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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14470

Title Uniqueness Verification System Using NLP for Ensuring Originality and Compliance

Anvee Deshpande[1], Suchita Kulkarni[2], Kaveri Ganesh[3], Swati Kamble[4], Prof. Shubhangi Pawar[5]

TSSM's Bhivarabai Savant College of Engineering and Research, Pune[1-5]

Abstract: Ensuring the originality and appropriateness of titles is crucial in academic research, project submissions, and business naming. The Title Uniqueness Verification System leverages Natural Language Processing (NLP) to provide an automated solution for title validation. Developed as a Python Flask-based web application, the system allows users to register, log in, and submit titles through an intuitive interface. Upon submission, the system employs the Cosine Similarity algorithm to compare the submitted title with a dataset of existing titles. If the similarity score exceeds a predefined threshold, the title is rejected, preventing duplication and potential plagiarism. Additionally, the system integrates a keyword filtering mechanism to identify and reject titles containing disallowed or restricted words, ensuring compliance with specific content standards. This real-time, automated verification method helps researchers, academic institutions, and organizations maintain originality and adhere to content guidelines, significantly reducing redundancy and enhancing the integrity of title submissions.

Keywords: Title Verification, Natural Language Processing (NLP), Cosine Similarity, Plagiarism Detection, Keyword Filtering, Flask Web Application, Title Originality, Automated Title Validation, Academic Integrity, Content Compliance.

I. INTRODUCTION

Oral In various fields such as academic research, project submissions, and business naming, ensuring the originality and appropriateness of titles is essential. Titles serve as the first impression of a work and play a crucial role in conveying the essence of the content. However, the increasing volume of submissions has led to challenges in maintaining uniqueness and preventing duplication. Manual verification methods are time-consuming and prone to errors, necessitating the development of an automated solution for title validation.

This paper introduces the Title Uniqueness Verification System, a Natural Language Processing (NLP)-based web application designed to assess the originality and compliance of submitted titles. The system is implemented using Python Flask and provides a user-friendly interface for users to register, log in, and submit titles for verification. The system utilizes the Cosine Similarity algorithm to compare the submitted title against a database of existing titles. If the similarity score surpasses a predefined threshold, the title is flagged as a duplicate and rejected, thereby reducing the risk of plagiarism and redundancy. Additionally, the system integrates a keyword filtering mechanism to identify and reject titles containing disallowed or restricted words, ensuring adherence to specific content guidelines.

II. RELATED WORK

The Title verification and originality assessment have been widely studied in the fields of Natural Language Processing (NLP), plagiarism detection, and text similarity analysis. Various techniques have been developed to address issues related to duplicate content detection, text matching, and compliance with predefined standards. This section reviews existing approaches and methodologies relevant to title uniqueness verification.

Several plagiarism detection systems have been implemented using NLP techniques, particularly Cosine Similarity, Jaccard Similarity, and Latent Semantic Analysis (LSA). Works such as Turnitin and Copyscape employ text-matching algorithms to compare input text with a large database of existing documents, helping in identifying duplicate content.

Similarly, Google Scholar and Cross Ref utilize text similarity techniques to detect overlapping content in academic papers. However, these systems primarily focus on full-text plagiarism detection rather than title uniqueness verification.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14470

In the domain of text similarity measurement, studies have demonstrated the effectiveness of vector space models and word embeddings in comparing textual data. Mikolov et al. (2013) introduced Word2Vec, an approach that represents words in continuous vector space, enabling more accurate semantic comparisons. Further advancements such as BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) have improved NLP-based similarity detection by capturing contextual meaning. While these deep learning techniques enhance semantic matching, they require significant computational resources, making them less suitable for lightweight web applications like the proposed system.

Keyword filtering has also been explored in various content moderation and compliance systems. Blacklisting and regular expression-based filtering methods are commonly used in applications like spam detection, comment moderation, and search engine filtering to restrict unwanted words or phrases. Studies in content moderation highlight the importance of such filtering techniques in maintaining ethical and regulatory standards.

Unlike the aforementioned approaches, the Title Uniqueness Verification System focuses on a lightweight yet effective solution by combining Cosine Similarity for duplication detection and keyword filtering for content compliance. By integrating these methods within a Flask-based web application, the system provides a real-time, user-friendly platform for verifying title originality. This approach fills the gap between large-scale plagiarism detection tools and simple text-matching techniques, offering a tailored solution for academic institutions, researchers, and businesses.

Thus, this work builds upon existing research in NLP-based text similarity analysis, plagiarism detection, and keyword filtering, while proposing an optimized framework specifically designed for title verification in various professional and academic domains.

III. METHODOLOGY

The Title Uniqueness Verification System is designed as a Python Flask-based web application that employs Natural Language Processing (NLP) techniques to verify the originality and compliance of submitted titles. The system follows a structured approach involving data preprocessing, similarity analysis, and keyword filtering to ensure accurate and efficient verification. The methodology consists of the following key steps.

3.1 System Architecture

The system architecture comprises three main components:

- **User Interface (Frontend):** A web-based interface allowing users to register, log in, and submit titles for verification.
- **Backend Processing:** A Flask-based application that handles title processing, similarity computation, and keyword filtering.
- **Database:** A structured dataset storing existing titles for similarity comparison and a predefined list of restricted words.

3.2 Data Preprocessing

Before conducting similarity analysis, the submitted title undergoes **preprocessing** to improve the accuracy of comparisons. The preprocessing steps include:

- **Lowercasing:** Converting the title to lowercase to maintain consistency.
- **Tokenization:** Splitting the title into individual words or phrases.
- **Stopword Removal:** Eliminating common words (e.g., "the," "and," "of") that do not contribute to meaning.
- **Stemming/Lemmatization:** Reducing words to their root form (e.g., "running" → "run").

3.3 Title Similarity Analysis Using Cosine Similarity

The system uses **Cosine Similarity**, a common NLP technique, to measure the similarity between the submitted title and existing titles stored in the database. The steps involved are:

1. **Text Vectorization:** Converting the preprocessed titles into numerical vectors using the **TF-IDF (Term Frequency- Inverse Document Frequency) method**.
2. **Cosine Similarity Calculation:** Measuring the cosine angle between the submitted title's vector and each stored title's vector.
3. **Threshold Comparison:** If the similarity score exceeds a predefined threshold (e.g., 0.8), the title is considered non- unique and is rejected.

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14470

3.4 Keyword Filtering for Content Compliance

To ensure titles adhere to specific guidelines, the system performs **keyword filtering** using a predefined list of **restricted or disallowed words**.

- The submitted title is scanned for the presence of any **blacklisted words** stored in the database.
- If a restricted word is detected, the title is automatically rejected, prompting the user to modify it.

3.5 Decision Module & User Feedback

- If the title is unique and free of restricted words, it is approved and stored in the database for future comparisons.
- If the title is too similar to an existing title, the system provides real-time feedback suggesting modifications.
- If the title contains restricted words, an alert is displayed, prompting the user to revise the submission.

3.6 Deployment & User Accessibility

- The system is deployed on a Flask web server, enabling real-time accessibility.
- Users can access the platform via a web interface, ensuring a seamless experience for academic institutions, researchers, and organizations.

IV. EXPERIMENTAL RESULTS

The To evaluate the effectiveness of the **Title Uniqueness Verification System**, a series of experiments were conducted using a dataset of existing titles from academic research, project submissions, and business naming conventions. The experiments focused on measuring the system's accuracy in detecting duplicate titles, identifying restricted words, and providing real-time feedback to users.

4.1 Dataset and Experimental Setup

- The system was tested using a dataset of **5,000 existing titles** collected from academic papers, project reports, and business name registries.
- A **test set** of **500 new titles** was used for validation, including both unique and intentionally duplicated or modified titles.
- The similarity threshold for duplicate detection was set at **0.8** (80%), meaning titles exceeding this threshold were considered too similar to existing ones.
- A list of **50 restricted words** was predefined to evaluate keyword filtering accuracy.

4.2 Performance Metrics

To assess the system's effectiveness, the following **performance metrics** were used:

- **Precision (P)** = TP / (TP + FP) → Measures how many detected duplicates were correctly identified.
- **Recall (R)** = TP / (TP + FN) → Measures how many actual duplicates were detected.
- **F1-Score** = 2 × (P × R) / (P + R) → Harmonic mean of precision and recall.
- **Execution Time** → Measures the average time taken for title verification.

4.3 Results and Analysis

Metric	Value
Duplicate Detection Accuracy	92.5%
Precision	91.8%
Recall	93.2%
F1-Score	92.5%
Keyword Filtering Accuracy	98.6%
Average Execution Time	0.87 seconds

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14470

- The system successfully identified **duplicate titles with an accuracy of 92.5%**, demonstrating high reliability in preventing title repetition.
- The **keyword filtering accuracy reached 98.6%**, correctly rejecting titles containing restricted words with minimal false positives.
- The average execution time of **0.87 seconds** ensured real-time processing, making the system practical for live web applications.

4.4 Case Study Examples

1. **Title Similarity Detection:**
   - **Submitted Title:** *"Machine Learning-Based Fraud Detection in E-Commerce"*
   - **Existing Title:** *"Fraud Detection in E-Commerce Using Machine Learning"*
   - **Cosine Similarity Score: 0.91** → Rejected as a duplicate

2. **Keyword Filtering Test:**
   - **Submitted Title:** *"Automated Content Moderation Using AI"*
   - **Restricted Word Found:** *"Moderation"*
   - **Result:** Rejected due to non-compliance

3. **Unique Title Acceptance:**
   - **Submitted Title:** *"A Novel Approach to Renewable Energy Optimization"*
   - **Cosine Similarity Score:** 0.45 → Accepted

4.5 Discussion

The experimental results demonstrate that the Title Uniqueness Verification System effectively prevents title duplication while maintaining high-speed performance. The Cosine Similarity algorithm proved reliable for text comparison, and the keyword filtering module ensured compliance with predefined content standards. The system provides real-time feedback, improving the efficiency of title submissions in academic, research, and business domains.

These results validate the practicality and efficiency of the system, confirming its usefulness for institutions and organizations aiming to maintain originality and compliance in title submissions.

V. CONCLUSION OF EXPERIMENTAL RESULTS

The experimental results validate the effectiveness of the Title Uniqueness Verification System in ensuring originality and compliance in title submissions. The system demonstrated high accuracy (92.5%) in detecting duplicate titles, effectively preventing redundancy and plagiarism. Additionally, the keyword filtering accuracy (98.6%) ensured that restricted words were correctly identified, maintaining adherence to content standards.

The precision (91.8%) and recall (93.2%) scores highlight the system's ability to balance correctly identifying duplicate titles while minimizing false positives and negatives. Furthermore, the real-time processing speed (0.87 seconds per query) makes the system highly practical for large-scale applications in academic, research, and business domains.

Case studies confirmed the system's capability to accurately reject duplicate or non-compliant titles while accepting unique ones, reinforcing its reliability. These results demonstrate that the Title Uniqueness Verification System is a robust, automated, and efficient solution for ensuring title originality, compliance, and standardization.

Future improvements could involve enhancing NLP models for better semantic understanding, integrating deep learning techniques, and expanding the database of existing titles to further refine accuracy and applicability.

REFERENCES

[1]. Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing (3rd ed.)*. Stanford University. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

[2]. Manning, C. D., Raghavan, P., & Sch□tze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

DOI: 10.17148/IJARCCE.2025.14470

[3]. Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.

[4]. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science*, **41**(6), 391□407.

[5]. Mihalcea, R., Corley, C., & Strapparava, C. (2006). "Corpus-based and Knowledge-based Measures of Text Semantic Similarity." *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 775□780.

[6]. Wu, Z., & Palmer, M. (1994). "Verb Semantics and Lexical Selection." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 133□138.

[7]. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hern□ndez, L. (2014). "Syntactic N-grams as Machine Learning Features for Natural Language Processing." *Expert Systems with Applications*, **41**(3), 853□860.

[8]. Singh, A., & Sorokin, A. (2020). "Using Cosine Similarity for Text Clustering and Duplicate Detection." *International Journal of Artificial Intelligence Research*, **4**(2), 12□18.

[9]. Hoad, T. C., & Zobel, J. (2003). "Methods for Identifying Versioned and Plagiarized Documents." *Journal of the American Society for Information Science and Technology*, **54**(3), 203□215.

[10]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135□1144.