Transformer Model Fine-Tuning for Indonesian Automated Essay Scoring with Semantic Textual Similarity

1st Abdul Hamid Nugroho
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
abdul.hamid.n@mail.ugm.ac.id

2nd Indriana Hidayah
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
indriana.h@ugm.ac.id

3rd Sri Suning Kusumawardani
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
suning@ugm.ac.id

2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-6654-5512-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISRITI56927.2022.10052865

***Abstract*—The quality of exam assessment is a very important part of education. Because the assessment process plays a role in various aspects of education. The results of the assessment are used to measure students' abilities, as a reference for achieving learning targets, evaluating educational curriculum, evaluating learning processes and others. In manual assessment, several problems arise when the amount of data that must be assessed is large. Manual assessments are time-consuming, subjective and will potentially lead to unbalanced assessments, especially if there are multiple raters involved. Automated Essay Scoring (AES) emerged as a new research field to address this problem. Many researchers have conducted research with various methods to overcome some of these problems. The most recent technology namely neural network has recently given fantastic results in NLP task. However, most of these AES systems use large datasets, so it's relatively common to get good results. Bidirectional Encoder Representations from Transformers (BERT)-based approach can improve NLP tasks with limited training data. But still has some drawbacks when used specifically in AES. In this paper, we propose a transformer-based AES model which is optimized by fine-tuning and hyperparameter-optimization methods to produce more accurate scoring. The results obtained based on the Quadratic Weighted Kappa (QWK) measurement are 93% and the accuracy is 92%.**

Keywords—automated essay scoring, deep learning, natural language processing, transformer, semantic textual similarity

I. Introduction

One of the important factors in the world of education is the assessment of test answers. The results of this assessment are used to measure the level of student ability in the learning that has been undertaken. If the quality of the assessment is good, it will produce accurate results for measuring student abilities. These results can be used as an evaluation material in terms of learning, student abilities, curriculum, or other things. So, with a good quality assessment can improve the quality of education [1].

The most frequently used answer assessment is the manual method with human assessment. A teacher must read the entire essay and grade one by one. This kind of assessment has several weaknesses, some of them are taking a long time and being prone to subjectivity if there are several people doing the assessment. Furthermore, if the number of essays assessed is large, the quality of the assessment will decrease further due to the inconsistency of the assessment [2].

Some researchers are trying to solve this problem by utilizing the capabilities of Artificial Intelligence (AI) by creating Automated Essay Scoring (AES). AES is a model that could assess essay answers based on the knowledge that has been learned. So, with this AES some of the shortcomings of manual assessment can be overcome. Several methods and technologies have been applied by researchers to build AES models including statistical calculations, machine learning, and deep learning [3].

Some of these methods, each has advantages and disadvantages. Statistical calculations, for example, can make a good assessment of data that does not pay much attention to word order. This is because statistical assessments only use statistical data, without considering other textual aspects such as coherence and semantics (text meaning) [3] [4] [5]. Then several other studies used machine learning techniques to accommodate textual assessments. The results from AES with machine learning techniques, especially with the classification task, give decent and accurate results [6] [7] [8] [9] [10]. Moreover, several AES research applies a neural network technique known as Deep Learning. There are several types of neural networks that are used to build AES models. In research [2] [11] [12] [13] [14] applied deep learning techniques using the Long Short-Term Memory (LSTM) model.

Based on several previous studies, it can be concluded that each model has advantages and disadvantages. But the ones that currently have the best performance are the models with transformer base [15]. One of the implementations of transformers is the Bidirectional Encoder Representation from Transformers (BERT) model. Some factors that make transformers have good results, in processing text, the transformer model processes text simultaneously, in contrast to LSTM which processes words sequentially. With this, in processing long text, transformers can access the encoding of all parts of the text. Another thing is that with the above mechanism, transformers can do more parallel training processes.

II. Related Work

In this study we focus on the method with a transformer-based model, because currently the AES model with the best assessment capability is built with a transformer-based model. One of the studies based on the same method is [1], The study proposes a hybrid model that can derive the interaction of words in an essay using BERT self-attention transformers, along with manually constructed syntactic features. The model is trained with the Automated Student Assessment Prize (ASAP) dataset, which has a large amount of data, so it can improve the quality of the assessment of the designed model. The evaluation metric used is Quadratic Weighted Kappa (QWK) with results in the range of 0.7 to 0.8.

In the study [16] built the AES model with the ensemble technique, which uses several techniques into one assessment model. The technique used is string-based similarity, semantic-based similarity, and embedding. Embedding models used include BERT, GloVe, and RoBERTa. This study shows that using simple methods such as String-based similarity or Semantic-based similarity or a combination of these techniques in a hybrid approach without the use of the BERT model can produce highly efficient correlation assessment results.

Another study [17] raised the problem of most AES models which were only designed on one essay topic. So, when it is used on several different topics the performance of the model decreases. To solve this problem, the research proposes a BERT model with Multi-Task Learning and fine-tune methods. The datasets used are ASAP and Chinese EFL Learners' Argumentation (CELA) datasets. Each dataset is used to experiment with a different type of essay. The test results on the ASAP dataset are a QWK score of 0.83, which is the highest compared to other models such as Memory-Augmented, LSTM-CNN, Two-Stage Learning Framework, Histogram String Kernel Intersection. And the test results on the CELA dataset produce a QWK score of 0.78.

In research [18], the researcher took a hybrid approach by applying embedding technique with the BERT model and the output used for the LSTM model as a decision regression. The dataset used is the ASAP dataset and the speech corpus dataset taken from the School of Information of Xiamen University. The experimental results on the ASAP dataset with the QWK evaluation metric of 0.8, while the results on the speech corpus dataset of 0.6. This high result can be obtained by utilizing the ability of the BERT model to extract semantic text information in depth.

Research using unsupervised learning techniques is applied by research [19] under the name Discourse Corruption. The AES model is designed with three main parts, namely the base document encoder, auxiliary encoder, and scoring function. The base document encoder using a pre-trained Longformer model, serves to generate vectors that represent each word in the essay, the auxiliary encoder which consists of an embedding layer and a Bi-directional Long Short-Term Memory (BiLSTM) layer, serves to generate vectors that represent additional information related to the essay. The results of the study were measured by an accuracy evaluation metric with a score range of 0.87 – 0.95 and Mean Squared Error (MSE) with a score of 0.16.

In research [20] designed a model with two approaches, namely Prompt-Unaware Components and Prompt-Unaware Components. The purpose of this study is to design an AES model that can assess an essay with a prompt or not. Furthermore, this study solves the shortcomings of transformer-based models, which require a very large number of parameters and a long training time. As an example, the BERT model requires at least half a million parameters for fine-tuning. This was fixed by adding adapter modules to the transformer model. The results of this study were measured by the QWK metric with an average score of 0.78.

Another transformer-based research was conducted by [21] using three word-embedding methods, Latent Semantic Indexing (LSI), Sentence-BERT (S-BERT), and word2vec. This model uses cosine similarity as an assessment of the essay. Based on the performance of each embedding technique, the LSI technique has good accuracy for the assessment of essays that have keywords from answers, Sentence-BERT provides a good assessment based on the semantic side of the text, even for long essays. While the word2vec method has a good assessment if the essay still contains related keywords, even though the essay does not have real keywords.

III. Dataset

The data used in this study is the "Quiz 1 Basic Programming Course" dataset taken from the Department of Electrical Engineering and Information Technology in 2020 with a total of 229 data. Each data contains five answers from five questions. The data is converted into xlsx format with three attributes, id for student ID, essay for answers, and rater_1 assessment score. The amount of data above is the amount of data after adding adversarial data. The type of adversarial data is a data with answers that have a low assessment score. The answer used is the answer to question number 1.

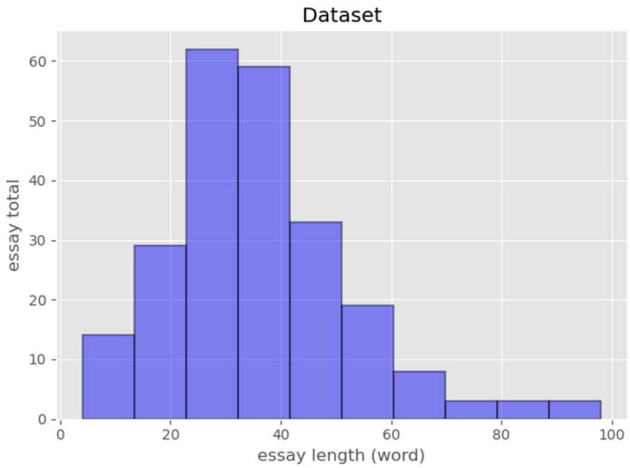

Fig 1 Dataset Length

Based on the data depicted in Figure 1, there are statistical data, such as the minimum number of words in one essay is 4 words, while the maximum number of words in one essay is 98 words, and the average number of words in one essay is 36 words. This statistical data is useful for hyperparameter-optimization. By considering these statistical data, we can obtain the model with the best performance. The language used in the essay is Indonesian with some common terms in programming and computers.

IV. Method

In this study, several answer essays were given to be assessed by the model automatically based on a predetermined answer key. The main purpose of this study that distinguishes it from previous research is that the model produced in this study can perform a good assessment with only one answer key that is used as training data. Our proposed AES model is a transformer-based model, namely BERT and S-BERT with fine-tuning and hyperparameter-optimization techniques. The architecture of the AES model is depicted in Figure 2.

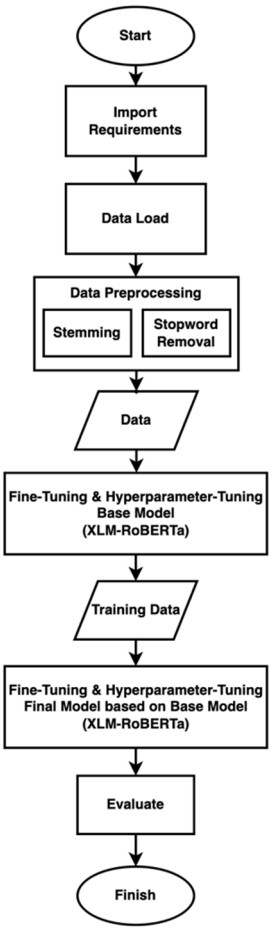

Fig 2 AES Architecture

A. Data Load & Data Preprocessing

The initial stage of the AES model is data loading. Data loading is the process of loading data from raw data which is still in docx or pdf format. This format cannot and inefficient to be processed or used directly in the modeling process. Therefore, it is necessary to convert the data into a format that is supported by the model. Based on the not-so-large amount of data, the data format used is xlsx format. Performance is not significantly different compared to other formats such as csv or tsv. Another consideration is that the xlsx format is not prone to data errors as in the csv format, if there is a comma in an inappropriate place, the system will read it as new data or rows.

Data preprocessing is needed to get the data that suits the needs. Because in the raw data prepared at the data loading stage there are many elements that are not actually needed by the model. What is meant by preprocessing here is data cleaning, by carrying out the stemming process, stop word removal, and deleting data with missing values. Stemming is a process that maps the variant form of a word to its root. The stemming process is usually done by removing suffixes and prefixes from words to get the root word. With data that only contains basic words, it can increase the efficiency of the training model and reduce the level of bias because there are unnecessary word affixes.

The next process is stop word removal. Stop words are general words that do not provide important information. Thus, removing stop words not only reduces the size of the text corpus but also eases the training process, reduces the index for each text and further improves the level of space and time complexity. For the stop word database, it is adjusted to the language used in this study, using libraries from nltk corpus.

TABLE I. DATA PREPROCESSING

Original Text	After Processing
'Tablet termasuk komputer karena tablet merupakan perangkat digital yang memiliki mesin dimana dapat diprogram atau diperintah untuk mengolah data dari suatu bentuk ke bentuk yang lain. Tablet dapat dikategorikan sebagai personal computer karena digunakan oleh satu orang. Mesin yang dapat diprogram dapat kita jumpai dalam tablet yang memiliki sistem operasi tertentu dan beberapa software/perangkat lunak yang digunakan untuk memproses perintah/komputasi dari pengguna. Pada tablet juga memiliki komponen memori Utama, Memori Sekunder, Central Processing Unit (CPU), input devices dan output devices.'	'tablet masuk komputer karena tablet rupa perangkat digital yang milik mesin mana dapat program atau perintah untuk olah data dari suatu bentuk ke bentuk yang lain tablet dapat kategori bagai personal computer karena guna oleh satu orang mesin yang dapat program dapat kita jumpa dalam tablet yang milik sistem operasi tentu dan beberapa software perangkat lunak yang guna untuk proses perintah komputasi dari guna pada tablet juga milik komponen memori utama memori sekunder central processing unit cpu input devices dan output devices'

The example in Table I is a comparison between the original text of the essay and the processed text. The difference in each word used is in the form of the basic word and the total number of words is less. With such a text composition, data processing at the next stage will become more efficient. And it can also reduce the level of bias caused by unnecessary words or suffixes.

B. Base Model Fine-Tuning

The data generated from the data preprocessing process will be used for fine-tuning the base model. Fine-tuning is done by using an answer key dataset, which is adaptive according to the amount of data available. Furthermore, this model will be used for encoding the answer essay text. Then the encoding results will be used to calculate the similarity value to the answer key. The base model is designed using a transformer-based pretrained model, specifically using the model developed from the BERT model, Cross-Lingual Model Representation (XLM-R) with the model name xlm-roberta-base. XLM-R is a model that has been trained with data in 100 types of languages, making this model has a good performance when used in texts that contain several types of languages [22]. The base model has one layer with 256 output features and uses the Gaussian Error Linear Unit (GELU) activation function.

C. Final Model Fine-Tuning

The final model is also used for encoding the answer essay text and uses the same architecture as the base model. The main difference is that the base model is fine-tuned with answer key data, while the final model is trained with the dataset generated from the base model. The structure of the encoding model can be seen in Figure 3. XLM-R works as a pretrained model for fine-tuning. Pooling layer serves to reduce the dimension features or parameters that will be used for training. This will reduce the number of computations performed in the model. Hidden layer is used for parameter activation function.

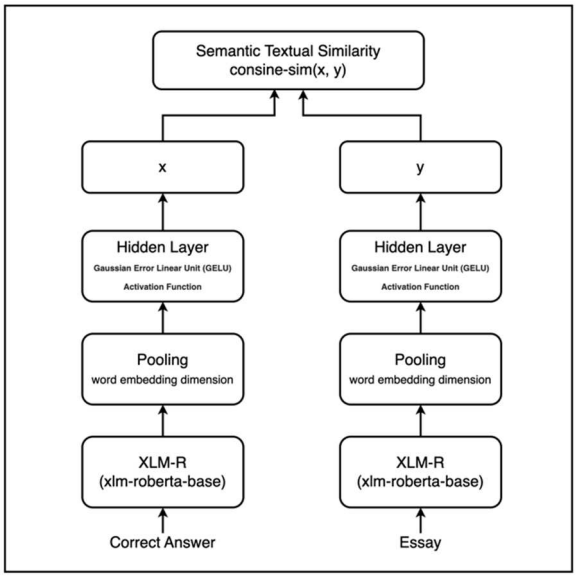

Fig 3 Encoding Model

In addition, the difference between the base model and the final model lies in the configuration of the hyperparameters used. Hyperparameter optimization is carried out based on the available statistical data. Some of these statistical data are the maximum length of sentence characters, word embedding. The following in Table II is the value of each hyperparameter used. The encoding results from the final model will be measured for similarity with the answer key using semantic textual similarity at the evaluation phase.

TABLE II. HYPERPARAMETER VALUE

Hyperparameter	Value
Max sequence length	328
Word embedding dimension	768
In feature	768
Out feature	256
Activation function	GELU

D. Evaluation

The measurement of semantic textual similarity is implemented using the cosine similarity method. Cosine similarity is a calculation of the degree of similarity obtained from the multiplication of the cosine angle of the two or more vectors being compared. Cosine 0° is 1 and less than 1 for other angles, the similarity value between two vectors can be said to be perfect or equal when the value of cosine similarity is 1. he following formula is for calculating cosine similarity.

$$Cos\ a = \frac{A\,.\,B}{|A|\,.\,|B|} = \frac{\sum_{i=1}^{n} A_i\,.\,B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\,.\,\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Where A is the weight of each feature of the vector and B is the weight of each feature of the vector B. The concept of the degree of similarity in cosine similarity is that the higher the degree formed from the two coordinate vectors, the lower the value of the similarity of the text. On the other hand, the lower the cosine degree, the higher the text similarity value. Cosine similarity is a technique for measuring the degree of similarity that is relatively easy and can be applied to several case studies.

Furthermore, the measurement of the model's performance is carried out with several evaluation metrics. Furthermore, the measurement of the model's performance is carried out with several evaluation metrics. QWK is a metric to measure the degree of agreement between a set of predictions and a set of labels. In the context of this research, QWK is used to measure the level of agreement between human assessment and the results of the AES model assessment. This metric is suitable to measure the performance of the AES model, because it must be measured based on the agreement of human assessment.

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (2)$$

$$k = 1 - \frac{\sum_{i,j} w_{i,j}\, O_{i,j}}{\sum_{i,j} w_{i,j}\, E_{i,j}} \quad (3)$$

The QWK assessment range is between 0 (no agreement) to 1 (completely agree). Each essay in each data set can be represented by a tuple (ea, eb) where ea refers to the score of the human rater and eb refers to the score predicted by the AES model. The N-by-NO histogram matrix is built over the essay rating, where Oi,j refers to the number of essays with an *i* grade by human rater and a grade *j* by the AES model. The *w* weighted N-by-N matrix, calculated based on the difference between the rater's score referring to Equation 2 and the Quadratic Weighted Kappa *k* found by Equation 3

V. RESULTS AND DISCUSSION

The result of the AES model in this study is an automatic assessment of each essay answer based on the answer key provided. The results of the assessment are obtained by measuring the degree of similarity using semantic textual similarity from each encoding result from the answer essay with the answer key essay encoding. The measurement results of the semantic textual similarity with the cosine similarity method are in the form of values ranging from 0 to 1. The following Table III is data sample from the assessment results in the decimal form with 2 values behind comma.

TABLE III. EXPERIMENT RESULT

Id	Essay	Manual	AES
456433	Karena tablet merupakan sebuah mesin yang dapat diprogram atau diperintah untuk melakukan sesuatu (mengolah data dari suatu bentuk ke bentuk yang lain) dan dibentuk oleh beberapa komponen seperti memori utama, memori sekunder, Central Processing Unit (CPU), input devices, dan output devices	9	8.70
456287	Karena Tablet merupakan sebuah alat/mesin yang memiliki fungsi dalam mengolah atau menjalankan beberapa data dari bentuk yang satu ke yang lainya melalui perintah atau pemograman. Tablet termasuk kedalam jenis personal komputer yang berarti	9	8.56

Id	Essay	Manual	AES
	hanya dapat digunakan oleh satu orang saja. Oleh karena itu tablet termasuk komputer		
456382	Intinya adalah komputer adalah mesin yang di program untuk melakukan sesuatu. Program tersebut dalam bentuk data yang diolah dari bahasa source ke bahasa komputer. Contoh mudahnya adalah kalkulator yang merupakan sebuah komputer uang diprogram untuk menghitung. Sama halnya dengan tablet yang merupakan suatu komputer karena memiliki program dan bisa diperintah.	8	8.08
400199	karena bisa digunakan untuk berhitung	3	2.75
456308	Karena tablet dapat diprogram /diperintah untuk melakukan sesuatu dibentuk oleh beberapa komponen utama seperti memori utama, memori sekunder, CPU, dan lain-lain	9	8.64
456407	Tablet termasuk dalam komputer karena tablet merupakan sebuah alat atau mesin yang dapat diberi perintah atau program (input) untuk mengolah sebuah data (proccess) dan mengeluarkan hasil dari data yang diolah dalam bentuk yang lain (output)	8	8.22
400193	bisa jadi sebuah mesin yang bisa digunakan	2	2.44
400278	komputer adalah mesin yang diciptakan	3	3.27

To measure the performance of the AES model of this research, the results of the assessment were measured by evaluation metrics. The following Table IV is the result of measuring model performance based on experiments with different number of epochs:

TABLE IV. EXPERIMENT RESULT

Base Model	Final Model	QWK	Accuracy	Recall	Precision
1	1	0.923	0.909	0.914	0.770
1	2	0.930	0.918	0.920	0.772
1	3	**0.933**	**0.922**	0.921	**0.776**
1	4	0.922	0.909	**0.944**	0.774
1	5	0.899	0.879	0.919	0.756
2	1	0.923	0.909	0.914	0.770

From the experimental results in Table IV, the best overall results are obtained with the number of epochs for the base model 1 and final model 3. There is one evaluation metric, namely recall, which has the best performance on the base epoch model 1 and the final epoch model 4. However, the best performance stick to the epoch number of base model 1 and final model 3, because most of the metrics show the best results on these epoch counts. These results also indicate that changes in the number of epochs in the base model do not have a significant effect. This is because the amount of data used for fine-tuning the base model is only 1 piece. The following.

TABLE V. RESULT COMPARISON

Method	Data	QWK
Hierarchical BERT Multi-Task Learning [17]	ASAP, CELA	0.83
Hybrid Approach and Feature Engineering [1]	ASAP, SAT, GRE	0.81
BERT-Based for Speech-Oriented Text Modality [18]	ASAP, SFL	0.808
Parameter-Efficient Transformer [20]	ASAP	0.785
Transformer-based Fine Tuning	**Quiz 1 Basic Programming Course**	**0.933**

Table V shows a comparison of the results of several previous studies with the results of this research. From the table it can be seen that most of the AES research uses ASAP data and other data sources. The ASAP dataset is a large dataset, which has 12948 rows of data. With such large data, it is natural that a deep learning model can have good performance. This is because the training process of deep learning models is highly affected by the amount of data used. The use of the QWK metric as a performance comparison measure is because QWK is the most relevant metric for AES tasks.

The best AES model resulting from this research is a standalone deep learning model, which can only be used independently. This means that it has not been integrated with an e-learning or other learning platform. In its use, it still must go through several processes and must know the basics of the python language. There is drawback in the data processing stage, precisely in the data conversion process. This drawback can be solved simultaneously with the integration process with a platform. Because converting the original data into a supported format manually is an inefficient process. Another drawback is that if there is answer data in the form of images, this AES model cannot yet assess the data. This can be solved by involving the image processing method.

VI. CONCLUSION

Research on automated assessment has been carried out using various methods such as statistical approaches, machine learning, and deep learning, most of the research using ASAP data which has a fairly large amount of data. So that it can give good results because the model is trained with sufficient data. This study proposes an AES model that can produce a more reliable and stable scoring performance even with very limited data. The main methods used are data preprocessing, fine-tuning and hyperparameter optimization.

Based on the result of this research, it can be concluded that fine-tuning and hyperparameter-optimization on transformer-based models have succeeded in generating an assessment model that has more reliable and stable performance. The success of improving the performance of the assessment is shown by the results of the model assessment that is close to human assessment. For the validity of the assessment results, the performance of the model is measured based on the evaluation metrics of QWK, accuracy, precision, and recall.

This study still has some limitations that need to be improved in future research. Here are some suggestions for further research:

1. Further research can explore other transformer-based models, with fine-tuning, hyperparameter-optimization, or other methods.
2. Build an end-to-end AES system, involving certain platforms (e-learning, MOOC, etc.), so that the AES model can be utilized in real cases.
3. Data integration with Extract Transform Load (ETL), so the data processing can be done more efficiently.
4. Improve AES ability for assessment on essay answers in the picture format.

REFERENCES

[1] S. Prabhu, K. Akhila and S. S, "A Hybrid Approach Towards Automated Essay Evaluation based on Bert and Feature Engineering," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, Pune, India., 2022.

[2] H. Chimingyang , "An Automatic System for Essay Questions Scoring Based on LSTM and Word Embedding," in *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, Shenyang, China, 2020.

[3] N. Chamidah, M. M. Santoni, H. N. Irmanda, R. Astriratma, L. M. Tua and T. Yuniati, "Word Expansion using Synonyms in Indonesian Short Essay Auto Scoring," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* , Jakarta, Indonesia, 2021.

[4] Gunawansyah, R. Rahayu, Nurwathi, B. Sugiarto and Gunawan, "Automated Essay Scoring Using Natural Language Processing And Text Mining Method," in *2020 14th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Bandung, Indonesia, 2020.

[5] R. Bhatt, M. Patel, G. Srivastava and V. Mago, "A Graph Based Approach to Automate Essay Evaluation," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Toronto, Canada, 2020.

[6] V. Prashanthi, T. N. P. Madhuri, V. Shailaja and S. Kanakala, "Automatic Valuation of Essay using Machine Learning," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, Karnataka, India, 2021.

[7] Y. Salim, V. Stevanus, E. Barlian, A. C. Sari and D. Suhartono, "Automated English Digital Essay Grader Using Machine Learning," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, Yogyakarta, Indonesia, 2019.

[8] A. A. P. Ratna, H. Khairunissa, A. Kaltsum, I. Ibrahim and P. D. Purnamasari, "Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and Latent Semantic Analysis," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Batam, Indonesia, 2019.

[9] H. K. Janda, A. Pawar, S. Du and V. Mago, "Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation," in *IEEE Access*, 2019.

[10] H. Thamrin, N. A. Verdikha and A. Triyono, "Text Classification and Similarity Algorithms in Essay Grading," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2021.

[11] C. Li, L. Lin, W. Mao, L. Xiong and Y. Lin, "An Automated Essay Scoring model Based on Stacking Method," in *2022 2nd IEEE International Conference on Software Engineering and Artificial Intelligence*, Xiamen, China, 2022.

[12] A. R. Arifin, P. D. Purnamasari and A. A. P. Ratna, "Automatic Essay Scoring for Indonesian Short Answers using Siamese Manhattan Long Short-Term Memory," in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, Kuala Lumpur, Malaysia, 2021.

[13] A. Wiratmo and C. Fatichah, "Assessment of Indonesian Short Essay using Transfer Learning Siamese Dependency Tree- LSTM," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2020.

[14] R. A. R. George, P. Yashasawi, N. A. A. Kumaran and V. K. Patnaik, "FACToGRADE: Automated Essay Scoring System," in *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, BALI, Indonesia, 2022.

[15] P. Wangkriangkri, C. Viboonlarp, A. T. Rutherford and E. Chuangsuwanich, "A Comparative Study of Pretrained Language Models for Automated Essay Scoring with Adversarial Inputs," in *2020 IEEE REGION 10 CONFERENCE (TENCON)*, Osaka, Japan, 2020.

[16] M. M. Saeed and W. H. Gomaa, "An Ensemble-Based Model to Improve the Accuracy of Automatic Short Answer Grading," in *2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, Cairo, Egypt, 2022.

[17] J. Xue, X. Tang and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," in *IEEE Access*, 2021.

[18] C. Zheng, L. Huang, H. Lin, Y. Guo and L. Huang, "BERT-Based Automatic Scoring Model for Speech-Oriented Text Modality," in *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, Changchun, China, 2022.

[19] F. S. Mim, N. Inoue, P. Reisert, H. Ouchi and K. Inui, "Corruption Is Not All Bad: Incorporating Discourse Structure Into Pre-Training via Corruption for Essay Scoring," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 2202 - 2215, 2021.

[20] A. Sethi and K. Singh, "Natural Language Processing based Automated Essay Scoring with Parameter-Efficient Transformer Approach," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2022.

[21] S. V and J. Narayanan, "Short descriptive answer evaluation using word-embedding techniques," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2021.

[22] M. Uto and M. Okano, "Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases," *IEEE Transactions on Learning Technologies,* vol. 14, no. 6, pp. 763 - 776, 2021.

[23] P. W. Foltz, W. Kintsch and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes,* pp. 285-307, 1998.