Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hybrid Deep Learning and USE Algorithm for Essay Scoring: Accuracy and Performance Analysis

Agus sriyanto[1)*], Kusrini [2)]
[1)2)]universitas amikom yogyakarta, Indonesia
[1)]agussriyanto17@students.amikom.com, [2)] kusrini@amikom.ac.id

Submitted :May 18, 2025 | **Accepted** : Jun 2, 2025 | **Published** : Jun 4, 2025

Abstract: The main challenge in digital education, particularly in the automatic assessment of essay answers in online learning systems, lies in the complexity of natural language understanding and semantic evaluation required to achieve the level of precision equivalent to human judgment. This study aims to develop and analyze the performance of a hybrid model that combines deep learning with a semantic similarity-based approach to essay auto-grading. The methods used include the collection of essay answer data from various disciplines, text processing to extract semantic representations, and the calculation of the degree of similarity between the participant's answers and the answer key using the similarity measure. The evaluation was carried out by comparing the results of automatic assessments with manual assessments by teachers. The results showed that the developed model achieved the highest accuracy level of 90.22% at 0.8 treshold, with a precision of 84.63%, a recall of 100%, and an F1 score of 91.68%. To strengthen the reliability of the findings, statistical validation was carried out using error evaluation metrics. RMSE value is 0.32 and RMAE value is 0.19. These findings show that the model is able to mimic human judgment reliably and consistently, and can distinguish linguistic variations that arise in different types of essay questions. This system offers an effective solution for the automation of assessments in an online learning environment, while maintaining the integrity and objectivity of the evaluation.

Keywords: Automated Essay Assessment; Cosine Similarity; Universal Sentence Encoder; Online Learning.

INTRODUCTION

In the last two decades, the integration of information systems into the world of education through the Learning Management System (LMS) has brought significant changes in the teaching and learning process, including in Indonesia. LMS enables efficient and systematic course management, distribution of materials, assignment collection, and learning assessment (Ahmad et al., 2024). This transformation is in line with the global trend towards the digitalization of education, which has been shown to increase student engagement as well as the effectiveness of learning outcomes (Slimi & Villarejo-Carballido, 2023).

Among the various evaluation methods used in education, essay questions remain one of the important instruments because they are able to describe students' conceptual understanding, critical thinking skills, and written communication skills (Aulianda et al., 2023). However, manual essay grading often poses challenges, such as inconsistencies between graders, long correction times, and high workloads—especially in the context of large-scale education such as distance learning or Massive Open Online Courses (MOOCs)) (Wu et al., 2022;Mahdi & Alkhateeb, 2025).

To answer this challenge, the use of Natural Language Processing (NLP) and artificial intelligence (AI) technology is a strategic solution. This technology enables objective, consistent, and efficient automated assessments, even against large numbers of essays (Wong & Bong, 2021). The key to the success of this system lies in the use of structured rubrics, as well as the model's ability to capture the semantic nuances and context of students' language (Kumar & Boulanger, 2021).

Various approaches have been developed, ranging from rule-based systems to machine learning and deep learning. Rule-based systems have limitations in terms of flexibility, while machine learning models such as Support Vector Machine (SVM) and Naïve Bayes tend to be less optimal in understanding complex semantic

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

relationships (Amalia et al., 2021). On the other hand, deep learning-based approaches such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) show better results, but have a high computational load (C. Zhang et al., 2020).

One promising alternative is the use of vector representation models such as the Universal Sentence Encoder (USE), which is capable of translating sentences or paragraphs into vector forms with fixed dimensions and semantically meaningful (Amalia et al., 2021). To assess the similarity between students' answers and answer keys, Cosine Similarity is used as a metric to measure the proximity of meanings in vector spaces (Lahitani, 2022). Although each of these approaches has been shown to be effective separately, there has not been much research that explicitly combines USE and Cosine Similarity in one integrated hybrid architecture for automated essay assessment, especially in the context of Indonesian students.

This research gap is the basis of this study. This study aims to design and evaluate an automatic scoring system for hybrid model-based essay questions between Universal Sentence Encoder and Cosine Similarity. This system is designed to be integrated directly into the LMS, with the aim of measuring the accuracy, efficiency of processing time, and the alignment of automated assessment results with lecturers' manual assessments (Fajari & A. Baizal, 2022; Mirda et al., 2022).

This research is important because it offers a scalable and adaptive evaluative approach in the digital era, especially in supporting distance learning and blended learning systems. With the application of cutting-edge NLP technology, this system not only improves lecturer efficiency, but also encourages transparency and fairness in the evaluation process (Mujianto et al., 2024). In addition, this contribution also supports the digital transformation of education in the era of the Industrial Revolution 4.0 and Society 5.0 (Arfandy & Musdar, 2020).

LITERATURE REVIEW

A literature review serves as a critical, analytical synthesis of existing scholarly work, enabling researchers to identify prevailing trends, theoretical and methodological approaches, as well as underexplored research gaps. Within the domain of automated essay scoring (AES), especially in the context of Learning Management Systems (LMS), the evolution from rule-based evaluation to machine learning and deep learning paradigms marks a significant shift. However, despite the technological advancement, many existing models still struggle with fully capturing semantic depth and contextual relevance in student essays.

For instance, the study by (Beseiso & Alzahrani, 2020) introduced a machine learning-based AES system designed to deliver consistent and reliable assessments. While this system shows promise in automation and accuracy, it lacks a comprehensive mechanism for semantic context representation—an essential factor in understanding subjective responses. This gap becomes more apparent when compared to studies that incorporate context-aware encoders. (Kumari & Seeja, 2021), in their critique of deep learning models in AES, focus on the interpretability issues surrounding these models, referring to them as “black boxes.” Despite raising a valid concern, their work did not provide concrete solutions, such as the application of interpretable techniques like cosine similarity or contextual embedding models such as the Universal Sentence Encoder (USE), to improve model transparency.

In another vein, (Shin & Gierl, 2021) attempted to streamline the development of AES frameworks through algorithmic efficiency. Their method prioritized speed and reduced computational overhead; however, semantic validation was notably lacking. This indicates a trade-off between system efficiency and deep content understanding, a compromise that may not be acceptable in contexts where nuance and argument quality are critical—such as in higher education essay evaluation.

The hybrid model approach by (Birla et al., 2022) combines classical NLP techniques with neural networks, offering a more balanced methodology. Yet, their model falls short in leveraging context-based sentence-level semantics, relying instead on general feature extraction that may miss the subtle meaning embedded in students' language. In contrast, (Y. Zhang et al., 2020) proposed a more semantically rich model by merging question and answer contexts into a unified framework. Although theoretically strong, their model’s complexity and need for extensive datasets limit its applicability in real-world LMS settings that often lack vast annotated corpora.

From a broader perspective, (Hussein et al., 2019) delivered a valuable literature review tracing the evolution of AES from rule-based approaches to contemporary deep learning systems. While this review provides historical and conceptual insight, it lacks the technical specificity required to guide the implementation of advanced models that incorporate semantic similarity measures like cosine similarity or embedding-based methods such as USE or SBERT.

In response to these limitations, recent advancements have begun to emphasize embedding models that capture the semantic essence of sentences. (Cer et al., 2018) introduced the Universal Sentence Encoder, demonstrating its capacity to generate robust semantic representations for a variety of NLP tasks, including sentence similarity. However, its performance diminishes in the face of long, complex essays with varied syntax. To overcome this, (Chamidah et al., 2023) applied Sentence-BERT (SBERT) for short essay scoring, achieving high accuracy (CaF1: 88.5%), surpassing traditional models like TF-IDF and Word2Vec. Meanwhile, (Wang et al., 2022) developed a

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

multi-scale BERT architecture that integrates both local and global representations, successfully improving performance (QWK: 0.82) on the ASAP dataset. Locally, (Pradani & Suadaa, 2023) used IndoBERT to evaluate Indonesian student essays and demonstrated strong results (MAE: 0.185, RMSE: 0.2001), yet their model is tailored to specific languages and may not generalize well across linguistic contexts.

Taken together, these studies reflect a growing interest in context-aware AES systems, yet significant gaps remain. Many models are either too complex for practical implementation, lack semantic depth, or offer limited interpretability. The proposed study seeks to address these issues by introducing a lightweight hybrid model that integrates the Universal Sentence Encoder (USE) with cosine similarity, providing both semantic accuracy and computational efficiency. This method not only simplifies implementation in LMS environments but also ensures meaningful evaluation through the use of contextual embeddings, thereby filling a critical gap in current AES research.

METHOD

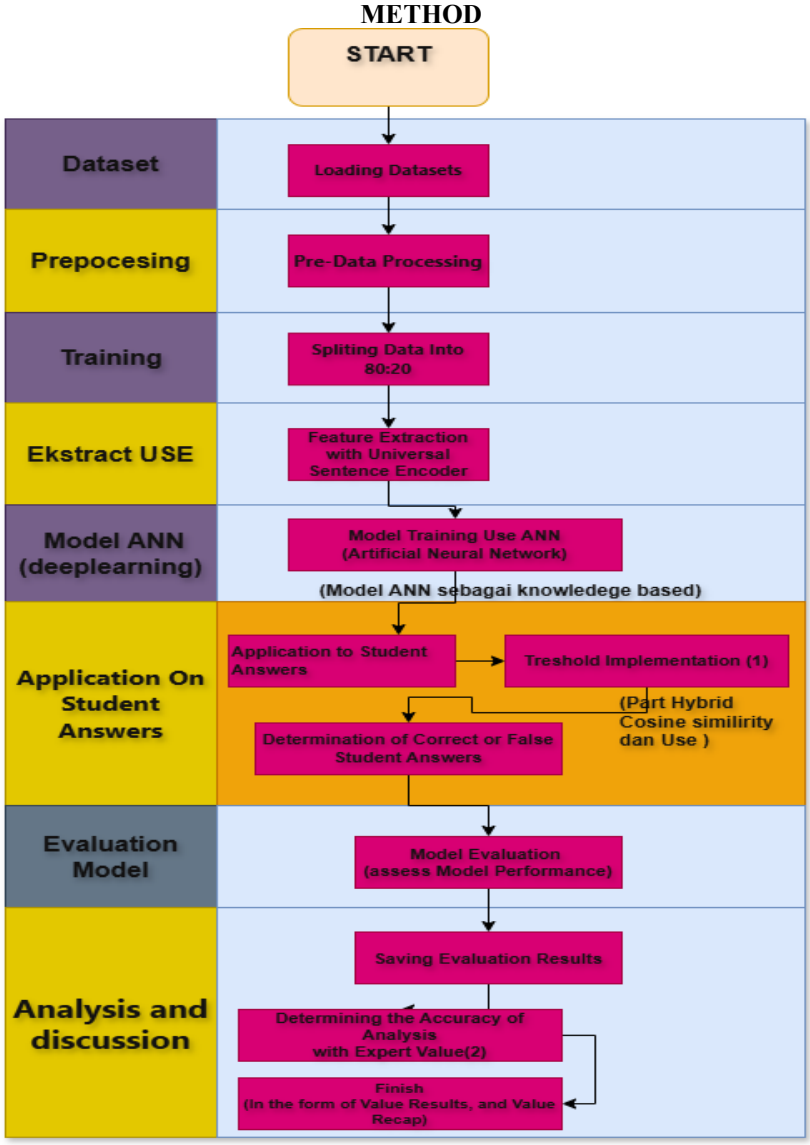

Figure 1. Research Flow for data analysis

This investigation comprises the primary phases as illustrated in figure 5 with the ensuing flow particulars, Steps: Start ,The procedure initiates with the initialization of the program or script that will be employed to process the data. This phase encompasses the preparation of the working environment, such as importing the requisite libraries (e.g., pandas, numpy, tensorflow-hub for USE, scikit-learn for data dissemination, etc.).

Loading Datasets Retrieve the question and answer dataset from an Excel file. The dataset comprises 2 tables consisting of a repository of questions and student responses acquired from 4500 student submissions. This dataset encompasses inquiries (questions) and correct responses (answer keys). The dataset is derived from public data available on kaggle.com. For an exemplification, reference can be made to Table 2.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Questions And Answers To Question Banks

Id	Question	Answer	Label
104	What are the three Hindu-Buddhist kingdoms in Indonesia!	Sriwijaya Kingdom, Majapahit Kingdom, Tarumanegara Kingdom, or ancient Kalingga or Mataram	1
105	Name three examples of physical and chemical changes!	Here are three examples of physical changes and chemical changes: Physical changes (not producing new substances): Ice melts into water, Paper is cut into small pieces, Water evaporates when heated. Chemical Changes (producing new substances): Wood burned into charcoal and ash, Milk that is stale due to the fermentation process, Iron that rusts due to the reaction with oxygen and water.	1
106	Translate into English from the sentence: 'I like to eat fried rice.'	I like eating fried rice	1

Load the comprehensive dataset comprising student responses from an Excel file that has been meticulously organized for academic analysis. This particular dataset encompasses the varied and nuanced answers that were provided by students in response to each individual question posed in the assessment. The structural format of this dataset exhibits similarities to that of the question dataset; however, it is important to note that the columns within this student response dataset are specifically designated to contain the individual responses articulated by each student. As illustrated in Table 3, one can observe a representative example of the answers submitted by students, which serves to highlight the diversity and range of their responses.

.

Table 2. Student Answers

Student's Name	Class	Id	Question	Student Answers
AFWAN MAULANA	X TKR 1	104	What are the three Hindu-Buddhist kingdoms in Indonesia!	HinduKingdom1.ancientmataram2.kendan3.sundaBuddhist Kingdom1.malay2.sriwijaya3.majapahit
shifa nur azizah	X TKJ 3	104	What are the three Hindu-Buddhist kingdoms in Indonesia!	Ancient Mataram Kingdom, Sigasari Kingdom, Majapahit Kingdom.

Data Pre-processing Removing Missing Values In the context of data analysis, it is not uncommon for datasets to exhibit the presence of empty cells, which are often represented as NaN (Not a Number) or NULL values, and this phenomenon can significantly disrupt and complicate the overall analysis process. Therefore, one widely accepted technique for addressing this issue involves either the complete deletion of entire rows that contain these empty values or alternatively, the substitution of the blank values with a predetermined default text string, such as "[NOT ANSWERED]", to ensure that the dataset remains as comprehensive and informative as possible.

Convert data to string types It is imperative to ensure that all textual elements within the dataset, which encompass the questions posed, the answers provided, as well as the responses given by students, are uniformly formatted as string data types in order to maintain a high level of consistency throughout the analytical process and to facilitate more accurate comparisons and interpretations of the data.

Sharing Data Separating data into training data and test data (80:20) The dataset utilized for the analysis is judiciously partitioned into two distinct subsets, wherein the training data comprises 80% of the total dataset and is specifically employed for the training of various predictive models, while the remaining 20% is allocated as test data, which serves the critical purpose of evaluating and assessing the performance and efficacy of the trained model in making accurate predictions.

Using train_test_split from scikit-learn: To facilitate this separation of data into training and testing subsets, the method known as train_test_split, which is part of the scikit-learn library, is effectively utilized, thereby enabling a methodologically sound approach to the division of the dataset into the respective components necessary for model training and evaluation.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Feature Extraction with Universal Sentence Encoder (USE) Convert text to numeric vector The Universal Sentence Encoder (USE), a sophisticated text embedding model that has been meticulously developed by Google, functions to transform sentences into high-dimensional numeric vectors, specifically of 512 dimensions, thereby allowing for a more nuanced representation of textual information that can subsequently be leveraged in various machine learning applications.

Model Training Neural Network Architecture In the endeavor of constructing a neural network model architecture that is tailored for the task of binary classification, which entails determining the veracity of statements as either True or False, it is essential to carefully delineate the function of each layer within the network, ensuring that it is designed to effectively process and analyze the text vectors derived from the Universal Sentence Encoder (USE), which are characterized by their 512-dimensional format.

Table 3. Neural Network Architecture Explained

Layer	Explanation
InputLayer(input_shape=(512,))	The input layer receives a vector with dimensions of 512 (output from USE).
Dense(256, activation='relu')	A hidden layer with 256 neurons and ReLU activation to capture non-linear patterns.
Dropout(0.3)	Dropout layer with a rate of 30% to prevent overfitting by shutting down random neurons during training.
Dense(128, activation='relu')	The second hidden layer with 128 neurons and ReLU.
Dropout(0.3)	The second layer dropout with a rate of 30%.
Dense(1, activation='sigmoid')	Output layer with 1 neuron and sigmoid activation for binary classification (0=False, 1=True).
optimizer='adam'	Adam optimizer for efficient weight adjustment.
loss='binary_crossentropy'	Loss function for binary classification problems.
metrics=['accuracy']	Evaluation metrics: the accuracy of the percentage of the prediction is correct.

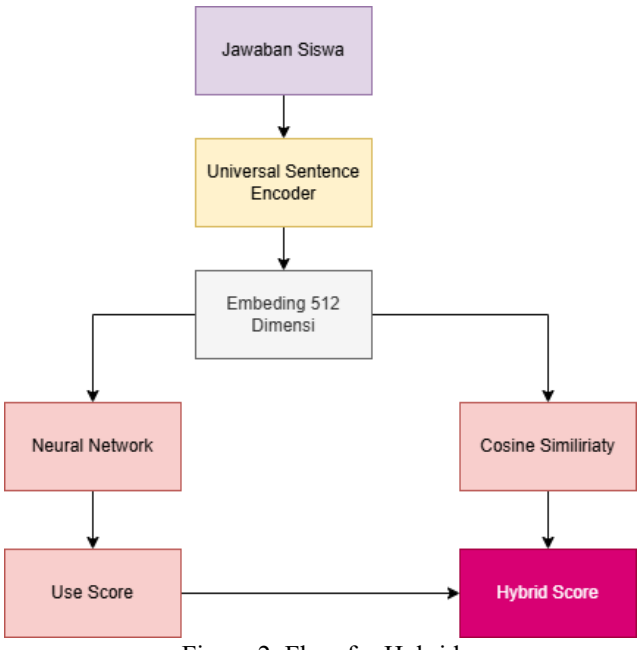

Figure 2. Flow for Hybrid

Figure 2 explains The process of assessing student answers begins with the text preprocessing stage, where the student's answers and answer keys are cleaned through several important steps. First, the text is tokenized into individual words, then stopwords such as "which", "and", or "in" are omitted to focus on the keyword. Furthermore, the words go through a stemming process using Sastrawi Stemmer to return the word to its basic form, for example "teaching" to "teaching".

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. Example Output from hybrid processing prediction

Student Name	Use_score (NN)	Cosine_similarity	Hybrid_score	Predictions (Threshold 0.8)
Andi	0.92	0.85	0.885	True (1)
Mind	0.75	0.82	0.785	False (0)

Application to Student Responses (Hybrid Utilization and Cosine Similarity). The Hybrid Cosine Similarity methodology is delineated as follows, Sentence Representation: The inquiry sentences, accurate answers, and student responses are transmuted into vector representations utilizing the Universal Sentence Encoder (USE) model. This vector encapsulates the semantics inherent to each sentence. The cosine similarity metric is computed between the vector representation of the correct answer and that of the student's response. The cosine similarity value serves as an indicator of the nearness between two vectors, where a value approaching 1 signifies a high degree of similarity (the student's response closely aligns with the correct answer), and a value nearing 0 indicates the contrary. Subsequent to the computation of cosine similarity, this resultant value is amalgamated with the predictions derived from the deep learning (USE-only) model to generate a hybrid score. The hybrid score is computed as the mean of the cosine similarity and the USE prediction score:
In the coding segment as follows:

```
dataset_siswa['use_score'] = use_scores
dataset_siswa['cosine_similarity'] = cosine_scores
dataset_siswa['hybrid_score'] = (np.array(use_scores) + np.array(cosine_scores))/2
```

The ensuing step involves the determination of a threshold to address the formulation of problem No. 1 in the phase of threshold establishment. Determine the optimal threshold that classifies students' responses as True or False through a comparison of the student answer vector and the answer key. Process: Compute the cosine similarity between the student answer vector and the answer key vector. Cosine similarity serves as a measure of semantic resemblance (range: -1 to 1, with values closer to 1 indicating greater similarity). Experimentation with a variety of threshold values (e.g., 0.6, 0.7, 0.8) and subsequent evaluation of the outcomes is essential. Example: If similarity ≥ 0.7 → True, otherwise → False. Select the threshold that yields the highest accuracy with respect to the validation dataset. Determining True or False Responses Employ the chosen threshold to transmute similarity scores into binary labels (1=True, 0=False). Contrast these with the original labels (if ground truth data from experts is available).

Model Evaluation Accuracy: The proportion of correctly classified answers is assessed. Confusion Matrix: Visualization of True Positives, False Positives, and other relevant metrics. Precision and Recall: These metrics are critical in scenarios involving class imbalance. Example Results: Accuracy: 85% → This indicates that the model aligns with 85% of the evaluations provided by experts.

The subsequent step is to ascertain the accuracy value to respond to the formulation of problem No. 2 in the phase of evaluating the accuracy between the system's responses and those of the experts. To validate the congruence between the system's assessment and the expert's evaluation. Step: Contrast the outputs of the classification system (y_pred) with the expert's labels (y_test). Calculate accuracy, kappa coefficient (to assess reliability), and F1-score (ensuring the formula and calculation method are elucidated). Conduct a divergence analysis (for example, examining instances of discrepancies between the system and expert evaluations). Next Steps: Integrate the results with the original dataset. Introduce new columns to the student answer dataset: Similarity Score, Prediction (True/False), Expert Correction (if applicable). Subsequently, save this information in an Excel file. Save the updated DataFrame to a new file:

Table 5. Correct Answer Output

Student's name	**class**	**cosine_similarity**	**use_score**	**hybrid_score**	**Answer_true**
MOHAMMAD KRISNA FEBRIAN	XII TITL 3	0,854485273	0,986285806	0,92038554	1
Moch Ilham Syahputra Kusuma	X TKR 2	0,511083543	0,992788553	0,751936078	0

RESULT

This research aims to develop and test an essay automatic correction system based on *Universal Sentence Encoder* (USE) and *Cosine Similarity* in the context of a Learning Management System (LMS). Based on these objectives, the results of the study show that the developed system can measure the semantic similarity between students' essay answers and answer keys with a high level of accuracy. The dataset used can be seen in table 2.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 6. Performance Results Of Each Treshold

Threshold	Accuracy	Precision	Recall	F1-Score	Note
0.6	57.43%	55.86%	100%	71.68%	The recall is very high (100%), but low precision → a lot of false positives.
0.7	68.51%	63.11%	100%	77.38%	F1-Score improves, precision balance and recall are better.
0.8	90.22%	84.63%	100%	91.68%	All metrics are very high → models are excellent.
0.9	80.99%	100%	64.71%	78.57%	Precision is perfect, but the recall drops → a lot of false negatives.

Analysis of Threshold Selection Based on Evaluation Metrics

In order to meticulously ascertain the most advantageous threshold within a classification system, it becomes paramount to conduct a comprehensive evaluation of various metrics, including but not limited to accuracy, precision, recall, and the F1-score, all of which provide critical insights into the model's performance and effectiveness. The table presented above illustrates the performance of the model across four distinct threshold values, specifically 0.6, 0.7, 0.8, and 0.9, thereby enabling a comparative analysis of how these thresholds influence the evaluation metrics.

Threshold 0.6, at the threshold value of 0.6, the recall metric achieves an impressive 100%, signifying that the model has successfully identified all instances of positive cases without exception, thus demonstrating its efficacy in this regard. Nevertheless, it is crucial to note that the accuracy metric is recorded at a mere 55.86%, suggesting a significant occurrence of false positives, which raises concerns about the reliability of the model's predictions in terms of overall correctness. Furthermore, the accuracy remains relatively low at 57.43%, indicative of numerous prediction errors that detract from the model's overall performance and reliability. The F1-score is calculated at 71.68%, highlighting a notable imbalance between the precision and recall metrics, which could potentially undermine the model's utility in practical applications. This threshold may be deemed appropriate in scenarios where the paramount concern is the non-miss of positive cases, such as in critical fields like fraud detection or disease diagnosis, even if this approach results in a substantial number of false positives that could complicate interpretations.

Threshold 0.7, When analyzing the threshold set at 0.7, it is noteworthy that the recall remains consistently high at 100%, indicating that the model continues to detect all positive cases with utmost efficacy. In conjunction with this, the precision metric has seen a commendable increase to 63.11%, and the F1-score has correspondingly risen to 77.38%, reflecting improvements in the model's ability to maintain a balance between correctly identifying positive cases and minimizing false positives. Additionally, the accuracy has also experienced an upward shift, now standing at 68.51%, which enhances the overall credibility of the model's predictive capabilities. This particular threshold demonstrates a notable enhancement in the equilibrium between precision and recall when contrasted with the earlier threshold of 0.6, while still maintaining the maximum level of recall.

Threshold 0.8, The threshold value of 0.8 emerges as the most favorable option, providing what can be described as the best overall performance across all evaluated metrics, which are critical in assessing the efficacy of the classification model. The accuracy at this level is an impressive 90.22%, while the precision is recorded at 84.63%, and recall remains at an outstanding 100%, thereby reinforcing the model's ability to identify all positive instances without fail. Furthermore, the F1-score reaches a notable 91.68%, underscoring the model's proficiency in achieving high reliability in predicting both positive and negative classes, accompanied by a remarkably minimal occurrence of errors that would otherwise detract from its performance.

Threshold 0.9, With the threshold elevated to 0.9, the accuracy metric achieves a perfect score of 100%, indicating that the model has successfully eliminated all instances of false positives, which is a desirable outcome in many contexts. However, this remarkable accuracy comes at a significant cost, as the recall metric experiences a drastic decline to 64.71%, suggesting that a considerable number of positive cases have unfortunately gone undetected, thereby raising concerns regarding the model's ability to identify all relevant instances. In addition, the accuracy has seen a decrease to 80.99%, and similarly, the F1-score has also diminished to 78.57%, reflecting a trade-off between the precision achieved and the risk of missing positive detections. This threshold may be deemed suitable for applications where the absolute avoidance of false positive errors is an imperative requirement, albeit with the understanding that this choice inherently tolerates the risk of failing to detect certain positive cases, which could have significant implications in critical decision-making scenarios.

It can be concluded that the test results with a treshold of 0.8 have the right limit with Accuracy: 90.22%, Precision: 84.63%, Recall: 100%, F1-score: 91.68%.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14784
e-ISSN : 2541-2019
p-ISSN : 2541-044X

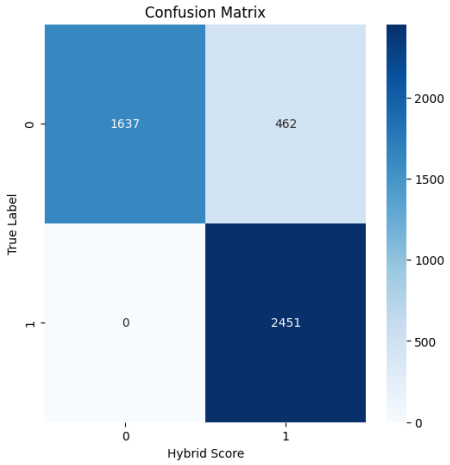

Figure 3. Confusion Matrix

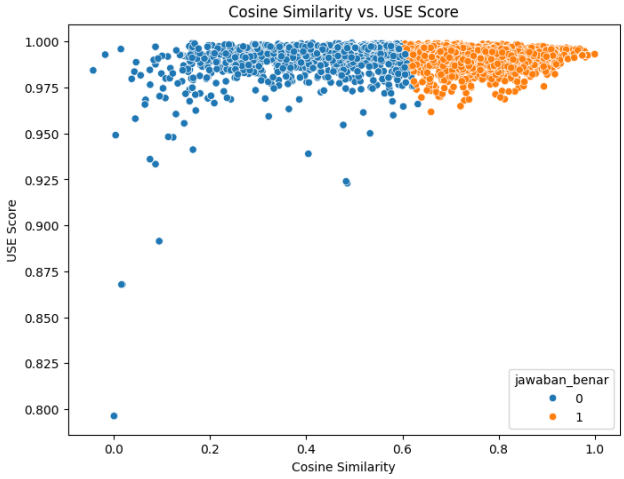

Figure 4. Cosine Equation Vs USE Score

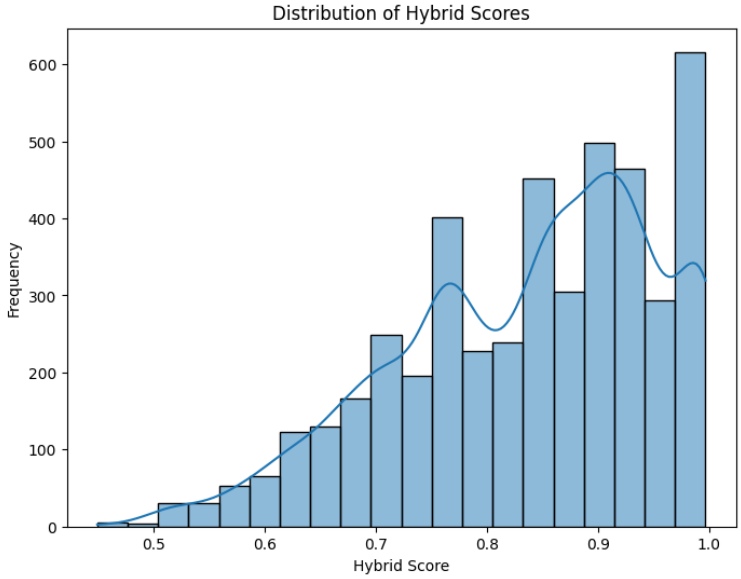

Figure 5. Distribution Of Hybrid Score Results

To strengthen the reliability of the findings, statistical validation was carried out using error evaluation metrics. That is, with an RMSE (Root Mean Square Error) test of 0.32, it means that the average model prediction error on the essay score is only 0.32 points from the actual value (on a scale of 0–1, for example). This shows minor deviations, meaning that the model is quite close in mimicking human judgment. In the context of an essay

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

score scale (e.g. from 0 to 1), this score is considered quite good because it is close to 0. An RMAE (Relative Mean Absolute Error) of 0.19 means that the average prediction error is only 19% of the actual value. This suggests that the model makes relatively small errors the predictions are consistent and not far from the original value. RMAE < 0.2 is usually considered good, especially in natural language-based automation systems.

Practically, the results of this research make a significant contribution in the field of *Educational Technology*. The implementation of an automated correction system based in the LMS can support teachers in managing essay grading tasks more quickly and objectively. This system allows teachers to provide faster feedback to students, while reducing the administrative workload associated with manual assessments. Therefore, this system can be integrated as an effective tool in improving the quality of learning and teaching in the higher education environment.

DISCUSSIONS

This study presents an analysis of how varying threshold values affect the classification model's performance concerning the metrics: accuracy, precision, recall, and F1-score. It is noted that increasing the value of the threshold from 0.6 to 0.8 continues to improve both accuracy and precision while maintaining recall (which flatlined at 100%) until finally at threshold 0.9 recall indiscriminately dropped while precision peaked. In summary Threshold 0.6–0.7 achieves recall 100% but precision is low (55.86%–63.11%), indicating a risk of over-predicting positive cases. Threshold 0.8 achieved optimal results with accuracy at 90.22%, precision 84.63%, and recall 100% yielding F1-score was 91.68%, the highest value. Threshold 0.9 precision becomes 100% but lowers recall to 64.71% indicating a high level of false negatives present.

In the context of mapping the F1 score during this study, the analysis of all existing threshold values, showed that at threshold 0.8 the optimal balance between precision and recall was obtained. In this case, the F1-score is the highest. It was also found in the benchmark carried out by , in the development of objects and re-identification of online fuel adaptive thresholding techniques. The model is claimed to increase the accuracy of the model by 12-45% if the thresold used is fixed at 0.3, 0.5, 0.7 or 0.7. On the other hand, it proposes an OTLP Framework that applies Mixed Integer Linear Programming to obtain an optimal threshold in classification, especially on classification problems with a very high level of class imbalance. With this approach, the performance model of detecting credit card fraud has improved. In terms of evaluation, in trying to generalize F1 metrics in their probabilistic approach and automation by rethinking what is in precision, recall, specificity and negative predictive values. Where the depiction of the limitations of the calculation on the confusion matrix is really extreme, the performance of the system is highly skewed.(Bohara, 2020)(Koseoglu et al., 2024)(Sitarz, 2022)

In an effort to validate the performance of the developed hybrid model, statistical evaluation metrics such as Root Mean Square Error (RMSE) and Relative Mean Absolute Error (RMAE) are used. The RMSE measures the square mean of the difference between the predicted value and the actual value, placing more emphasis on large errors. Meanwhile, RMAE provides a perspective relative to the average absolute error compared to the actual value, allowing for a more contextual interpretation of the model's performance. In this study, an RMSE value of 0.32 indicates that the average deviation between the score predicted by the model and the actual score is quite low, indicating that the model is able to make predictions with minimal error. An RMAE value of 0.19 indicates that the average absolute error is only 19% of the actual value, confirming the model's ability to provide judgments that are close to human evaluation. These findings are in line with recent research that emphasizes the importance of using metrics such as RMSE and MAE in the evaluation of automated scoring models. For example, a study by (Chansanam et al., 2021) highlights the successes and challenges in Massive Open Online Courses (MOOCs), where automated evaluations play a crucial role at scale. Additionally, research by (Wu et al., 2022) shows that the use of these metrics can improve accuracy in automated essay assessment. Thus, the integration of RMSE and RMAE metrics in the evaluation of automated essay scoring models not only provides statistical validation of the model's performance, but also reinforces confidence in the application of these systems in the broader educational context.

Choosing the exact threshold is critical within a classification model, especially in the case of an imbalanced dataset. threshold adjustment can greatly improve model quality, and tools such as ClassificationThresholdTuner can aid in this matter. In addition, the study by emphasizes considering broader evaluation metrics, such as probabilistic F1, to obtain more accurate estimates of model performance. For real-world scenarios like fraud detection or medical diagnosis, optimal threshold selection can have a substantial effect on outcome. Hence, flexible approaches and comprehensive evaluation metric integration are strongly recommended.(Sitarz, 2022)

This study mimics the call of model power with the model user scenario in the real world by implementing it, we hope that future research can experiment with context-influenced dynamics rather than just using the convex metrics of the ROC – AUC and Precision Recall Curve to set more flexible limits on threshold setting limits. The prime flaw in this paper is that the case study does not apply to a wider scope of applications:Threshold were evaluated statically; they have not been tested on approaches using adaptive thresholds that consider the data context in real-time. Metric results were not accompanied by any analysis of statistical significance (confidence

*name of corresponding author
This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

interval or hypothesis testing, for example), which is important for generalization. The research still offers great insights toward determining the optimal threshold, but it also leaves great room for further exploration in improving the classification system's adaptability and precision across various contexts.

CONCLUSION

Conclusions were drawn based on the research objective, which aims to identify the optimal threshold value within a classification system to achieve optimal model performance in terms of accuracy, precision, recall, and F1-score. From the analysis results, it was discussed that a threshold of 0.8 achieved the highest accuracy at 90.22%, precision at 84.63%, recall at 100%, and F1-score at 91.68%. This demonstrates the model's ability to detect all positive cases without sacrificing the accuracy of positive predictions while minimizing overall classification errors. On the other hand, overly low thresholds (0.6) do maintain 100% recall, but low precision leads to a large number of erroneous predictions. While overly high thresholds (0.9) drastically reduce recall, meaning that many positive cases go undetected. This study recommends implementing a threshold of 0.8 for systems requiring an equilibrium between recall and precision. Average Deep Learning Execution Time (USE only): 131.09 ms ms, Average Hybrid Execution Time (Cosine Similarity + USE): 1.94 ms. To strengthen the reliability of the findings, statistical validation was carried out using error evaluation metrics. That is, with an RMSE (Root Mean Square Error) test of 0.32, it means that the average model prediction error on the essay score is only 0.32 points from the actual value (on a scale of 0–1, for example). This shows minor deviations, meaning that the model is quite close in mimicking human judgment. In the context of an essay score scale (e.g. from 0 to 1), this score is considered quite good because it is close to 0. An RMAE (Relative Mean Absolute Error) of 0.19 means that the average prediction error is only 19% of the actual value. This suggests that the model makes relatively small errors—the predictions are consistent and not far from the original value. RMAE < 0.2 is usually considered good, especially in natural language-based automation systems. As a result, implementing the correct threshold has the potential to significantly boost the model's effectiveness and reliability, especially in critical applications such as detection systems, automated evaluation, and machine learning-based decision-making. Further research is encouraged to test these parameters using different datasets and contexts while implementing an adaptable data-driven threshold determination strategy.

In the future, this research paves the way for various further explorations. It is recommended that this model be tested on different types of data, fields of study, and languages to test its adaptability. In addition, a dynamic and data-driven threshold approach needs to be developed so that the system can automatically adapt to the context of its use. Integrating AI technology that can provide explanations (explainable AI) is also an important step so that this system is not only intelligent, but also transparent and trustworthy. Through these advanced measures, automated evaluation systems in the future will not only replace manual assessment tasks, but also expand the horizons of a more equitable, fast, and meaningful digital education.

REFERENCES

Ahmad, I. F., Setiawati, F. A., Prihatin, R. P., Fitriyah, Q. F., & Thontowi, Z. S. (2024). Technology-Based Learning Effect on the Learning Outcomes of Indonesian Students: A Meta-Analysis. *International Journal of Evaluation and Research in Education (Ijere)*. https://doi.org/10.11591/ijere.v13i2.25383

Amalia, E. L., Jumadi, A. J., Mashudi, I. A., & Wibowo, D. W. (2021). Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA). *Jurnal Teknologi Informasi Dan Ilmu Komputer*. https://doi.org/10.25126/jtiik.2021824356

Arfandy, H., & Musdar, I. A. (2020). Rancang Bangun Sistem Cerdas Pemberian Nilai Otomatis Untuk Ujian Essai Menggunakan Algoritma Cosine Similarity. *Inspiration Jurnal Teknologi Informasi Dan Komunikasi*, *10*(2), 123. https://doi.org/10.35585/inspir.v10i2.2580

Aulianda, N., Wijayati, P. H., Ebner, M., & Schön, S. (2023). Analysis of Learning Management System Towards Students' Cognitive Learning Outcome. *International Journal of Emerging Technologies in Learning (Ijet)*. https://doi.org/10.3991/ijet.v18i23.36443

Beseiso, M., & Alzahrani, S. (2020). An Empirical Analysis of BERT Embedding for Automated Essay Scoring. In *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2020.0111027

Birla, N., Kumar Jain, M., & Panwar, A. (2022). Automated assessment of subjective assignments: A hybrid approach. *Expert Systems with Applications*, *203*, 117315. https://doi.org/10.1016/j.eswa.2022.117315

Bohara, B. (2020). *Adaptive Threshold for Online Object Recognition and Re-identification Tasks*. http://arxiv.org/abs/2012.14305

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder*. http://arxiv.org/abs/1803.11175

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Chamidah, N., Yulianti, E., & Budi, I. (2023). Evaluating the Impact of Sentence Tokenization on Indonesian Automated Essay Scoring Using Pretrained Sentence Embeddings. *Revue d'Intelligence Artificielle*, *37*(5), 1101–1108. https://doi.org/10.18280/ria.370502

Chansanam, W., Poonpon, K., Manakul, T., & Detthamrong, U. (2021). Success and Challenges in MOOCs: A Literature Systematic Review Technique. *TEM Journal*, 1728–1732. https://doi.org/10.18421/TEM104-32

Fajari, A. N., & A. Baizal, Z. K. (2022). Chatbot-Based Culinary Tourism Recommender System Using Named Entity Recognition. *Jipi (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*. https://doi.org/10.29100/jipi.v7i4.3210

Hussein, M., Hassan, H., & Nassef, M. (2019). Automated Language Essay Scoring Systems: A Literature Review. In *Peerj Computer Science*. https://doi.org/10.7717/peerj-cs.208

Koseoglu, B., Traverso, L., Topiwalla, M., Kraev, E., & Szopory, Z. (2024). *OTLP: Output Thresholding Using Mixed Integer Linear Programming*. http://arxiv.org/abs/2405.11230

Kumar, V., & Boulanger, D. (2021). Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, *31*(3), 538–584. https://doi.org/10.1007/S40593-020-00211-5

Kumari, P., & Seeja, K. R. (2021). A novel periocular biometrics solution for authentication during Covid-19 pandemic situation. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 1–17. https://doi.org/10.1007/S12652-020-02814-1

Lahitani, A. R. (2022). Automated Essay Scoring Menggunakan Cosine Similarity Pada Penilaian Esai Multi Soal. *Jurnal Kajian Ilmiah*. https://doi.org/10.31599/jki.v22i2.1121

Mahdi, H. S., & Alkhateeb, A. A. (2025). Revolutionising Essay Evaluation. *International Journal of Computer-Assisted Language Learning and Teaching*. https://doi.org/10.4018/ijcallt.368226

Mirda, M., Sarjan, M., & Khairat, U. (2022). Aplikasi Ujian Essay Koreksi Otomatis Menggunkan Metode Cosine Similarity. *Journal Peqguruang Conference Series*. https://doi.org/10.35329/jp.v4i1.2344

Mujianto, A. H., Aswin, I. B., & Susanto, E. S. (2024). Implementasi Algoritma Cosine Similarity Untuk Koreksi Jawaban Ujian Essay Berbasis Website. *Jurnal Informatika Teknologi Dan Sains*. https://doi.org/10.51401/jinteks.v6i3.4214

Pradani, K. A., & Suadaa, L. H. (2023). Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(6), 1177–1184. https://doi.org/10.25126/jtiik.2023107338

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247–272. https://doi.org/10.1177/0265532220937830

Sitarz, M. (2022). *Extending F1 metric, probabilistic approach*. https://doi.org/10.54364/AAIML.2023.1161

Slimi, Z., & Villarejo-Carballido, B. (2023). Systematic Review: AI's Impact on Higher Education - Learning, Teaching, and Career Opportunities. *Tem Journal*. https://doi.org/10.18421/tem123-44

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3416–3425. https://doi.org/10.18653/v1/2022.naacl-main.249

Wong, W. S., & Bong, C. H. (2021). Assessing Malaysian University English Test (MUET) Essay on Language and Semantic Features Using Intelligent Essay Grader (IEG). *Pertanika Journal of Science and Technology*. https://doi.org/10.47836/pjst.29.2.12

Zhang, C., Li, Q., & Cheng, X. (2020). Text Sentiment Classification Based on Feature Fusion. In *Revue D Intelligence Artificielle*. https://doi.org/10.18280/ria.340418

Zhang, Y., Lin, C., & Chi, M. (2020). Going deeper: Automatic short-answer grading by combining student and question models. *User Modeling and User-Adapted Interaction*, *30*(1), 51–80. https://doi.org/10.1007/s11257-019-09251-6

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.