Pinnacle Academic Press
Proceedings Series

Vol. 6 2025

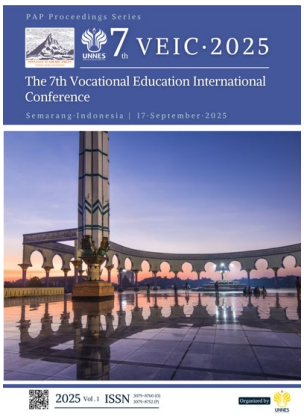
Article **Open Access**

The 7th Vocational Education International Conference (VEIC 2025)

Improving Automatic Essay Assessment Through Cosine Similarity Leveraging a Semantic Corpus

Fitria Ekarini [1], Septian Eko Prasetyo [1,*], Anan Nugroho [1], Alfian Ardhiansyah [1], Clarita Aprilliani [1] and Fakhri Ahmda Kurnia [1]

[1] Universitas Negeri Semarang, Semarang, Indonesia
* Correspondence: Septian Eko Prasetyo, Universitas Negeri Semarang, Semarang, Indonesia



Abstract: Automatic Essay Scoring (AES) represents an effective solution for facilitating automated evaluation of written essays by mitigating evaluator subjectivity and accelerating the assessment process. Nonetheless, a persistent challenge lies in achieving high accuracy due to limitations in semantic understanding. This study employs Cosine Similarity as a baseline approach and further integrates a semantic data corpus to enhance the representation of textual meaning. Empirical results demonstrate that relying solely on Cosine Similarity captures predominantly lexical-level similarities, yielding limited correlation with human scoring. The incorporation of a semantic corpus substantially improves the system's capacity to recognize synonyms and linguistic variations, thereby enhancing the sensitivity and reliability of the scoring process. Despite these improvements, the findings underscore the necessity for further corpus refinement, evaluation on larger and more diverse datasets, and assessment using multiple correlation metrics. Overall, this study provides a substantive contribution to the development of AES systems that are more accurate, consistent, and closely aligned with human judgment, thereby advancing the field of automated educational assessment.

Keywords: Automatic Essay Scoring; Cosine Similarity; semantic data; text processing

Received: 05 November 2025
Revised: 21 November 2025
Accepted: 25 December 2025
Published: 27 December 2025

1. Introduction

The rapid advancement of information technology and Artificial Intelligence (AI) has had a significant impact across various sectors, including education [1,2]. The integration of technology in learning processes has fostered numerous innovations, such as online learning, Learning Management Systems (LMS), and various digital educational tools that facilitate access and interaction within educational settings. One aspect that has undergone substantial transformation is digital-based evaluation systems. These technologies not only simplify the distribution of assessments to students but also enable automation in the grading process, reducing subjectivity, while providing faster, more efficient, and accurate analysis of results. With the increasing demand for adaptive, data-driven evaluation systems, the application of intelligent technologies in academic assessment has become highly relevant to enhance both the effectiveness and quality of learning [3].

However, although digital systems have proven effective in handling multiple-choice and short-answer assessments, automatic evaluation of essay responses still faces

complex challenges. Manual essay grading is often time-consuming and prone to subjectivity, bias, and inter-rater inconsistencies [4,5]. Differences in interpretation among evaluators can result in inconsistent scores, reducing the objectivity of academic assessment. These challenges become particularly significant when the volume of essays to be graded within a period is high. For instance, studies show that instructors may spend substantial time grading student essays, potentially compromising the efficiency and precision of the evaluation process [6,7]. High workloads can also induce cognitive fatigue, ultimately affecting grading quality. Consequently, there is a pressing need for technology-based solutions that can enhance efficiency and consistency in essay evaluation without compromising fairness and accuracy.

To address these challenges, Automatic Essay Scoring (AES) systems have been developed as an alternative solution leveraging Natural Language Processing (NLP) and AI to automatically evaluate essay responses [8]. AES enables faster, more consistent, and objective assessment compared to manual grading, thereby reducing instructors' workload and increasing educational efficiency. One widely used method in AES is Cosine Similarity, which measures the degree of similarity between student responses and reference answers based on vector representations of words in multidimensional space. This method calculates the angle between two text vectors, where higher values indicate greater similarity. While effective in measuring lexical similarity, Cosine Similarity has limitations in capturing semantic meaning and contextual variations in student responses. It primarily relies on explicit word matches, making it less capable of recognizing synonyms, semantic relationships between phrases, or differences in writing style that convey similar meaning [9,10].

Previous research has explored various approaches to enhance AES accuracy using more advanced natural language processing techniques [11]. Commonly applied methods include Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA), which extract word meaning by representing text in a mathematical semantic space, allowing systems to recognize latent relationships between words within a document [12]. Additionally, Word Embeddings, such as Word2Vec and GloVe, have been employed to capture contextual meaning by mapping words to multidimensional vectors based on their usage patterns in large corpora [13]. Although these techniques improve semantic understanding, they still have limitations in capturing higher-level semantic relationships, such as implicit meanings in a text or stylistic variations used by students in essays. These shortcomings can reduce AES sensitivity to differences in language style, synonyms, and sentence structures that convey the same meaning, highlighting the need for more adaptive and context-aware approaches to improve automatic scoring performance.

Research indicates that Cosine Similarity exhibits an error rate of 59.49% in evaluating essays in Indonesian, suggesting that this approach is not sufficiently accurate for a functional AES system. The high error rate stems from Cosine Similarity's limitation in capturing deeper semantic meaning, as it measures only lexical similarity without understanding relationships between words in various contexts [14]. Consequently, the system often fails to identify synonyms, expression variations, and sentence structures that convey equivalent meaning. This limitation underscores the need for further development of AES systems using more sophisticated approaches, such as leveraging semantic data corpora to enrich contextual understanding of text, alongside integrating adaptive NLP methods to capture nuanced meanings across diverse student language usage.

Previous studies in Automatic Essay Scoring (AES) have widely adopted vector-based similarity methods to evaluate essays automatically. Pribadi et al. employed a word-overlap approach, which is effective for short answers but fails to capture deeper semantic meaning, making it less suitable for complex essays. Fauzi et al. developed an AES system using N-Grams and Cosine Similarity within a gamified e-learning platform, showing optimal results with unigrams, though it still struggled to understand semantic

context fully [15]. Pramukantoro et al. compared string similarity with corpus-based similarity using Latent Semantic Analysis (LSA), demonstrating that LSA outperforms string-based methods with a correlation of 59.7%, albeit requiring larger corpora and higher computational resources [16]. Another study found that Cosine Similarity combined with LSA outperformed k-Nearest Neighbors (k-NN), confirming the effectiveness of semantic vector-based methods, though limitations remain in capturing syntactic and grammatical nuances. Sitikhu et al. showed that Cosine Similarity with TF-IDF vectors performed best for short news texts, indicating its continued relevance in measuring text similarity, but it is less optimal for longer, more complex texts like essays. Therefore, this study aims to combine Cosine Similarity with semantic corpus utilization to improve AES accuracy, addressing limitations of previous research [17].

This study aims to enhance the accuracy of Cosine Similarity-based AES by utilizing a semantic data corpus. By enriching the scoring model with extensive semantic information, the system is expected to more effectively comprehend context and linguistic variations in student essay responses. The anticipated contribution of this research is the development of an automatic essay scoring system capable of better understanding the meaning and context of student answers, thereby supporting a more fair and accurate learning evaluation process.

2. Materials and Methods

To ensure the effectiveness of the Automatic Essay Scoring (AES) system, the research follows a structured sequence of text preprocessing and analysis steps before applying the main similarity and semantic techniques [18]. The preprocessing stage is crucial as it transforms raw essay input into a standardized and analyzable form, reducing noise and improving the accuracy of subsequent similarity computations. The overall research process can be seen in Figure 1.

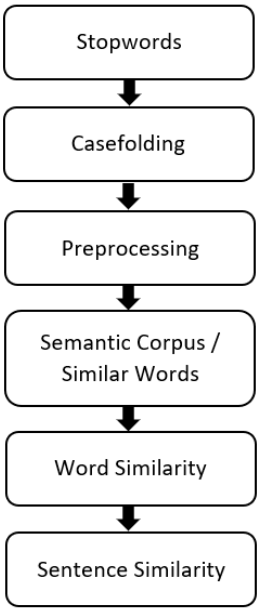


Figure 1. Research Step.

The research methodology follows a structured sequence of text processing and similarity measurement steps to ensure accurate and meaningful comparison of student essays. The stages are as follows:

1) Stopword Removal

Common words that carry little semantic weight (e.g., dan, yang, adalah) are removed. This reduces noise in the dataset and ensures that only meaningful terms contribute to the similarity analysis.

2) Case Folding

All text is converted into lowercase form. This normalization prevents discrepancies caused by capitalization (e.g., Microsoft vs. microsoft).

3) Preprocessing

At this stage, sentences are cleaned by removing punctuation and splitting text into tokens. The preprocessing ensures uniform input for the similarity functions.

4) Semantic / Corpus-based Word Mapping

A semantic corpus is developed to capture relationships between words. Synonyms and contextually related terms are stored in a structured corpus. This allows the system to identify semantic equivalence even when different word choices are used.

5) Word Similarity Measurement

Each word from a student essay is compared to words in the reference essay. If two words are identical or appear as synonyms in the corpus, they are considered semantically similar.

6) Sentence Similarity Computation

The similarity between two sentences is computed by aggregating the word-level similarity scores. Each word in the first sentence is matched with the most semantically similar word in the second sentence [19]. The final similarity score is calculated as the average of these maximum values, producing a semantic similarity measure between the two sentences.

This study employs Cosine Similarity in conjunction with a semantic data corpus to enhance the scoring performance of an Automatic Essay Scoring (AES) system. Cosine Similarity is a vector-based metric that quantifies the degree of similarity between two textual representations by analyzing the distribution of words in a multidimensional feature space. In the context of AES, each essay is encoded as a vector, and the cosine of the angle between the essay vector and the corresponding reference vector is computed. Cosine Similarity scores range from 0 to 1, where higher values indicate greater semantic alignment between the texts.

While Cosine Similarity is effective in capturing lexical and syntactic similarities, it exhibits inherent limitations in detecting deeper semantic relationships. Sentences conveying equivalent meanings but expressed with different word choices may yield low similarity scores, thereby potentially compromising the accuracy of automated scoring. Consequently, a context-aware approach that incorporates semantic information is essential for more reliable assessment of textual content. Figure 2 illustrates the application of Cosine Similarity in measuring vector-based relationships between words within a sentence. Cosine similarity is employed to measure the similarity between two document vectors, A and B, which represent the textual content numerically (e.g., word frequency, TF-IDF, or word embeddings). It is defined as:

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

Figure 2. Cosine Similarity Measure.

To improve scoring results, it is necessary to utilize a data corpus that can provide an understanding of words with similar meanings within a sentence, allowing the system to capture nuances beyond simple word matching. Therefore, this study will develop a semantic corpus that can be directly populated by users, leveraging supporting data relevant to the given prompts. The corpus will not only include synonyms but also

contextual relationships between words and phrases, enabling the AES system to recognize semantic patterns and infer meaning more accurately. By utilizing such a semantic data corpus, the system is expected to interpret essays in a more contextually aware manner, rather than relying solely on explicit word similarity or surface-level text features. This approach also allows for continuous improvement of the corpus over time, as users contribute new data that reflect evolving language usage and domain-specific terminology. The basic concept of the corpus design is illustrated in Figure 3 below.

```
Text 1 {
    Data Semantic 1, Data Semantic 2, Data Semantic 3
};
Text 2{
   Data Semantic 1, Data Semantic 2, Data Semantic 3
}
```

Figure 3. Semantic Corpus Data Concept.

By combining these two approaches—vector-based similarity measures and a dynamically enriched semantic corpus—this study aims to enhance the accuracy and reliability of the AES system, making its assessments more consistent with human grading. Furthermore, the integration of semantic understanding is expected to improve the system's ability to evaluate complex essay structures, infer implied meanings, and provide more meaningful feedback. Ultimately, this research contributes to the broader development of technology for automated essay evaluation, offering a foundation for future studies in educational assessment and natural language understanding.

3. Results and Discussion

The performance of the proposed Cosine Similarity-based Automatic Essay Scoring (AES) system was evaluated in two scenarios: using pure Cosine Similarity and integrating a semantic data corpus. Initial evaluation using only Cosine Similarity revealed that the system primarily captured lexical-based similarity, i.e., identical words in form. Consequently, the correlation with manual scoring was relatively low, ranging between 0.4 and 0.5. This finding aligns with previous studies indicating that Cosine Similarity has inherent limitations in assessing essays written in natural languages, particularly in Indonesian, due to variations in synonyms, writing styles, and sentence structures.

After integrating the semantic data corpus, significant improvements were observed. The system became more sensitive to synonyms and contextual variations. For example, sentences containing "pendidikan" (education) and "pembelajaran" (learning), which were previously scored differently, were now recognized as semantically equivalent. Figures 4 and Figure 5 illustrate the difference in similarity measurements. Figure 4 shows similarity scores using pure Cosine Similarity, whereas Figure 5 shows scores after semantic corpus integration. The semantic-enhanced system produced higher similarity scores, with semantically equivalent sentences approaching a similarity value of 1, indicating near-perfect alignment with human judgment.

Kalimat 1	Microsoft word adalah software pengelola kata
Kalimat 2	Ms word merupakan aplikasi pengolah kata
Similarity Index	0.4

Figure 4. Similarity Index Using Cosine Algorithm.

Kalimat 1	Microsoft word adalah software pengelola kata
Kalimat 2	Ms word merupakan aplikasi pengolah kata
Semantic Similarity	0.76

Figure 5. Similarity Index Using Cosine Algorithm with Semantic Corpus.

These results confirm that pure lexical-based Cosine Similarity alone is insufficient for accurately evaluating essays, particularly for text containing varied expressions and synonyms. The integration of a semantic corpus substantially improves the system's capacity to recognize equivalent expressions, enhancing the overall similarity assessment.

The findings highlight the importance of semantic enrichment in Automatic Essay Scoring. By incorporating a semantic data corpus, the system can bridge the limitations of purely lexical similarity and provide a more context-aware analysis. For instance, the identification of "pendidikan" and "pembelajaran" as semantically similar demonstrates that the AES system can now capture meaning beyond word overlap, which is crucial for more accurate and fair assessment.

Despite these improvements, several challenges remain. First, the corpus coverage is currently limited to certain academic texts, meaning that semantic representation may not fully capture the linguistic diversity of Indonesian essays. Second, the dataset size used for evaluation is relatively small, limiting generalizability. Third, the correlation analysis between system and manual scores has so far employed simple measures, and further evaluation using more comprehensive statistical metrics—such as Pearson Correlation, Spearman Rank Correlation, and Cohen's Kappa—is required to objectively assess alignment with human judgment.

The implications of these findings are significant for AES development. By integrating a semantic corpus, the system can go beyond word-form similarity and capture meaning embedded in text, potentially producing more fair, consistent, and human-aligned scoring. This is especially important in the Indonesian language, which exhibits substantial variation in vocabulary and expression. Moreover, the results suggest that AES development should not rely solely on classical lexical approaches but also incorporate advanced Natural Language Processing (NLP) techniques, such as Word Embeddings or transformer-based models, to further enhance semantic understanding.

Overall, the preliminary results of this study indicate a positive direction in improving AES accuracy. With broader semantic data enrichment, larger datasets, and more robust evaluation, a Cosine Similarity-based AES system enhanced with a semantic corpus has strong potential to become an effective automated scoring tool for supporting educational assessment and learning evaluation.

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

4. Conclusions

This study demonstrates that pure Cosine Similarity is primarily effective in capturing lexical-based similarity, resulting in a relatively low correlation with manual scoring. This underscores the limitations of word-overlap approaches in evaluating Indonesian essays, which often exhibit diverse vocabulary and writing styles. The integration of a semantic data corpus significantly improved scoring quality, particularly in recognizing synonyms and variations in language expression, allowing the system to provide fairer and more consistent assessments even when authors use different linguistic styles. Overall, these findings indicate that semantic-based approaches can bridge the shortcomings of Cosine Similarity and support the development of Automatic Essay Scoring (AES) systems that more closely approximate human judgment. For future development, it is recommended to expand and refine the semantic corpus to better represent the diversity of the Indonesian language, test the system on larger and more varied datasets to ensure stability and consistency, and employ comprehensive evaluation metrics such as Pearson Correlation, Spearman Rank Correlation, and Cohen's Kappa for more objective comparison with manual scoring. Additionally, integrating advanced NLP techniques, such as word embeddings or transformer-based models, could further

enhance the system's ability to capture complex semantic meanings, ultimately contributing to more reliable and context-aware automated essay evaluation in educational settings.

References

1. L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access*, vol. 8, 2020.
2. A. Alam, "Possibilities and apprehensions in the landscape of artificial intelligence in education," in Proc. *Int. Conf. Computational Intelligence and Computing Applications (ICCICA),* 2021.
3. K. Ernawati, B. S. Nugroho, C. Suryana, A. Riyanto, and E. Fatmawati, "The advantages of digital applications in public health services on automation era," *Int. J. Health Sci. (Qassim*), vol. 6, no. 1, 2022.
4. D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artif. Intell. Rev.,* vol. 55, 2021.
5. E. E. Hall, "A user-centered design approach to evaluating the usability of automated essay scoring systems," M.S. thesis, Virginia Tech, Blacksburg, VA, USA, 2023.
6. Z. Berezvai, G. D. Lukáts, and R. Molontay, "Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching," *Assess. Eval. High. Educ.*, vol. 46, no. 5, 2021.
7. W. Stroebe, "Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis," *Basic Appl. Soc. Psych.,* vol. 42, no. 4, 2020.
8. C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, "A comprehensive review of automated essay scoring (AES) research and development," *Pertanika J. Sci. Technol.,* vol. 29, no. 2, 2021.
9. B. D. Wijanarko, Bachtiar, R. B. Hassan, D. F. Murad, R. B. Ihsan, and Y. Heryadi, "AI-based feature extraction and cosine similarity for automation of student learning assessment," in *Proc. Int. Arab Conf. Information Technology (ACIT),* 2023.
10. J. Y. H. Bai et al., "Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education," in *Pan-Commonwealth Forum* 10 (PCF10), 2022.
11. V. Wagh, S. Laddha, and P. Kadam, "Detecting plagiarism using latent semantic analysis and cosine similarity approach," in *Proc. IEEE Int. Conf. Blockchain and Distributed Systems Security (ICBDS)*, 2024.
12. S. Ahmad and M. Laroche, "Extracting marketing information from product reviews: A comparative study of latent semantic analysis and probabilistic latent semantic analysis," *J. Marketing Analytics,* vol. 11, no. 4, 2023.
13. E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "The accuracy comparison among Word2Vec, GloVe, and FastText towards convolution neural network (CNN) text classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 2, 2022.
14. F. Rahutomo, T. A. Roshinta, R. Erfan, and I. Siradjuddin, "Open problems in Indonesian automatic essay scoring system," *Int. J. Eng. Technol*., vol. 7, no. 4, 2018.
15. F. Pribadi, T. B. Adji, A. E. Permanasari, and A. Mulwinda, "Automatic short answer scoring using words overlapping methods," in *Proc. 5th Int. Conf. Education, Concept, and Application of Green Technology*, 2017.
16. M. A. Fauzi, D. C. Utomo, B. D. Setiawan, and E. S. Pramukantoro, "Automatic essay scoring system using n-gram and cosine similarity for gamification-based e-learning," in *Proc. Int. Conf. Advances in Image Processing (ICAIP),* 2017.
17. E. S. Pramukantoro and M. A. Fauzi, "Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification," in *Proc. Int. Conf. Advanced Computer Science and Information Systems (ICACSIS),* 2016.
18. A. A. Ewees, M. Eisa, and M. M. Refaat, "Comparison of cosine similarity and k-NN for automated essays scoring," *Int. J. Adv. Res. Comput. Commun. Eng*., vol. 3, no. 12, 2014.
19. P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A comparison of semantic similarity methods for maximum human interpretability," in *Proc. Int. Conf. Artificial Intelligence for Transforming Business and Society*, 2019.