Automatic essay scoring for discussion forum in online learning based on semantic and keyword similarities

Bachriah Fatwa Dhini
Department of Multimedia Teaching Material Production Center,
Universitas Terbuka, Tangerang Selatan, Indonesia

Abba Suganda Girsang
Computer Science Department,
BINUS Graduate Program - Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia, and

Unggul Utan Sufandi and Heny Kurniawati
Faculty of Sains and Technology, Universitas Terbuka,
Tangerang Selatan, Indonesia

Abstract
Purpose – The authors constructed an automatic essay scoring (AES) model in a discussion forum where the result was compared with scores given by human evaluators. This research proposes essay scoring, which is conducted through two parameters, semantic and keyword similarities, using a SentenceTransformers pre-trained model that can construct the highest vector embedding. Combining these models is used to optimize the model with increasing accuracy.

Design/methodology/approach – The development of the model in the study is divided into seven stages: (1) data collection, (2) pre-processing data, (3) selected pre-trained SentenceTransformers model, (4) semantic similarity (sentence pair), (5) keyword similarity, (6) calculate final score and (7) evaluating model.

Findings – The multilingual paraphrase-multilingual-MiniLM-L12-v2 and distilbert-base-multilingual-cased-v1 models got the highest scores from comparisons of 11 pre-trained multilingual models of SentenceTransformers with Indonesian data (Dhini and Girsang, 2023). Both multilingual models were adopted in this study. A combination of two parameters is obtained by comparing the response of the keyword extraction responses with the rubric keywords. Based on the experimental results, proposing a combination can increase the evaluation results by 0.2.

Originality/value – This study uses discussion forum data from the general biology course in online learning at the open university for the 2020.2 and 2021.2 semesters. Forum discussion ratings are still manual. In this survey, the authors created a model that automatically calculates the value of discussion forums, which are essays based on the lecturer’s answers moreover rubrics.

Keywords Discussion forum, Essay scoring, Semantic similarity, Keyword similarity

1. Introduction
Online education provides students with a wide choice of amenities that is no longer purely hypothetical (Pavan Kumar, 2021). Numerous universities have noted the benefits of online
learning worldwide. Online learning is designed to support the learning process in a distance learning system, generally presented using a Learning Management System (LMS) platform that allows lecturers and students to interact in an asynchronous method (Antoro and Sudilah, 2016). All kinds of LMS support lectures, activities, practices and project work. The online discussion forum is one of the most widely utilized LMS communication tools.

Universitas Terbuka (UT) has more than 300,000 students. One of the learning modes is discussion forums through online learning. UT has more than 20,000 online classes based on courses where one online class has a maximum of 50 students. Discussion forum assessment is currently still manually done. Several problems were found, such as the delay of tutors or lecturers in assessing forum discussions, giving biased scores, subjective assessments or variance from predetermined assessment standards. Manually grading essays is a complex and time-consuming task. Even if there is a set scoring guide, individual elements, such as mood and personality, impact the scoring process, making the results of the scoring procedure subjective and unreliable. Academics must conduct academic labor, such as teaching and learning, to raise the institution’s ranking and place it on one of the world’s best university lists. Higher education institutions in this situation must enhance organizational performance to satisfy the necessary quality criteria (Sulartopo et al., 2022).

It has been demonstrated that collaborative learning is an efficient method of learning. Numerous instances exist where students’ participation in a discussion forum is properly considered while assessing their performance (Pawade et al., 2020). Discussion forums apply educational technology to improve accessibility to learning resources, meet students’ diverse needs and provide synchronous and asynchronous interactions that make it easier for students (Onyema et al., 2019). The process of assessing one’s writing skills and giving an essay a grade without human intervention is automatic essay scoring (AES) (Jong et al., 2022). AES systems had as high a level of agreement with human graders as human graders had with each other (Shermis and Hamner, 2012).

According to Chong et al. (2021), data extraction and categorization are the two fundamental system elements in data mining. For data extraction, the information will be mined from numerous documents with varied structures, such as free text or tables. Natural Language Processing (NLP) will then be used for data classification or extraction. To classify and even extract the pre-processed text into categories, a machine learning model called Bidirectional Encoder Representations from Transformers (BERT) is included as a system engine (Chong et al., 2021).

It is similar to the AES pattern, which is divided into two ways: data classification and data extraction. Data classification in AES is like assessing essay answers by classifying them into the entailment, contradiction or neutral category using data belonging to Stanford Natural Language Inference (SNLI). An example of an automatic essay assessment by extracting semantic words and comparing them with reference answers (Amalia et al., 2019). AES system generally assesses the content by extracting essay processes ranging from equating words as keywords (keyword similarity) (Gunawansyah et al., 2020; Setiadi Citawan et al., 2018) and sentences (sentence similarity) (Hasanah et al., 2019; Putri Ratna et al., 2019).

The original AES system was developed in 1966 along with Project Essay Grade (PEG) and rate essays using some linguistic characteristics. Furthermore, machine learning began to be massively used in several cases. Machine learning’s primary function in developing and automating text analytics is to advance and construct such parts of speech as named entity recognition (NER) and sentiment analysis (Srinivasan et al., 2021). They used commonly used machine learning tools, such as Support Vector Machine (SVM), Random Forest classifier, extreme gradient boosting (XGBoost), logistic regression, Word2Vec, Gensim, etc. that are available in Python. The semantics-related research of AES caught the
interest of certain researchers in the 1980 and 1990s since it was rapidly expanding. Subsequently, adapted Intelligent Essay Assessor (Foltz et al., 1999) was suggested, and the degree of semantic similarity between texts was measured using LSA. The development of massive AES research generally does not only equate words (similarity) but also seeks semantic meaning.

The same core characteristics apply to them as to word embedding; for example, they capture a range of semantic relationships between sentences, such as similarity, contradiction and entailment. Sentence embedding that could handle text from several languages (Aponyi, 2021). They might be used to calculate the cosine similarity between two sentences’ vectors, which expresses how semantically linked two sentences are to one another recently published survey on semantic similarity (Chandrasekaran and Mago, 2022). According to Chandrasekaran, surveys evaluating the semantic similarity of different text components, such as words, phrases or documents, play a significant part in many NLP activities. Chandrasekaran’s work is categorized into four semantic similarity methods, two of which are deep neural network-based techniques and corpus-based techniques.

Corpus-based semantic similarity techniques were the first approach to measuring semantic similarity between phrases using data from sizable corpora. It has been used in several related studies of AES, such as Latent Dirichlet allocation (LDA) (Jin et al., 2017) and LSA (Setiadi Citawan et al., 2018; Jin et al., 2017), which presented the usefulness of distributed semantic representation to AES. New features based on word embedding or combined with standard text features to increase the effectiveness of the AES system. Using the semantic generate feature of Word2Vec, Bag of Words (BOW) and LDA. Semantic features achieve better effectiveness based on word embedding (Jin et al., 2017). Several studies conveyed to AES combine its model with word embedding, including Word2Vec (Google), Glove (Stanford) and BERT (Rodriguez et al., 2019; Mayfield and Black, 2020; Ormerod et al., 2021). The word-embedding feature of BERT outperforms previous algorithms such as Word2Vec and GloVe.

The second technique utilized deep neural networks and semantic similarity techniques, which took advantage of recent advancements in neural networks to improve performance. This technique related to AES includes Long Short-Term Memory (LSTM) (Dasgupta et al., 2018), Siamese Bidirectional Long Short-Term Memory Architecture (SBLSTMA) (Liang et al., 2018), XLNet (Rodriguez et al., 2019), MobileBERT (Ormerod et al., 2021) and Transformers (Mayfield and Black, 2020; Ndukwew et al., 2020), which still exploited word embedding built using huge datasets. In recent years, language models using pre-trained teachers have shown breakthroughs in NLP (Devlin et al., 2019). Standard NLP research suggests transformers-based (Mayfield and Black, 2020). They have evaluated the pre-trained performance of several enhanced NLP models with several simple parameters on the AES data set. Ormerod et al. (2021) combined Electra and Mobile-BERT with 38 parameters, training 1.5x speed and 1.0x inference time speed. Electra and Mobile-BERT showed higher performance than BERT.

Semantic similarity between two documents could be done by using word embedding. Embeddings were vector representations of text where words or sentences with the same meaning or context had similar representations. Word embedding converts a word into a vector or array of numbers (Adam, 2019). Modern Sentence Embedding Techniques: SentenceTransformers (SBERT) (Reimers and Gurevych, 2019) This framework combines the strength of transformer topologies and twin-neural networks to produce high-quality sentence representations. It is based on the well-known BERT model.

The SentenceTransformers method mapped a sentence to a vector space using conjoined and triplet networks, which could create semantic meaningful sentence embeddings (Reimers and Gurevych, 2019). The focus of SentenceTransformers was comparing large-scale semantic similarity, grouping and finding information through semantic search, and SBERT also produces vector embedding of 768 elements for each input sentence. The transcendence
of SentenceTransformers in terms of time is finding the most similar pair of 10,000 sentences from 65 h by BERT to 5 s. It could also be compared with cosine similarity. The model is pre-trained by BERT and RoBERTa and then fine-tuned on SNLI to produce fixed-sized sentence embeddings. Ndukwe et al. (2020) proposed the SBERT language model on AES and showed an average value of 0.70 Quadratic Weighted Kappa (QWK).

Some AES assessment techniques include research using keyword similarity, which is also used to get essay scoring scores. Gunawansyah et al. (2020) compared the AES assessment results between system scoring without keyword synonyms and system scoring using keyword synonyms. The scoring system that uses keyword synonyms is closer to human scoring. Word similarity is obtained by calculating the number of reference keywords in the input essay. Its limitation was that it had not applied the keyword extraction approach. AES on e-learning with the LSA method produces n-gram features and compares them with unigram, bigram and trigram (Setiadi Citawan et al., 2018), resulting in a higher unigram accuracy evaluation value.

According to Hendre et al. (2021), assessing the similarity of keywords in AES is insufficient. Utilizing sentence similarity and keyword parameters, Hasanah proposed an automatic assessment using the Longest Common Subsequence (LCS), Cosine Coefficient (CC), Jaccard Coefficient (JC) and Dice Coefficient (DC) methods to assess the essay model (Hasanah et al., 2019). Combining two value sentences and keyword similarity can increase the correlation. The shortage of this research was that sentence similarity is not linked to semantic similarity, and keyword similarity has not developed an algorithm that could automatically assess and extract keywords.

Reported from the website toward data science, Yang (2020) analyzed the keyword extraction with five approaches: TextRank, TopicRank, Term Frequency–Inverse Document Frequency (TF-IDF), Yet Another Keyword Extractor (YAKE) and KeyBERT. Keyword extraction is one of the necessary steps in text mining: for instance, the extraction method in a document needs to find a collection of words that best describe the argument. According to Yang, no model performs well on every document. Performance results vary depending on the type of document, context and corpus used for the pre-trained model. The combination of the TF-IDF will be used to weigh words in short essays, and the results will be input to the Support Vector Machine (SVM), which extracts words based on the topic to reduce unrelated answers and then assesses using LSA. LSA was adopted in AES with a reasonable % accuracy value of 72% (Putri Ratna et al., 2019).

Elaborated keyword extraction used deep learning by adopting the BERT method, one of which is KeyBERT (Maarten Grootendorst, 2020). KeyBERT continues to improve its performance in terms of accuracy by leveraging BERT embeddings. KeyBERT excels in lightweight because it works very well with CPU configurations and is powerful because it supports the latest and best-performing embedding models, such as Flair, Spacy and Gensim. Another significant advantage is that KeyBERT can even be used with models of pre-trained SentenceTransformers.

Essay scoring research is not so massively developed in Indonesia. Data limitations are an obstacle to developing the AES system in Indonesia. This is contrary to the abundance of English public data availability. The development of AES is well developed with different algorithms, including machine learning, which is considered its core component (Mahana and Apte, 2012). Since its first usage in other research, neural network models have enhanced performance without requiring feature engineering. The performance of the various neural networks employed for AES has improved over time. The majority of recent works have utilized AES neural network models. Now, neural network pre-trained model transformers such as BERT are proven to produce high word vectors.

The use of artificial intelligence (AI) in education has dramatically changed during the past ten years. AES is yet another problematic issue in the educational field. The task of
automatically calculating an exact or close score for an essay answer is AES (Chassab et al., 2021). A thorough examination of the answer’s textual features would be necessary to determine an appropriate score. To perform such an analysis, it is frequently recommended in the literature to obtain a reference answer (also known as a model or template answer) and compare it with the student’s response. This study aims to create a model that generates semantic and keyword similarities well to optimize model performance. The data source for this study includes responses to discussion forums, scores and reference answers used as a rubric, obtained from the online learning. This study also experimented with selecting a pre-trained multilingual model, which can be a recommendation for which the model performs more accurate results and which the model is speedy in the training process. The format of this paper is as follows: Section 3 depicts the methodology of the study, and Sections 4 and 5 provide a result and discussion that implements the proposed approach. Moreover, the evaluation metrics used are Pearson correlation and Mean Absolute Error (MAE). The correlation test measures the level of agreement or value of proximity between the score generated by the human and the score generated by the system language model. MAE is used to measure model errors.

2. Research method

Generally, the development of the model in the study is divided into seven stages (Figure 1): (1) data collection, (2) pre-processing data, (3) selected pre-trained Sentence Transformers model, (4) semantic similarity (sentence pair), (5) keyword similarity, (6) calculate final score and (7) evaluating model. The first step is obtaining data; the data source is responses to discussion forums, scores and reference answers as a rubric. The data were obtained from the LMS application at the Online Tutorial (TUTON) of Terbuka University. The forum data collection is obtained by querying the database using DBeaver from the TUTON application. A maximum of 50 students attend one online class. Each session has one discussion forum task that is different from each session. The online classes taken are general biology online classes at the faculty of science and technology. The forum data are taken in two stages: in 2020, in the even semester (2020.2), there were 546 responses from 18 classes, and in 2021, in the even semester (2020.2), there were 482 responses from 12 classes. Thus, the total number of datasets used is 1,028.

Scores are categorized by class in Table 1. Figure 2 shows the statistics on the distribution of values. The dataset in Figure 2 shows a class imbalance with a tendency toward grades 4.

Figure 1. Diagram stage of study
and 5 compared to grades 1, 2 and 3. If the proportion is calculated, classes other than 4 and 5 are around 20.4% of the entire dataset. Rajagede (2021) states that balanced data indicate increasing accuracy. It can be seen in Figure 2 that classes 1 to 3 have a small amount of data. One of the biggest problems encountered when processing machine learning is the problem of unbalanced training data. Researchers face many challenges where they only have an unequal representation of data where the minority class is usually more important and therefore requires methods to increase its recognition rate.

The second stage is pre-processing, where data are processed using five stages of Indonesian pre-processing techniques: (a) remove HyperText Markup Language (HTML) tags, (b) case folding (lowercase, remove special characters), (c) stopword, (d) stemming and (e) tokenize. Clean data is used for fine-tuning the model. In this process, an experiment was conducted to compare the multilingual pre-trained SentenceTransformers model on huggingface.com with the sentence similarity model category. Hence, the model with the best evaluation will be chosen in the proposed AES system. The selection of the multilingual model was based on research by (de Vargas Feijó and Moreira, 2020), which stated that although it only resulted in a difference in evaluation values below 5%, the multilingual model performed better than the monolingual model.

Essays are a crucial part of traditional exams, and it can be difficult for lecturers to grade them correctly, quickly and effectively. AES is a challenging job that uses technology to help teachers score. Based on the grading criteria, traditional AES approaches only pay attention

<table>
<thead>
<tr>
<th>No</th>
<th>Range score</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100–91</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>90–81</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>80–71</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>60–41</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>20–40</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Score categories

Source(s): Table by authors

Figure 2. Distribution score

Source(s): Figure courtesy of (Dhini and Girisan, 2023)
to shallow language variables and ignore the impact of deep semantic features. In contrast, contextually aware sentence embeddings are produced by deep learning. Instead of just averaging out the word vectors in the phrase, this enables SBERT to capture the contextual meaning of sentences and include word order and dependencies, leading to more accurate embeddings (Li et al., 2023). To determine how semantically similar two sentences are, utilize Sentence Transformers (SBERT). Embedding texts in a high-dimensional space and measuring their cosine similarity provides a more precise measure of similarity, enhancing sentence vectorization and deepening our grasp of phrase semantics (Reimers and Gurevych, 2019). The time complexity analysis demonstrates that traditional AES requires significant time. Conversely, the sentence-level feature extraction framework is already more lightweight when compared to other pre-trained models (Li et al., 2023).

The SBERT approach transfers a sentence to a vector embedding space using conjoined and triplet networks and can offer sentence embeddings in semantic meaning (Reimers and Gurevych, 2019). The main focus of Sentence Transformers is comparing large-scale semantic similarities, grouping and finding information through semantic search. SBERT also produces vector embedding of 768 elements for each input sentence. Sentence embedding then calculates the cosine similarity between the two insertion sentences. The cosine similarity metric determines how similar two documents are by calculating the cosine of the angle between two embeddings. After calculating the keyword similarity value, keyword extraction is performed to extract keywords from the responses. Keyword extraction generally extracts keywords and key phrases from a document, assigning weights to each word to indicate its importance in the document and the more exhaustive corpus. KeyBERT will exploit to extract from responses of forum discussions, assigning weights to each word to indicate its significance in the document. Cosine similarity will also be used to calculate the similarity of the response keywords to the rubric keywords.

This study focuses on obtaining an automatic assessment model based on two phases: semantic and keyword similarities. The final score calculation uses a fixed-composition experimental approach referring to Hasanah et al. (2019), with a ratio of 50:50 and a ratio of param1 and param2 in multiples of 10 (Table 2). The technique for evaluating the Pearson correlation and MAE will be carried out in the last stages. The expected success criteria in the automatic scoring system based on the correlation value are perfect correlation (r > 0.81), strong correlation (r = 0.61–0.80), medium correlation (r = 0.41–0.60) or less correlation (r < 0.40). Hence, MAE is used to measure the error rate by finding the absolute difference between the scores generated by the lecturer and the system.

<table>
<thead>
<tr>
<th>Composition comparison (%)</th>
<th>Param1</th>
<th>Param2</th>
<th>Final score</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>90</td>
<td></td>
<td>(param1<em>10)+(param2</em>90)</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td></td>
<td>(param1<em>20)+(param2</em>80)</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
<td></td>
<td>(param1<em>30)+(param2</em>70)</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td></td>
<td>(param1<em>40)+(param2</em>60)</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td></td>
<td>(param1<em>50)+(param2</em>50)</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td></td>
<td>(param1<em>60)+(param2</em>40)</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td></td>
<td>(param1<em>70)+(param2</em>30)</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td></td>
<td>(param1<em>80)+(param2</em>20)</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
<td></td>
<td>(param1<em>90)+(param2</em>10)</td>
</tr>
</tbody>
</table>

Table 2. Fix composition comparison
3. Results

Based on the number of datasets, the data were divided into three types, namely training data, validation data and testing data (Figure 3). Data training this type of data are used in the model. The model evaluates the data repeatedly to learn more about the behavior of the data and then adjusts itself to meet its intended goals. The algorithm remembers all the inputs and outputs in the training data set during the training process. Validation data are used to produce hyper-parameters while fine-tuning the model. The test data are used for testing the model as a simulation of using the model.

Henceforward, resampling approaches combined with data balancing can boost model accuracy. The resampling dataset frequently solves unbalanced data issues, including undersampling and oversampling. Majority class undersampling takes a random draw of the dominating class from the dataset to match the number of non-dominating classes. As a general rule, the drawback of engineering is that it causes the loss of some valuable data by wasting it. Still, having a large data set may prove computationally better to reduce the sample. The minority class oversampling is the opposite.

The undersampling dataset technique processes training and validation data by eliminating the majority dataset and equating it with the minority data. Table 3 shows the number of comparison results of training and validation datasets using undersampling. The oversampling technique was used in the training dataset. The training data and validation data are processed by adding synthesis data with the oversampling method resampling technique, which makes the amount of data equal to the majority dataset. Table 3 displays the results of comparing the training and validation datasets using oversampling. In comparison, the dataset testing uses the discussion forum dataset for the semester 2021.2. Research experiments were conducted by comparing the results of balanced data in fine-tuning the model to select pre-trained SentenceTransformers by comparing multilingual models.

<table>
<thead>
<tr>
<th>Undersampling dataset</th>
<th>Quantity</th>
<th>Oversampling dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>256</td>
<td>1,033</td>
</tr>
<tr>
<td>Validation set</td>
<td>97</td>
<td>476</td>
</tr>
</tbody>
</table>

Table 3

Undersampling and oversampling dataset

Source(s): Table by authors

Figure 3.
Data process flowchart

Source(s): Figure by authors
3.1 Pre-processing

Data cleaning, transformation and reduction are the procedures that makeup data pre-processing. Data cleaning is a pre-processing step that transforms raw data into an understandable format. Pre-processing data are essential because it can provide functions or benefits to data mining. First, text was analyzed through pre-processing techniques, explicitly removing HTML tags, including links, case-folding (lowercase), removing punctuation, removing special characters (such as emoticons, symbols and some signs found in addition to punctuation) and stopwords. Remove the selected frequent words from the dataset in the stopwords processing step. The word omitted for both question types is greetings (“selamat pagi”, “siang”, “sore”, “malam”, “asalamualaikum” and “bismillah”), introduction (“ijin menanggapi” and “maaf menggangu waktunya”), calls to lecturers or tutors and closing remarks (“terima kasih”).

The process of changing data from one format to another, often from the source system’s format to the destination system’s format, is known as data transformation. In this case, the value in the score column is normalized into a range of numbers 0–1. The normalization of these values will then become a reference in measuring the similarity of embedding sentence1 with sentence2 results. Data reduction frequently results in a loss of 1%–15% of the raw data’s variability (depending on how many components or characteristics are stored). The data reduction categories in this study were responses in the form of attachments, responses with a word length of more than 250 words and responses containing sub-responses.

3.2 Selected model

The accuracy of the model evaluation results between monolingual and multilingual models has been tested with several NLP tasks, concluding that the multilingual model’s performance is superior (de Vargas Feijó and Moreira, 2020). Previous researchers uploaded the pre-trained SentenceTransformers model he developed on the HuggingFace model hub website. In this process, an experiment was conducted to compare the multilingual pre-trained SentenceTransformers model on huggingface.com with the sentence similarity model category. A total of 11 pre-trained multilingual models, which have been trained in more than 50 languages, including Indonesian, are available. The multilingual paraphrase-multilingual-MiniLM-L12-v2 and distilbert-base-multilingual-cased-v1 models got the highest scores from comparisons of 11 pre-trained multilingual models of SentenceTransformers with Indonesian data (Dhini and Girsang, 2023). Both multilingual models were adopted in this study. Table 4 shows Pearson’s evaluation of the comparison of resampling with oversampling data getting a higher value of 0.63.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pearson Train</th>
<th>Pearson Val</th>
<th>Pearson Test</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undersampling data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual paraphrase-m</td>
<td>0.72</td>
<td>0.70</td>
<td>0.59</td>
<td>00:15:04</td>
</tr>
<tr>
<td>multilingual-MiniLM-L12-v2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distilbert-base-m</td>
<td>0.72</td>
<td>0.69</td>
<td>0.59</td>
<td>00:20:18</td>
</tr>
<tr>
<td>multilingual-cased-v1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oversampling data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual paraphrase-m</td>
<td>0.62</td>
<td>0.59</td>
<td>0.61</td>
<td>01:17:21</td>
</tr>
<tr>
<td>multilingual-MiniLM-L12-v2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distilbert-base-m</td>
<td>0.66</td>
<td>0.58</td>
<td>0.63</td>
<td>01:44:45</td>
</tr>
<tr>
<td>multilingual-cased-v1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Comparison of the multilingual model for undersampling and oversampling data

Source(s): Table courtesy of Dhini and Girsang (2023)

Note(s): The values are italicized based on Pearson -Test, i.e. the highest value is the best result
### 3.3 Generate semantic similarity

Semantic similarity will be generated using the *distillate-base-multilingual-cased-v1* model based on the outcomes of the multilingual comparison experiment. The model encodes *sentence1* and *sentence2* into vector embedding. SentenceTransformers convert words into sentence embedding in 1x768 dimensions. The semantic similarity between the students’ answers (*sentence1*) and reference answers (*sentence2*) can be established by measuring the document vector distance with an equal to the student’s answer document vector and B equal to the reference answer document vector. The cosine similarity equation can determine the distances between vector documents as follows (1).

$$
cos(\theta) = \frac{A \cdot B}{||A|| ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} 
$$

The angle between the vector embeddings of two text data sets is calculated using the cosine similarity metric. Cosine similarity between vectors as it performs pretty well in high dimensionality. When there is no angle between two embeddings, we obtain a $C(\theta)$ value of 1, which indicates that the embeddings are identical to one another. Table 5 displays the results of measuring the cosine similarity of *sentence1* and *sentence2*, which is the value of *param1*.

### 3.4 Generate keyword similarity

BERT is a transformer-based model for NLP. The pre-trained model can transform sentences or words into language representations of numbers. Words or sentences with similar

<table>
<thead>
<tr>
<th>No</th>
<th>Sentence1</th>
<th>Sentence2</th>
<th>cos_predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fermentasi digolongkan kedalam salah bioteknologi . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>fermentasi proses tahapan penguarian zat molekul . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>fermentasi berasal latin ferment enzim fermentasi . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>fermentasi proses pengawetan makanan alami dim . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>fermentasi tahapan tahapan fermentasi fermentasi . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.7</td>
</tr>
<tr>
<td>418</td>
<td>fermentasi respirasi anaerob respirasi oksigen . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.8</td>
</tr>
<tr>
<td>419</td>
<td>fermentasi membebaskan energi penerima elektron . . .</td>
<td>fermentasi membebaskan energi penerima elektro . . .</td>
<td>0.8</td>
</tr>
<tr>
<td>420</td>
<td>fermentasi proses repirsasi oksigen anaerob mudah . . .</td>
<td>fermentasi membebaskan energi penerima elektron . . .</td>
<td>0.8</td>
</tr>
<tr>
<td>421</td>
<td>fermentasi proses sel menghasilkan energi atp . . .</td>
<td>fermentasi membebaskan energi penerima elektron . . .</td>
<td>0.7</td>
</tr>
<tr>
<td>422</td>
<td>fermentasi respirasi anaerob fermentasi jalur . . .</td>
<td>fermentasi membebaskan energi penerima elektron . . .</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 5. Semantic similarity score

**Source(s):** Table by authors
representations (embedding) should mean anything semantically. Utilizing KeyBERT, this method can extract keywords from a text. The way keyword extraction works by extracting keywords and key phrases from a document and assigning weights to each word to indicate its importance in the document and the more comprehensive corpus. The stemming process is carried out at this stage before extracting the keywords in the response. The stemming process uses a Sastrawi library StemmerFactory.

The advantage of KeyBERT is that it adapts SentenceTransformers models so that the KeyBERT keyword extraction process can use the fine-tuned distillate-base-multilingual-cased-v1 model. The additional parameter keyphrase_ngram_range contains the N-gram range that should be considered when extracting keywords and key phrases. Only unigrams will be extracted using only the value (1, 1) (the default value) and the keyword. Another parameter determines the number of candidate keyword sequences needed (top_n). Then the model extracts the keywords to get the candidate keywords from the response (Table 6). The results obtained keywords in their order and the acquisition value of the essential keyword sequences in each sentence. The same is done in the rubric as a reference answer to get candidate keywords.

Keyword extraction using KeyBERT produces candidate keywords in responses (candidatesRespon) and rubrics (candidatesRubrik). The following process calculates the similarity of the candidates’ responses and rubrics with the cosine similarity values that become param2 (Table 7). At this stage, through the pre-processing stage, the keyword is extracted from the responses and rubric using KeyBERT. The implementation of the keyword similarity process between candidatesRespon and candidatesRubrik can be described in Figure 4.

### 3.5 Evaluation
After the pre-processing stage and getting each parameter value (param1 and param2), these values are combined into the final value. Hasanah calculates the essay value using two similarity methods of essay assessment scores and keyword matching with a ratio of 50:50 (Hasanah et al., 2019). Regarding previous research, this combination will give the final score composition, as shown in Table 2.
Correlation with manual scores: We determined the relationship between the similarity score and the actual grades provided by the subject matter experts in this section. The correlation is calculated using the Pearson correlation coefficient (equation 2). In addition, MAE is used to measure the error rate by finding the absolute difference between the score generated by the lecturer (X) and the score generated by the model (Y) and dividing it by the number of datasets obtained using formula (3).

\[
r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}}
\]  

(2)
4. Discussion

An experimental approach in the evaluation process of this research, two experimental approaches were made. (1) The first approach: this approach only compares student responses with rubrics by generating semantic similarity to get \( \text{param1} \) as a score. (2) Second approach: the combination of two parameters, namely \( \text{Param1} \) and \( \text{param2} \), which are obtained by comparing the response of the keyword-extraction responses with rubric keywords.

The evaluation shows that the oversampling dataset is better than the results of the undersampling dataset. The results of the evaluation with \( \text{param1} \) identify semantic similarity to get a correlation value of 0.63. In contrast, the evaluation with \( \text{param2} \) identifies the similarity keyword to get a correlation value of 0.61. Based on the experimental results, proposing a combination of \( \text{param1} \) and \( \text{param2} \) can increase the evaluation results by 0.2. The final value calculation using the composition fix gets the highest correlation, 0.65. These results are included in the strong correlation category based on the Pearson correlation value category. Tables 8 and 9 show the outcomes of the fixed-composition experiment using undersampling and oversampling datasets, respectively.

Compared with the monolingual model, experiments were carried out using IndoBERT, which had been trained from the enormous data corpus and Indonesian hygiene data collection (Indo4B, such as news, social media, blogs and websites). The results of the resulting model get a low value of 0.43 and take quite a lot of time (00:58:02), which is lower than the results of the multilingual model. SentenceTransformers are a technique for NLP problems with a more straightforward approach but more capabilities. In particular,
multilingual pre-trained models can perform very effectively. The latest innovation, xlm-RoBERTa (xlm-R), supports 100 languages (Moberg, 2020), while the remaining are competitive with monolingual alternatives. Thanks to current research, cross-language transmission is expected to improve. There are several reasons why this direction in research is important: as more attention is paid to power-efficient computing for use on small devices, the deep learning community will likely see a greater emphasis on smaller efficient models in the future.

Most studies employed statistical criteria like word count, number of sentences, sentence length, etc. However, 32% of the systems used content-based characteristics for the short answer and essay grading, per the data collection results (Ramesh and Sanampudi, 2022). Kaggle Automated Student Assessment Prize (ASAP) competition in 2012 is the essay grading tool utilized in 90% of English datasets. The limited Indonesian language dataset is an obstacle to AES’s development, so essay-scoring research is not so massively developed in Indonesia. Dhini and Girsang (2023) analyze the outcomes of various AES investigations conducted in Indonesia. Most studies on the classification employed the Ukara public dataset, which contains information from the automated short answer system, and they found that the SBERT or SentenceTransformers method could classify sentences with an accuracy of 80%. However, the results for Indonesian language instances in regression tasks are typically below 0.70.

A research study (Hasanah et al., 2019) got a Pearson evaluation AES score of 0.65. In contrast, in this study, the Pearson evaluation with a 50:50 assessment composition got 0.64, which can be said to be similar values, and the difference is insignificant. However, in the calculation results, measuring the model error rate using the MAE study (Hasanah et al., 2019) got a value of 0.90, while in this study, it was less than 0.2 with an error value of 0.70. The comparison can be considered disproportionate because the dataset is a short answer with several words below 20 (Hasanah et al., 2019). In contrast, the brief answer responses from discussion forums that make up the dataset for this study have an average word count of 86 and a maximum of 250 after pre-processing.

Student performance in the educational system is significantly affected by assessment. The current evaluation system uses human evaluation. Automated-essay scoring will be helpful when evaluating answers on a vast scale because manual correction can lead to several problems (Rajagede, 2021). If the model is used or implemented, the results can facilitate the development of a scoring system that can be adopted to enhance teaching and learning outcomes. It can be applied to assessing discussion forums in online tutorials, where tutors currently carry out manual assessments. Based on this background, the results of this research can be a solution as well as several other problem findings, such as not all tutors deposit the value of the discussion forum on time or according to schedule. This can impact the process of processing students’ final grades.

Both linguistics and machine learning are interested in the automated-essay scoring model. The approach can be used in education and large industrial enterprises to increase operational efficiency because it systematically categorizes writing quality. This research is not yet perfect and needs to be further investigated by adding more comprehensive data for all categories of teaching materials and increasing the amount of response data to train models with more data. We need a sizable amount of data to train the models to deal with the problem of text mining, particularly in the machine learning and deep learning domain (Rajagede, 2021). This study contributes to selecting a multilingual pre-training sentence modifier model that performs well in generating evaluation scores. These results can be a reference for further research using multilingual sentence modifier models, especially the Indonesian language dataset.
5. Conclusion

This study obtained an automatic assessment model for discussion forums with semantic similarity parameters using SentenceTransformers with the results of the correlation between the value of the model and the actual value (from the lecturer) of 0.63 and the error value of the model using an MAE of 0.70, and the study proposed SentenceTransformers by combining semantic similarity and keyword similarities in AES with the Indonesian language dataset to get the optimal model with an increase in the correlation of 0.2, but the value of the model error rate increases by 0.1. The resulting model has a Pearson correlation category, namely a strong correlation with a value between 0.60 and 0.80.

Based on the literature review and research results, many obstacles and limitations were found in development, the biggest obstacle being data collection for training data, which was insufficient and a model library specifically for Indonesians. The research is not yet said to be perfect and still needs to be further investigated by adding more comprehensive data for all categories of teaching materials and increasing the amount of response data to train models with more data. The SentenceTransformers model requires extensive training data and target assignment adjustments to achieve competitive performance. This contradicts the general case; very little training data are available. There are still many obstacles in running the AES system, some of which are the limitations of the dataset class (unbalanced data). For further studies in improving performance through small datasets, an effective data-augmentation method using SentenceTransformers known as Augmented-SBERT can improve the model’s accuracy.

References


**Corresponding author**
Bachirah Fatwa Dhini can be contacted at: riri@ecampus.ut.ac.id