released the Ukara Automatic Short Answer System and the Ukara Automatic Essay Scoring Challenge (Ukara 1.0 Challenge). By [23], the UKARA dataset was utilized using single, ensemble, and deep learning (LSTM). Ukara has two datasets, A and B, published publicly to classify the short essay scoring category. The label score used is true and false. The study on Ukara has been conducted and followed the development of the NLP algorithm. For dataset A, a random forest with unigram+SVD, and dataset B, logistic regression with TF-IDF is the best single model [23]. The combined F1 score for the 'models' forecast was 0.812. In [24] the authors proposed the combination of the Fasttext neural network model, stacking model, and XGBoost. Moreover, they suggested to apply Synthetic Minority Over-sampling Technique (SMOTE) to manage the optimization procedure for imbalanced classes and hyperparameters, solve various issues with automated short answer scoring, and enhance model performance.

The F1-score of 0.821 outperformed earlier research attempts using the same dataset. In 2021, Ukara used a well-researched sentence transformer to produce vector embedding for its classification task. In this case, SentenceTransformers have more features and benefits such as processing contextual inserts while paying attention to words. To improve the accuracy, a pre-trained multilingual "*paraphrase-xlm-r-multilingual-v1*" sentence transformer was proposed in [25]. There are many challenges in building an AES model, some of which are data restrictions and imbalanced classes [25]. The new model's F1-score is higher than the 0.829 of the previous models. Several parameter settings employ dropout, decoupled weight decay, incremental batch size, and SMOTE to enhance model performance and decrease overfitting.

Deep neural network-based methods have exploited recent developments in semantic similarity for performance improvement. The expansion of this method is massively used generally for the case of public datasets in English. Processes related to AES include LSTM, LSTM Siam, XLNet, MobileBERT, and Transformers. Deep neural network research, specifically the transformer-based models made famous by BERT, has increasingly dominated the NLP research community in recent years [26].

SentenceTransformers is a popular library for performing the NLP task. It is a clear and concise library that quickly calculates solid vector representations for text. It includes numerous state-of-the-art pre-trained models that have been fine-tuned for various applications. Word embedding provides vector representations of words wherein these vectors retained the underlying linguistic relationship between the observations [27]. The work presented in [28] modified a pre-trained BERT network using Siamese and triplet networks. A structure derives semantically meaningful sentence embeddings that can be compared using cosine similarity. On typical STS tasks and transfer learning tasks, SBERT and SRoBERTa outperformed the previous method. SBERT is well-known and has influenced modern sentence embeddings that could handle text from several languages [29].

SBERT's success with its ease and openness in model development has made a state-of-the-art sentence embedding method. Work in [28] proposed a fast and efficient method to expand the number of supported sentence embedding models. The driving idea behind multilingual NLP models is to create a single model that can comprehend many languages instead of training a single model for each language. An enticing idea is that building such a model necessitates training on a substantial multilingual corpus.

This method makes it possible to modify previously monolingual models into multilingual equivalents. Notably, Google and Facebook have made their respective multilingual versions of the multilingual BERT and XLM-R models available. Given that these two models have already been pre-trained in 100 different languages, there is a good chance that a specific language, like Indonesian, will be included. Despite supporting 100 languages, XLM-R is competitive with monolingual models on a monolingual benchmark [30]. The average performance of XLM-R is 91.5%, whereas BERT, XLNet, and RoBERTa each achieve 90%, 92%, and 92.8%, accordingly. Additional evidence regarding the performance of the multilingual model is superior to LASER and LaBSE [9].

SentenceTransformers' multilingual model can be downloaded from the Hugging Face webpage. Several pre-trained multilingual models were selected with particular criteria like semantic similarity tasks for the model training in 50+ languages, including Indonesian (shown in Table 4), such as *paraphrase-xlm-r-multilingual-v1* was chosen [25] to solve the essay scoring problem in the Ukara. SentenceTransformers was deeply explored in monolingual and multilingual models by [31], comparing the two models with some natural language understanding tasks such as STS, semantic analysis, classification cases, *etc.* The evaluation results show that multilingual BERT outperformed other alternatives. The comparison of evaluation results only differed by 5%. It achieved better results in a larger number of tasks. However, few research studies have begun introducing contextual pretrained language models on languages other than English.

In addition to comparing all pre-trained multilingual sentence transformer models, this study compares monolingual models for IndoBERT and cross-encoder in SentenceTransformers. IndoBERT is a resource model for training, validating, and benchmarking Indonesian NLU [32]. IndoBERT can complete NLU tasks such as single-sentence sequence tagging, single-sentence classification, pairwise sentence classification, and pairwise sentence sequence labelling.

3. Methodology

The model development in the study consists of four stages: (1) literature review, (2) data preprocessing, (3) model selection and model construction, and (4) model evaluation focused on accuracy. Figure 1 illustrates the flow of essay assessment architecture in this study. The literature review using desk research method reviews the literature related to an automatic essay scoring model that can be used. Some

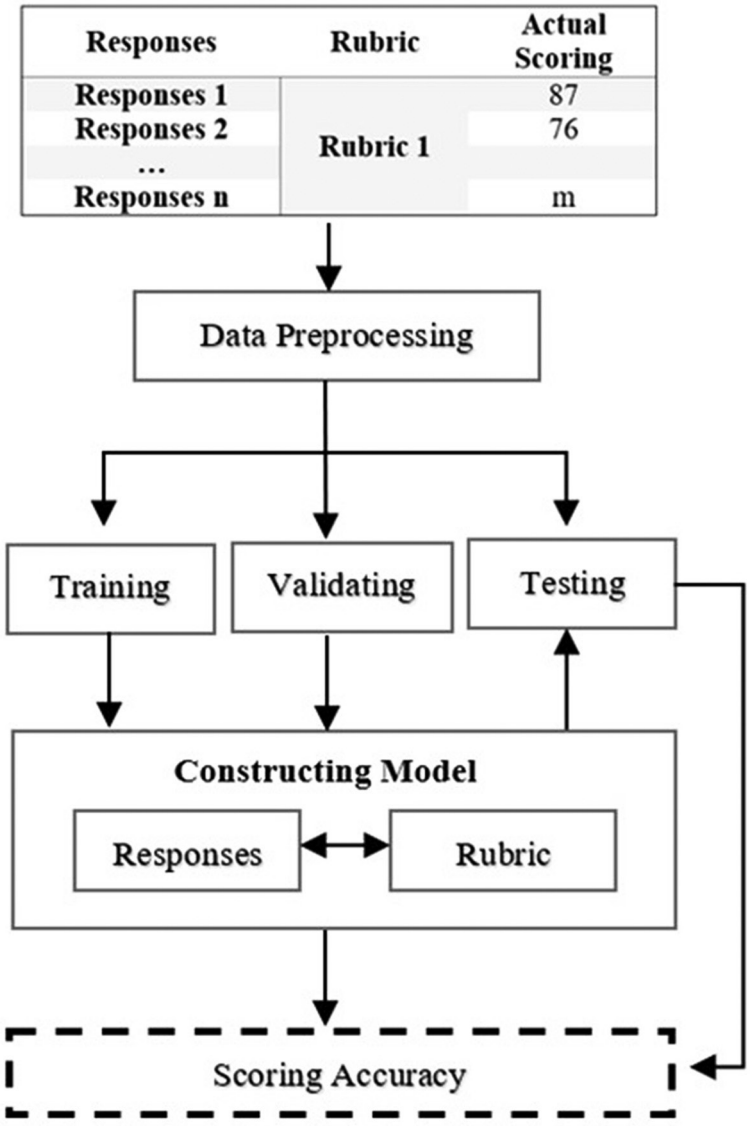

Figure 1. Essay scoring architecture.

keywords in the literature search include online learning, online discussion forums, automatic essay scoring, automatic assessment, NLP, preprocessing, BERT, SentenceTransformers, bi-encoder, cross-encoder, and text embedding. Through the literature review, NLP problem-solving algorithms, especially paired sentences, were obtained and pre-trained models developed for Indonesian language corpus were identified. The initial process after collecting data is preprocessing. Preprocessing is the first step in processing the dataset, where the data is processed using five preprocessing techniques for the Indonesian language: (a) HTML tags removal, (b) case folding (lowercase, remove special characters), (c) stopword removal, (d) stemming, and (e) tokenization.

The data used in the study are student responses in the rubric discussion forum. The data was obtained from the LMS application, *i.e.*, Tutorial Online (TUTON) provided by Universitas Terbuka. The forum data was collected through the queries on the database using DBeaver from the TUTON application. Referring to the scope of the research, the discussion forum responses used are definitive responses or understanding of a particular object, instead of science class employing formulas or notations. A maximum of 50 students attend one online class. Each

session has one different forum discussion task. The online class taken is a general biology class in the Faculty of Science and Technology. This class is a mandatory course for students. Data from the forum was collected in two stages, in the even semester of 2020 and 2021. The size of datasets is shown in Table 1.

Table 1. Discussion Forum Datasets.

		Number of Classes	Number of Responses	Number of Scores
1	2020.2	18	546	546
2	2021.1	12	482	482

The data is divided into three parts: training data, validation data, and testing data. The division of training and validation data uses the Python data splitting module with a ratio of 70:30, Table 2. Training data is used in the model, and the model evaluates the data repeatedly to learn more about the behavior of the data and then adjusts to meet the intended goals. The algorithm analyzes the training data, classifies input and output, and then re-analyzes it. The algorithm remembers all input and output in the training data set during the training process. Completion of the training process is followed by the validation process. The training and validation processes are carried out sequentially in every epoch or iteration. While fine-tuning the model, hyperparameters are adjusted using validation data. Testing data simulates model utilization while testing the model. The model must never be trained on the testing data and it should contain previously unseen datapoints.

In literature, several techniques were used to measure the quality of AES systems, including Pearson and Spearman's correlation which is widely known as a practical evaluation measure for the AES systems [12], [17], [26], [33]. The agreement between the score provided by the AES system and the actual score is assessed using Pearson and Spearman correlation. A perfect score of 1.0 is granted when the predictions and the actuals are identical. The lowest possible score is -1, given when the predictions are furthest away from actuals. The expected success criteria in the automatic scoring system, according to the correlation value, the condition is Excellent ($r > 0.75$), Adequate ($r = 0.40–0.75$) or Poor ($r < 0.4$) [34]. By calculating the absolute difference between the marks produced by the lecturer and the system, MAE measure is used to calculate the error rate.

This research is focused on finding paired sentences that have semantic overlap between them. The best way to do this is to get a vector matrix between two sentences and then compare it with the cosine distance measure. A task-specific multi-dimensional vector representation of data, such as a word, image, or document, is called embedding. A wide variety of words might be used in a discussion forum. Therefore, the dimensions of each vector may be 1×10000 or even 1×100000. The well-known sentence transformer model uses embedding to create high-dimensional feature vectors. Sentence embedding has wide applications in NLP, such as semantic search, semantic textual similarity, and automated essay grading. This new property resulted in the emergence of a pre-training model for NLP. Official pre-trained models for over 500 sentence modifier models on the Hugging Face Hub can be found. A substantial amount of training data was used to build a pre-trained sentence transformer model. Sbert.net pioneers produced multilingual pre-trained models. With the distillation approach, multilingual knowledge [9] is extended through BERT (SBERT) sentences, so one of its advantages is that this embedding model can be developed to support additional languages, including Indonesian. Recently, 11 models were trained multilingually on Hugging Face. One of which is the distilroberta-base-paraphrase-v1 multilingual model, which is trained using parallel data for more than 50 languages, including Indonesian.

Table 2. The size of the training, validation, and testing sets.

	Number of Responses
#samples training set	253
#samples validation set	97
#samples testing set	101

3. Results

The data was analyzed using descriptive statistical methods, which help describe and understand the features of a particular dataset by providing a summary of the sample and data size. Four attributes are needed in the paired task sentence data, namely: *sentence1* as the response answer, *sentence2* as the reference answer, the *score* as the actual discussion score in the range 0-100, and the *length* to calculate the character length of the answer. From these data, the distribution of discussion response lengths and discussion scores were analyzed. The distribution of values and length of discussion responses is as follows. The descriptive statistics of scores show the maximum score is 100, the minimum score is 20, the mean is 74, the standard deviation is 14, and the median is 75. Furthermore, the length statistics show that the maximum length is 532 words, the minimum is 15 words, the average is 109 words, the standard deviation is 48 words, and the median is 112 words. The descriptive statistics of scores and lengths are shown in Table 3, while Figure 2 shows the score distribution.

Table 3. Descriptive statistics of score and length.

Task	Score	Length
Samples	451	451
Mean	74	109
Median	75	112
STD	14	48
Minimum	20	15
Maximum	100	256

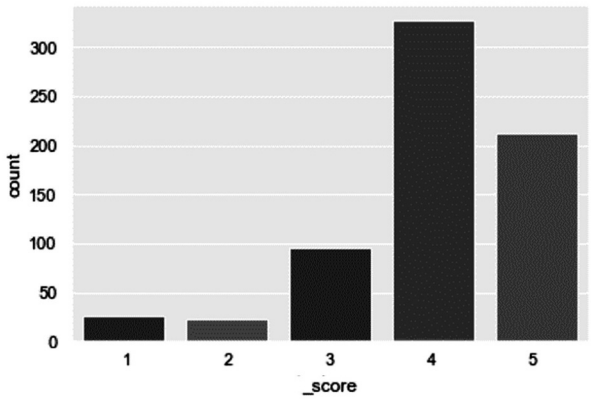

Figure 2. Score Distribution.

Data preprocessing is a technique applied to the database to remove noise and missing values. Raw data can occasionally be redundant, unbalanced, and incomplete. Several procedures are used in data preprocessing to convert unprocessed data into processed data. The data preparation stages include cleaning, transformation, and reduction. Preprocessing data frequently produces inconsistent, insufficient, and incomplete data used in real-time systems.

As a result, the findings of data mining become less accurate. Therefore, it is vital to perform data preprocessing activities to improve the quality of the analyzed data. It is in line with choosing the proper preprocessing technique to increase effectiveness and accuracy [35]. The results of the data analysis show the distribution of HTML tags, symbols, punctuation marks, numbers, emojis, capital letters, and nonmeaningful words. Nonmeaningful words include conjunctions in Indonesian such as *dan*, *yang*, *akan*, *etc.* The words also include the results of the analysis of responses in the discussion of writing style, which consists of opening, introductory, repeated questions, explanations, and closing greetings. Therefore, factors such as special characters, numbers, and punctuation marks, slow down the computational speed, interfering with the transfer of learning. Therefore, the data cleansing process will remove unique characters using numpy. char and HTML tags using *BeautifulSoup* html. parser, symbols, non-breaking spaces, emojis, and distinctive characters like (*), (&), (?), *etc.* The data cleansing processes include tokenization to get individual words from the text, cessation of word deletion to remove general words, and unique word sets based on findings of the analysis of response patterns considered unimportant or unrelated to the topic. Tokenization is a fundamental task in natural language processing (NLP), dividing a given text into smaller units known as tokens. These tokens can be words, phrases, or characters depending on the specific tokenization approach used. Tokenization is typically the first step in many NLP projects because it is a foundation for building effective models and gaining a better understanding of the text. Stemming and tokenization have been constructed using *nltk.stem* and *nltk.tokenize* library.

Table 4. Custom stopword removal.

Categories	Samples
Greetings	*Assalamualaikum*, *bismillah*, *maaf menggangu*, *minal aidin wal faidzin*, *etc.* Assalamualaikum, bismillah, sorry to disturb, minal aidin wal faidzin, *etc.*
Introductions	*program studi*, *jurusan*, *nama kota*, *upbjj*, *perkenalkan*, *etc.* study program, department, city name, *upbjj*, introduce, *etc.*
Closing	*Mohon maaf*, *izinkan*, *wasalamualaikum*, *mohon koreksinya*, *mohon masukkanya*, *terima kasih*, *referensi*, *sumber*, *revisi*, *etc.* Sorry, allow me to, *wassalamualaikum*, please correct it, suggestion please, thank you, references, sources, revisions, *etc.*

The stopword removal uses *stopWordRemoverFactory* from the *Pysastrawi* library. The response analysis results show that the writing style consists of opening greetings, introduction, repeated questions, explanations, and closing greetings. Therefore, the module eliminates customized words by adding a collection of non-meaning category words including greetings, introduction, rewriting questions, and closure (Table 4). However, not all frequent words are omitted since some of them are answers or keywords in the assessment.

Data transformation shapes the data into a form suitable for constructing the model. Normalization is one of the processes of data transformation, scaling the data on an interval of 0 to 1. In this case, the attribute score is normalized using equation (1). The descriptive statistical distribution shown in Table 2 indicates the min score is 20, and the max score is 100. Where Z_i is the normalized attribute value, X_i is the attribute value, min(x) is the minimum value, and max(x) is the maximum value.

$$Z_i = \frac{X_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Normalization of data is used to compare values with the results of matrix comparisons or paired vector sentences through cosine similarity, which is a measurement that quantifies the similarity between two or more vectors. The concept of cosine similarity determines the angle between two points or vectors to compare them. It is a value bounded by a finite range of 0 and 1. In this case, the cosine equality has a value of 0; this means that the two vectors are orthogonal or perpendicular to each other. The greater the cosine similarity value is close to 1, the smaller the angle between the two vectors A and B. Finally, data reduction removes categorical data including discussion forum data that use attached files.

Table 5 shows some collected pre-trained sentence transformer models of Hugging Face. In this case, the pre-trained model of "paraphrase-xlm-r-multilingual-v1", available on the SBERT homepage, was used. This pre-trained model has been trained on 50+ languages, including Indonesian. The model has been trained by paraphrasing the data to see similarities in sentences. It produces a vector embedding of 768 elements for each input sentence; the maximum sequence length is 250 words. If the word exceeds the maximum limit, then the word is not considered or truncated. The nearly multilingual model based on BERT has spawned several model derivatives such as DistilBERT, RoBERTa, and XLNet, which cover tasks such as semantic similarity, essay grading, Q&A, and more.

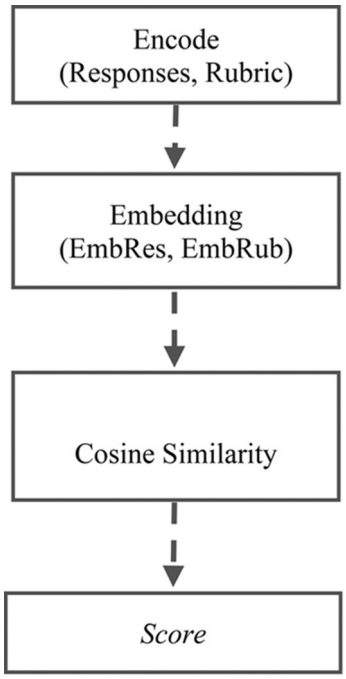

Figure 3. Generated score framework.

After defining our model, sentence pairwise will be computed to get both similarity scores. As was mentioned in the introduction, the method entails using the model to encode the text pairs before determining how close the two embeddings are by measuring their cosine similarity. The semantic similarity score is the outcome. The text data must be transformed into its numerical vector representation to calculate semantic similarity. The vector representation is built as a vector embedding. Let A be the vector embedding of the response's answer, and let B be the vector embedding for the essay from the reference answer. The similarity between pairs of sentences is determined using the cosine similarity measure. The cosine similarity metric calculates the angle between the vectors representing two text data sets. Cosine similarity between vectors performs well in high dimensional spaces. Equation (2) provides the formula for calculating the cosine similarity.

Table 5. Pretrained multilingual models of SentenceTransformers.

NO	MODEL	BASED MODEL	MAX LENGTH	DIM
1.	sentence-transformers/paraphrase-multilin gual-MiniLM-L12-v2	BertModel	250	384
2.	sentence-transformers/distiluse-base-multilingual-cased-v2	DistilBert-Model	250	768
3.	sentence-transformers/paraphrase-xlm-r-multilingual-v1	XLMRoberta-Model	250	768
4.	sentence-transformers/distilbert-multilingual-nli-stsb-quora-ranking	DistilBert-Model	250	768
5.	sentence-transformers/paraphrase-multilingual-mpnet-base-v2	XLMRoberta-Model	250	768
6.	sentence-transformers/distiluse-base-multilingual-cased	DistilBert-Model	250	768
7.	sentence-transformers/distiluse-base-multilingual-cased-v1	DistilBert-Model	250	768
8.	sentence-transformers/stsb-xlm-r-multilingual	XLMRoberta-Model	250	768
9.	sentence-transformers/clip-ViT-B-32-multilingual-v1	DistilBert-Model	128	768
10.	sentence-transformers/quora-distilbert-multilingual	DistilBert-Model	128	768
11.	sentence-transformers/use-cmlm-multilingual	BertModel	256	768

When the value of cos(θ) is 1, the two vector embeddings are perfectly comparable. If the value is 0, then there is presumably no similarity between them. This procedure works primarily by searching through all the reference responses used to fine-tune the model (all pre-trained multilingual models) to find the one with the most significant similarity and to forecast a score for semantic similarity for each provided response (Figure 3).

$$cosine\ similarity = \frac{A \cdot B}{\|A\| \|B\|} \quad (2)$$

Previous traditional algorithms, such as Logistic Regression, LSTM, and Transformer, require searching or tuning hyperparameters for the best accuracy value. The process is arduous and time-consuming. The model hyperparameters, a characteristic that controls the entire training process, are the subject of this study's experiments. The variables commonly set up before model training are listed in Table 6. Hyperparameters are crucial because they directly affect how the training algorithm functions and significantly impact how well the trained model performs. The training process without hyperparameter tuning has resulted in less favorable performance values [36]. Hyperparameter tuning in this study uses Optuna. The results of hyperparameter tuning and Optuna values can be seen in Table 6. The training performance results differ significantly between tuned hyperparameters and without hyperparameter tuning.

Table 6. Hyperparameter and Optuna Values.

Hyperparameter	Optuna Values
Epochs	3
Batch size	4
Learning rate	1e-04
Weight decay	0.01

Batch size, learning rate, weight decay, and epoch are parameters that need to be fine-tuned. Batch size is the number of data samples simultaneously passing through the neural network. The batch size determines the number of samples that must be processed before updating the internal model parameters. In this study, because the dataset is small, the batch size determination interval starts from 4 and is increased by 2, with a maximum of 10. The higher the batch size, the higher the computational demands, so batch size determination also needs to take into account computational resources. The learning rate is one of the training parameters used to calculate the weight correction during the training process. The learning rate value is in the range of 0 to 1. The tuning of the learning rate in this study starts from the smallest value of 1e-7 up to 1e-4.

Weight decay or L2 Regularization is a technique used in machine learning to reduce model complexity and prevent overfitting. According to sber.net, the weight decay value starts from a high value of 0.1. Therefore, in this study, the weight decay interval is between 1e-1 and 1e-2. Epochs represent the number of iterations that must be performed on the data set. Epochs indicate one cycle of the deep learning algorithm learning on the entire training dataset. The training process in SentenceTransformers with a small dataset is generally below five epochs. If the number of epochs is increased, the evaluation results will decrease. In this study, the interval of epochs used is 2-3.

Evaluation is done to gauge how well the model works, beginning with the training, validating, and testing phases. Measuring the final performance of the model comes to the correlation between the actual score and the semantic score described by the Pearson correlation. Table 7 shows the performance evaluation of the eleven multilingual models through Pearson correlation.

The multilingual model of *paraphrase-multilingual-MiniLM-L12-v2* got the highest score of 0.59 and a faster training time. Then *distiluse-base-multilingual-cased-v1* is in the following sequence with the same value of 0.56 and with a difference in training of five minutes. Findings in [26] stated that although the evaluation difference in other multilingual models through Pearson correlation is only below 5%, it can still be said that the models are considered equivalent. In addition to evaluation, paying close attention to the model choice made during training is critical. They were precisely estimating the time needed to train

a machine-learning model. This is particularly true when training models in a cloud setting use enormous amounts of data. *Stsb-xlm-r-multilingual* model became the first sequence of training time, reaching almost one hour of training. *Paraphrase-xlm-r-multilingual-v1 l* model is in the following sequence with forty-five minutes of training.

The determination of the data is based on a "score" distribution to balance the data (Figure 2). Authors in [25] highlighted the importance of having balanced data as an indicator for improving accuracy. Unbalanced data in a dataset typically occurs when class distribution is unequal. Figure 2 shows that the distribution of scores for classes 1 to 3 has very few data points. This study then conducted experiments using oversampled and undersampled datasets. Balancing the training data with resampling techniques can improve model accuracy performance. Several methods are available for oversampling or undersampling data in imbalanced datasets.

Table 7. Evaluation of the Eleven Multilingual Models.

Model	Pearson			Time Consumed
	Train	Val	Test	
1.	**0.72**	**0.70**	**0.59**	**00:15:04**
2.	0.65	0.59	0.55	00:27:26
3.	0.66	0.65	0.52	00:49:11
4.	0.62	0.60	0.51	00:25:16
5.	0.71	0.65	0.55	00:47:25
6.	0.64	0.63	0.56	00:26:06
7.	**0.72**	**0.69**	**0.59**	**00:20:18**
8.	0.69	0.68	0.55	00:51:00
9.	0.68	0.61	0.56	00:23:49
10.	0.60	0.62	0.51	00:24:27
11.	0.75	0.60	0.55	01:39:01

Table 8. Experimental Undersampling and Oversampling Results.

Model	Pearson			Waktu
	Training	Validation	Test	
Undersampling				
paraphrase-multilingual-MiniLM-L12-v2	**0.72**	**0.70**	**0.59**	00:15:04
distiluse-base-multilingual-cased-v1	0.72	0.69	0.59	00:20:18
Oversampling				
paraphrase-multilingual-MiniLM-L12-v2	0.6268	0.599	0.6183	01:17:21
distiluse-base-multilingual-cased-v1	**0.6683**	**0.5803**	**0.6395**	01:44:45

Undersampling the majority class involves randomly selecting data points from the dominating class of the dataset to match the number of the minority class. The downside of this technique is the loss of valuable data by discarding it. However, reducing the sample size may be computationally beneficial when working with large datasets. Oversampling the minority class is the opposite approach, where random data points are duplicated to match the number of cases in the dominant class. In this case, it essentially creates duplicate data to train the model. The results show that the oversampled dataset performed better than the undersampled dataset, as outlined in Table 8 and Table 9. However, the Pearson correlation coefficient shows an increase of only 0.2, which was marginal. On the other hand, paraphrase-multilingual-MiniLM-L12-v2 outperformed other approaches when it comes to model processing time.

Table 9. Evaluation Results.

#	Experiment	Pearson	MAE
1	*Undersampling*	0.59	1.03
2	*Oversampling*	0.63	0.70

4. Discussion

Previous studies on Automated Essay Scoring (AES) algorithms presented algorithms used for general English and specific datasets, such as Indonesian datasets. Table 10 summarizes the comparison of results from several AES studies in Indonesia. Regarding classification, most studies used the Ukara public dataset, which includes automated short answer system data, achieving an accuracy of 80% using the SBERT or SentenceTransformers algorithm. However, for regression tasks, the results for Indonesian language cases are generally below 0.70. The dataset used in this study consists of short answer responses from discussion forums, with an average length of 86 words and a maximum of 250 words after preprocessing.

Compared to a previous study [19], this study achieved a Pearson evaluation score of 0.64, similar to the previous study's score of 0.65. However, the MAE error value of this study was better, with a value of 0.70 compared to the previous study's value of 0.90. It should be noted that the previous study used a short answer dataset with responses consisting of less than 20 words, while this study used an average of 86 words and a maximum of 250 words per response. Therefore, the comparison may not

Table 10. Multilingual model accuracy.

Model	Task	Evaluation	Score
LSTM + No Transfer learning	Regression	QWK	48.26
LSTM + Transfer learning		QWK	64.58
LCS, CC, JCD, and DC		Pearson	0.65
Word2Vec & CBOW		Pearson	0.70
LSA		-	72%
TF-IDF & SVD	Classification	F1 Score	0.812
Fastext model, stacking model & xgboost		F1 Score	0.821
SBERT + NN		F1 Score	0.829

be entirely fair. Table 11 shows the evaluation of the Pearson correlation monolingual model using IndoBERT which has been trained on the enormous data corpus and Indonesian cleaning data collections (Indo4B), such as news, social media, blogs, and websites. The results of the generated model show a low value below 0.50.

Table 11. Monolingual model accuracy.

Model	Pearson	Time Consumed
IndoBERT	0.43	00:58:02

SentenceTransformers is the method that has a more straightforward model but with richer features for NLP tasks. Multilingual models can be pretty effective. The most recent innovation, XLM-R, supports 100 languages [30] while still being competitive with monolingual alternatives. The current research shows multilingual transmission is expected to continue becoming better. There are several reasons these directions in research are essential: seeing more attention given to power-efficient computing for usage on small devices, predictably, and the deep learning community will set a greater emphasis on smaller efficient models in the future.

This study shows a multilingual sentence transformer model could perform well in generating evaluation scores and achieve faster model processing time. Based on the comparison, the multilingual paraphrase-multilingual-MiniLM-L12-v2 model performed better in both resampling processes. These results can be a reference for future studies using multilingual sentence transformer models, especially for Indonesian language datasets.

5. Conclusion

This paper explores a semantic similarity approach to automatic essay scoring. We believe this paper makes two significant contributions. First, while previous studies have used related case classifications and limitations of Indonesian-specific models such as IndoBERT, the method we propose uses the multilingual model, a state-of-the-art NLP model with the capability of a faster model processing time. Second, the results of comparing eleven multilingual models are available on Hugging Face where these models have been pre-trained in various languages, including Indonesian. This paper can be used as a reference for multilingual models, especially in Indonesian. Making Automatic Essay Scoring models is challenging for numerous reasons, including data limitations and unbalanced class sizes. In the future, we intend to improve the model's performance through a small dataset with an effectual data-augmentation method known as Augmented SBERT, where it stacks a cross-encoder to label pairwise sentences of a larger corpus to augment data training for the bi-encoder.

References

[1] T. Belawati, *Online Learning*. Jakarta: Universitas Terbuka, 2020.

[2] S. D. Antoro and S. Sudilah, "Enhancing Learning Interaction through Inter-Forum Group Discussion in Online Learning: A Case Study on Online Teaching of Research in English Language Teaching Course", *Ahmad Dahlan J. English Stud.*, vol. 3, no. 2, p. 64, 2016. http://dx.doi.org/10.26555/adjes.v3i2.4994

[3] D. F. Murad et al., "Text Mining Analysis in the Log Discussion Forum for Online Learning Recommendation Systems", I*nt. Semin. reserach Inf. Technol. Intell. Syst.*, 2018.

[4] N. Wanas et al., "Automatic Scoring of Online Discussion Posts", 2008.

[5] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?", *Int. J. Artif. Intell. Educ.*, vol. 31, pp. 538–584, 2021.

[6] K. Taghipour, "Robust Trait-Specific Essay Scoring Using Neural Networks and Density Estimators", 2017.

[7] M. D. Shermis and B. Hamner, "Contrasting state-of-the-art automated scoring of essays: Analysis", *Annu. Natl. Counc. Meas. Educ. Meet.*, pp. 14–16, 2012.

[8] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, 2019. http://dx.doi.org/10.48550/arXiv.1810.04805

[9] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525. http://dx.doi.org/10.18653/v1/2020.emnlp-main.365

[10] P. Foltz et al., "The Intelligent Essay Assessor: Applications to Educational Technology", *Interact. Multimed. Electron. J. Comput. Learn.*, 1999.

[11] F. Dong and Y. Zhang, "Automatic Features for Essay Scoring – An Empirical Study", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1072–1077. http://dx.doi.org/10.18653/v1/D16-1115

[12] J. Cancan et al., "A Study of Distributed Semantic Representations for Automated Essay Scoring", *Knowl. Sci. Eng. Manag. 10th Int. Conf. Melbourne, VIC, Aust.*, 2017.

[13] G. Liang et al., "Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture", *MDPI Symmetry*, 2018.

[14] T. Dasgupta, A. Naskar, L. Dey, and R. Saha, "Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring", 2018.

[15] L. Zhang et al., "An Automatic Short-answer Grading Model for Semi-open-ended Questions", *Interact. Learn. Environ.*, vol. 30, no. 1, pp. 177–190, 2022. http://dx.doi.org/10.1080/10494820.2019.1648300

[16] Kaggle, "Automated Student Assessment Prize: The Hewlett Foundation: Short Answer Scoring", Available: https://www.kaggle.com/c/asap-sas

[17] P. U. Rodriguez et al., "Language Models and Automated Essay Scoring", arXiv.org, 2019.

[18] C. M. Ormerod et al., "Automated Essay Scoring Using Efficient Transformer-based Language Models", arxiv.org, 2021. http://dx.doi.org/10.48550/arXiv.2102.13136

[19] U. Hasanah et al., "A Scoring Rubric for Automatic Short Answer GradingSystem", *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 17, no. 2, p. 763, 2019. http://dx.doi.org/10.12928/telkomnika.v17i2.11785

[20] F. F. Lubis et al., "Automated Short-Answer Grading using Semantic Similarity based on Word Embedding", *Int. J. Technol.*, vol. 12, no. 3, p. 571, 2021. http://dx.doi.org/10.14716/ijtech.v12i3.4651

[21] A. A. Putri Ratna et al., "Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and Latent Semantic Analysis", in *Proc. of the 2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2019, pp. 363–367. http://dx.doi.org/10.1109/ICECOS47637.2019.8984528

[22] A. Amalia et al., "Automated Bahasa Indonesia Essay Evaluation with Latent Semantic Analysis", *J. Phys. Conf. Ser.*, vol. 1235, no. 1, p. 012100, 2019. http://dx.doi.org/10.1088/1742-6596/1235/1/012100

[23] A. A. Septiandri and Y. A. Winatmoko, "UKARA 1.0 Challenge Track 1: Automatic Short-Answer Scoring in Bahasa Indonesia", 2020, [Online]. Available: http://arxiv.org/abs/2002.12540

[24] R. A. Rajagede and R. P. Hastuti, "Stacking Neural Network Models for Automatic Short Answer Scoring", 2020, http://dx.doi.org/10.1088/1757-899X/1077/1/012013

[25] R. A. Rajagede, "Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature", *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 11–18, 2021. http://dx.doi.org/10.22219/kinetik.v6i1.1196

[26] E. Mayfield and A. W. Black, "Should You Fine-Tune BERT for Automated Essay Scoring?", in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 151–162. http://dx.doi.org/10.18653/v1/2020.bea-1.15

[27] T. Schnabel et al., "Evaluation Methods for Unsupervised word Embeddings", in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 298–307. http://dx.doi.org/10.18653/v1/D15-1036

[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", arxiv.org, 2019. http://dx.doi.org/10.48550/arXiv.1908.10084

[29] A. Aponyi, "What are Sentence Embeddings and Their Applications?", blog.taus.net, 2021. https://blog.taus.net/what-are-sentence-embeddings-and-their-applications

[30] J. Moberg, "A Deep Dive Into Multilingual NLP Models Min Read", 2020. Available: https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models

[31] D. de Vargas Feijó and V. P. Moreira, "Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks", ArXiv, vol. abs/2007.0, 2020.

[32] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding", in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857, [Online]. Available: https://aclanthology.org/2020.aacl-main.85

[33] V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis", *Int. J. Comput. Appl.*, vol. 131, no. 4, pp. 30–36, 2015. http://dx.doi.org/10.5120/ijca2015907309

[34] I. G. Ndukwe et al., "Automatic Grading System Using Sentence-BERT Network", A*rtif. Intell. Educ.*, vol. 12164, pp. 224–227, 2020.

[35] J. L. Fleiss et al., *Statistical Methods for Rates and Proportions.* Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003.

[36] E. Elgeldawi et al., "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis", *Informatics*, vol. 8, no. 4, p. 79, 2021. http://dx.doi.org/10.3390/informatics8040079

Received: October 2022
Revised: April 2023
Accepted: June 2023

Contact addresses:
Bachriah Fatwa Dhini
Department of Computer Science
BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
Jakarta
Indonesia
e-mail: bachriah.dhini@binus.ac.id

Abba Suganda Girsang
Department of Computer Science
BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
Jakarta
Indonesia
e-mail: agirsang@binus.edu

Bachriah Fatwa Dhini is a master's degree student of computer science at Bina Nusantara University, Jakarta, Indonesia. She is currently working on learning technology distance education in Terbuka University. Specifically, she is involved in a variety of related activities concerning media development in learning using a variety of cutting-edge tools or applications as well as deep learning for Learning Management Systems (LMS). In addition to media development activities, the author has also conducted research and written articles for national seminars.

Abba Suganda Girsang obtained a PhD degree from the Institute of Computer and Communication Engineering, Department of Electrical Engineering and National Cheng Kung University, Tainan, Taiwan, in 2014. He was a staff consultant programmer in Bethesda Hospital, Yogyakarta, in 2001 and worked as a web developer in the period 2002–2003. He then joined the faculty of the Department of Informatics Engineering in Janabadra University as a lecturer in the period 2003–2015. He also taught some subjects at some universities in the period 2006–2008. His research interests include swarm intelligence, business intelligence, machine learning and media social text mining.