International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-6

# A Comparative Investigation of Text Similarity Techniques for Human-Written and AI-Generated Scientific

**S.Venkateswara Rao[1], G. Varalaxmi[2]**

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

***Abstract* – Significantly transformed the way scientific documents are generated, summarized, and paraphrased. Modern AI systems are capable of producing highly fluent and contextually meaningful scientific abstracts that often resemble human-written content,Natural Language Processing (NLP), plagiarism detection, academic integrity, and automated content verification. text similarity methods for evaluating A large dataset consisting of approximately 36,500 scientific abstracts is categorized into three groups to facilitate detailed similarity analysis. The proposed evaluation framework incorporates multiple categories of similarity techniques, including lexical similarity methods, character-based algorithms, statistical text representation approaches, semantic embedding models, and transformer-based language models. Traditional , Dice Coefficient, Smith-Waterman Algorithm, Levenshtein Distance, and TF-IDF are evaluated alongside modern embedding techniques including Word2Vec, FastText, GloVe, and BERT. To measure the effectiveness of these approaches, statistical indicators Experimental analysis demonstrates that contextual embedding models consistently achieve superior semantic similarity performance compared with conventional lexical methods. The findings also indicate that AI-generated abstracts preserve semantic meaning more effectively than lexical structure, highlighting the importance of contextual language models in scientific text evaluation. The proposed comparative analysis provides valuable insights for plagiarism detection systems, AI content verification, scientific publishing, and future NLP applications involving automated text similarity assessment.**

*Keywords* - **Text Similarity, AI-Generated Text, Human-Written Text, Natural Language Processing (NLP), Scientific Text Analysis, Machine Learning**

## I. INTRODUCTION

Unlike simple keyword matching, modern text similarity aims to identify lexical, syntactic, semantic, and contextual relationships that exist between different documents. Accurate similarity measurement has become an essential requirement for numerous intelligent applications, including plagiarism detection, document retrieval, duplicate document identification, semantic search, automatic summarization, machine translation, recommendation systems, and question-answering platforms. As the volume of digital textual information continues to grow rapidly, reliable text similarity techniques have become increasingly important for efficiently organizing and analyzing textual data.

Despite significant progress in NLP, measuring text similarity remains a challenging task because natural language is highly complex and context-dependent. Two documents may express identical ideas using completely different vocabulary, sentence structures, or writing styles, while documents sharing many common words may actually convey different meanings. Factors such as synonym usage, paraphrasing, contextual dependencies, ambiguity, grammatical variations, and domain-specific terminology further complicate similarity measurement. Consequently, traditional lexical comparison methods are often insufficient for accurately evaluating semantic relationships among scientific documents.

including ChatGPT and similar AI systems, has revolutionized automatic text generation. These models are capable of generating high-quality scientific abstracts, summaries, technical reports, translations, and research articles that closely resemble human writing. Their impressive language generation capabilities have significantly improved productivity across academic and industrial domains. However, the increasing availability of AI-generated content has introduced new concerns regarding scientific originality, plagiarism detection, research authenticity, copyright protection, and academic ethics. As AI-generated scientific documents become more sophisticated, distinguishing them from genuine human-written documents becomes increasingly difficult.

To address these challenges, researchers have proposed numerous text similarity techniques based on lexical matching, statistical representations, semantic embeddings, and deep learning. Traditional algorithms Longest Common Subsequence, and TF-IDF evaluate similarity using surface-level textual characteristics. More recent approaches including Word2Vec, FastText, GloVe, DSSM, and BERT generate contextual vector representations capable of capturing deeper semantic meaning and relationships between documents. These contextual models significantly improve similarity estimation by understanding word meaning within different linguistic contexts rather than relying solely on exact word overlap.

this work contribute toward improving AI-generated content evaluation while providing valuable guidance for future research in plagiarism detection, semantic similarity analysis, and trustworthy AI-based scientific writing.

## II. LITERATURE SURVEY

Text similarity has been extensively studied over the past several decades because of its importance in Numerous researchers have proposed lexical, syntactic, semantic, statistical, and deep learning approaches to estimate textual similarity across different application domains. Early studies primarily focused on measuring lexical similarity through exact word matching and edit-distance algorithms. Although these methods achieved reasonable performance for simple textual comparisons, they were unable to accurately capture semantic relationships between documents that conveyed similar meanings using different vocabulary.

Character-based similarity algorithms such as Levenshtein Distance, Smith-Waterman Algorithm, Ratcliff–Obershelp Algorithm, Longest Common Subsequence (LCS), and Jaro-Winkler Similarity became popular because they effectively identify textual modifications and structural differences. These approaches evaluate similarity by measuring insertion, deletion, substitution, and matching operations between character sequences. While computationally efficient, these algorithms primarily focus on surface-level textual similarity and generally fail to recognize conceptual relationships between semantically equivalent sentences.

To overcome these limitations, statistical document representation methods such as Bag-of-Words (BoW), Topic Modeling techniques were introduced. These methods convert textual documents into numerical feature vectors that allow similarity computation using mathematical distance measures. Although statistical representations improve semantic understanding compared to character-based algorithms, they still struggle to model contextual relationships and long-range dependencies within scientific documents.

These methods outperform traditional statistical models because semantically related words are represented by nearby vectors within high-dimensional embedding spaces. However, static embedding models still assign identical vector representations to words regardless of context, limiting their effectiveness for polysemous words.

embedding methods, BERT analyzes both preceding and succeeding words simultaneously, enabling comprehensive contextual understanding of language. Several recent studies have demonstrated that BERT consistently achieves superior performance in semantic similarity, paraphrase detection, document matching, and AI-generated content evaluation. Researchers have also investigated using ROUGE, BLEU, Word2Vec, FastText, and BERT, concluding that contextual embedding models provide significantly better semantic discrimination than conventional lexical similarity algorithms.

Overall, the literature indicates that traditional lexical methods remain useful for detecting exact textual overlap, whereas contextual embedding models are considerably more effective for evaluating semantic and transformer-based techniques therefore provides a comprehensive framework for evaluating modern AI-generated scientific content.

## III. PROPOSED METHODOLOGY

The proposed system introduces a comprehensive framework for evaluating textual similarity amongmetric, the framework integrates lexical, syntactic, statistical, semantic, and transformer-based approaches to provide a more reliable and comprehensive assessment of document similarity. The overall objective is to identify which similarity techniques most accurately capture semantic relationships while effectively distinguishing AI-generated scientific content from genuine human-written abstracts.

The framework begins with dataset preparation, where approximately 36,500 scientific abstracts are collected from biomedical literature. These abstracts are organized into three separate categories: are retrieved from PubMed, AI-paraphrased abstracts are generated by rewriting the original content using GPT, and AI-generated abstracts are created by providing extracted keywords to the language model. This structured dataset enables detailed pairwise similarity analysis across multiple document categories.

Before similarity computation, the collected documents undergo preprocessing to improve consistency and reduce irrelevant textual variations. Standard preprocessing operations include text normalization, lowercase conversion, punctuation removal, tokenization, stop-word elimination, and optional stemming or lemmatization. These preprocessing steps ensure that similarity algorithms focus primarily on meaningful semantic information rather than formatting differences.

The processed abstracts are subsequently analyzed using several categories of similarity methods. Lexical similarity is evaluated through Cosine Similarity, Jaccard Similarity, Dice Coefficient, and TF-IDF representations. Character-level similarity is measured using Levenshtein Distance, Smith-Waterman Algorithm, Longest Common Subsequence, Ratcliff–Obershelp, and Jaro-Winkler algorithms. Statistical representation employed to capture document-level relationships, while semantic similarity is evaluated using Word2Vec, FastText, GloVe, DSSM, and transformer-based BERT embeddings. Each method analyzes textual similarity from a different perspective, providing complementary information regarding lexical overlap, semantic consistency, contextual understanding, and structural similarity.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-6

The similarity scores generated by each algorithm are statistically analyzed using mean, median, and standard deviation to evaluate performance consistency across different datasets. In addition, comprehensive evaluation is performed metrics to measure lexical correspondence, semantic preservation, contextual fidelity, and paraphrasing quality. Comparative analysis of these evaluation results enables identification of the most reliable similarity techniques for AI-generated scientific text assessment.

Experimental observations demonstrate that contextual embedding models consistently outperform conventional lexical algorithms when evaluating semantic similarity between scientific abstracts. Word2Vec, FastText, GloVe, and BERT achieve the highest similarity accuracy because they effectively capture The proposed framework therefore provides a reliable methodology for evaluating AI-generated scientific documents while supporting plagiarism detection, scientific integrity verification, academic publishing, and future research involving trustworthy Artificial Intelligence systems.

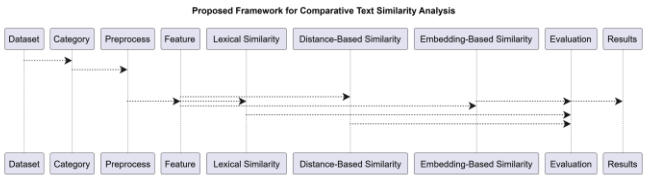


Fig. 1. Proposed Framework for Comparative Text Similarity Analysis of Human-Written

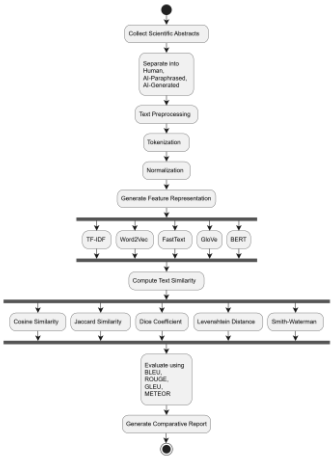


Fig. 2. Workflow of Scientific Abstract Evaluation

## IV EXPERIMENTAL ANALYSIS

performance of multiple text similarity techniques across three categories of scientific abstracts: human-written, AI-paraphrased, and AI-generated documents. Rather than relying on a single similarity algorithm, the study performs a comprehensive comparison of lexical, statistical, semantic, and contextual similarity methods to determine their capability in measuring textual resemblance under different writing scenarios. The experiments are conducted using a large collection of approximately 36,500 scientific abstracts, enabling statistically reliable observations and meaningful performance comparisons.

The evaluation framework investigates several categories of similarity methods, including string comparison algorithms, edit-distance techniques, vector space models, and embedding-based semantic approaches. Lexical similarity is computed using algorithms such as Cosine Similarity, Jaccard Similarity, Dice Coefficient, Levenshtein Distance, Smith–Waterman, Ratcliff–Obershelp, and Longest Common Subsequence (LCS). Statistical document representations are generated through TF-IDF, while semantic representations are obtained using Word2Vec, FastText, GloVe, and BERT embeddings. Each method evaluates textual similarity from a different perspective, allowing a comprehensive assessment of both lexical overlap and semantic consistency among scientific abstracts.

To ensure objective comparison, the similarity scores produced by individual algorithms are analyzed using These statistical indicators provide valuable insights into the consistency, stability, and variability of each similarity technique across different document categories. Furthermore, the generated similarity scores are evaluated using established

Experimental observations indicate that traditional lexical similarity methods perform effectively when comparing documents containing substantial word overlap but exhibit reduced performance when evaluating paraphrased or AI-generated content that preserves meaning while altering vocabulary and sentence structure. Conversely, semantic embedding techniques consistently achieve higher similarity scores because they capture contextual relationships rather than relying solely on exact word matching. In particular, Word2Vec, FastText, and BERT demonstrate superior capability in identifying semantic equivalence highlighting their effectiveness for modern NLP applications involving contextual language understanding.

The comparative analysis further reveals that transformer-based language models provide more reliable semantic representations than conventional statistical approaches. BERT, owing to its bidirectional contextual encoding mechanism, effectively models complex linguistic relationships and demonstrates improved robustness when analyzing AI-generated scientific documents. Similarly, FastText benefits from subword representations that improve semantic understanding for rare and domain-specific scientific terminology, whereas Word2Vec offers efficient distributed representations suitable for large-scale similarity analysis.

Overall, the experimental evaluation confirms that embedding-based similarity methods significantly

outperform conventional lexical algorithms in measuring semantic similarity among scientific abstracts. The findings demonstrate that contextual language models provide more accurate evaluation of AI-generated content while supporting applications such as plagiarism detection, scientific document verification, academic integrity assessment, semantic search, and intelligent document retrieval. These results further emphasize the growing importance of transformer-based Natural Language Processing techniques for future scientific text similarity research.

## V. DISCUSSION

The comparative investigation demonstrates that no single similarity technique is universally optimal for every text comparison scenario. Traditional lexical algorithms remain effective for identifying direct textual overlap and minor textual modifications. However, these methods become less reliable when documents are paraphrased or generated using advanced language models because lexical variations often conceal underlying semantic similarity. Consequently, relying exclusively on lexical comparison may lead to inaccurate similarity estimation for modern AI-generated scientific documents.

Statistical document representation techniques, including TF-IDF, provide improved performance by considering the relative importance of words within documents. Nevertheless, these approaches continue to treat documents primarily as collections of independent terms and therefore fail to capture deeper contextual relationships among words and sentences. Their effectiveness decreases when semantically equivalent information is expressed using different vocabulary or grammatical structures.

Embedding-based approaches significantly improve semantic similarity estimation by representing textual information within continuous high-dimensional vector spaces. Word2Vec, FastText, and GloVe successfully preserve semantic relationships among words, while BERT further enhances similarity analysis through contextual bidirectional language modeling.because they recognize semantic equivalence even when lexical overlap is relatively low.

The study also demonstrates algorithm performance than relying on a single similarity measure. Statistical indicators together with ROUGE, BLEU, GLEU, and METEOR collectively evaluate lexical correspondence, semantic preservation, contextual consistency, and overall language quality. Such comprehensive evaluation enables more objective assessment of similarity algorithms across different document categories.

Overall, the experimental findings emphasize the increasing importance of semantic and contextual similarity techniques for evaluating AI-generated scientific content. As Large Language Models continue to improve, contextual embedding methods plagiarism detection, scientific integrity verification, document authentication, semantic retrieval, and intelligent academic publishing systems.

## VI. CONCLUSION

Initially, the personalized medicine recommendation problem is formally introduced, highlighting how patient-specific clinical information can be utilized to generate appropriate medication recommendations under different healthcare scenarios.

The survey systematically reviews existing approaches by categorizing them into four major groups: medicine recommendation for multi-disease patients, recommendation based on medicine combination patterns, recommendation enhanced with additional medical knowledge, and recommendation utilizing patient feedback. Each category is analyzed with respect to its underlying methodology, advantages, and existing limitations, providing a structured understanding of current research developments.

Furthermore, commonly adopted evaluation strategies are discussed by emphasizing two essential performance aspects: recommendation effectiveness and medication safety. Rate, are summarized to illustrate how personalized medicine recommendation models are assessed in practical healthcare applications.

Although significant progress has been achieved, several research challenges remain open, including improving medication safety, enhancing model interpretability, supporting dynamic and fine-grained treatment recommendations, and effectively integrating lifelong healthcare records into intelligent recommendatio reliable, transparent, and patient-centered personalized medicine recommendation frameworks capable of improving healthcare quality and clinical outcomes.

**Future Enhancement**

Although existing text similarity techniques have achieved considerable success in evaluating lexical and semantic relationships among scientific documents, several opportunities remain for improving their effectiveness in the era of content while preserving semantic consistency. The continuous advancement of Large Language Models (LLMs) will require similarity evaluation methods that are more context-aware, interpretable, and robust across diverse scientific disciplines.

One promising direction is the integration of advanced transformer-based language models such as RoBERTa, DeBERTa, SciBERT, and domain-specific biomedical language models. These contextual embedding techniques have demonstrated improved language understanding and may provide more accurate semantic similarity estimation

than conventional embedding methods. Hybrid approaches that combine lexical similarity, statistical representations, contextual embeddings, and knowledge graph information may further improve the identification of subtle semantic differences between human-authored and AI-generated scientific abstracts.

Future studies may also investigate multilingual text similarity by extending the evaluation framework to scientific documents written in different languages. Such an extension would enable cross-lingual semantic comparison and facilitate international scientific collaboration. Additionally, incorporating explainable Artificial Intelligence (XAI) techniques could improve the interpretability of similarity scores by identifying which textual components contribute most significantly to semantic similarity or dissimilarity.

Another important research direction involves constructing larger and more diverse benchmark datasets containing abstracts from multiple scientific disciplines, publication venues, and AI language models. Evaluating similarity algorithms using continuously updated datasets will improve model generalization and ensure reliable performance under evolving text-generation technologies. Furthermore, integrating real-time AI content detection with plagiarism detection systems, digital libraries, academic publishing platforms, and peer-review systems may strengthen scientific integrity and support trustworthy scholarly communication.

Overall, future advancements in contextual language understanding, hybrid similarity models, explainable AI, and multilingual document analysis are expected to significantly improve automated text similarity assessment and contribute to more reliable evaluation of AI-generated scientific literature.

## VII. CONCLUSION

The comparative analysis demonstrates that traditional lexical similarity methods remain effective for detecting direct textual overlap but are less capable of identifying semantic equivalence in paraphrased or AI-generated scientific documents. In contrast, contextual embedding techniques, particularly Word2Vec, FastText, and BERT, consistently provide superior semantic representations by capturing contextual relationships beyond exact word matching. These methods exhibit greater robustness when evaluating AI-generated scientific abstracts, making them more suitable for modern Natural Language Processing applications.

The experimental findings further emphasize the importance of combining statistical analysis with multiple evaluation metrics to obtain a comprehensive assessment of similarity algorithms. Rather than relying on a single similarity measure, integrating lexical, semantic, and contextual approaches provides a more reliable understanding of textual relationships among scientific documents. Such comprehensive evaluation is particularly valuable for plagiarism detection, scientific document verification, academic integrity assessment, semantic search, and AI-generated content analysis.

As Artificial Intelligence continues to transform scientific writing, accurate text similarity evaluation will become increasingly important for maintaining research authenticity and supporting trustworthy scholarly communication. The insights presented in this study provide useful guidance for selecting appropriate similarity techniques while establishing a foundation for future research involving intelligent document comparison, AI-generated content verification, and advanced Natural Language Processing systems.

## REFERENCES

1. "Comparative analysis of text similarity methods for human and AI-generated scientific abstracts," T. Lin, Y. Wang, and H. Chen, Scientific Reports, vol. 14, no. 1, pp. 1–18, 2024.
2. "BERT: Pre-training of deep bidirectional transformers for language understanding," J. Devlin, K. Lee, M. W. Chang, and K. Toutanova, Proceedings of NAACL-HLT, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.
3. "Efficient estimation of word representations in vector space," T. Mikolov, K. Chen, G. Corrado, and J. Dean, Proc. Int. Conf. Learning Representations (ICLR), 2013.
4. "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017; P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov.
5. "GloVe: Global vectors for word representation," J. Pennington, R. Socher, and C. Manning, Empirical Methods in Natural Language Processing (EMNLP), Proceedings 1532–1543, Doha, Qatar, 2014.
6. "Term-weighting approaches in automatic text retrieval," G. Salton and C. Buckley, Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.
7. "A vector space model for automatic indexing," G. Salton, A. Wong, and C. S. Yang, Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.
8. V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707–710, 1966.
9. W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage," U.S. Census Bureau Tech. Rep., Washington, DC, USA, 1990.
10. "Identification of common molecular subsequences," T. F. Smith and M. S. Waterman, Journal of Molecular Biology, vol. 147, no. 1, pp. 195–197, 1981.

11. Introduction to Information Retrieval by C. D. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, Cambridge, UK, 2008.
12. Speech and Language Processing, D. Jurafsky and J. H. Martin, Upper Saddle River, NJ, USA: Pearson, 2024, 3rd ed.
13. S. Venkateswara Rao, "Use Machine Learning to Predict the Probability of Being Admitted to a University," International Journal of Computer Networks and Wireless Communications (IJCNWC), vol. 15, no. 2, 2025.
14. S. Venkateswara Rao, "An AI-based Method for Analyzing the Relationships Between Financial Variables and Stock Price Changes," International Journal of Biology, Agriculture and Healthcare Research (IJBAR), vol. 15, no. 2, 2025.