



Behavioral Analysis of Semantic Similarity Metrics under Transformer-Based Representations

Musthofa Galih Pradana^{1*}, Nindy Irzavika², Nurhuda Maulana³, Syaila Ananta Karenina⁴, Salma Ashiila Rabbani⁵

^{1,2}Data Science Department, Computer Science Faculty, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta, Indonesia.

^{3,4,5}Informatics Department, Computer Science Faculty, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta, Indonesia.

*musthofagalihpradana@upnvj.ac.id

Abstract

Knowledge extraction has several approaches such as traditional approaches that rely on lexical representation capabilities, one of which is TF-IDF whose implementation can be combined with several classic similarity metrics such as cosine similarity and dice coefficient similarity. In addition to applying the lexical representation approach, this study tries to apply it to a more modern type of representation, namely transformer-embedding-based contextual representation. The data used in this study is abstract document data of students' theses. The findings of the study show that contextual embedding changes the behavior of similarity values. The results of the analysis showed an average ranking shift of 6.70 positions. The test results showed a weak rating correlation value (Spearman = 0.22; Kendall = 0.146), and the high-ranking alignment measured with NDCG (0.97), which shows structural differences in the order of similarity between lexical and contextual representations. Other findings also show that the gap in ranking produced by the two representations used is quite far due to the difference in the mechanism and working pattern of the two representations that are far different, the selection of the type of representation must be on the characteristics of the data to be processed, if looking at the character of the text data in academic documents, the selection of contextual representations based on transformer embedding will be more suitable with contextual understanding to avoid the use of variations in words avoid plagiarism detection when applying a semantic-based representation approach.

Keywords: cosine similarity; knowledge extraction; rank shift analysis; semantic similarity; transformer-based embeddings

How to Cite: M. G. Pradana, N. Irzavika, N. Maulana, S. A. Karenina, and S. A. Rabbani, "Behavioral Analysis of Semantic Similarity Metrics under Transformer-Based Representations", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 332 - 340, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.7490>

Received: January 14, 2026

Accepted: April 5, 2026

Available Online: April 20, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

The growth of scientific article data globally is quite significant. The large growth of this article shows the massive development related to the production of scientific articles in the world. The volume and quantity of scientific articles that continue to increase certainly have two interpretations, the first is about the good productivity of citizens about writing scientific articles, the second is that this can also be a challenge to make truly original scientific papers. The ease of access to data, information and communication is also a driving factor for the emergence of a growing number of scientific articles. This can have implications for plagiarism. The act of plagiarism is a serious problem and can threaten academic and research integrity, which if left unchecked will cause massive corrosion in the

world of education. There are many preventive steps that can be taken for plagiarism in Indonesia, referring to this quite complex problem, technology that on the one hand can give rise to plagiarism because of its ease of access, can also be used on the other hand to be able to read and detect plagiarism early [1]. Concrete steps that can be taken are to check the initial stages of the research to be submitted, where the important part is abstract. Overall, starting from the problem, the solution offered or the method applied to the final result will be described in the abstract, where all important information is contained in the abstract. Detection can be done by making a flow of students having to submit an abstract first to check the level of similarity of the document. Detection of similarities in documents can utilize computational methods, to strengthen the

position of the following research are some references to previous research that apply identical techniques and methods or with similar cases. A paper survey that identifies several approaches that can be done for the detection of document similarity, according to this study, in the detection of document similarity has many approaches and each approach has its own advantages and disadvantages, such as the corpus method is more optimal for cross-language, semantic approach has good results but requires more resources [2], While the graph structure method needs to rely on learning good graph representation to have good performance. Based on this study, the semantic approach is the most optimal result [3-5]. Abstracts have an important role because they contain and represent the objectives, methodologies and contributions of research. Therefore, the abstract part becomes the part where the level of similarity can be seen.

Measuring semantic similarity is one of the things that can be used, but this also still contains problems in natural language processing. The semantic approach so far still relies on lexical representations, such as the TF-IDF model and the bag of words combined with the cosine similarity value [6]. This creates sparse feature space and can create overlapping terms. One of the novelties that can cover this gap is the use of transformer-based models such as BERT with various variants capable of showing strong performance [7-8]. However, from the progress in the use of transformers, there is still a gap with studies that have many stated semantic similarities or groupings as the ultimate goal, not as a mechanism for knowledge discovery. In addition, the characteristics of thesis abstract data are different from journals, so that the method of direct adoption without contextual evaluation becomes less effective [9-10].

In line with these developments, the current research in the measurement of semantic similarities is beginning to shift from a lexical-based approach to context-based representation. Transformer models such as BERT, Sentence-BERT, and their derivatives have become the dominant approach [11] in semantic textual similarity, information retrieval, and document clustering tasks because of their ability to capture contextual meaning across sentences and documents better than traditional models [12]. Evaluation of these models is generally carried out using vector geometry-based similarity metrics, specifically cosine similarity, which is considered compatible with high-dimensional and dense embedding spaces [13].

Nevertheless, most studies still position semantic similarity measurements as the end goal, which are measured through an increase in similarity scores or correlations against annotation labels, without examining how representational changes affect the behavior of the similarity metric itself [14], especially in terms of ranking stability and interpretability of results [15]. This approach has the potential to mask a fundamental shift in the way systems assess the

proximity between documents, particularly as representations move from sparse feature spaces (sparse) to dense contextual embedding spaces [16]. On the other hand, overlap-based similarity metrics such as the Dice coefficient and Jaccard index are still widely used in text similarity measurement and document grouping, especially on TF-IDF-based representations and bag-of-words, due to their simplicity and interpretability [17-19]. This metric is built on the assumption of non-negative vectors and the properties of the set [20-21], so that similarity is seen as a degree [22]. Slice between features [23], however, some recent studies have implicitly begun to apply such overlap metrics to transformer embedding without evaluating the suitability of its mathematical assumptions to the dense and negative characteristics of contextual embedding [24].

This condition indicates a significant research gap, namely the lack of studies that explicitly analyze the behavior of classical similarity metrics (cosine and Dice) when applied to lexical and contextual representations [25] simultaneously [26]. In addition, there is still limited research examining semantic similarities [27] as a mechanism of knowledge discovery [28] such as the structuring of academic abstracts and exploration of research themes, especially in the corpus of thesis abstracts which have different characteristics from journal articles in terms of length, methodological depth, and terminology variations. Based on the state of the art, this study tested the cosine similarity and Dice coefficient in the two paradigms of TF-IDF representation as lexical representation and transformer-based embedding as a contextual representation to understand not only the difference in the similarity value produced [29], but also changes in ranking behavior and stability of similarity assessments [30]. This approach will try to provide an overview of the suitability of the matrix with the representation category used, in addition to analyzing the context of the thesis abstract document as a research contribution. Furthermore, it highlights the implications of representation metric alignment in the analysis of thesis abstracts, particularly in supporting structured exploration, thematic organization, and academic knowledge discovery. Conceptually, semantic similarity measurement cannot be understood only as a score output, but rather as the result of an interaction between the representation space and the metric function used to measure the proximity between documents. Recent literature in text matching and text similarity modeling shows that representational changes (e.g. from lexical to neural embedding) often have consequences for the way proximity is defined and interpreted, so evaluation should not stop at increasing scores alone, but also examining the resulting ranking behavior [31].

In terms of representation, lexical approaches such as TF-IDF construct sparse and non-negative vectors, where the dimension of the feature represents explicit terms in the corpus. In contrast, transformer-based

models map text to fixed-dimensional dense embedding that captures latent semantic relationships and context, so that documents can be considered close even if they don't have large word overlaps. This difference in sparse-dense nature is also widely discussed in modern information retrieval research which highlights the challenges and consequences when systems move from index-based representations (sparse) to vector-based (dense) searches [32]. The similarity metric also has a

different mathematical foundation. Cosine similarity is a geometry-based metric that measures the proximity of a vector's direction and is commonly used in high-dimensional embedding spaces. However, recent studies also warn that cosine in embedding does not always represent similarity intuitively in all conditions, so its interpretation and use require methodological caution shown on Table 1.

Table 1. Research Gap

Aspect	Previous Studies	Identified Limitation	Proposed Contribution
Research Focus [5]	Focus on performance measurement.	Semantic similarity as a score-based goal.	Similarity analysis as a behavioral phenomenon analysis.
Representation Analysis [11]	Focus only on one type of representation	There is no analysis of the metric's behavior representation.	Evaluate changes in the behavior of the similarity of the metrics.
Ranking Behavior [26]	Evaluation based on the value of similarity.	The stability of the rating has not been studied.	Using rank shift analysis and rank correlation
Metric Suitability [22]	Metrics are applied directly.	Not doing an evaluation math assumption.	Analysis of the compatibility of similarity metrics.

2. Methods

The methods that will be discussed in this section include the stages of the research, and some detailed points such as processing and evaluation that will be carried out as part of the scenario of this study.

2.1. Research Methodology

The research methodology and stages carried out in this research are shown in Figure 1.

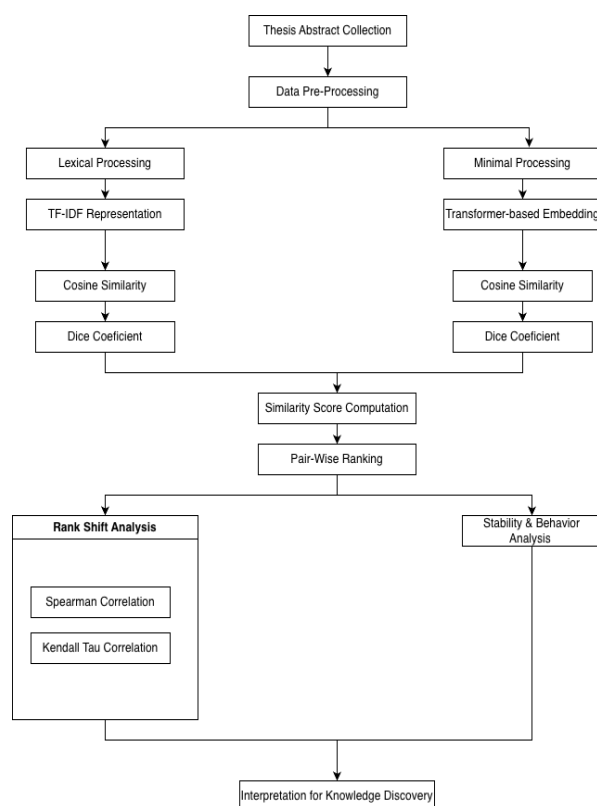


Figure 1. Research Stage

Figure 1 describes the research flow used in this study to analyze the behavior of semantic similarity metrics on various representations of texts. The research

process began with the collection of student thesis abstracts obtained from the institutional repository of universities as the main source of data. After the data collection stage, preprocessing is carried out tailored to the type of text representation used, where lexical preprocessing is applied to the TF-IDF representation, while Word Piece tokenization-based minimal preprocessing is applied to the transformer-based representation.

The preprocessed text is then transformed into two different representation paradigms. The first paradigm uses TF-IDF to generate sparse and non-negative lexical vectors, while the second paradigm uses a transformer model to generate dense contextual embedding through mean pooling and vector normalization. These two types of representations are the basis for the process of calculating semantic similarity. The analysis process was carried out to identify similarities using cosine similarity and dice coefficient similarity. Both similarity metrics apply TF-IDF and also contextual embedding, in this case using BERT. More specifically in the Dice coefficient similarity on the TF-IDF representation will evaluate the feature-slice-based similarity. Application to contextual embedding analysis by assessing the suitability of mathematical assumptions to vector characteristics that have negative values.

The similarity matching process will generate a similarity value and the calculation is performed at the document pair level. Each pair of abstract documents will be compiled and sorted based on their similarity values. This process will form a rank shift that measures changes and rank shifts from document similarity. The stability values of the ratings were evaluated using the correlation coefficients of Spearman and Kendall's Tau ratings. In the final part of the research stage, the analysis of the stability and behavior of similarity metrics is carried out. It is used to obtain information about the impact of different representations of text on an abstract document through semantic similarity.

These findings can be used as evidence to support the process of discovering knowledge from a thesis abstract document. All stages of this research are to seek an understanding of the behavior of similarity metrics and their suitability in the type of text representation used, not only making the similarity score the main benchmark.

The focus and approach of the selected research is quantitative experimental with a focus on the behavioral analysis of semantic similarity metrics. This is different from previous relevant research, where most of them focused on evaluating semantic similarities that make the similarity score the end point of the research findings. The difference in this study lies in the value of semantic similarity placed in the analytical mechanism used as a basis for understanding and evaluating patterns formed from similarities between abstract documents. The analysis process was arranged in a controlled manner by paying attention to the value of each pair of abstract documents to observe the pattern and results of the rank shift that occurred. This is the main contribution of the findings of this study.

2.2. Dataset Description

The data used in this study are abstract documents of specific student thesis at the faculty of computer science which are used as research objects. The total data used was 631 thesis abstract documents. The data collection was taken from the repository of one of the higher education institutions. The retrieved documents are saved in spreadsheet format (XLSX). The choice of this format is based on the reason for the integrity of multi-line text, which often has parsing problems in CSV types. The abstract document is represented as a complete text document without any labeling or categorization process and without any annotation technique.

The mechanism for using abstract document data is formed into 60 pairs derived from 120 selected abstract documents. This will be used as a parameter for the change in the ranking that occurs to be observed and analyzed further as a result of the research findings. The reason for the selection of these 120 abstract documents is a sufficient number and represents the entire document used. The selection was based on several technical stages such as filtering abstract documents based on the completeness and integrity of the text, then also evaluated based on the length of the text and the density of information, as well as based on the results of normalization related to language and character consistency.

2.3. Material and Technical Tools

Information related to supporting materials and technical devices in this study is important to be conveyed, so that the configuration of this supporting environment can be observed by other researchers who will conduct similar research. The computing environment used is cloud-based and local. The process carried out is the implementation, evaluation, and

analysis of semantic similarity models. The type of configuration used is also detailed for the framework process. The details of the materials and technical devices used are as the platform used as the main tool in this study to run the experiment is Google Collaboratory. The details of the use of this platform are used to execute Python code in test scenarios that have been applied to the research. This makes it easier because using a Google Collaboratory environment can use GPU resources without using a dedicated local GPU.

Operating Environment with the experiment was conducted using the Python runtime version 3.10 in a Linux-based virtual machine environment provided by Google Collaboratory. Libraries and Frameworks are using to research is several open-source libraries and frameworks are used to support each stage of the experiment, namely: Transformers (Hugging Face, version $\geq 4.40.0$) for the implementation of transformer-based embedding models. Scikit-learn (version $\geq 1.3.0$) for TF-IDF vectorization, cosine similarity calculation, Dice coefficient, and rating-based evaluation metrics. NumPy and Pandas for numerical computing and data preprocessing. Matplotlib for visualization of similarity distributions, rank shifts, and correlation analysis. NLTK and Sastrawi for Indonesian text preprocessing, including tokenization, stopword removal, and stemming on lexical representations

2.4. Text Preprocessing

The stages of preprocessing are adjusted to the type of text representation used lexical representation (TF-IDF). Abstract text is processed through the stages of case folding, removal of non-alphabetic characters, and tokenization. TF-IDF representations are built using unigrams and bigrams with maximum feature limits to maintain efficiency and reduce extreme sparsity. Contextual representation (Transformer-based embedding).

Transformer-based models, preprocessing is minimal, following the Word Piece tokenizer standard. The text is cropped to a certain maximum length to maintain input consistency, without stemming or removing stop words, as contextual information is captured implicitly by the model.

2.5. Semantic Similarity Modeling and Analysis

To evaluate semantic similarity behavior, this study tested two commonly used similarity metrics cosine similarity is used to measure angular similarity between vectors and is applied to TF-IDF representations as well as contextual embedding. This metric was chosen because of its non-vector magnitude-dependent nature and is compatible with high-dimensional embedding spaces. Dice coefficient is used as an overlap-based metric and is applied to the TF-IDF representation to evaluate the feature-slice-based similarity. The application of Dice to contextual embedding was analyzed exploratory to assess the suitability of metric

mathematical assumptions against dense representations containing negative values.

Documents in subsets are arranged into pairs sequentially, so that each pair consists of two abstracts that stand independently. This pair-based approach allows direct observation of differences in similarity values and ranking changes between metrics and representations, without being influenced by the global distribution of the entire corpus.

To assess the change in similarity behavior, this study conducted a rank shift analysis with the following steps. The first is calculate the similarity value for each document pair using each combination of representations and metrics, sort document pairs based on similarity values in descending, calculate the absolute ranking difference between the lexical and contextual representations for each pair, perform an analysis of the rank shift distribution. A detailed approach with descriptive statistics. The resulting rank stability results will be analyzed using Spearman's rank correlation and Kendall's Tau. This result measures the consistency level of each document's rating.

2.6. Evaluation Focus

The evaluation in this study does not aim to optimize accuracy or achieve the best performance, but to focusing on identifying and analyzing the results of behavioral changes. The similarity metric refers to the use of different text representations. Assessment of the results of the stability of the resulting rating by interpreting the results of the similarity produced. Evaluate the suitability of the assumptions of the resulting similarities to the type of representation used in mathematical assumptions. The results of the evaluation produced by this study are expected to be able to provide important information about the in-depth research methodological flow in the selection and suitability of the type of text representation that is suitable for each technique. Overall, the methodology of this study combines lexical and contextual representations with classical similarity metrics to analyze changes in semantic similarity assessment behavior. By focusing on the analysis of rating and metric stability, this study provides an alternative perspective on the use of transformer models in the measurement of semantic similarity, especially in the abstract corpus of thesis. The overall summarize of the methodology of this study is a combination of lexical and contextual representations with a similarity metric approach. The focus of this research in general is the analysis of the results produced by rank shift. Thus, it is hoped that this study will be able to provide a perspective on the suitability of text representation with the right techniques and models in the thesis abstract corpus document.

3. Results and Discussions

The results and discussion section will discuss the predetermined test scenario, starting from information

about the availability of data, the standard experiment used. Testing is also presented in this section to reinforce the results of the discussion of the understanding and analysis of the results.

3.1. Dataset Overview and Experimental Validation

The abstract corpus data of 631 abstract documents was used, a controlled subset process was carried out using 120 abstract documents divided into 60 document pairs. The selection of controlled subsets aims to select document pairs that can describe the behavior of the rating change with the design of the scenario performed. The description of the dataset and experiments carried out is shown in Figure 2.

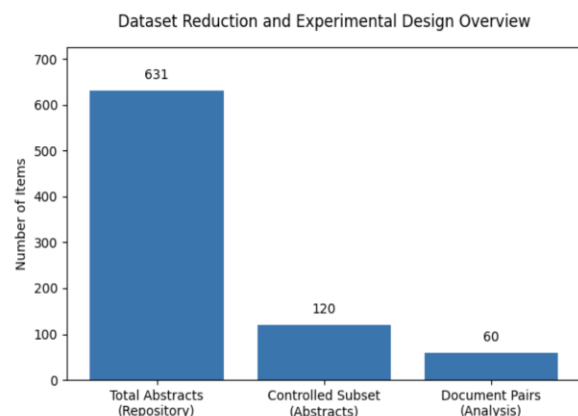


Figure 2. Experimental Design

The results of this dataset configuration will be used to describe the distribution of the resulting score for each corpus document and the analysis of the resulting values.

3.2. Similarity Score Distribution Across Representations

The use of two representations, namely lexical and contextual, shows the results of differences in the measurement of similarity values. The results of lexical representation with the TF-IDF combination using cosine similarity had a fairly low result, as well as the dice similarity result. This result is shown in Figure 3.

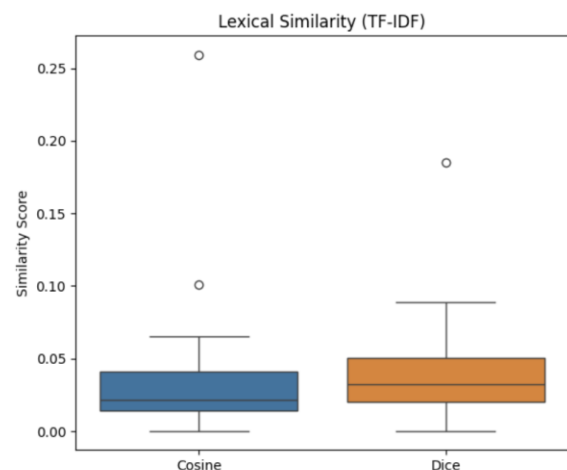


Figure 3. Distribution Scores Lexical

In contextual representations using transformers, the result of cosine similarity gets a high value. This indicates that a contextual approach with transformers or embedding is able to detect semantic proximity better. On the other hand, the application of dice coefficient similarity shows a much different value. The resulting values have a much wider range. This will be discussed on a separate point. This distribution is visualized on Figure 4.

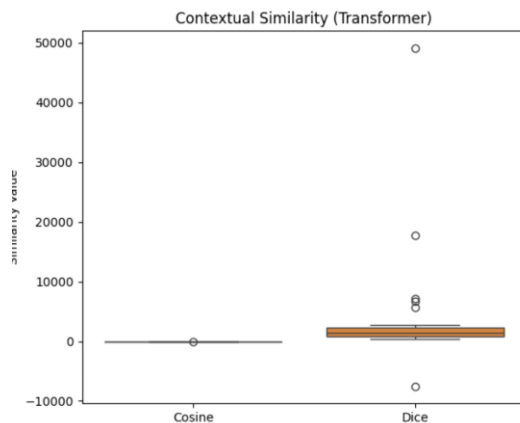


Figure 4. Distribution Scores Contextual

3.3. Descriptive Statistics of Similarity Scores

These results can be described and grouped into 3 parts. Observations from the results of using TF-IDF representations with cosine and dice show that the results of the similarity value are quite low and have a

narrow range of values. This is relevant to the sparse feature space adopted by the TDF-IDF.

The second result is that embedding transformers with cosine similarity results in relatively stable data variation values. The third point of the combination of transformer with dice shows extreme results. These results have a large data range and contain negative values. This will be discussed separately at the point of failure dice with transformers.

Table 2 contains statistical testing information of similarity values. The information about the results of this statistical test is to provide a basic and preliminary overview of the scale of the data, the variation of the data and its direction to the stability of the similarity score of the abstract document.

These results can be described and grouped into 3 parts. Observations from the results of using TF-IDF representations with cosine and dice show that the results of the similarity value are quite low and have a narrow range of values. This is relevant to the sparse feature space adopted by the TDF-IDF.

The second result is that embedding transformers with cosine similarity results in relatively stable data variation values. The third point of the combination of transformer with dice shows extreme results. These results have a large data range and contain negative values. This will be discussed separately at the point of failure dice with transformers.

Table 2. Descriptive Statistics

Representation	Similarity Metric	Mean	Std Dev	Min	Max
TF-IDF	Cosine	0.038	0.054	0.000	0.259
TF-IDF	Dice	0.041	0.039	0.000	0.185
Transformer	Cosine	0.613	0.080	0.483	0.790
Transformer	Dice	4356.600	10686.813	-7581/316	49165.117

3.4. Rank Shift Analysis Between Lexical and Contextual Similarity

To assess whether the difference between lexical and contextual representations is simply a shift in the value scale or reflects a more fundamental behavioral change, a rank shift analysis was performed between TF-IDF-based cosine similarity and transformer embedding-based cosine similarity. The results of the analysis showed an average rating shift of 6.70 positions with a median of 5.00 positions. The maximum shift reached 16 positions, which indicates a significant change in the order of similarity in a number of document pairs. The document pairs with the largest rank shift values are presented in Table 3, which shows a concrete example of how changes in text representation can result in a substantial rearrangement of similarity rankings.

Document on Table 3 shows the results of the movement of document ranking. These results can be observed and show that a significant change in ranking occurs between the two approaches. This is natural given the different ways the two work, the transformer-

based approach understanding the contextual and the approach with semantic representation having a word frequency-based approach.

Table 2. Rank Shift Result

Document	Pair Id	TF-IDF Rank	BERT Rank	Rank Shift
10	P011	6	22	16
14	P015	4	19	15
5	P006	16	3	13
20	P021	22	9	13

3.5. Rank Correlation and Stability Analysis

Evaluation of the ranking was carried out by applying the Spearman and Kendall correlation coefficients. The results of this correlation coefficient show that Spearman's rho value is 0.22 and Kendall's tau is 0.146. Reflecting on these results, the indication that emerges is that the correlation of ratings produced by the lexical and contextual approaches shows weak values. This can be represented as a fairly significant difference between these two approaches.

The results obtained confirm the previous findings that the process mechanisms carried out by these two approaches are quite significant, so these results can confirm that these differences in mechanisms have real implications for the results obtained. Especially for the embedding transformer approach that assesses similarities not only based on the occurrence of sentence frequencies, but also more deeply through deepening and contextual meaning. These results are corroborated in Figure 5.

Figure 5 visualize the relationship between TF-IDF-based and transformer-based similarity ratings. The wide spread of the points away from the diagonal indicates a weak ranking agreement between the two approaches. This visual pattern is consistent with the low rho Spearman (0.22) and tau Kendall (0.146), which confirm substantial ranking instability rather than small score fluctuations.

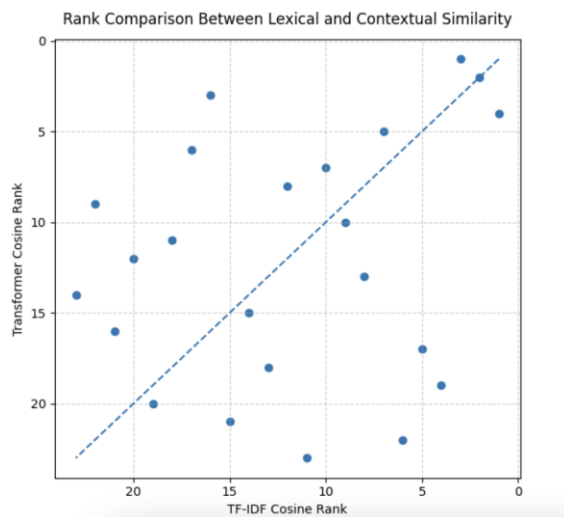


Figure 5. Rank Comparison

Other test results used Normalized Discounted Cumulative Gain (NDCG) which focuses on measuring the quality of document sequences based on their relative relevance. The results of the evaluation using NDCG resulted in a value of 0.9715. This represents that the ranking with TF-IDF is still consistent and relevant to the similarities based on contextual embedding. This result is not in conflict with the rank shift and the correlation of the rank obtained with weak results. The findings confirm that the use of transformer embedding results in structural changes in the similarity assessment, not just changes in the scale of values.

3.6. Qualitative Interpretation of Rank Shift Behavior

Qualitative analysis based on the results obtained from the rank shift that occurred can be interpreted as the shift that occurred still shows a consistent pattern of shift. The pattern of shifts that occur in these two representations has similarities, namely in the ranking results that are shown are rankings that are far from each other between types of representations.

The results reflected from the test, for the generated ranking semantic representation results in a low rank

value, the transformer representation results in a high rank value. The consistency of these results is shown by the two representations used and tested in this study. Thus, it can be concluded that from these results, semantic and lexical similarities are not always in line due to the differences in the way in which these two types of representation work.

3.7. Failure of Dice Similarity on Contextual Embeddings

An important methodological finding in this study is the failure of Dice similarity when applied to contextual embedding. In contrast to its application to TF-IDF representations that produce stable and interpretable similarity values, the use of Dice coefficients in transformer-based embedding results in extreme and inconsistent values, including very large values as well as negative values. This phenomenon occurs because the Dice coefficient is built on the assumption of non-negative vectors and the overlapping properties of features, where similarity is seen as the degree of wedge between vector components. These assumptions are not fulfilled in contextual embedding that is dense and contains negative values, so Dice's calculations no longer reflect the concept of semantically meaningful similarity. This failure is not caused by an implementation error, but rather is a direct consequence of the mismatch of the metric's mathematical assumptions to the characteristics of the representation used.

An important point that can be summarized is the failure of the application of dice similarity, so that the application of dice similarity directly to contextual embedding results in extreme and leads to poor interpretation of similarity. The reason for dice similarity, which theoretically applies sets with vectors, and has different mathematical assumptions, is still tested in this study as an exploratory approach, where many recent studies have emphasized the importance of the suitability of text representations and similarity metrics to be used. Thus, the exploratory approach with dice coefficient similarity will remain relevant in this study.

3.8. Summary of Findings

The conclusion from a series of test scenarios and analyses is that the implementation of transformer-based embedding differs significantly from the semantic-based approach. This difference is reflected in the processing mechanisms employed by these two representations, as evidenced by the results obtained. The resulting gap reflects the contrasting processes of these two representations. Another interpretation is that the failure of Dice Similarity when paired with transformer embedding is due to the different mathematical assumptions underlying overlap-based matrices and contextual embeddings, which are dense and contain negative values. This conclusion emphasizes that the selection of a similarity matrix must be appropriate and not arbitrary, depending on the

type of text representation being used, as reflected in the similarity values generated in this research experiment.

Methodological findings indicate that the choice of mechanism and type of text representation should be tailored to the characteristics and analytical objectives of the text. As in academic text analysis, the wide variety of equivalent words can distort the meaning or content of the research if the assessment is based on the frequency of word occurrences. If it is based on the contextual sentence, then this use will be more relevant using an embedding transformer because an abstract document will be interpreted as a complete unit, not a document that can be confused with the use of various equivalent words.

4. Conclusions

The research aims to uncover the behavior of semantic similarity with transformer-embedding-based representations. The result is that these two types of representations produce opposite values due to the difference in the details of the mechanism. This is shown from the results of the rankings produced. In addition, the test results also revealed that the application of Dice coefficient similarity is also not suitable to be combined with the transformer embedding representation due to the mismatch of the mathematical assumptions of the two. It can be concluded that the selection of similarity metrics must pay attention to the suitability of mathematical assumptions to be able to produce an optimal similarity score. This selection must also be based on the characteristics of the data to be processed, if looking at the character of the text data in an academic document, the selection of transformer-based embedding-based contextual representation will be more suitable with contextual understanding to avoid the use of word variations in avoiding plagiarism detection if applying a semantic-based representation approach. Suggestions for the development of advanced research, can test the impact of behavior change on both approaches with practical implementation.

Acknowledgements

The authors declare no conflict of interest. This research was funded by the Institute for Research and Community Service (LPPM), Universitas Pembangunan Nasional Veteran Jakarta, through the Scopus-Based Research (RISCOP) scheme in 2025.

References

- [1] M. Botto-Tobar, A. Serebrenik, and M. G. J. Van den Brand, "Cross-Language Plagiarism Detection: Methods, Tools, and Challenges: A Systematic Review," *Int J Adv Sci Eng Inf Technol*, vol. 12, no. 2, pp. 589–599, Apr. 2022, doi: <https://doi.org/10.18517/ijaseit.12.2.14711>.
- [2] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "Modeling evidential grids using semantic context information for dynamic scene perception," *Knowl Based Syst*, vol. 215, p. 106777, Mar. 2021, doi: <https://doi.org/10.1016/j.knosys.2021.106777>.
- [3] B. Klimova, M. Pikhart, A. D. Benites, C. Lehr, and C. Sanchez-Stockhammer, "Neural machine translation in foreign language teaching and learning: a systematic review," *Educ Inf Technol (Dordr)*, vol. 28, no. 1, pp. 663–682, Jan. 2023, doi: <https://doi.org/10.1007/s10639-022-11194-2>.
- [4] X. Huang, C. Wang, X. Tian, Z. Li, C. Zhao, and M. Xiao, "A Multi-Hop Semantic Communication System Enhanced by Semantic Importance," *IEEE Access*, vol. 13, pp. 140685–140693, 2025, doi: <https://doi.org/10.1109/ACCESS.2025.3593341>.
- [5] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity A Survey," *ACM Comput Surv*, vol. 54, no. 2, pp. 1–37, Mar. 2022, doi: <https://doi.org/10.1145/3440755>.
- [6] P. Yang, H. Wang, J. Yang, Z. Qian, Y. Zhang, and X. Lin, "Deep Learning Approaches for Similarity Computation: A Survey," *IEEE Trans Knowl Data Eng*, vol. 36, no. 12, pp. 7893–7912, Dec. 2024, doi: <https://doi.org/10.1109/TKDE.2024.3422484>.
- [7] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsaifi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif Intell Rev*, vol. 58, no. 6, p. 166, Mar. 2025, doi: <https://doi.org/10.1007/s10462-025-11162-5>.
- [8] M. Galih Pradana, N. Irzavika, N. Maulana, J. Mu, and K. Wari, "Performance Improvement of Cosine Similarity Algorithm with Bidirectional Encoder Representations from Transformers on Abstract Document Similarity Detection," *JOIV: International Journal on Informatics Visualization*, vol. 9, no. 2, pp. 824–830, Mar. 2025, doi: <http://dx.doi.org/10.62527/joiv.9.2.2853>.
- [9] S. Zoupanos, S. Kolovos, A. Kanavos, O. Papadimitriou, and M. Maragoudakis, "Efficient comparison of sentence embeddings," in *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, New York, NY, USA: ACM, Sep. 2022, pp. 1–6, doi: <https://doi.org/10.1145/3549737.354975>.
- [10] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 6894–6910, doi: <https://doi.org/10.18653/v1/2021.emnlp-main.552>.
- [11] K. Abdalgader, A. A. Matroud, and K. Hossin, "Experimental study on short-text clustering using transformer-based semantic similarity measure," *PeerJ Comput Sci*, vol. 10, p. e2078, May 2024, doi: <https://doi.org/10.7717/peerj-cs.2078>.
- [12] H. Licht, "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings," *Political Analysis*, vol. 31, no. 3, pp. 366–379, Jul. 2023, doi: <https://doi.org/10.1017/pan.2022.29>.
- [13] F. F. Ijebu, Y. Liu, C. Sun, and P. U. Usip, "Soft cosine and extended cosine adaptation for pre-trained language model semantic vector analysis," *Appl Soft Comput*, vol. 169, p. 112551, Jan. 2025, doi: <https://doi.org/10.1016/j.asoc.2024.112551>.
- [14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990, doi: <https://doi.org/10.18653/v1/D19-1410>.
- [15] K. Nassiri and M. Akhloufi, "Transformer models used for text-based question answering systems," *Applied Intelligence*, vol. 53, no. 9, pp. 10602–10635, May 2023, doi: <https://doi.org/10.1007/s10489-022-04052-8>.
- [16] P. Jiang and X. Cai, "A Survey of Text-Matching Techniques," *Information*, vol. 15, no. 6, p. 332, Jun. 2024, doi: <https://doi.org/10.3390/info15060332>.
- [17] L. Zhang, "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification," *PLoS One*, vol. 20, no. 7, p. e0327347, Jul. 2025, doi: <https://doi.org/10.1371/journal.pone.0327347>.
- [18] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU Journal of Science and*

- Technology, vol. 4, no. 1 & 2, pp. 27–33, Aug. 2022, doi: <https://doi.org/10.56471/slujst.v4i.266>.
- [19] S. Akuma, T. Lubem, and I. T. Adom, “Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets,” *International Journal of Information Technology*, vol. 14, no. 7, pp. 3629–3635, Dec. 2022, doi: <https://doi.org/10.1007/s41870-022-01096-4>.
- [20] F. Lan, “Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method,” *Advances in Multimedia*, vol. 2022, pp. 1–11, Apr. 2022, doi: <https://doi.org/10.1155/2022/7923262>.
- [21] D. Meidelfi, - Yulherniwati, I. Rahmayuni, T. Hidayat, and D. Chandra, “TF-IDF Implementation for Similarity Checker on The Final Project Title,” *International Journal of Advanced Science Computing and Engineering*, vol. 3, no. 1, pp. 40–52, 2021, doi: <https://doi.org/10.30630/ijasce.3.1.3>.
- [22] V. Raina et al., “Tackling Bias in the Dice Similarity Coefficient: Introducing NDSC for White Matter Lesion Segmentation,” in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Apr. 2023, pp. 1–5. doi: <https://doi.org/10.1109/ISBI53787.2023.10230755>.
- [23] U. ur Rehman and T. Mahmood, “The generalized dice similarity measures for bipolar complex fuzzy set and its applications to pattern recognition and medical diagnosis,” *Computational and Applied Mathematics*, vol. 41, no. 6, p. 265, Sep. 2022, doi: <https://doi.org/10.1007/s40314-022-01948-6>.
- [24] A. Levy, B. R. Shalom, and M. Chalamish, “A guide to similarity measures and their data science applications,” *J Big Data*, vol. 12, no. 1, p. 188, Jul. 2025, doi: <https://doi.org/10.1186/s40537-025-01227-1>.
- [25] B. Pavlyshenko and M. Stasiuk, “SEMANTIC SIMILARITY ANALYSIS USING TRANSFORMER-BASED SENTENCE EMBEDDINGS,” *Electronics and Information Technologies*, vol. 30, p. 803, 2025, doi: <https://doi.org/10.30970/eli.30.4>.
- [26] D. Wang and S. Zhang, “Large language models in medical and healthcare fields: applications, advances, and challenges,” *Artif Intell Rev*, vol. 57, no. 11, p. 299, Sep. 2024, doi: <https://doi.org/10.1007/s10462-024-10921-0>.
- [27] S. Agarwala, A. Anagawadi, and R. M. Reddy Guddeti, “Detecting Semantic Similarity of Documents Using Natural Language Processing,” *Procedia CIRP*, vol. 189, pp. 128–135, 2021, doi: <https://doi.org/10.1016/j.procs.2021.05.076>.
- [28] B. Ihnaini, B. Abuhaija, E. A. Mills, and M. Mahmuddin, “Semantic similarity on multimodal data: A comprehensive survey with applications,” *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, p. 102263, Dec. 2024, doi: <https://doi.org/10.1016/j.jksuci.2024.102263>.
- [29] J. Guan, B. Chen, and S. Yu, “A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data,” *Expert Syst Appl*, vol. 249, p. 123700, Sep. 2024, doi: <https://doi.org/10.1016/j.eswa.2024.123700>.
- [30] S. W. Brown, J. Bonn, G. Kazeminejad, A. Zaenen, J. Pustejovsky, and M. Palmer, “Semantic Representations for NLP Using VerbNet and the Generative Lexicon,” *Front Artif Intell*, vol. 5, Apr. 2022, doi: <https://doi.org/10.3389/frai.2022.821697>.
- [31] A. M. Elkhataf, “Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities,” *International Journal for Educational Integrity*, vol. 19, no. 1, p. 15, Aug. 2023, doi: <https://doi.org/10.1007/s40979-023-00137-0>.
- [32] W. Mu and K. H. Lim, “Modelling Text Similarity: A Survey,” in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, New York, NY, USA: ACM, Nov. 2023, pp. 698–705. doi: <https://doi.org/10.1145/3625007.3627305>.