IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

A Machine Learning and NLP Framework for Automatic Title Similarity Detection and Preliminary Verification

Miss. Suvarna Bhoi[1], Dr. Sarkarsinha H. Rajput[2], Dr. Shital A. Patil[3], Dr. Dinesh D. Puri[4]

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra, India[1]

Asso. Professor, Department of Computer Engineering, SSBT's COET, Jalgaon, Maharashtra, India[2-4]

Abstract: Title verification becomes difficult when a large collection of publication titles has to be checked for similarity before a new title is accepted. Exact string matching is not sufficient because titles may contain spelling variations, reordered words, prefixes or suffixes, periodicity changes, phonetic variations, or different words with similar meanings. This research proposes a practical Natural Language Processing (NLP) and Machine Learning framework for automatic preliminary title verification. The proposed approach first normalizes and preprocesses title text and then represents titles using Term Frequency–Inverse Document Frequency (TF-IDF). Cosine Similarity is used as the main lexical similarity measure to rank existing titles against a newly submitted title. Character-level and phonetic features can be added to improve detection of spelling and sound-based variations. Logistic Regression, Support Vector Machine (SVM), and Random Forest are considered for classifying title pairs as similar or non-similar. The system will return the most similar existing titles, similarity scores, and a verification recommendation for human review. Evaluation will consider accuracy, precision, recall, F1-score, false-positive and false-negative behaviour, and processing time. The study aims to provide an interpretable, scalable, and practical approach that reduces repeated manual comparison while keeping final verification under human control.

Keywords: Title Similarity Detection, Natural Language Processing (NLP), TF-IDF, Cosine Similarity, SBERT, Machine Learning, Semantic Similarity, Preliminary Verification.

I. INTRODUCTION

The registration and verification of publication titles is an important administrative activity when a large number of newspapers, magazines, journals, or other publications submit new titles. Before a title is accepted, it is useful to check whether the same or a closely related title already exists in the available collection. When the database contains a large number of records, manual comparison becomes time-consuming and difficult to perform consistently. An automated preliminary checking system can help a verifier identify the closest existing titles before making a final decision.

Traditional exact string matching has an important limitation: two titles can be different as character strings but still be highly similar in content. For example, "Maharashtra Daily News" and "Maharashtra News Daily" contain the same important terms although their word order is different. Small spelling changes can also create false differences. In addition, the SIH1782 problem requires attention to phonetic similarity, prefixes and suffixes, disallowed words, combinations of existing titles, periodicity modifications, similar meanings across languages, and efficient database interaction. These requirements make the problem suitable for a combination of NLP, similarity measures, Machine Learning, and database search techniques.

NLP provides a systematic way to clean and represent title text. TF-IDF is a useful baseline for short titles because it assigns higher importance to informative terms, while Cosine Similarity provides an easy-to-understand measure of closeness between two title vectors. Levenshtein Distance can capture small character changes, and phonetic techniques can support sound-based comparison. Word2Vec and transformer models such as IndoBERT can further improve semantic matching when different words express related ideas. However, more advanced models generally increase implementation and computational requirements.

The motivation of this research is to study a practical combination of methods that balances accuracy, interpretability, scalability, and response time. The proposed framework starts with a simple TF-IDF and Cosine Similarity baseline and then examines additional character, phonetic, and Machine Learning features. The system is designed as a decision-support tool: it highlights potentially similar titles and recommends whether a case needs further review, while the final administrative decision remains with the authorized verifier.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

II. LITERATURE REVIEW

Although existing studies demonstrate useful title-similarity techniques, most focus on academic thesis or research titles and primarily return a similarity score. The SIH1782 problem requires a broader preliminary verification workflow that can handle a large title database, spelling and phonetic variations, word-order changes, multilingual titles, candidate ranking, and human review. The proposed research therefore extends the existing work through a staged architecture: text normalization, fast lexical retrieval, semantic retrieval, character and phonetic comparison, feature fusion, machine-learning reranking, explainable Top-K results, and final human verification. This combination is proposed as a research framework; its effectiveness must be established experimentally on the selected dataset.

Research Gap and Proposed Direction

The reviewed studies show a clear progression from simple lexical similarity toward hybrid and contextual approaches. TF-IDF with Cosine Similarity provides an interpretable and computationally practical baseline [1], [2], [10]. Levenshtein Distance adds character-level information for spelling variation [4]. Word2Vec and other word embeddings provide additional word-level relationships [1], [3]. SBERT and IndoBERT provide contextual semantic representations that can identify related meanings even when titles do not share many exact words [5]–[8]. Recent work further suggests that a scalable title-verification system can combine lexical and semantic retrieval with Top-K candidate selection and machine-learning reranking [9].

Synthesis of the Literature

Sarimuddin and co-authors developed a web-based thesis-title similarity system using TF-IDF and Cosine Similarity and discussed its use for early detection of duplicate or highly similar titles. Their study also identifies semantic similarity integration and automated indexing as useful future directions. This is relevant to SIH1782 because a real verification system must continue to work as the title database grows. Automated indexing and efficient candidate retrieval can therefore be included as scalability considerations in the proposed framework [10].

10. Sarimuddin et al. (2026) – TF-IDF/Cosine Similarity with Continuous Database Growth [10]

Al Imran, Ardiansyah and Abidin proposed a recent thesis-title similarity system that combines TF-IDF lexical retrieval, SBERT semantic retrieval, Top-K candidate selection, and machine-learning reranking using LightGBM. The architecture is highly relevant to the scalability problem in SIH1782 because it avoids treating the entire database as one undifferentiated comparison set. The proposed research can adopt the same two-stage idea: first retrieve a manageable candidate set using fast lexical and semantic methods, then use richer features and a classifier or reranker to order the most relevant candidates [9].

9. Al Imran, Ardiansyah and Abidin (2026) – Hybrid Retrieval and ML Reranking [9]

Hossain et al. studied semantic textual relatedness across English, Marathi, Telugu, and Spanish using transformer-based models including Sentence Transformers and language-specific models. Their work is particularly relevant to the Indian context because Marathi and Telugu were included in the evaluation. The study demonstrates that semantic similarity performance can vary across languages and that transformer-based methods can be useful for multilingual relatedness. This motivates a multilingual extension of the proposed title-verification framework for English and Indian-language titles [8].

8. Hossain et al. (2024) – Multilingual Semantic Relatedness [8]

Koto et al. introduced IndoBERT together with the IndoLEM benchmark for Indonesian Natural Language Processing. Their work demonstrates the usefulness of language-specific pretrained models for semantic and language-understanding tasks. Although the SIH1782 problem is not limited to Indonesian, this study demonstrates the value of language-specific pretrained transformer models for semantic and language-understanding tasks. This supports considering language-specific or multilingual models when the proposed system handles regional or multilingual titles [7]. For the proposed framework, IndoBERT can be treated as an advanced semantic model where the dataset language makes it appropriate [7].

DOI: 10.17148/IJARCCE.2026.15927

7. Koto et al. (2020) – IndoBERT for Indonesian NLP [7]

Reimers and Gurevych introduced Sentence-BERT (SBERT), a Siamese and triplet-network adaptation of BERT designed to generate semantically meaningful sentence embeddings that can be compared efficiently using Cosine Similarity. The work is important for large-scale title verification because it makes semantic similarity search substantially more practical than directly comparing every sentence pair with a full BERT model. SBERT therefore provides a strong candidate for the semantic-retrieval stage of the proposed system [6].

6. Reimers and Gurevych (2019) – Sentence-BERT for Semantic Search [6]

Ishak and Bengnga investigated research-title similarity using IndoBERT embeddings and Ditto Whitening. Their study shows that transformer embeddings can suffer from anisotropy, where semantically different titles may receive unnecessarily high similarity scores. They evaluated embedding quality using cosine-similarity distributions, isotropy, hubness, t-SNE, UMAP, and heatmaps. This provides an important research direction for the proposed framework: semantic embeddings should be evaluated and calibrated rather than treated as automatically reliable similarity measures [5].

5. Ishak and Bengnga (2026) – IndoBERT Embedding Optimization [5]

Sunandar and Muiz proposed a web-based thesis-title similarity system using Levenshtein Distance and Cosine Similarity. Their approach is useful for this research because Levenshtein Distance can identify character-level changes such as insertion, deletion, substitution, and spelling variation, while Cosine Similarity provides a broader lexical comparison. Their study also includes normalization of older spelling forms, showing that text normalization can directly affect title matching performance [4].

4. Sunandar and Muiz (2025) – Character and Lexical Similarity [4]

Maharani, Syah and Saputra studied thesis-title similarity using term weighting and word embedding. Their work highlights the difference between lexical importance and relationships between words. A term-based method may identify titles that share important words, while embeddings can capture relationships between words that do not exactly match. This supports the proposed use of multiple feature types rather than depending only on TF-IDF [3].

3. Maharani, Syah and Saputra (2022) – Term Weighting and Word Embedding [3]

Baba, Lukman and Wahyuni investigated thesis-title similarity using preprocessing, TF-IDF representation, and Cosine Similarity on a dataset of 1,000 thesis titles. Their study demonstrates that a transparent similarity pipeline can be used for early title verification and evaluated with standard classification metrics. The reported precision and recall also show why a title-verification system should not rely only on a single similarity score. This supports the use of precision, recall, and F1-score in the proposed research [2].

2. Baba, Lukman and Wahyuni (2025) – Cosine Similarity for Thesis Titles [2]

Baksir, Rosihan and Albaar developed an Android-based research title similarity system that combines TF-IDF, Word2Vec, and Cosine Similarity. Their work is directly relevant because it combines lexical term importance with word-level semantic information for title comparison. The study shows that a hybrid lexical-semantic approach can be more informative than relying on exact word overlap alone. This provides a useful baseline for the proposed SIH1782 framework, while also leaving scope for stronger contextual embeddings and machine-learning reranking [1].

1. Baksir, Rosihan and Albaar (2026) – Hybrid Lexical-Semantic Baseline [1]

Study	Main method	Main contribution	Relevance to SIH1782
Baksir et al. (2026) [1]	TF-IDF + Word2Vec + Cosine	Hybrid lexical-semantic title matching	Strong baseline for hybrid title similarity
Baba et al. (2025) [2]	TF-IDF + Cosine	Transparent thesis-title similarity pipeline	Supports baseline Evaluation and metrics

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

Maharani et al. (2022) [3]	Term weighting + embeddings	Combines term importance with word relationships	Supports multi-feature similarity
Sunandar & Muiz (2025) [4]	Levenshtein + Cosine	Handles character and lexical variation	Useful for spelling variations
Ishak & Bengnga (2026) [5]	IndoBERT + Ditto Whitening	Improves embedding quality and similarity analysis	Advanced semantic similarity direction
Reimers & Gurevych (2019) [6]	SBERT	Efficient semantic sentence embeddings	Suitable for semantic candidate retrieval
Koto et al. (2020) [7]	IndoBERT	Language-specific contextual representation	Supports ultilingual/regional title processing
Hossain et al. (2024) [8]	Sentence Transformers + multilingual models	Semantic relatedness across English, Marathi and Telugu	Supports Indian-language extension
Al Imran et al. (2026) [9]	TF-IDF + SBERT + Top-K + LightGBM	Hybrid retrieval with ML reranking	Strong model architecture for scalable verification
Sarimuddin et al. (2026) [10]	TF-IDF + Cosine	Web-based duplicate-title screening and indexing direction	Supports scalable database verification

III. METHODOLOGY

The research follows an experimental and applied design to develop and evaluate a prototype for automatic title similarity detection. The methodology consists of five main stages.

1. **Data Collection and Preprocessing:** Existing title records will be collected from the available dataset and suitable sources. Titles will be normalized by handling case differences, punctuation, extra spaces, tokenization, and suitable normalization. If supervised learning labels are unavailable, a documented annotation procedure will be used to prepare similar and non-similar title pairs.
2. **Feature Extraction and Similarity:** TF-IDF will represent important terms and Cosine Similarity will provide the main similarity score. Character similarity, normalized Levenshtein Distance, word overlap, and phonetic similarity may be added to handle spelling and sound-based variations.
3. **Machine Learning Classification:** Logistic Regression, Support Vector Machine (SVM), and Random Forest will be trained on title-pair features to classify pairs as similar or non-similar. Training and test data will be separated, with validation data used for selecting thresholds or model settings.
4. **Verification Output:** For each new title, the system will rank the closest existing titles and display their similarity scores. A preliminary recommendation such as "Likely Similar – Review Required" or "Low Similarity – Proceed for Verification" will be generated. The recommendation supports human review and is not a final legal or administrative decision.
5. **Evaluation and Tools:** Accuracy, precision, recall, F1-score, confusion matrix, false-positive/false-negative cases, and average processing time will be measured. Python, Pandas, NumPy, Scikit-learn, NLTK or spaCy, and a database such as SQLite, MySQL, or PostgreSQL can be used. A simple Flask/FastAPI interface may be developed for the prototype.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

Fig. 1. Overall architecture

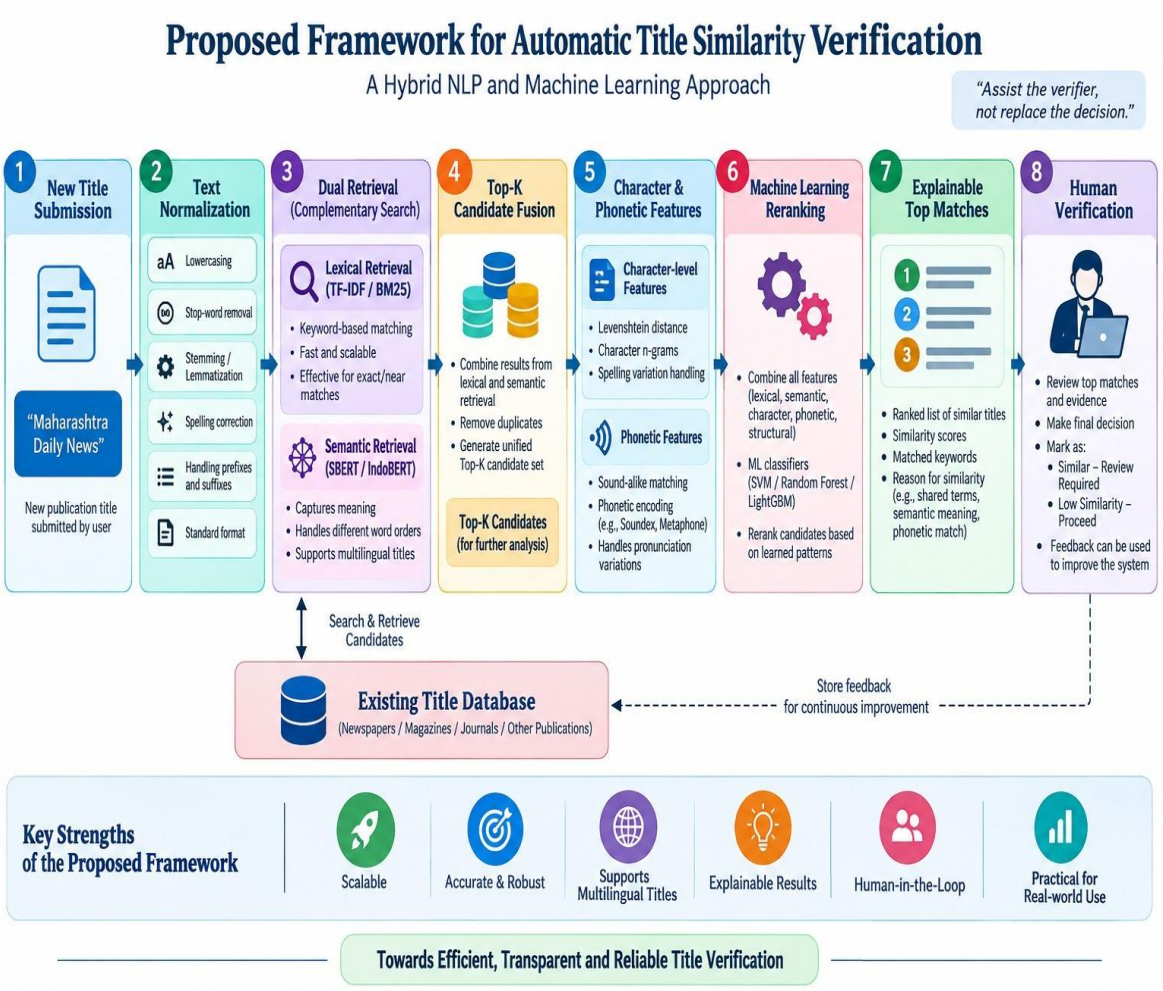

Fig. 1. Overall architecture of the proposed automatic title similarity detection and preliminary verification system.

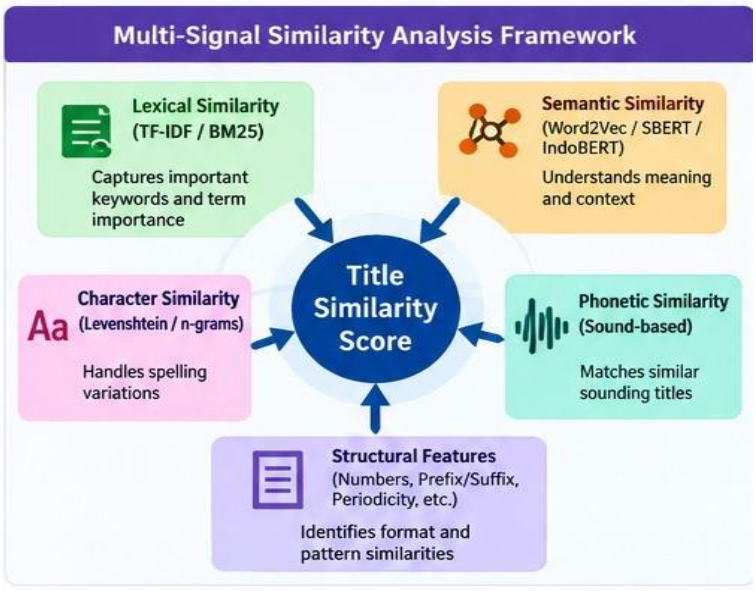

Fig. 2. Multi-signal similarity analysis framework for title matching.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

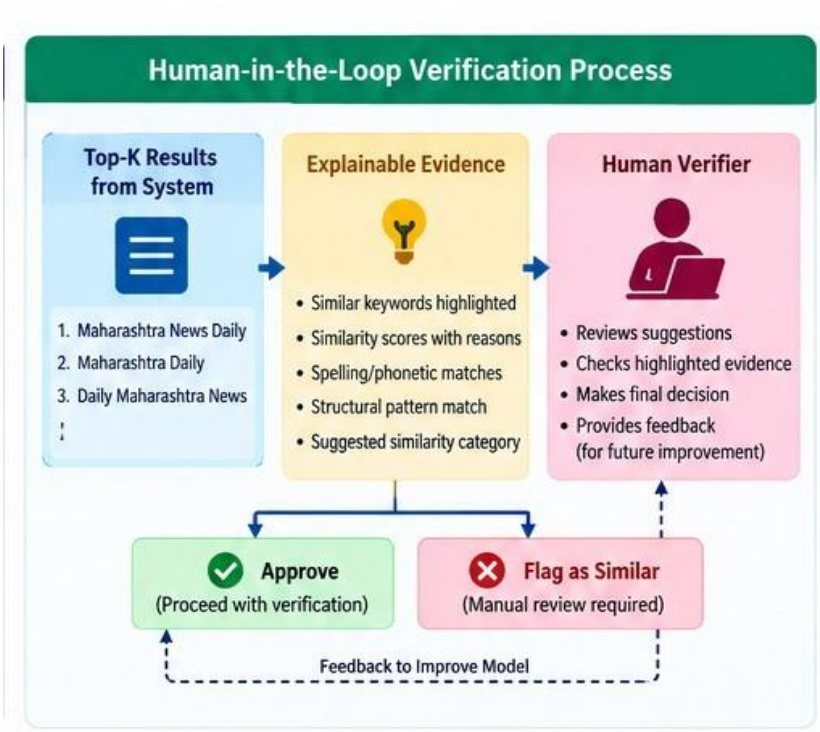

Fig. 3. Human-in-the-loop verification and decision support process.

The proposed framework is illustrated through five complementary diagrams. The architecture combines fast lexical retrieval with semantic retrieval, character and phonetic features, machine-learning reranking, explainable Top-K results, and human verification. The evaluation framework further considers classification performance, retrieval quality, error cases, processing time, and scalability. These figures represent the proposed system design and evaluation plan; they do not claim experimental results before testing.

IV. RESULTS

The research is designed to produce results that can be interpreted directly by a verifier. Since the final experiments depend on the actual dataset and labelled title pairs, numerical performance values are not assumed in advance. The result analysis will compare the baseline TF-IDF with Cosine Similarity against the enhanced feature and Machine Learning approaches.

Evaluation Aspect	Measure / Output	Purpose
Classification	Accuracy, Precision, Recall, F1-score	Measures how correctly the system identifies similar and non-similar title pairs.
Ranking	MAP@K, NDCG@K, Top-K similar titles	Evaluates whether the most relevant existing titles appear near the top of the ranked results.
Similarity	Cosine similarity, character similarity, phonetic similarity	Shows different levels of textual similarity between the submitted title and existing titles.
Error Analysis	False positives and false negatives	Identifies cases where the system incorrectly flags or misses similar titles.
Efficiency	Processing time and database response time	Checks whether the framework can provide results efficiently for a growing title database.
Verification Support	Top matching titles, scores and explanation	Provides evidence to the human verifier for the final title-approval decision.

The primary result will be a ranked list of existing titles for every submitted title, together with similarity scores. A high similarity score will indicate that the title should receive closer human review, while a low score will indicate a lower priority for detailed comparison. The final threshold will be selected from validation results rather than being fixed

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15927

without evidence.

Model performance will be summarized using accuracy, precision, recall, and F1-score. Processing time will also be reported because SIH1782 requires efficient interaction with a large title database. False positives and false negatives will be examined separately because both types of error have different practical consequences.

The figure below summarizes how the final output can be interpreted. It is a result-interpretation framework, not a claim of experimental accuracy. Actual numerical findings will be inserted after the system is tested on the selected dataset.

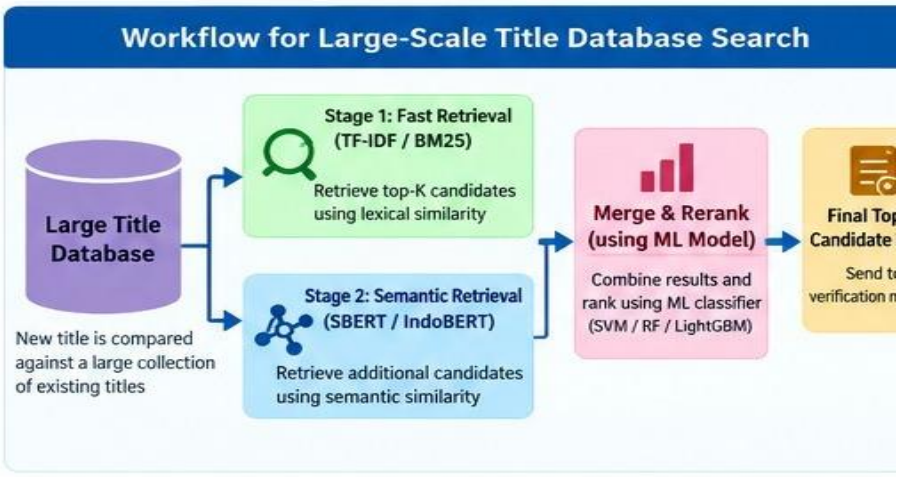

Fig. 4. Two-stage retrieval and reranking workflow for efficient database sea

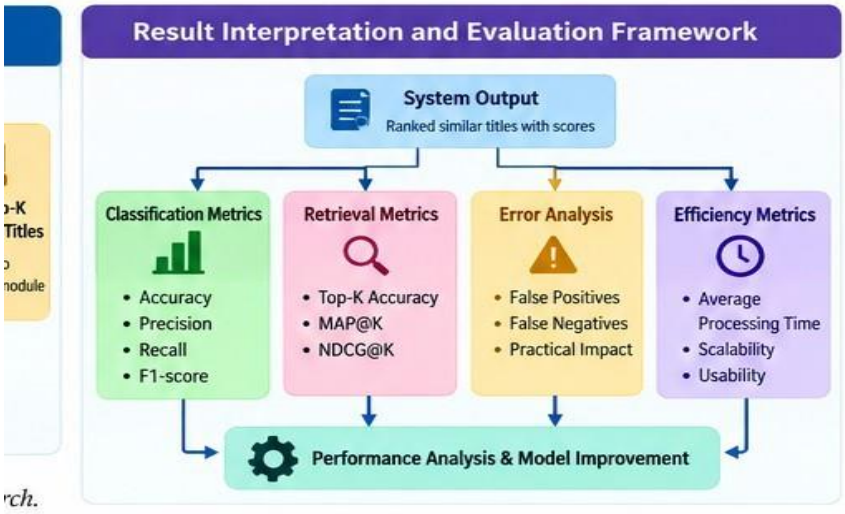

Fig. 5. Evaluation framework for result analysis and model improvement.

V. DISCUSSION

The proposed framework combines methods that operate at different levels of similarity. TF-IDF and Cosine Similarity provide a strong and understandable baseline for lexical similarity and are suitable for short titles. Character and phonetic features can address spelling and sound-based variations, while Word2Vec or IndoBERT can provide richer semantic information when word overlap is insufficient.

The most important practical issue is the balance between false positives and false negatives. Too many false positives may cause unnecessary manual reviews, whereas false negatives may fail to flag a potentially similar title. Precision, recall, F1-score, and error-case inspection are therefore necessary. The threshold should be selected using validation evidence.

The quality of the database and labels will directly affect reliability. Multilingual titles, abbreviations, periodicity changes, combined titles, and domain-specific terms may require additional rules or multilingual semantic models.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15927

Candidate retrieval and database indexing will also be important for scalability. Therefore, the system is best viewed as an intelligent preliminary screening tool that improves consistency and reduces manual workload while leaving the final decision to an authorized human reviewer.

VI. CONCLUSION

This research proposes a practical NLP and Machine Learning framework for automatic title similarity detection and preliminary verification. The approach begins with text normalization and TF-IDF representation, uses Cosine Similarity to rank existing titles, and examines character, phonetic, and Machine Learning features for additional improvement.

The literature indicates that simple lexical methods are transparent and efficient, while embeddings and transformer models can provide stronger semantic information at greater computational cost. The staged methodology therefore provides a reasonable balance between accuracy, interpretability, and practical implementation.

The final system will present the closest existing titles, similarity scores, and a verification recommendation so that human reviewers can focus on potentially similar cases. Evaluation using classification metrics, error analysis, and processing time will determine which combination of methods is most suitable for the available dataset.

REFERENCES

[1] A. D. Baksir, Rosihan, and M. R. Albaar, "Android-Based Research Title Similarity Detection Using a Combined Word2Vec and TF-IDF with Cosine Similarity Score," Teknika, vol. 15, no. 2, pp. 273–282, 2026, doi: 10.34148/teknika.v15i2.1487.

[2] H. Baba, Lukman, and T. Wahyuni, "Determining the Similarity Level of Students' Thesis Titles at the Faculty of Teacher Training and Education, Unismuh Makassar, Using the Cosine Similarity Method," Ainet: Jurnal Informatika, vol. 7, no. 2, pp. 127–137, 2025, doi: 10.26618/zzptwc89.

[3] S. Maharani, F. Syah, and N. Saputra, "Sistem Deteksi Kemiripan Judul Skripsi dengan Metode Term Weight and Word Embedding," Jurnal Dinamika Informatika, 2022.

[4] D. Sunandar and A. Muiz, "Thesis Title Similarity Detection System Using Levenshtein Distance and Cosine Similarity," bit-Tech, vol. 8, no. 1, pp. 1131–1139, 2025, doi: 10.32877/bt.v8i1.2864.

[5] R. Ishak and A. Bengnga, "Optimizing of IndoBERT Embedding with Ditto Whitening for Measuring Research Title Similarity," Jambura Journal of Electrical and Electronics Engineering, vol. 8, no. 1, pp. 46–54, 2026, doi: 10.37905/jjeee.v8i1.35554.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proc. 28th International Conference on Computational Linguistics, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[8] M. S. Hossain, A. I. Paran, S. H. Shohan, J. Hossain, and M. M. Hoque, "SemanticCUETSync at SemEval-2024 Task 1: Finetuning Sentence Transformer to Find Semantic Textual Relatedness," in Proc. 18th International Workshop on Semantic Evaluation, pp. 1222–1228, 2024, doi: 10.18653/v1/2024.semeval-1.178.

[9] A. M. Al Imran, M. Ardiansyah, and A. A. Zainal Abidin, "Development and Evaluation of an Indonesian Thesis-Title Similarity Detection System Using Hybrid TF-IDF–SBERT Retrieval and LightGBM Reranking," Jurnal Media Elektrik, vol. 23, no. 3, 2026, doi: 10.59562/metrik.v23i3.14414.

[10] S. Sarimuddin, N. Azlina, A. Anggun, V. N. Khalifah, and N. Nasarudin, "Analisis Kemiripan Judul Skripsi Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Untuk Mencegah Duplikasi," SemanTIK: Teknik Informasi, vol. 12, no. 1, pp. 33–42, 2026, doi: 10.55679/semantik.v12i1.259.