A Survey and Hybrid Similarity–Based Framework for Automated Publication Title Verification

Srinidhi S N [1], Sumanth A N [2], Tharun V [3], Uday S [4], and Dr. Thirumagal Mohan [5]
[1,2,3,4]*Student, REVA University, Bangalore, India*
[5]*Assistant Professor, REVA University, Bangalore, India*

***Abstract*—The rapid growth of print, digital, and multilingual media has significantly increased the number of publication title registration requests handled by regulatory bodies such as the Press Registrar General of India (PRGI). Traditional methods that rely on manual review or simple keyword matching are no longer sufficient, as they often fail to capture similarities arising from pronunciation differences, spelling variations, transliteration, or changes in wording. This can lead to duplication of titles, inconsistencies, and delays in the approval process. In this work, a hybrid similarity-based framework is proposed to improve the accuracy and reliability of publication title verification. The system combines phonetic similarity, string-based comparison, and semantic similarity using transformer-based sentence embeddings, allowing titles to be evaluated from multiple perspectives. In addition, rule-based validation is incorporated to ensure compliance with regulatory constraints. To further enhance usability, the system also provides a mechanism for suggesting alternative titles with lower similarity. These suggestions are generated using a combination of structured approaches and language-based generation techniques, helping users choose more unique and acceptable titles. The experimental evaluation demonstrates that the hybrid approach performs better than individual techniques, achieving an F1-score of 0.76 in representative test cases. Overall, the proposed system offers a practical and scalable solution for supporting automated decision-making in publication title verification.**

***Index Terms*—Hybrid AI, Phonetic Similarity, String Similarity, Semantic Similarity, Natural Language Processing, Sentence-BERT, Publication Title Verification, PRGI**

I. INTRODUCTION

A. Background and Motivation

The Press Registrar General of India (PRGI) acts as a regulatory authority in the process of registering, regulating, and approving publication titles in India. It ensures the uniqueness of titles, their authenticity, and compliance with regulatory guidelines in a manner that does not promote public misinformation, duplication of titles, or the misuse of sensitive terminology. With the proliferation of print media, digital media, online publications, and journalism in various regional languages, the number of title registration requests handled by the PRGI has increased manifold in recent years. Traditionally, various title verification processes have relied on human assessment and search mechanisms based on keywords. Publication titles may vary in terms of phonetic pronunciation, linguistic transliteration, and language usage.

B. Problem Statement

The existing mechanisms of verifying the titles through the usage of keywords and manual inspections have a few underlying drawbacks. The usage of keywords does not consider phonetic variations, misspellings, and transliteration of words. Manual inspections are subjective and rely on the expertise of the individual performing the task. This results in a lack of consistency in the decision-making process.

The above-mentioned drawbacks of the existing mechanisms indicate the need for a new automated system that can incorporate the results of phonetic, lexical, and semantic analysis.

C. Objectives of the Study

It is possible to outline the primary goals of this research. Firstly, the paper aims to conduct a critical review of some of the similarity detection methods including phonetic similarity, string similarity, and semantic similarity using Natural Language Processing (NLP). Secondly, the proposed study would conduct a critical analysis of the techniques, particularly in the context of their use in title verification particularly within a PRGI context. Lastly, the paper aims at suggesting a hybrid model that will combine different similarity methods to come up with a single system of title verification. This study is intended to serve as a decision-support tool, especially in enhancing title verification processes by PRGI.

D. Contributions of This Paper

The key contributions of this paper are as follows:

(i) a systematic and comprehensive survey of phonetic, string, and semantic similarity detection approaches pertinent to the verification of publication titles.

(ii) a domain-specific study on the issues of similarity detection in the context of PRGI systems and title approval systems in the regulatory domain.

(iii) the development of a hybrid similarity-based approach that incorporates phonetic encoding, string similarity scoring, and semantic comparison.

(iv) the development of a prototype implementation that validates the feasibility of the proposed approach as a scalable decision support tool for the automated verification of publication titles.

II. RELATED WORK

Several approaches have been explored in the literature for measuring similarity in textual data, especially for short texts and name matching tasks. Phonetic techniques such as Soundex and Metaphone are commonly used to identify similarities in pronunciation, even when the spelling differs. In contrast, string-based methods like Levenshtein distance and TF-IDF focus on lexical and character-level similarities, making them useful for detecting spelling variations and minor text differences. However, these approaches are limited in capturing deeper meaning and are often sensitive to word order and vocabulary changes.

III. COMPARATIVE ANALYSIS OF SIMILARITY TECHNIQUES

A. Evaluation Criteria

The reviewed similarity techniques are evaluated based on accuracy, linguistic robustness, computational cost, and scalability, which collectively reflect their suitability for large-scale regulatory title verification.

Technique-wise Comparison

Table I presents a comparative summary of phonetic, string-based, semantic, and hybrid similarity techniques based on the defined evaluation criteria.

Table I: Comparison of Similarity Detection Techniques for Title Verification

Techniqu ue Category	Representati ve Methods	Accurac y	Linguistic Robustness	Computation al Cost
Phonetic Similarit y	Soundex, Metaphone, Double Metaphone	Modera te	High for pronunciati on variants, low for semantics	Low
String-Based Similarit y	Levenstein Distance, TF-IDF + Cosine	Modera te	Low to moderate	Low to moderate
Semantic Similarit y	Word2Vec, BERT, SBERT	High	High	High
Hybrid Similarit y	Combined phonetic, lexical, and semantic scoring	High	High	Moderate

B. Observations and Insights

The analysis also reveals that there is no similarity technique for finding the similarity between the publication title and the given title, which can effectively cover all the cases of similarity. The results of the study also show that the use of phonetic similarity is effective for the similarity based on pronunciation and transliteration, but it ignores the meaning of the words. The results also show that the use of string similarity is effective for the similarity based on spelling, but it is affected by the vocabulary and the order of the words. The results also show that

the use of semantic similarity is effective, but it is computationally complex and affected by short texts.

The above observations indicate that the reliance on a single similarity approach might lead to incomplete or inconsistent results for the verification of publication titles. The hybrid approaches to similarity, which include phonetic, string, and semantic similarity, appear to provide a fair and balanced approach to the problem of publication title verification by compensating for the limitations of the individual approaches. Thus, the above approaches to similarity appear to emerge as a viable solution for the development of automated decision support systems. However, most of the existing studies appear to have focused on the evaluation of the individual similarity approaches, while the hybrid approaches involving phonetic, lexical, and semantic similarity for the purpose of regulatory title verification have not been sufficiently explored.

IV. CHALLENGES IN MANUAL AND SEMI-AUTOMATED TITLE VERIFICATION SYSTEMS

A. Scalability and Efficiency Issues

Such manual and semi-automated title verification approaches are limited in terms of their scalability with the increased number of applications. With the increase in the number of registered titles, as well as pending titles for publication, the human verification process needs to check more titles in a limited amount of time, causing delays in the title verification process. In the case of large data, the probability of skipping certain matches increases, causing reduced accuracy in the verification process.

B. Multilingual and Transliteration Challenges

This linguistic variation creates a high level of complexity while dealing with title verification. There is a possibility that, while dealing with publication titles, there could be multiple language, script, or transliteration variations. There is a possibility that, even though there is phonetic similarity between words, lexical similarity is not present. There is also a possibility that, due to inconsistent transliteration, there could be multiple textual forms of a single word. Manual verification methods face difficulties while dealing with these issues, which could result in duplication.

C. Subjectivity and Inconsistency

Manual decision-making in title verification is based on the experience, interpretation, and understanding of the human reviewers. In this regard, there are chances that the same case could be decided differently based on the interpretation of the reviewer. This leads to a lack of consistency and transparency in the decision-making process in title verification. In the absence of similarity assessment metrics, the decision could vary with time as well as the personnel involved in the process, which could lead to a lack of trust in the regulatory process itself. Semi-automated systems, which are based on keyword-based static rules, could only offer limited benefits in the title verification process, as the final decision remains with the human reviewers.

D. Regulatory Constraint Enforcement

Regulatory guidelines also provide restrictions on the use of certain specific words, especially when it refers to government authority or national affiliation. Maintaining the enforcement of restrictions through manual checks is a challenging task, especially when it refers to the use of indirect phrases or references in the title. Keyword restrictions for rule enforcement may also result in incorrect flags for certain legitimate cases. The absence of rule-based validation also results in a cognitive overload.

V. SYSTEM ARCHITECTURE AND METHODOLOGY

A. Overall System Architecture

The proposed system is designed as a modular hybrid similarity-based decision support system for automated verification of publication titles. The system architecture is designed as a sequential workflow in which an input title to be verified is analysed against a repository of registered titles through a series of similarity assessment modules. Each module has been conceived to measure similarity from a particular point of view.

The process starts with the ingestion of the input title, followed by the pre-processing of the input title in the form of text normalization. The pre-processed input title is fed as input to the phonetic similarity module, string similarity module, and semantic similarity module individually. Each module calculates the similarity score between the input title and the titles

available in the database. The similarity scores are then calculated using the weighted hybrid scoring method. In parallel, the input title is fed as input to the rule-based validation module, which validates the input title for the presence of restricted terms. The end similarity score is then employed to classify the input title into some levels of risk that are presented to the user.

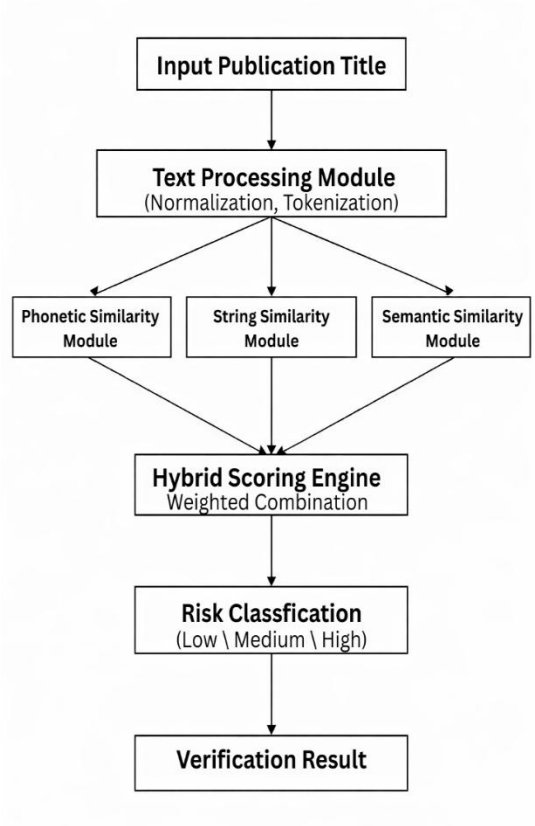

Fig. 1. System Architecture of the Hybrid Title Verification Framework.

B. Text Preprocessing Module

Text processing involves normalization, tokenization and removal of stop words to guarantee similarity computability.

C. Phonetic Similarity Module

The phonetic similarity module will be used in the identification of similar sounding and transliteration variations between the publication titles. For this purpose, the Double Metaphone algorithm is used. This algorithm generates the primary and secondary phonetic code of the words used in the title. The reason why the Double Metaphone algorithm has been selected is that it is more effective in addressing pronunciations and spelling variations than other algorithms.

Phonetic encodings are generated and compared against each other to detect common phonetic representations. Phonetic similarity score is determined based on similarity of the encoded forms of input titles as compared to other titles. This module specifically works well in establishing similarities because languages and scripts differ in pronunciation.

D. String Similarity Module

The string similarity module is dedicated to the lexical similarity and typographical variations. The distance measures used are the edit distance measures, including Levenstein distance, which are used to quantify the character differences between the titles. This is helpful in detecting minor spelling differences.

In addition, the calculation of lexical similarity based on the vector approach utilizes a mix of Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity. The title is represented as a weighted term vector, and the cosine similarity is used to compute the angular similarity. Although the string-based approach is vocabulary- and order-dependent, it provides significant information, which can also be used with phonetics and semantics.

E. Semantic Similarity Module

The semantic similarity module is responsible for capturing contextual similarity between titles. This module uses a Sentence BERT (SBERT) model, which is a transformer-based model that generates fixed-length embeddings optimized for sentence-level similarity comparison. Each title is converted to a dense vector representation.

The cosine similarity is calculated for the embeddings of the input title and the other titles. This helps with the detection of semantically similar titles. Although this task is computationally more expensive than other modules, it provides important information for the detection of meaning-equivalent short text titles.

F. Hybrid Similarity Score Computation

To utilize the strengths of individual similarity modules, a hybrid scoring approach is adopted. Each module generates a normalized similarity score over

a range of values. The scores of phonetic, string and semantic similarity are assigned weights based on the relative importance and ability to be used empirically to verify titles.

The last similarity score is determined as a weighted sum of the scores of the individual modules. Such a hybrid method enables a reasonable combination of the pronunciation-based, lexical, and semantic levels. The scores are normalized so that the scores are easier to interpret and compare. To support scalability, the semantic embeddings of the registered titles are calculated in advance and re-used to evaluate the similarity.

G. Rule-Based Validation and Risk Classification

Besides the similarity, there is also a rule-based validation module incorporated into the system. This module includes a verification of the existence of some restricted keywords, which interest in particular based on the regulatory guidelines. The titles that contain such words are determined to be investigated further.

Depending on the end hybrid similarity score and the rule-based validation findings, the system goes ahead to classify the title under a few pre-defined risk levels that include Low, Medium and High risk. Low-risk titles have low similarity and no rule violations, whereas high-risk titles have high similarity or possible rule violations. This is a decision support feature for human approvers to help them make consistent approval decisions.

H. Hybrid Formula Equation

$$Hybrid = W_p S_p + W_s S_s + W_e S_e$$

Where:

- S_p= Phonetic similarity score
- S_s= String similarity score
- S_e= Semantic similarity score
- W_p, W_s, W_e= weights

VI. IMPLEMENTATION DETAILS

A. Technology Stack

The proposed framework for calculating the hybrid similarity is coded using the Python programming language due to its support for Natural Language Processing and machine learning tasks. Python is a widely used language for Natural Language Processing tasks due to its extensive support for them.

NLP-related functionalities such as text preprocessing, tokenization, and stop words are implemented by utilizing general-purpose NLP libraries. Phonetic similarity calculation is achieved by utilizing libraries that support Double Metaphone encoding schemes, and string similarity calculation is achieved by utilizing optimized string matching and vectorization libraries. To calculate semantic similarity, transformer-based sentence embedding models are utilized from general-purpose deep learning libraries. The semantic similarity module uses the Sentence-BERT model 'all-MiniLM-L6-v2' due to its "trade-off between computational efficiency and semantic representation quality."

The modular structure of the tech stack makes it easy to update or replace individual components without affecting the overall workflow of the system.

The system also provides a feature for suggesting alternative publication titles that are less similar to existing ones. Instead of relying on fixed patterns alone, it uses a combination of structured approaches and language-based generation techniques to create suggestions that are both meaningful and relevant to the input. This helps users explore better title options and increases the chances of selecting a unique title, thereby reducing the likelihood of rejection during the registration process.

B. Dataset Description

The system was evaluated using a dataset of approximately 10,000 registered publication titles, which act as the reference base for comparison. Whenever a new title is submitted, it is checked against this dataset to identify possible similarities or duplication. The dataset contains a wide range of title variations, including differences in spelling, word order, phrasing, and structure. This helps in simulating realistic conditions where publication titles may not always follow a consistent format. In addition to the main dataset, a smaller set of test cases was created to represent common scenarios such as spelling errors, phonetic similarities, semantic resemblance, and the use of restricted terms. Although the dataset is not extremely large when compared to industrial-scale systems, it is sufficiently diverse to reflect real-world title variations. This

makes it suitable for evaluating the effectiveness of both individual similarity techniques and the proposed hybrid approach. The increased dataset size compared to initial experiments also improves the reliability of the results and provides better insight into system performance.

C. System Configuration

The system setup comprises of similar thresholds and weights that are applied to the computation of the hybrid similarity score. The similarity thresholds are predetermined with references to empirical evidence and apply to define the low, medium, and high similarity based on the system behaviour. Moreover, weights applied in the phonetic, string and semantic similarity modules are modified so that accuracy and efficiency are balanced.

The implementation is carried out on a standard desktop computing environment, which has a modern multi-core processor and enough memory support for the computation of the transformer-based embedding. It also contains a Python runtime environment and the necessary third party libraries. This demonstrates the fact that the proposed system can be successfully used in practice as regulatory application because it can be operated without any special hardware support.

VII. EXPERIMENTAL RESULTS AND EVALUATION

A. Evaluation Metrics

The evaluation criterion of the proposed system comprises both of the qualitative and quantitative indication that is suitable to make an assessment of the short text similarities. The evaluation criteria involve the similarity score which is obtained by combining the phonetic, string and semantic similarity. The similarity score is a quantitative parameter that shows the degree of similarity of the input title to registered titles.

In addition, if applicable, precision and recall are also considered to evaluate the system in terms of the capacity of the system in identifying the titles correctly and reducing the rate of false detection. Precision, in this context, is associated with the rate of correctly identified similar titles, whereas recall is associated with the rate of relevant similar titles that are correctly identified. Since the evaluation set is of a relatively small size, these metrics are considered to gain insights on the system rather than for validation.

B. Case Study Examples

To demonstrate the application of the proposed framework, some specific case studies of publications are analysed using actual and simulated publication title examples. The publication title examples include some with minor spelling differences, some with phonetic similarity, and some with semantic similarity. In some of these cases, it was noticed that the verification methods based on keywords could not detect similarity due to lexical differences, while the proposed hybrid system was able to detect similarity based on meaningful resemblance.

The risk classification mechanism also assists in the improvement of the interpretability of the system since the titles are classified as Low, Medium, or High-risk. Those titles that have the least similarity and have no regulatory concerns are considered low risk, whereas those titles that have a high similarity or limiting terms are considered to be high risk.

The prototype is also useful in the generation of alternative low similarity title suggestions thus assisting the applicants in choosing the appropriate title to be published.

C. Performance Analysis

The experimental outcomes indicate that the hybrid similarity approach is more efficient as compared to the similarity approaches applied independently. The system is able to detect more similarity patterns when phonetic similarity, string similarity, and semantic similarity are used in combination to ensure that the keyword that is typed by the user is verified. The optimized system has the capability of giving close real time response, which is required in interactive applications in the regulation sphere.

The proposed system is also more sensitive to subtle similarity situations, compared to the keyword-only verification method, in addition to meeting computational efficiency. While it is true that the computation of semantic similarities has a greater computational overhead, it is also clear that the selective application of the concept in the proposed system has achieved a balance in terms of efficiency. It can thus be concluded that the proposed system, as a hybrid solution in publication title verification,

offers a more reliable solution compared to traditional keyword-based approaches.

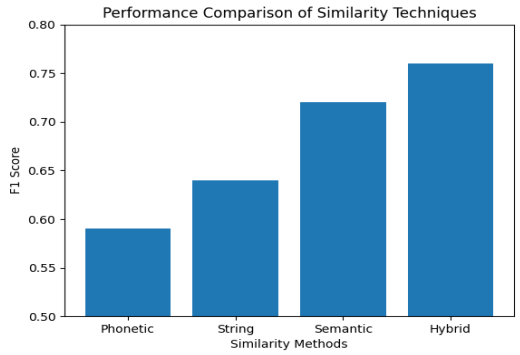

Fig. 2. Performance comparison of similarity techniques based on F1-score

D. Results

TABLE II Performance Comparison of Similarity Methods

Method	Precision	Recall	F1 Score
Phonetic Only	0.61	0.58	0.59
String Only	0.66	0.63	0.64
Semantics Only (SBERT)	0.74	0.70	0.72
Proposed Hybrid	0.73	0.80	0.76

VIII. DISCUSSION

A. Key Findings

The study demonstrates that the verification of publication title is a multidimensional issue that requires the assessment of similarities, which cannot be done using a single technique. The results obtained in the experiments show that the phonetics-based techniques and the strings-based techniques and semantics-based techniques can recognize different aspects of title similarities. Phonetics-based techniques can recognize phonetic similarities, the strings-based techniques can recognize typographical similarities, and the semantics-based techniques can recognize semantic similarities. The application of a hybrid technique that combines all three techniques can recognize all aspects of title similarities. The application of rule-based validation can increase the suitability of the system to assist in regulatory compliance.

B. Advantages of Hybrid Approach

The hybrid similarity framework has several advantages when applied in the context of regulatory title verification. The combination of different similarity paradigms minimizes the possibility of false matches, which is a common problem with other similarity methods that rely on keywords alone. The risk-based classification approach is also beneficial, as it provides the human reviewer with information that is not just binary, which is an advantage in the context of transparency. The flexibility of the framework is an additional advantage, as it is easily adaptable without major changes in the architecture.

C. Limitations of the Current System

Despite the advantages, the proposed system also has some disadvantages, as the effectiveness of the proposed hybrid approach relies on the weights and thresholds, which are empirically tuned. There is also an additional computational overhead of the semantic similarity approach, especially when dealing with large title repositories. The evaluation of the proposed approach has also been done on a small set of data, limiting the ability to generalize the performance of the approach over various publication domains. The multilingual and cross-lingual capabilities of the proposed approach also need to be improved to cover the linguistic diversity of India. The performance of the proposed approach also relies on the empirically tuned weights for the hybrid approach, and the effectiveness may vary across the languages or scripts that are not well represented in the embedding model.

IX. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper proposed a survey and hybrid similarity approach to the automated process of verifying publication titles. The proposed approach combines phonetic, lexical, and semantic approaches to offer a more comprehensive approach to verifying publication titles than each of the individual approaches on its own. The proposed approach also shows the potential of AI-assisted governance

systems, where hybrid AI models can be used to make regulatory decisions.

B. Future Enhancements

The future of this work will be focused on furthering this framework for addressing more comprehensive linguistic and operational needs. One area of extension for this framework is furthering the multilingual support of this title processing system. This means that it should be possible for the system to process titles in other languages and scripts used in India. This could be further enhanced through the inclusion of cross-lingual embedding models for more effectively determining semantic similarity. Furthermore, the framework could be further optimized for wider applicability by including more efficient mechanisms for handling increased title sets.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[2] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in Proc. IJCAI Workshop on Information Integration on the Web

[3] D. Jurafsky and J. H. Martin, Speech and Language Processing, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2009.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.

[6] R. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin, Germany: Springer, 2012.

[7] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[8] K. Kukich, "Techniques for automatically correcting words in text," ACM Computing Surveys, vol. 24, no. 4, pp. 377–439, 1992.

[9] A. L. Phillips, "Hanging on the metaphone," Computer Language, vol. 7, no. 1, pp. 12–15, 1990.

[10] M. Zaharia et al., "Accelerating the machine learning lifecycle with MLflow," IEEE Data Engineering Bulletin, vol. 41, no. 4, pp. 39–45, 2018.