Jurnal Teknologi Pendidikan, Agustus 2025, 27 (2), 751-764
DOI: http://dx.doi.org/10.21009/JTP2001.6
p-ISSN: 1411-2744 e-ISSN: 2620-3081
Accredited by Directorate General of Strengthening for Research and Development
Available online at http://journal.unj.ac.id/unj/index.php/jtp

Correlation Between Automatic Short Answer Scoring and Manual Scoring by Teacher on Indonesian Assessments

Ravika Ayu[1], Dade Nurjanah[2]
[1,2]Informatics, School of Computing, Telkom University, Bandung, Indonesia

Received:
Revised:
Accepted:

Abstract

Assessment is one tool evaluation in the learning teaching process to determine quality of learning. One of method being assessed Enough complicated in the assessment process is essay test. The essay test requires more time lots in the proofreading process as well as low validity and reliability because possible essay assessment influenced element subjective. Therefore, needed application used for correct essay answers expected automatically can help teachers to correct answer with fast and more results objective namely Automated Short Answer Scoring (ASAS). This research use sentence embedding method for measure similarity meaning between answer key and students answer. Data sets used consists of two data. One data consists of 1200 pairs of answer keys and students answer used for train the model. Two data totaling 250 words is used for evaluate models. Before enter the sentence embedding process, answering will through the pre-processing stage is remove stop words, remove empty, case folding, delete number, delete punctuation. Testing this system will done with method compare assessment carried out system with assessment carried out by teachers conventional use coefficient correlation. The result of test coefficient correlation Pearson of 0.81 and concluded reached 81 % similar with human rater assessment. This study can help teachers to be more efficient in the assessment.

Keywords: Automated Short Answer Scoring, Sentence Embedding, Sentence Transformers

(*) Corresponding Author: ravika@student.telkomuniversity.ac.id

How to Cite: Ayu, R., & Nurjanah, D. (2025). Correlation Between Automatic Short Answer Scoring and Manual Scoring by Teacher on Indonesian Assessments. *JTP - Jurnal Teknologi Pendidikan*, *27*(2), 751–764. https://doi.org/10.21009/jtp.v27i2.48378

INTRODUCTION

Along development digital technology, the world of education also does it significant adaptation to development technology. Evaluation is important stages in the learning process for measure achieved objective learning. Some institutions education have using computer assisted assessment (CAA) (Conole & Warburton, 2005), (Stephens, 2001), (Chalmers & McAusland, 2002). CAA refers to use computers and technology for facilitate and improve the assessment and evaluation process knowledge, skills, and performance student. The aim of the CAA is for simplify the assessment process, provide feedback fastly, and make evaluation more efficient and effective. Evaluation is important stage in the evaluation process learning (Tim Pusat Penilain Pendidikan, 2019), (Zupanc & Bosnić, 2017). Evaluation can done with various method like multiple choice, essay, true false, matching etc. Every test have the advantages and disadvantages (Black & Wiliam, 2018). E-learning assessment usually using multiple choice

because easy in the assessment process. However multiple choice tests have opportunity correct answer with method guess. From that, assessment with method multiple choice tests can't measure power ability students and easy to cheating (Tim Pusat Penilain Pendidikan, 2019), (Susongko, 2010).

Problems arise when there is parts need student's understanding where test should do in essay form. According to guide written test (Tim Pusat Penilain Pendidikan, 2019) essay test is test which the answer demand participant educate for remember and organize ideas or things that have been studied with method put forward or express idea in form description written. The form essay test provides freedom to every student. For express his mind, so the answer will show ability think. However in carry out There is an essay test constraint in the assessment process like need relatively long time more difficult in the process of correction, so difficult used For test scale big. Essay assessment can be done influenced by nature subjective evaluator so that tests graded by different people can produce different results. Essay assessment is also experienced problem low validity and reliability even tests graded by the same person can produce different values.

For overcome problems the An automatic essay assessment computer program is needed For become solution to the essay assessment process automatic so that time correction become more fast and objective. Assessment system in a way automatic used For help implementation tests and assessments so become effective and efficient. Assessment using an automation system This possible student answer question through the program and can direct obtain values (Ratna et al., 2007), (Cerratto Pargman et al., 2023), (Yongqi GU & LAM, 2023), (Dadi & Sanampudi, 2022).

For building a ASAS system can used various method such as sentence embedding, word embedding, context embedding, sense embedding. Based on the ASAS state of art (M. Putnikovic & J. Jovanovic, 2023), (B. Wang & C. . -C. J. Kuo, 2020), (H. Choi et al., 2021), (Conneau & Kiela, 2018) sentence embedding is the right choice in represent sentences and phases. This matter used For detect similarity sentence students answers and keys answer keys. The more tall level similarity student answer with answer key so the more tall score obtained student. Based on matter the An effective essay assessment system is needed. For realize matter the done studies case about Automated Short Answer Scoring. This study will do for essays basic lessons in the Computer and Network Engineering Program class XI at SMKN 1 Rao Selatan, Pasaman Regency, West Sumatera.

Related Research

Many methods are used for build assessment assay automatic (Cerratto Pargman et al., 2023) (Ramnarain-Seetohul et al., 2022) such as word embedding, contextual embedding, sense embedding and sentence embedding. Every method own the advantages and disadvantages of each are several study use combination from a number of method the. Word embedding represents words into vectors so that not enough complex compared to representation sentence. No word embedding either notice word order Although so word embedding is more simple in its implementation. Contextual embedding requires big computing. Needed around 3.4 billion words so If used on relatively small sheets Possible difficult For get context from the sentence. Sense embedding makes it possible representation

of words with precision usually done with utilise source Power lexical like wordnet. Sense embedding requires induction meaning as a first step and for can do matter the need expensive costs.

The sentence embedding method is choice best in represent sentence. Sentence represented to in form vector numeric so that can catch meaning semantics from sentence For measure similarity text. Use representation vector possible more comparison Good between text. Even though sentence embedding has Lots advantages, there are also several necessary deficiencies considered that is need source Power more Lots rather than word embedding. Sentence embedding is an improvement on word embedding. So sentence embedding is good method in represent sentence (Cerratto Pargman et al., 2023).

G. Herwanto designing Ukara (Herwanto et al., 2018) is an Automatic Short Answer Scoring System for Indonesian which proposes a combination of Natural Language Processing (NLP) and supervised machine learning techniques to produce F1 score above 97% and 70% on dichotomous and polytomous scoring types respectively.

AAP Ratna et al built Simple O (Ratna et al., 2007) using Latent Semantic Analysis with Learning Vector Quantization and Word Similarity Enhancement produces the average increase in accuracy after the addition of word similarity function is 5.4% from 90.9% to 96.3

Gomaa et all uses Ans2Vec (Gomaa & Fahmy, 2020) as system evaluation short answer. Ans2Vec is a scoring model short answer, effective and easy to used. Reference and representation vectors students answer made using the skip-thinking model. For predict score students answer, products based on components and differences absolute calculated, combined, and entered to in regression logistics. Ans2Vec achieved 0.89 RMSE and 0.63 Pearson Correlation for the dataset from the University of North Texas.

Rajagede et al (Rajagede, 2021) This is studies about enhancement system evaluation automatic use pre-trained BERT sentence embedding method .

Lubis et all (Lubis et al., 2021) is studies about system evaluation automatic For evaluate answer in Indonesian using word embedding and synthetic analysis methods with coefficient correlation 0.7085 and Mean absolute Error 0.7009

G. Herwanto, Ratna, Gomaa, Rajagede and Lubis have solution change key answer and reply student into vector and search similarity the data pair. This research will look for semantic similarity results between students answer and answer key use sentence embedding and measuring methods coefficient correlation with evaluation Manually by teacher .

METHODS

Datasets

This research using sets daily data lesson basics of skills programs technique class X computer network at SMKN 1 Rao Selatan. Data used 2 datasets. Data 1 consists answer from 120 students on 10 questions so that there are 1,200 pairs. It scored by 2 teachers and calculate average score teacher 1 and teacher 2. This data is used For training the sentence transformer model. Data 2

consists of 250 pairs of sentences key answer key and students answer obtained of 25 students answer the next 10 questions assessed by 2 teachers. This data used in the evaluation process.

Table 1. Data 1 Example Partner Answers key and students answer who have given score

Question	Key Answer	Student Answer	Score (average teacher 1&2)
Jelaskan apa itu jaringan komputer?	Jaringan Komputer adalah saling terkoneksinya dua perangkat komputer atau lebih sehingga bisa saling berbagi pakai data, informasi dan perangkat yang ada dalam jaringan tersebut	Jaringan komputer mengacu pada perangkat komputasi yang saling terhubung serta dapat bertukar data dan berbagi sumber daya satu sama lain. Perangkat jaringan ini menggunakan sistem aturan, yang disebut sebagai protokol komunikasi, untuk mentransmisikan informasi melalui teknologi fisik atau nirkabel.	9
Jelaskan apa itu jaringan komputer?	Jaringan Komputer adalah saling terkoneksinya dua perangkat komputer atau lebih sehingga bisa saling berbagi pakai data, informasi dan perangkat yang ada dalam jaringan tersebut	Jaringan komputer adalah jaringan yang saling bertukar data satu sama lain	5

Table 2. Data 2 Examples Partner Answers key and students answer who have given scores by 2 teachers

Question	Key answer	Student Answer	Score by Teacher 1	Score by Teacher 2
Jelaskan apa itu jaringan komputer?	Jaringan Komputer adalah saling terkoneksinya dua perangkat komputer atau lebih sehingga bisa saling berbagi pakai data, informasi dan perangkat yang ada dalam jaringan tersebut	Jaringan komputer mengacu pada perangkat komputasi yang saling terhubung serta dapat bertukar data dan berbagi sumber daya satu sama lain. Perangkat jaringan ini menggunakan sistem aturan, yang disebut sebagai protokol komunikasi, untuk mentransmisikan informasi melalui teknologi fisik atau nirkabel.	8	10
Jelaskan apa itu jaringan komputer?	Jaringan Komputer adalah saling terkoneksinya dua perangkat komputer atau lebih sehingga bisa saling berbagi pakai data, informasi dan perangkat yang ada dalam jaringan tersebut	Dua atau lebih komputer yang yang saling terhubung sehingga bisa saling bertukar data sumber daya agar bisa saling memperoleh informasi.	10	8

1. Sentence Transformer

   Sentence transformer uses transformer architecture that implements self-attention. In ASAS this allows the model to focus on the considered part relevant, and necessary deep understanding about context or connection between sentence. Transformers have ability to parallelized with more efficient, which is possible training fast on larger datasets. Transformers have ability for processing all over context in one time, so can evaluate long answer or complex.

   This method produce representation sentence with project token to vector. Models that have trained previously (pretrained) on very large data is more processing efficient. This model depends on the large model and requires training for reach best results. In applying this model

2. Coefficient Correlation

   Coefficient correlation is statistical measurements for measure strength and direction linear relationship between two variables. Coefficient correlation give size numeric indicating how much strong two variable related. An absolute value close to 1 indicates strong relationship, whereas values close to 0 indicate weak relationship.

Table 3. Intervert Coefficient Correlation (Sugiyono, 2019)

Range	Degree of Relationship
0,00 – 0,19	Extremely low
0,20 – 0,39	Low
0,40 – 0,59	Medium
0,60 – 0,79	High
0,80 – 1,00	Very High

Architecture

Evaluation short answer consists three parts: (1) preprocessing; (2) change answer text become the representation vector; and (3) similarity. In preprocessing done tokenization, leminatization. Answer changed become representation vector use sentence embedding method with sentence transformer and measuring similarity. Modeling flow in this research can illustrated in the chart following:

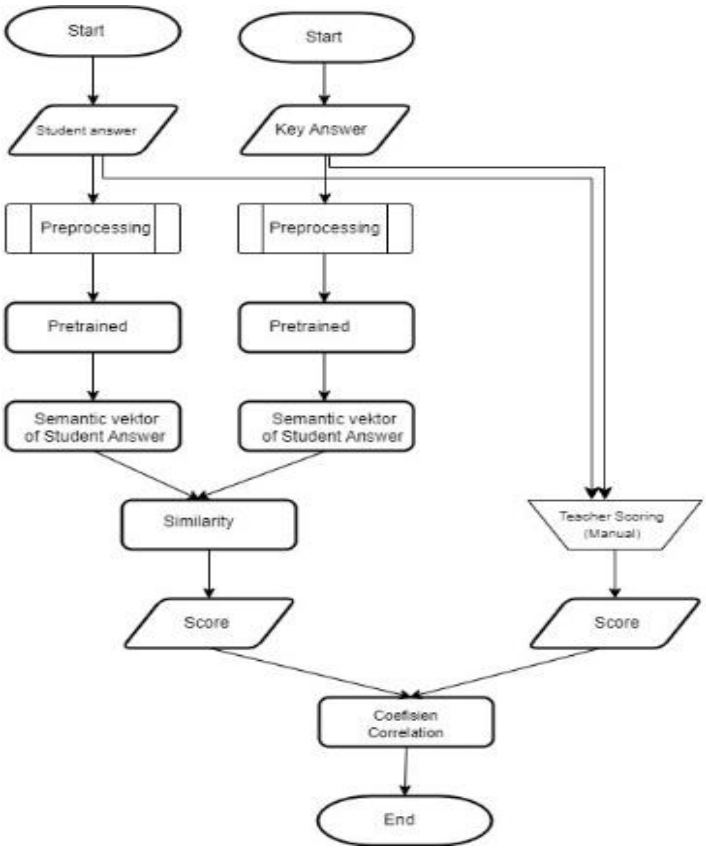

Figure 1. Automated short answer scoring

1. Data Input (Student Answer and Answer key)

   Data used in this study obtained from question exam student class XI majoring in Computer and Network Engineering (TKJ). Data consists from question, answer key, student answer and scores. Selected questions covers basic material for the Computer Engineering.

2. Preprocessing

   The purpose of preprocessing is simplify input data to make it more easy processed. Pre- processing stages (U. Hasanah et al., 2018), (Hickman et al., 2022).

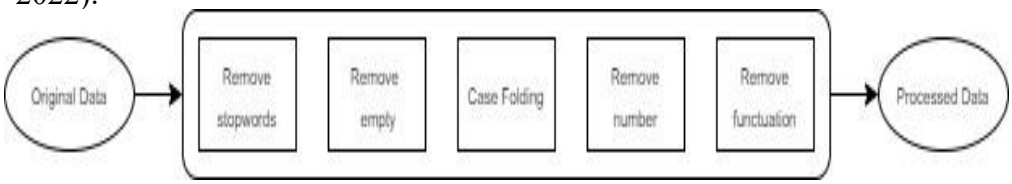

Figure 2. Preprocessing

   a. Remove stopwords

      Stopwords Removal is step for throw out words that don't need like a conjunction. Stopwords are the most frequent words appear in text but hasn't significant meaning. Delete these words will reduce data input size and can increase sentence embedding results . Stop word will use Indonesian with the NLTK library.

b. Remove empty
   At stage cleaning process is carried out or delete empty or no mark
c. Case Folding
   At stage for make lowercase.
d. Delete Numbers
   At stage preprocessing for delete numbers in the column references answer and student answer
e. Delete punctuation
   At stage preprocessing for delete punctuation like ., , and others punctuation in the column references answer and student answer.

3. Sentence Transformer Model
   a. Pretrained

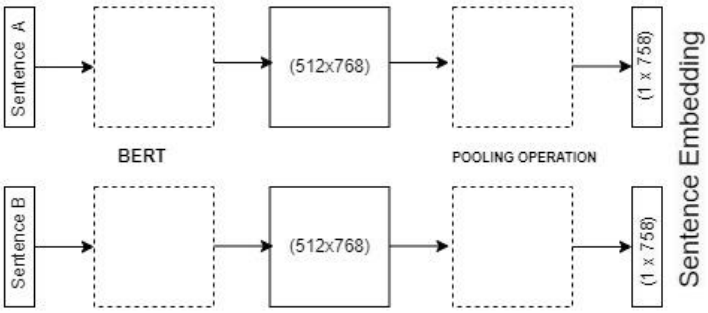

Figure 3. Sentence Transformer Process Flow

Sentence transformer utilizes trained transformer architecture unsupervised. This Architectur understand context bidirectionally, so can consider the words before and after. The result of this process represent text become richer . The sentence transformer process begins with input Sentence A and Sentence B into in the BERT model. Then the model generates vector representation for every token in sentence, with covers information of the words before and after. Each token is represented as vector with high dimensions, for example, 768 dimensions. Representation allows the model to understand meaning of the word sentence context. The next process the pooling process is carried out for combine information from all tokens in sentence become one vector. The final result from pooling operation is a vector of mentioned sentence as sentence embedding. Type of pooling carried out namely mean pooling. Sentence embedding loads semantic information from sentences (Reimers & Gurevych, 2019), (B. Wang & C. . -C. J. Kuo, 2020), ("Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," 2023), (Pires et al., 2019), (H. Choi et al., 2021).

After pretraining, a model is generated arranged for fine-tuning on tasks specifically, namely automated short answer scoring. Arrangement hyperparameters done. For optimizing deep model performance task (Sung et al., 2019), (Ormerod, 2022). In some task processing natural language, BERT is the most advanced learning model (Devlin et al., 2019). There are

2 steps in framework BERT works, namely pretrained and fine-tuning. In the pretrained stage, the model is trained on data that is not have a label. At the finetuning stage, the BERT model parameters are initialized moreover formerly with parameters and trained using data that have labels. This research using the existing sentence transformer model trained with IndoBERT for Indonesian.

b. Testing Models

Evaluation is done to measure how much near mark predictions with mark actual in a way continuous so this model use metric evaluation namely Mean Absolute Error (MAE), (Shcherbakov et al., 2013), (A. A. Mamun et al., 2020).

Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |x_i y_i|}{n}$$

4. Cosine Similarity

Cosine similarity is used to calculate the similarity vector 2 sentence of vector key answer and vector student answer (Rahutomo et al., 2012), (A. R. Lahitani et al., 2016), (Yunanda et al., 2022). Each vector represents sentence being compared. Cosine Similarity measurement using the equation:

$$Similarity(A,B) = \cos(A,B) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (Ai)^2} x \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

If the results show a similarity value of 1 means the text is the same and if the results show a similarity of 0 means the text is very different.

5. Score Output

This similarity score reflect the extent of the student answer suitable with criteria that have been set by the teacher. The similarity results will be become output scores ("Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," 2023), (Li & Han, 2013), (Lubis et al., 2021). to obtain the following equation for the short essay's score, $score$:

$$score = sim(A, \mathrm{B})$$

6. Analysis Method

Evaluation similarity between system evaluation with man evaluation is based on the Pearson Correlation Coefficient, a metric statistics that measure level linear relationship between two variables (D. Chicco et al., 2021). In context assessment, the Pearson Correlation Coefficient value is close to 1 indicating positive linear relationship perfect between evaluation systems and assessments humans, meanwhile values close to -1 indicate negative linear relationship perfect. With using the Pearson Correlation Coefficient, we can in a way quantitative evaluate extent of assessment system correlated with evaluation man. The more tall mark coefficient correlation, increasingly similar system evaluation with man evaluation. This result evaluation give more understanding in about how much Good system can catch substance students answer and suitability with criteria man evaluation. Therefore, assessment

based on Pearson Correlation Coefficient gives strong foundation for measure validity and reliability evaluation system automatic.

Pearson Correlation Coefficient (r)

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

RESULTS & DISCUSSION

Experiment

Experimental process is done to student class XI Computer Engineering Network (TKJ) with carry out a direct assessment with web platforms. In assessment here, students answer ten questions that cover material base technique computer network, with limitation time 30 minutes. In this process student can quick know score they after finish exam. Therefore, students can evaluate performance they with hurry up and get it bait come back in a way direct. Next, results score obtained student will analyzed and compared with assessment carried out by the teacher.

How to determine student score:

- key answer 1

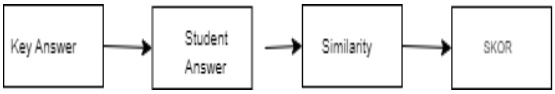

Figure 4. processed 1 answer

- key answer 2

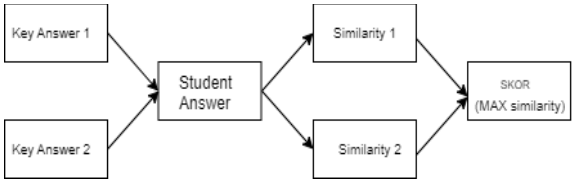

Figure 5. processed 2 answer

Evaluation

Evaluation done with manual assessment by 2 teachers. The first step count correlation assessment between teachers. Furthermore compare assessment results of each teacher using the ASAS system and average results teacher assessment using the ASAS system. The results evaluation were also compared with results study previously with ASAS that uses use and Word Embedding combined with Syntactic Analysis

Table 4. Comparison of Correlation Results Coefficient

No	ASAS Comparison	Coefisien Correlation	Mean Absolute Error (MAE)
1	Teacher 1 (Manual) vs Teacher 2 (Manual)	0,8481	0.2920

2	Teacher 1 (Manual) vs proposed ASAS (Automated used Sentence Transformer)	0,8232	0.5339
3	Teacher 2 (Manual) vs proposed ASAS (Automated used Sentence Transformer)	0,7372	0.6093
4	Average Teacher 1&2 (Manual) vs proposed ASAS (Automated used Sentence Transformer)	**0,8132**	**0.5096**

Results and Mean Absolute Error

Comparison grades produced by the system, teacher 1 and teacher 2 can manually assess seen in the graph below This:

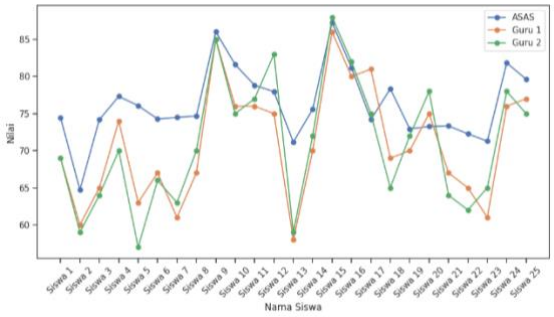

Figure 6. Comparison of Correlation Coefficient

ASAS uses a sentence transformer to have excess that Evaluation automatic capable give consistent assessment without influenced by factors external like tiredness, atmosphere heart, or possible personal bias experienced by teachers. Evaluation Algorithm automatic hasn't a possible bias subjective owned by the teacher. Algorithm evaluate based on existing patterns and data, so decrease the potential for bias is not realized.

This model also has lack. It is not enough capable understand context in a way deep as can be carried out by the teacher. so that system skip bullet points important as possible valued in manual assessment. Manual grading by teachers involves consideration to analysis whereas system depends on the model. Teachers tend to give tolerance to small errors and focused on student effort, while system possible evaluate based on more strict rules automatically

CONCLUSION

This research analyzed automatic essay answer assessment using semantic similarity based on sentence embedding. This research produces a Pearson correlation coefficient of 0.81. Based on the interpretation table coefficient correlation by sugiyono this number describe relationship between evaluation using the ASAS system with manual teacher assessment. It concluded that assessment system 81% agree with human rater assessment. This matter expected can help teachers in carry out essay exam and become consideration for school will use essay evaluation automatically.

The next research can Improve Correlation Automatic Short Answer Scoring (ASAS) and Manual Scoring by Teacher on Indonesian Assessments. Automatic Short Answer Scoring (ASAS) can process numbers, math symbols, images and can used in the longer essays.

ACKNOWLEDGMENT

This research full funded by Indonesian Education Scholarship (BPI) program between Center for Higher Education Fund ministry of education, culture, research, and technology of the republic Indonesia (PUSLAPDIK) and The Indonesia Endowment Funds for Education (LPDP).

REFERENCES

Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai. (2023). Jurnal Teknologi Informasi dan Ilmu Komputer, 10(6), 1177–1184. https://doi.org/10.25126/jtiik.1067338

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice, 25(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Cerratto Pargman, T., Lindberg, Y., & Buch, A. (2023). Automation Is Coming! Exploring Future(s)-Oriented Methods in Education. Postdigital Science and Education, 5(1), 171–194. https://doi.org/10.1007/s42438-022-00349-6

Chalmers, D., & McAusland, W. (2002). Computer-assisted assessment. The Handbook for Economics Lecturers, 1–20.

Chicco, D., Starovoitov, V., & Jurman, G. (2021). The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. IEEE Access, 9, 47112–47124. https://doi.org/10.1109/ACCESS.2021.3068614

Choi, H., Kim, J., Joe, S., & Gwon, Y. (2021). Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. 2020 25th International Conference on Pattern Recognition (ICPR), 5482–5487. https://doi.org/10.1109/ICPR48806.2021.9412102

Conneau, A., & Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. ArXiv, abs/1803.05449. https://api.semanticscholar.org/CorpusID:3932228

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. ALT-J, 13(1), 17–31. https://doi.org/10.1080/0968776042000339772

Dadi, R., & Sanampudi, S. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55, 1–33. https://doi.org/10.1007/s10462-021-10068-2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North

American Chapter of the Association for Computational Linguistics. https://api.semanticscholar.org/CorpusID:52967399

Gomaa, W. H., & Fahmy, A. A. (2020). Ans2vec: A Scoring System for Short Answers. Dalam A. E. Hassanien, A. T. Azar, T. Gaber, R. Bhatnagar, & M. F. Tolba (Eds.), The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) (hlm. 586–595). Springer International Publishing.

Gu, P. Y., & Lam, R. (2023). Developing Assessment Literacy for Classroom-Based Formative Assessment. 46(2), 155–161. https://doi.org/10.1515/CJAL-2023-0201

Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., & Pambudi, R. A. (2018). An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian. 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), 230–234. https://doi.org/10.1109/ICITISEE.2018.8720957

Herwanto, G. B., Sari, Y., Prastowo, B. N., Bustoni, I. A., & Hidayatulloh, I. (2018). UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia. ICEAP Proceeding Book Vol 2. https://api.semanticscholar.org/CorpusID:209097879

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. Organizational Research Methods, 25(1), 114–146. https://doi.org/10.1177/1094428120971683

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. 2016 4th International Conference on Cyber and IT Service Management, 1–6. https://doi.org/10.1109/CITSM.2016.7577578

Li, B., & Han, L. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. Dalam H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, & X. Yao (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2013 (hlm. 611–618). Springer Berlin Heidelberg.

Lubis, F. F., Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., Rosmansyah, Y., & others. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. International Journal of Technology, 12(3), 571–581.

Mamun, A. A., Sohel, M., Mohammad, N., Sunny, M. S. H., Dipta, D. R., & Hossain, E. (2020). A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models. IEEE Access, 8, 134911–134939. https://doi.org/10.1109/ACCESS.2020.3010702

Ormerod, C. M. (2022). Short-answer scoring with ensembles of pretrained language models. ArXiv, abs/2202.11558. https://api.semanticscholar.org/CorpusID:247058701

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? ArXiv, abs/1906.01502. https://api.semanticscholar.org/CorpusID:174798142

Putnikovic, M., & Jovanovic, J. (2023). Embeddings for Automatic Short Answer Grading: A Scoping Review. IEEE Transactions on Learning Technologies, 16(2), 219–231. https://doi.org/10.1109/TLT.2023.3253071

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012, Oktober). Semantic Cosine Similarity. [Tipe publikasi tidak jelas, diasumsikan Prosiding Konferensi].

Rajagede, R. A. (2021). Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(1), 11–18. https://doi.org/10.22219/kinetik.v6i1.1196

Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. Education and Information Technologies, 27(4), 5573–5604. https://doi.org/10.1007/s10639-021-10838-z

Ratna, A. A. P., Astato, A. W., Budiardjo, B., & Hartanto, D. (2007). SIMPLE-O: Web Based Automated Essay Grading System Using Latent Semantic Analysis method for Indonesian Language considering weighted word and. Chairman of 10th International Conference on QIR 2007.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing. https://api.semanticscholar.org/CorpusID:201646309

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., Kamaev, V. A., & others. (2013). A survey of forecast error measures. World Applied Sciences Journal, 24(24), 171–176.

Stephens, D. (2001). Use of computer assisted assessment: Benefits to students and staff. Education for Information, 19(4), 265–275. https://doi.org/10.3233/EFI-2001-19401

Sugiyono, P. (2019). Metode penelitian pendidikan (kuantitatif, kualitatif, kombinasi, R&D dan penelitian pendidikan). Metode Penelitian Pendidikan, 67.

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-Training BERT on Domain Resources for Short Answer Grading. Dalam K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (hlm. 6071–6075). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1628

Susongko, P. (2010). Perbandingan Keefektifan Bentuk Tes Uraian dan Teslet dengan Penerapan Graded Response Model (GRM). Jurnal Penelitian Dan Evaluasi Pendidikan, 14. https://api.semanticscholar.org/CorpusID:142555426

Tim Pusat Penilain Pendidikan. (2019). Panduan Penilaian Tes Tertulis. Pusat Penilaian Pendidikan.

Wang, B., & Kuo, C. .-C. J. (2020). SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2146–2157. https://doi.org/10.1109/TASLP.2020.3008390

Yunanda, G., Nurjanah, D., & Meliana, S. (2022). Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods. Building of Informatics, Technology and Science (BITS), 4(1), 277−284. https://doi.org/10.47065/bits.v4i1.1670

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. Knowledge-Based Systems, 120, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006