BIS INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

Volume 3 2026 | DOI: 10.31603/bistycs.485
7th Borobudur International Symposium on Science and Technology (BIS-STE) 2025

UNIMMA PRESS
Universitas Muhammadiyah Magelang

Evaluation of Word2Vec and FastText models for text similarity measurement assessment

Tukino[1,2*], Eko Sediyono[2], Hendry Hendry[2], Agustia Hananto[1], Elfina Novalia[1], and Fitria Nurapriani[1]

[1] Buana Perjuangan Karawang University, Karawang, Indonesia
[2] Satya Wacana Christian University, Salatiga, Indonesia
[*]Corresponding author email: tukino@ubpkarawang.ac.id

Abstract

Measuring text similarity assessment is crucial in the field of Education in the digital age, such as automated question evaluation, content alignment, and mapping learning outcomes, but estimating semantic similarity accurately for short and specific texts is challenging. Existing approaches often lack systematic comparisons across embedding models and weighting schemes. We evaluated Word2Vec and FastText embeddings (CBOW and Skip-gram) combined with TF-IDF, POS weighting, and BM25, to calculate cosine similarity using 112 sentence pairs and evaluated the models using Pearson and Spearman correlations as well as RMSE and MAE to compare the scores from the models with those from experts. The best performing configurations were FT+CBOW+TFIDF (highest Pearson's α = 0.7493) for semantic agreement with experts and W2V+CBOW+TF-IDF (lowest mean error, MAE = 0.91, RMSE = 1.10, overall error = 1.01) for prediction accuracy; The BM25-based variant produced significantly higher errors. These findings indicate that CBOW with TF-IDF provides the most stable similarity estimates for short educational texts, which supports automated evaluation tools in learning environments.

Keywords

Word2Vec, FastText, Cosine similarity, Text similarity, TF-IDF, RMSE, MAE

Published:
May 04, 2026

Selection and Peer-review under the responsibility of the 7th BIS-STE 2025 Committee

Introduction

Text similarity measurement plays a crucial role in diverse academic contexts including aligning learning outcomes, evaluating assessment questions, and designing digital learning evaluations. Accurate text similarity calculations enable automated measurement models to produce assessment scores comparable to expert-referenced assessments, as demonstrated in recent research on automated evaluation and embedding-based semantic analysis [1]. However, evaluating semantic similarity in short, specific educational texts remains an area of potential improvement because

lexical and word-level features often fail to fully represent contextual meaning and semantic nuances [2].

Previous research has introduced various approaches, ranging from lexical similarity methods such as TF-IDF to embedding-based models such as Word2Vec, FastText, BERT, and GPT-based language models [3]. Word2Vec and FastText have been widely used due to their ability to capture semantic relationships between words in vector representations [4]. However, most existing research focuses on general datasets, while fewer studies investigate similarity measurements in real-world educational assessment texts, particularly in Indonesian.

Several research gaps can still be identified: (1) the application of text similarity in academic evaluation and learning outcomes remains underexplored; (2) comparative analysis of various weighting schemes (TF-IDF, POS, BM25) for short educational texts is still limited; and (3) systematic evaluations using expert judgment as the gold standard are rare. This study addresses these limitations by evaluating several configurations of the Word2Vec and FastText models on real-world assessment statements written by experts [5].

The purpose of this study is to evaluate the performance of various Word2Vec and FastText configurations in measuring text similarity and to identify the best-performing model for short educational texts. Model evaluation is performed using correlation-based metrics and error-based deviation metrics to ensure accuracy and agreement with expert benchmark assessments.

Related Work

Text similarity measurement has been extensively studied in various NLP applications such as information retrieval, question and answer, and automated scoring [6]. Early methods based on lexical overlap and word weighting are still widely used to perform various retrieval tasks, but tend to fail to capture semantic similarity effectively in short and specific texts where synonyms and morphological variations are common. Recent reviews and empirical studies have shown that although lexical approaches are computationally efficient, they are inferior to vector-based methods in semantic tasks due to limitations in representing contextual relationships and latent meanings [2]. Furthermore, the choice of weighting scheme (TF-IDF, BM25, Part-of-Speech (POS)) can substantially affect performance, depending on the task and text length [7], [8].

Embedding-based representations starting from static vector models such as Word2Vec and FastText have become the dominant approach for modeling semantic similarity. Comparative evaluations show that static embeddings (Word2Vec, FastText) provide strong performance on many similarity tasks at relatively low computational cost. Meanwhile, contextual models often yield higher semantic fidelity but at the expense of computational and annotation requirements; importantly, benchmarking work on Indonesian language corpora and short texts highlights domain-specific trade-offs and

the need to evaluate models with subword (FastText) and weighting strategies (TF-IDF, POS) for short educational texts [9].

Recent empirical studies that systematically compare embedding and weighting combinations reinforce the value of multi-configuration evaluation: they show that (1) combining TF-IDF with appropriate embeddings can improve sentence-level similarity for short texts [10], (2) BM25 while powerful for document retrieval may be less effective for semantic similarity of sentence pairs, and (3) evaluations incorporating expert judgment provide a more realistic assessment of the model's utility in educational settings [11], [12]. Based on these observations, this study is motivated to systematically compare Word2Vec and FastText across TF-IDF, POS, and BM25 weighting schemes using expert-addressed educational statements [13].

Method

Dataset

The dataset used consists of 112 unique short educational text pairs. Each pair includes an evaluation response from an academic (D1) and a reference statement or ideal answer as assessed by an expert (D2–D4). A similarity score is calculated by comparing the academic response (D1) to each reference statement (D2, D3, and D4). The use of expert-annotated references as the gold standard is consistent with current practice in semantic similarity evaluation to ensure human-aligned benchmarks [13].

Modeling and weighting

We evaluate four static word embedding configurations: Word2Vec (using CBOW and Skip-Gram) and FastText (using CBOW and Skip-Gram). Static embeddings remain competitive for tasks involving short texts and limited-domain data due to their lower computational requirements, as demonstrated in a comparative embedding study [14]. To complement the embeddings, three weighting schemes are applied before computing the text similarity measure:

1. Term Frequency-Inverse Document Frequency (TF-IDF) Weighting: This is a lexical weighting technique widely used in measuring text similarity because it emphasizes important words in documents and reduces the influence of common words. This approach has proven effective for short text similarity in various fields, including education and information retrieval [15], [16], [17].
2. Part-of-Speech (POS)-based weighting: This approach utilizes syntactic category information to strengthen the semantic role of nouns, verbs, and adjectives in measuring text similarity. This approach helps highlight highly meaningful and frequently used words in hybrid models that combine embedding and weighting to improve the accuracy of short text similarity [18], [19].
3. BM25 Weighting: Although it is an effective ranking-based method for Information Retrieval, especially for long documents, it is still considered a useful technique. However, at sentence-level semantic similarity, its performance is often weaker than

embedding-based methods because BM25 does not capture semantic depth or contextual relationships between words in depth [20], [21].

Finally, the semantic similarity between text pairs is calculated using Cosine similarity on the embedding vectors or weighted embeddings. This approach was chosen because cosine similarity effectively measures the orientational proximity of vectors, making it suitable for models based on numerical representations of words or sentences. This method is also a common standard in semantic similarity research because it is robust to variations in text length and able to represent relationships in a semantically stable manner [22], [23]. Thus, cosine similarity remains the primary method in evaluating modern embedding models [21].

Evaluation metrics

We used a combination of correlation-based and error-based metrics to evaluate model performance:

1. Pearson correlation (p) and Spearman correlation (rs) to measure the linear correlation and rank correlation between model-predicted similarities and expert scores. These metrics are standard in semantic similarity research [13].
2. RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) to quantify the magnitude of deviation from expert scores, which captures the precision and error of predictions. Using these two metrics offers a more comprehensive evaluation of model reliability [13].

This combined evaluation strategy aligns with best practices in recent embedding-based and hybrid similarity assessment models, which combine semantic, lexical, and syntactic features to improve the accuracy of short text representation. This approach has been shown to be effective in various recent studies [24].

Proposed method

The proposed model configuration is designed to leverage the complementary strengths of embedding architectures and weighting schemes. Word2Vec, in both its CBOW and Skip-gram variants, offers semantic representations based on contextual co-occurrence, making it effective for capturing latent relationships between words in short educational statements [25], as demonstrated in a comparative study of short embeddings [2] and a Word2Vec-based stacking-ensemble study [4]. FastText extends this capability by incorporating subword information, enabling the model to handle morphological variations and out-of-vocabulary terms more effectively a crucial advantage for Indonesian texts, which are rich in derivational and inflectional forms. This subword-based advantage is further reinforced by the FastText model's performance, which shows significant improvements in processing morphological variations [26], [22], [27].

The TF-IDF integration serves as a lexical weighting mechanism that improves term-level discrimination, allowing embeddings to focus on more informative words. POS-based weighting provides syntactic rationale by highlighting linguistically salient categories

especially nouns, verbs, and adjectives that are crucial in interpreting evaluative statements [19]. Meanwhile, BM25 offers a probabilistic relevance-based weighting scheme that has traditionally excelled in information retrieval tasks. Its use provides a rigorous benchmark to check whether IR-oriented weighting aligns with sentence-level semantic similarity. The overall rationale for combining these methods is to build a hybrid model that balances semantic, lexical, and structural cues, thus offering a more holistic framework for short text similarity measurement. Text similarity measurement model shown in Table 1.

Table 1. Text similarity measurement model.

Embedding	TF-IDF-Weighting	BM25-Weighting	POS-Weighting
W2V + CBOW	W2V-CBOW + TFIDF	W2V-CBOW + BM25	W2V-CBOW + POS
W2V + SG	W2V-SG + TFIDF	W2V-SG + BM25	W2V-SG + POS
FT+ CBOW	FT-CBOW + TFIDF	FT-CBOW + BM25	FT-CBOW + POS
FT + SG	FT-SG + TFIDF	FT-SG + BM25	FT-SG + POS

Results and Discussion

Text similarity analysis using the Word2Vec Model

The results in Table 2 show significant differences in the average similarity scores and stability between the embedding + weighting configurations of the Word2Vec model. Specifically, the W2V+CBOW+TF-IDF configuration stands out as having the lowest Overall Average Similarity Score of 1.87 and the lowest Overall Average Standard Deviation of 0.81. In the context of measuring text similarity for short texts and specific domains such as educational evaluation statements, a relatively low and stable similarity score indicates that the model tends to provide conservative but consistent similarity scores, meaning it does not "guess" semantic similarity when compared to expert references. This is important for applications such as content alignment or question evaluation, where over-calculation of similarity can lead to misjudgments. On the other hand, the W2V+SG+POS configuration shows the highest similarity score of 3.01 with a standard deviation of 0.83, despite its large average similarity score, its stability is relatively good. This shows that the Skip-Gram architecture with POS weighting is able to capture similarities between texts aggressively, perhaps because POS weighting gives weight to certain word tokens that are considered syntactically/semantically important.

However, this high average similarity value may indicate that the model tends to overgeneralize similarity, which can pose a risk when texts differ in meaning despite using similar vocabulary.

Furthermore, the BM25-based configuration (both CBOW and SG) showed the highest Overall Mean Standard Deviation of 1.9, indicating highly volatile model predictions. This high variability suggests that BM25 as a weighting scheme is not suitable for short-sentence text similarity tasks in educational settings consistent with research findings

showing BM25 to be more effective for document retrieval than for semantic similarity of short sentences or texts with limited vocabulary [13].

Overall, the data suggest that the combination of W2V+CBOW with TF-IDF provides the best balance between conservatism and score consistency a highly desirable trait when models are used for automated evaluation in learning systems. This suggests that, compared to other configurations, this model is more reliable in producing similarity estimates that are close to expert judgment, without the risk of high fluctuation or similarity overestimation.

Table 2. W2V Model similarity score and standard deviation

Text Similarity Measurement Model	Similarity Score			Standard Deviation			Average Standard Deviation
	P1	P2	P3	P1	P2	P3	
W2V+CBOW+TFIDF	1.98	1.98	1.65	0.97	0.73	0.72	0.81
W2V+SG+TFIDF	2.90	2.91	2.74	0.88	0.84	0.91	0.88
W2V+CBOW+BM25	1.41	1.33	1.47	1.94	1.87	1.96	1.92
W2V+SG+BM25	1.42	1.33	1.47	1.92	1.87	1.95	1.91
W2V+CBOW+POS	2.43	2.45	2.19	1.05	0.83	0.84	0.91
W2V+SG+POS	3.05	3.06	2.92	0.87	0.80	0.81	0.83

These results support the hybrid approach (embedding+weighting) widely recommended in the modern literature for similarity tasks in the short text domain. Previous studies such as "HyEWCos: A Comparative Study of Hybrid Embedding and Weighting Techniques for Text Similarity in Short Subjective Educational Texts" demonstrated that the combination of embedding (Word2Vec/FastText) with weighting (TF-IDF, POS, BM25) offers a reproducible and competitive baseline compared to embedding alone or lexical methods. Other studies comparing Word2Vec and Doc2Vec for document similarity also found that traditional embedding remains superior for documents or short texts, especially when combined with normalization methods such as cosine similarity.

Thus, these findings strengthen the claim that traditional embedding (Word2Vec) combined with statistical weighting (TF-IDF) remains relevant and effective for the educational domain, especially when computational or data resources are limited, where practical contributions become important as many educational institutions use short texts and the use of relatively stable text similarity measurement models.

Text similarity analysis using the FastText Model

Analysis of the six FastText configurations reveals consistent differences in model performance across assessment text pairs. According to Table 3, the FT+SG+POS model yielded the highest average similarity score of 3.46 with the lowest standard deviation among all configurations, at 0.67. This indicates that this configuration is able to capture semantic relationships more robustly while providing stable scores between text pairs. This pattern indicates that the Skip-Gram approach with linguistic structure-based (POS)

weighting is more sensitive in modeling the lexico-semantic features of short and specific educational evaluation texts.

This finding is consistent with contemporary NLP research reporting that FastText models excel at representing words with morphological variations and narrow contexts, particularly when used on natural language datasets with limited text length [28]. However, the FT+CBOW+TF-IDF model also exhibited strong performance characteristics, namely score stability with a low total standard deviation of 0.81. This shows that TF-IDF is able to provide a more consistent and consistent distributional representation on narrow domain texts, as also shown in recent research on domain text matching and educational analysis [29], [30].

In contrast to the configuration above, the FastText model combined with BM25 produces the highest standard deviation, especially in FT+SG+BM25 with a value of 1.58. This high score variability indicates that BM25 tends to be less effective when used for measuring the similarity of short sentence texts. This finding is supported by studies in the field of information retrieval which concluded that BM25 is more suitable for document retrieval contexts than calculating the semantic similarity of short sentences [31]. Thus, the pattern that emerged in FastText strengthens the results in Word2Vec, that BM25 does not provide stable similarity estimates for educational evaluation contexts.

Table 3. FastText similarity scores and standard deviations

Text Similarity Measurement Model	Similarity Score			Standard Deviation			Average Standard Deviation
	P1	P2	P3	P1	P2	P3	
FT+CBOW+TFIDF	2.38	2.35	2.11	0.84	0.82	0.76	0.81
FT+SG+TFIDF	3.33	3.29	3.23	0.78	0.86	0.84	0.83
FT+CBOW+BM25	0.80	0.63	0.73	0.84	0.87	0.99	0.90
FT+SG+BM25	2.23	1.92	2.17	1.28	1.77	1.70	1.58
FT+CBOW+POS	2.74	2.70	2.51	0.81	0.77	0.74	0.77
FT+SG+POS	3.49	3.47	3.41	0.65	0.70	0.67	0.67

Overall, these results demonstrate that the FastText configuration performs optimally when combined with TF-IDF and POS. These findings provide an important contribution to previous research that has typically been conducted on general domains or large datasets, while validating the effectiveness of FastText on short sentence-based text similarity tasks and the educational domain, a context rarely analyzed in previous literature. The practical implication of this study is that using FastText with an appropriate weighting configuration can improve the accuracy of automated assessment systems and text-based learning competency alignment.

Statistical result validation

Table 4 presents the statistical performance of six Word2Vec configurations evaluated using correlation- and error-based metrics. The results indicate that W2V+CBOW+TFIDF demonstrated the strongest agreement with expert similarity scores, achieving the

highest Pearson correlation values of 0.62 and Spearman correlation values of 0.55, combined with the lowest mean error (Mean Error of 1.01). This pattern reflects the model's ability to capture semantic closeness between short educational statements and maintain ranking consistency when compared to expert judgments. The relatively low RMSE and MAE values further confirm that the model's predictions are closer to the expert similarity scores, demonstrating its effectiveness in representing educational and domain-specific evaluation texts.

In contrast, the W2V configuration combined with BM25 yielded the lowest correlation values (Pearson 0.10 and Spearman 0.11) and the highest error score (Mean Error 1.85), indicating that BM25 is not optimal for sentence-level similarity. The high error values indicate significant deviations from expert scores, reinforcing the finding of a recent study that BM25 is better suited for longer documents or retrieval tasks than for short semantic matching [18]. Meanwhile, the POS-weighted model showed moderate performance, indicating that syntactic cues contribute positively, but do not outperform TF-IDF when applied to short educational assessment texts.

These findings align with recent research showing that W2V combined with frequency-based weighting tends to outperform other weighting schemes for semantic similarity in specific domains [4]. More importantly, the results of this study extend previous research by showing that performance validation using expert-based ground truth and combined metrics (correlation+error) provides a more reliable measurement framework, especially for educational evaluation. While previous studies have mostly relied on benchmark datasets or automated annotation, this study integrates expert judgment as a standard, which strengthens the ecological validity of the results and aligns with recent recommendations in the evaluation of text similarity models.

Table 4. W2V Model performance summary

Text Similarity Measurement Model	Pearson Corr	Spearman Corr	RMSE Score	MAE Score	Avg. Error
W2V+CBOW+TFIDF	0.62	0.55	1.10	0.91	1.01
W2V+SG+TFIDF	0.48	0.59	1.39	1.26	1.33
W2V+CBOW+BM25	0.10	0.11	1.91	1.78	1.85
W2V+SG+BM25	0.10	0.11	1.90	1.77	1.84
W2V+CBOW+POS	0.54	0.55	1.24	1.06	1.15
W2V+SG+POS	0.45	0.58	1.50	1.35	1.43

Overall, the statistical results highlight that the integration of the CBOW architecture with TF-IDF weighting yields the most stable and accurate similarity predictions for educational text similarity measurement. The implications of these findings are significant for real-world applications, as the model can be applied to automate assessments, validate learning materials, and support decision-making in academic evaluation systems.

Performance evaluation of word embedding model FastText

Table 5 presents the comparative performance of six FastText configurations evaluated using correlation-based and error-based metrics. The results indicate that the FT+CBOW+TFIDF configuration achieved the most balanced performance, indicated by relatively high Pearson correlations of 0.60 and Spearman correlations of 0.57, as well as the lowest mean error (Mean Error 1.03). These patterns indicate that FastText combined with TF-IDF is able to maintain semantic consistency and closeness to the similarity scores assigned by experts. The relatively high correlation and low error indicate that this configuration can effectively capture lexical variation and semantic nuances in short educational statements.

The results also show that FT+SG+POS achieves the highest similarity prediction error (Mean Error 1.75), despite a relatively strong Spearman correlation (0.60). This difference indicates that Skip-Gram with POS can correctly maintain the relative similarity ranking between text pairs, but produces larger absolute deviations in numerical predictions. This behavior has also been identified in a recent study, which highlighted that FastText Skip-Gram tends to produce larger embedding space variance when applied to short texts due to the highly frequent fragmentation of subword characters [28]. In contrast, FT+CBOW+TFIDF exhibits lower deviation volatility, confirming its suitability for short text similarity estimation.

Furthermore, configurations incorporating BM25 consistently exhibit lower performance, with higher RMSE and MAE values, regardless of whether CBOW or Skip-Gram is used. This trend aligns with findings in recent literature that emphasize the limitations of BM25 for sentence-level semantic similarity tasks, and its better suitability for document retrieval and ranking [30], [32]. The higher error values suggest that BM25 may over-penalize lexical sparsity in short educational statements, resulting in unstable predictions across evaluation pairs.

Table 5. FastText Model performance summary

Text Similarity Measurement Model	Pearson Corr	Spearman Corr	RMSE Score	MAE Score	Avg. Error
FT+CBOW+TFIDF	0.60	0.57	1.11	0.94	**1.03**
FT+SG+TFIDF	0.39	0.61	1.69	1.57	1.63
FT+CBOW+BM25	0.60	0.63	1.40	1.20	1.30
FT+SG+BM25	0.51	0.64	1.49	1.24	1.37
FT+CBOW+POS	0.53	0.55	1.27	1.01	1.14
FT+SG+POS	0.35	0.60	1.80	1.70	1.75

These results reinforce the idea that the advantage of the FastText architecture in its ability to incorporate subword information is most effective when combined with statistical weighting such as TF-IDF or POS. Compared with previous studies that primarily evaluated FastText using large benchmark datasets or domain-general corpora, this study provides empirical evidence based on expert-validated similarity scores in a real-world educational evaluation context. This methodological contribution

is significant because it demonstrates that FastText can be used to reliably approximate the decision-making patterns of domain experts when processing short-form assessment texts, offering practical implications for the automation of educational evaluation systems.

Comparison of model performance and expert assessment

Table 6 summarizes the comparative performance across selected similarity pairs evaluated by Word2Vec and FastText. The results show a clear pattern where the W2V+CBOW+TFIDF and FT+CBOW+TFIDF configurations consistently produce similarity scores closest to the expert evaluations. For example, the P1-28 pair (W2V+CBOW+TFIDF) exhibits the lowest error and nearly identical similarity scores between the model and expert assessments (1.01 vs. 1.00) with a standard deviation of only 0.01. This evidence strengthens the conclusion that TF-IDF weighting significantly contributes to stabilizing similarity score predictions and reducing fluctuations across text pairs. Meanwhile, the FT+CBOW+TFIDF configuration displays similar behavior with very small deviations (e.g., P3-22 with SD=0.03), indicating that FastText subword modeling is advantageous when capturing lexical variation at the sentence level.

In contrast, configurations involving BM25 or Skip-Gram exhibited larger prediction errors. The P2-2 pair evaluated by W2V+SG+BM25 yielded the largest deviation (RMSE/MAE equivalent error of 1.84), confirming that BM25 contributes to greater numerical instability when applied to short sentences. Similar behavior was also found with FT+SG+BM25, showing a large gap between model and expert similarity (e.g., P2-12 with 2.23 vs. 1.5). These results indicate that while Skip-Gram tends to preserve relative order (as reflected by moderate Spearman correlation), it is less effective in producing accurate similarity measurements. This is in line with recent studies reporting that SG-based embeddings have higher variance in short text similarity because word distribution in short sentences is sparse and highly context-dependent [28], [33].

Table 6. Model Performance vs. Expert Judgment

Text Similarity Measurement Model	Text Pair ID	Avg. Error	Pearson Corr	Model Score	Expert Score	Standard Deviation
W2V+CBOW+TFIDF	P1-28	1.01	0,62	1.01	1.00	0.01
W2V+CBOW+TFIDF	P3-8	1.01	0,62	2.10	1.00	0,10
W2V+CBOW+POS	P3-21	1.15	0,54	1.40	2.50	1,10
W2V+SG+BM25	P2-2	1.84	0,10	3.99	15	2.49
FT+CBOW+TFIDF	P3-22	1.03	0,6	2.97	3.00	0,03
FT+CBOW+POS	P3-28	1.14	0,53	0.94	2.50	1.56
FT+SG+BM25	P2-12	1.37	0,51	3.73	1.50	2.23

A key advantage of this study compared to previous research is the explicit use of expert-validated similarity scores as the ground truth. In studies related to text similarity measurement, text validation is considered more ecologically valid and domain-appropriate, particularly when evaluating texts for assessment and instructional alignment [34], [35]. Therefore, the results provide a more realistic reflection of

semantic correspondence between educational statements than approaches that evaluate models solely based on automated annotations.

Overall, these findings demonstrate a consistent pattern: (1) CBOW combined with TF-IDF produces the most reliable similarity estimates, (2) FastText improves performance in capturing morphological and lexical variations, and (3) BM25 consistently performs worse for short semantic similarities. These results have important practical implications, particularly for automating assessment review and text validation in educational contexts. By aligning similarity scores with expert judgment, the models used in this study can support real-world applications such as question bank evaluation, content recommendation, and automated outcome mapping.

Evaluation of W2V Model

Figure 1 shows a comparison of RMSE and MAE values for six Word2Vec configurations. The performance of the W2V+CBOW+TFIDF model produces the lowest RMSE and MAE values among all models, with an RMSE of 1.10 and an MAE of 0.91. This indicates that this configuration has the smallest prediction error. Meanwhile, the two BM25-based models produce the highest RMSE and MAE values, with an RMSE above 1.90 and an MAE above 1.75. These values indicate that BM25 is less suitable for measuring text similarity in the context of short texts such as self-evaluation statements.

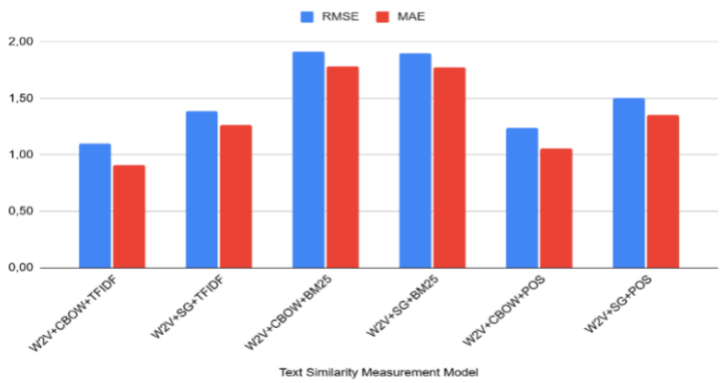

Figure 1. Evaluation chart of Word2Vec Models based on RMSE and MAE

The standard deviation between text pairs also shows that the W2V+CBOW+TFIDF and W2V+CBOW+POS models are more stable in predicting text similarity. Based on these results, it can be concluded that: CBOW performs better than Skip-Gram for short texts, TF-IDF improves the quality of vector representation, and BM25 is not suitable for sentence-level similarity due to its weighting calculation characteristics.

Evaluation of FastText Model

The evaluation results of the FastText model using TF-IDF, BM25, and POS weighting for the CBOW and Skip-Gram architectures are presented in Table 1. Furthermore, the performance evaluation of the FastText model based on RMSE and MAE values is shown in Figure 2. This chart provides a visual comparison of the performance for the 6 text similarity measurement models, making it easier to identify the model with the best performance based on the lowest error value.

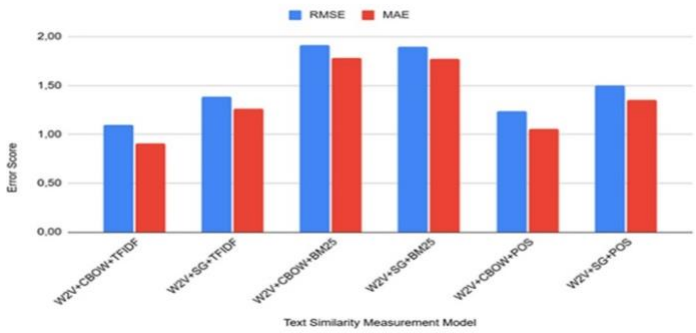

Figure 2. Evaluation Chart of FastText Models Based on RMSE and MAE

Based on Figure 2, it can be observed that the FastText–SkipGram model with the TF-IDF weighting approach obtained the lowest RMSE and MAE values when compared to the POS and BM25 approaches. This indicates that this model is closest to expert evaluation in predicting the similarity score of the assessment text. In contrast, the FastText–CBOW model combined with BM25 weighting showed the lowest performance with the highest error value. This finding confirms that the implementation of FastText with the SkipGram architecture is more effective in capturing semantic context in relatively short and varied texts.

Conclusion

This study evaluated several Word2Vec and FastText embedding configurations for measuring text similarity in accreditation assessments. Experimental results showed that the W2V+CBOW+TFIDF configuration achieved the lowest error score of 1.01, while the FT+CBOW+TFIDF configuration produced the highest correlation with expert judgment (Pearson's coefficient = 0.7493). These findings indicate that embedding-based similarity models are capable of representing semantic proximity in short academic texts with high accuracy. In contrast, BM25 consistently produced the highest error, indicating that this weighting scheme is less suitable for sentence-level similarity in short, domain-specific educational texts.

Overall, the results of this study confirm that embedding-based similarity measurement offers a reliable approach to support automated evaluation, learning material alignment, and assessment validation in educational settings. The use of expert judgment as ground truth strengthens the validity of the findings and highlights the potential application of these models in real-world academic evaluation systems. Future research can explore expanded datasets, incorporate transformer-based models, and develop hybrid similarity frameworks to improve prediction precision and model generalization across educational domains.

Acknowledgement

We would like to thank Buana Perjuangan University, Karawang, for providing the resources and facilities for conducting the research, as well as the editors and reviewers for their time and in-depth comments during the preparation of this article.

Abbreviations

P1-28	Pair #28 Text Pair D1 vs. D2
P3-8	Pair #8 Text Pair D1 vs. D4
P3-21	Pair #21 Text Pair D1 vs. D4
P2-2	Pair #2 Text Pair D1 vs. D3
P3-22	Pair #22 Text Pair D1 vs. D4
P3-28	Pair #28 Text Pair D1 vs. D4
P2-12	Pair #12 Text Pair D1 vs. D3
P1	Text Pair D1 vs. D2
P2	Text Pair D1 vs. D3
P3	Text Pair D1 vs. D4
TF-IDF	Term Frequency–Inverse Document Frequency
POS	Part-of-Speech Weighting
BM25	Probabilistic Relevance Weighting
CBOW	Continuous Bag of Words
SG	Skip-Gram
FT	Fastext

References

1. Z. Li, Y. Tomar, and R. J. Passonneau, "A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading," Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. Punta Cana, Dominic. Republic, 7–11 Novemb. 2021;, pp. 6030–6040, 2021, doi: 10.18653/v1/2021.emnlp-main.487.
2. K. Babić, F. Guerra, S. Martinčić-Ipšić, and A. Meštrović, "A comparison of approaches for measuring the semantic similarity of short texts based on word embeddings," J. Inf. Organ. Sci., vol. 44, no. 2, pp. 231–246, 2020, doi: 10.31341/jios.44.2.2.
3. M. Thapa, P. Kapoor, S. Kaushal, and I. Sharma, "A Review of Contextualized Word Embeddings and Pre-Trained Language Models, with a Focus on GPT and BERT," Proc. 1st Int. Conf. Cogn. Cloud Comput. Jaipur, India,1–2 August 2024, no. IC3Com 2024, pp. 205–214, doi: 10.5220/0013305900004646.
4. S. Subba, B.; Kumari, "A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings," Comput. Intell., vol. 38, no. 2, pp. 530–559, 2022, doi: https://doi.org/10.1111/coin.12478.
5. A. Allahim and A. Cherif, "Advancing Arabic Word Embeddings: A Multi-Corpora Approach with Optimized Hyperparameters and Custom Evaluation," Appl. Sci., vol. 14, no. 23, p. 11104, Nov. 2024, doi: 10.3390/app142311104.
6. J. Wang and Y. Dong, "Measurement of text similarity: A survey," Inf., vol. 11, 421., no. 9, pp. 1–17, 2020, doi: 10.3390/info11090421.
7. S. Das, A. Dutta, T. Lindheimer, M. Jalayer, and Z. Elgart, "YouTube as a Source of Information in Understanding Autonomous Vehicle Consumers: Natural Language Processing Study," Transp. Res. Rec., vol. 2673, no. 8, pp. 242–253, 2019, doi: 10.1177/0361198119842110.
8. C. Deng, G. Lai, and H. Deng, "Improving word vector model with part-of-speech and dependency grammar information," CAAI Trans. Intell. Technol., vol. 5, no. 4, pp. 260–267, 2020, doi:

10.1049/trit.2020.0055.
9. X. Li, A. Henriksson, M. Duneld, J. Nouri, and Y. Wu, "Evaluating Embeddings from Pre-Trained Language Models and Knowledge Graphs for Educational Content Recommendation," Futur. Internet, vol. 16, no. 1, 2024, doi: 10.3390/fi16010012.
10. J. Yang, S.; Huang, G.; Ofoghi, B.; Yearwood, "Short text similarity measurement using context-aware weighted biterms," Concurr. Comput. Pr. Exp., p. 34.e5765., 2020, doi: https://doi.org/10.1002/cpe.5765.
11. L. Xiao, Q. Li, Q. Ma, J. Shen, Y. Yang, and D. Li, Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec. PLoS ONE 2024, vol. 19,. e0305095. doi: 10.1371/journal.pone.0305095.
12. S. Ramadhani, M. A. Hariyadi, and C. Crysdian, "The Evaluation of Computer Science Curriculum for High School Education Based on Similarity Analysis," Int. J. Adv. Data Inf. Syst., vol. 4, no. 2, pp. 201–213, 2023, doi: 10.25008/ijadis.v4i2.1307.
13. H. Hendry, T. Tukino, E. Sediyono, A. Fauzi, and B. Huda, "HyEWCos: A Comparative Study of Hybrid Embedding and Weighting Techniques for Text Similarity in Short Subjective Educational Text," Inf., vol. 16, no. 11, pp. 1–28, 2025, doi: 10.3390/info16110995.
14. D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," J. Teknol. Inf. dan Ilmu Komput., vol. 12, no. 1, pp. 133–144, 2025, doi: 10.25126/jtiik.2025129143.
15. T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. Punta Cana, Dominic. Republic, 7–11 Novemb. 2021;, pp. 6894–6910, doi: 10.18653/v1/2021.emnlp-main.552.
16. R. Rani et al., "A Survey of Numerous Text Similarity Approach," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 10, no. November, pp. 132777–132785, 2021, doi: 10.1109/ACCESS.2022.3230592.
17. N. H. Hameed, A. M. Alimi, and A. T. Sadiq, "Short Text Semantic Similarity Measurement Approach Based on Semantic Network," Baghdad Sci. J., vol. 19, no. 6, pp. 1581–1591, 2022, doi: 10.21123/bsj.2022.7255.
18. K. Zhang, Y. Liu, F. Mei, G. Sun, and J. Jin, "IBGJO: Improved Binary Golden Jackal Optimization with Chaotic Tent Map and Cosine Similarity for Feature Selection," Entropy, vol. 25, 1128., no. 8, pp. 1–23, 2023, doi: 10.3390/e25081128.
19. C. Sánchez-Antonio et al., "A Short-Text Similarity Model Combining Semantic and Syntactic Information," Mathematics, vol. 12, no. 22, p. 3126, Nov. 2024, doi: 10.3390/electronics12143126.
20. D. Chandrasekaran and V. Mago, "Evolution of Semantic similarity a survey," ACM Comput. Surv., vol. 54, no. 2, pp. 1–35, 2021, doi: 10.1145/3440755.
21. M. R. A. H., M. Ilham, D. F. Surianto, and A. M. Mappalotteng, "Semantic Similarity Measurement Evaluation of KBBI Synonyms Using a Word Embedding Approac," J. Nas. Tek. Elektro dan Teknol. Inf., vol. 14, no. 2 SE-Articles, pp. 112–120, 2025, [Online]. Available: https://jurnal.ugm.ac.id/v3/JNTETI/article/view/17117
22. A. Pertiwi, A. Azhari, and S. Mulyana, "Fast2Vec, a modified model of FastText that enhances semantic analysis in topic evolution," PeerJ Comput. Sci., vol. 11, pp. 1–36, 2025, doi: 10.7717/peerj-cs.2862.
23. D. C. Kendaraan, "Analisis Sistem Pendeteksi Posisi Plat Kendaraan Dari Citra Kendaraan," J. Ilm. SPEKTRUM, no. July 2016, 2015, [Online]. Available: https://ojs.unud.ac.id/index.php/spektrum/article/view/20008
24. J. L. Xianming Li, "AoE - Angle-optimized Embeddings for Semantic Textual Similarity.," Proc. ofthe 62nd Annu. Meet. ofthe Assoc. Comput. Linguist. August 11-16, 2024, vol. 1, pp. 1825–1839, 2024, doi: 10.18653/v1/2024.acl-long.101.
25. A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. C. da Fonseca, "Semantic Sensitive TF-IDF to Determine Word Relevance in Documents," in In Advances in Computing and Network Communications; Lecture Notes in Electrical Engineering; Springer: Singapore, 2021, pp. 327-337. doi: 10.1007/978-981-33-6987-0_27.
26. S. Chawla, R. Kaur, and P. Aggarwal, "Text classification framework for short text based on TFIDF-FastText," Multimed. Tools Appl., vol. 82, no. 26, pp. 40167–40180, 2023, doi: 10.1007/s11042-023-15211-5.
27. M. Umer et al., "Impact of convolutional neural network and FastText embedding on text classification," Multimed. Tools Appl., vol. 82, no. 4, pp. 5569–5585, 2023, doi: 10.1007/s11042-022-13459-x.
28. Y. Wang, B. Zhang, W. Liu, J. Cai, and H. Zhang, "STMAP: A novel semantic text matching model augmented with embedding perturbations," Inf. Process. Manag., vol. 61, no. 1, p. 103576, 2024, doi:

10.1016/j.ipm.2023.103576.
29. D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques," Artif. Intell. Rev., vol. 56, no. 11, pp. 13407–13461, 2023, doi: 10.1007/s10462-023-10472-w.
30. A. M. Priyatno, M. R. A. Prasetya, P. Cholidhazia, and R. K. Sari, "Comparison of Similarity Methods on New Student Admission Chatbots Using Retrieval-Based Concepts," J. Eng. Sci. Appl., vol. 1, no. 1, pp. 32–40, 2024, doi: 10.69693/jesa.v1i1.2.
31. P. Gong, J. Liu, Y. Xie, M. Liu, and X. Zhang, "Enhancing context representations with part-of-speech information and neighboring signals for question classification," Complex Intell. Syst., vol. 9, no. 6, pp. 6191–6209, 2023, doi: 10.1007/s40747-023-01067-7.
32. T. Paryono, E. Sediyono, Hendry, B. Huda, A. Lia Hananto, and A. Yuniar Rahman, "Intelligent classification and performance prediction of multi-text assessment with recurrent neural networks-long short-term memory," IAES Int. J. Artif. Intell., vol. 13, no. 3, pp. 3350–3363, 2024, doi: 10.11591/ijai.v13.i3.pp3350-3363.
33. Y. Ma, X. Liu, L. Zhao, Y. Liang, P. Zhang, and B. Jin, "Hybrid embedding-based text representation for hierarchical multi-label text classification," Expert Syst. Appl., vol. 187, p. 115905, 2022, doi: 10.1016/j.eswa.2021.115905.
34. M. R. Islam, A. Ahmad, and M. S. Rahman, "Bangla text normalization for text-to-speech synthesizer using machine learning algorithms," J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 1, p. 101807, 2024, doi: 10.1016/j.jksuci.2023.101807.
35. M. O. Gani, R. K. Ayyasamy, S. M. Alhashmi, A. Sangodiah, and Y. T. Fui, "ETFPOS-IDF: A Novel Term Weighting Scheme for Examination Question Classification Based on Bloom's Taxonomy," IEEE Access, vol. 10, no. November, pp. 132777–132785, 2022, doi: 10.1109/ACCESS.2022.3230592.