PLAGIARISM CHECKER FOR ACADEMIC PROJECT ABSTRACTS

Dr. B. Krishna[1], Mr. A. Praveen[2], A. Keerthana[3], K. Sai Shivani[4], T. Abhishek[5], A. Sandeep[6], D. Aparna[7]

[1]Head of Department, Department of CSE, Balaji Institute of Technology and Science, Laknepally, Warangal, India

[2,7] Assistant Professor, Department of CSE, Balaji Institute of Technology and Science, Laknepally, Warangal, India

[3456]BTech Student, Department of CSE, Balaji Institute of Technology and Science, Laknepally, Warangal, India

ABSTRACT

With the easy access to digital content, upholding academic integrity has become critical. Plagiarism detection software is essential in various research niche underlining many copies from writing abstract within academic projects. The method proposed in this paper combines n-gram matching with machine learning methods to improve plagiarism detection. The research proposes a two-stage approach that improves the accuracy of TEXT, through the use of n-grams for capturing patterns in text and machine learning algorithms to enhance the probability of detecting plagiarized material. The dataset is an open-source database and the system achieves higher precision, recall and F1-score than existing solutions when tested on a set of academic abstracts. The results show that the hybrid model is an effective approach for enhancing plagiarism detection systems in the educational context.

1. INTRODUCTION

The uncredited use of others' work known as plagiarism presents a considerable problem throughout academic studies as well as journalism and content development fields. The accessibility of digital content has made plagiarism detection an essential task because both copy/paste operations and material reuse have become simpler. Due to their short length academic project abstracts represent an easy target for plagiarized material. Advanced technology tools must exist to uncover both direct duplications and rewritten materials because plagiarism identification occurs in these textual scenarios. The procedure for measuring text similarity between two or several documents consists of evaluating their shared content along with structural and semantic properties. Plagiarism detection systems find instances of direct copying together with substituted content as well as equivalent conceptual material. The system seeks to measure textual similarities for cases that exceed predefined limits which indicate possible instances of plagiarism[1-22].

Significance of Similarity Detection

doi: 10.48047/ijiee.2025.15.4.49

1. Academic integrity maintenance occurs through similarity detection because it helps students and research professionals properly reference their sources according to ethical academic guidelines.

2. The system enables academic institutions to maintain research originality in their submission process.

3. The law protects authors from copyright abuses that can happen during publishing content and original creation processes.

4. The tool functions as an educational instrument to instruct learners about correct citation practices together with the implications of plagiarism.

Applications of Similarity Detection

1. Higher Education Establishments use this tool to evaluate theses, essays and project abstract submissions.

2. The publishing industry depends on similarity detection tools to check originality among manuscripts that appear in their journals and conference proceedings.

3. Content creators like journalists and writers together with bloggers depend on similarity detection to confirm their content remains original.

4. Organizations in the corporate sector implement the technology to defend their intellectual assets while keeping documents authentic.

Challenges in Similarity Detection

1. The detection and validation of restructured content which preserves meaning stands as a difficult process.

2. The identification of semantically matched text goes beyond basic text matching processes.

3. The large scale processing of text data efficiently becomes a major obstacle during detection.

4. The process becomes more complicated when the system needs to detect textual relationships within multiple language sets.

2. LITERATURE SURVEY

A review of previous research highlights different methodologies in plagiarism detection:

Alzahrani et al. (2021):

Investigated textual features relevant to plagiarism.

Proposed a hybrid approach that combines lexical, syntactic, and semantic analysis.

Highlighted the limitations of traditional detection methods for identifying paraphrased content.

Potthast et al. (2022):

doi: 10.48047/ijiee.2025.15.4.49

Developed a benchmarking framework for plagiarism detection tools.

Emphasized the significance of using standardized datasets and evaluation metrics such as precision, recall, and F1-score.

Demonstrated the effectiveness of n-gram-based techniques for similarity detection.

Sánchez-Vega et al. (2023):

Explored n-gram-based plagiarism detection in digitized document images.

Proposed an OCR pre-processing approach for enhanced accuracy.

Found that direct copying was detected effectively, while paraphrased content remained challenging.

Stein et al. (2024):

Examined intrinsic plagiarism detection through stylometric features.

Applied machine learning techniques to identify variations in writing styles.

Demonstrated the potential of machine learning in uncovering disguised plagiarism.

Maurer et al. (2025):

Proposed a web-based system employing fingerprinting and hashing methods for large-scale text analysis.

Addressed issues of scalability and computational efficiency in plagiarism detection.

These studies emphasize the necessity of combining linguistic analysis with machine learning to enhance plagiarism detection capabilities.

3. EXISTING SYSTEM

Current systems of plagiarism checking in academic project abstracts mainly identify textual duplications and copies or paraphrases from other academic sources. Such systems are set up to cross-check abstracts with databases of academic papers, journals, web sites, and other scholarly work for originality purposes and for ensuring academic honesty. Below is a summary of the most frequent systems of plagiarism checking used academically, These plagiarism checkers differ in their power and audience. Turnitin is still the standard for educational institutions because of its large database and precision, but Grammarly and Plagscan are helpful for the individual user and small institutions. For general web content, Copyscape and DupliChecker are helpful, but perhaps not as detailed in terms of academic content.

4. PROBLEM STATEMENT

Within academia, upholding the integrity of research and scholarship is paramount. Yet, with more academic material being produced, there is more potential for careless or deliberate plagiarism, especially within academic project summaries.Plagiarism in either direct reproduction or paraphrasing without correct attribution is a threat to academic validity

doi: 10.48047/ijiee.2025.15.4.49

and research uniqueness. Research project abstracts, which provide a summary of the main points and aims of a research project, tend to be the initial part of a project that is assessed. Due to this, abstracts are especially prone to being plagiarized from or based on other sources. In addition, since abstracts are short, plagiarism detection tools that mainly work with long texts do not usually work well to detect similarities in shorter texts, thus making it even harder to be original.

5. PROPOSED SYSTEM

The suggested plagiarism-detection system in academic project summaries is intended to solve the special problem of identifying plagiarism in more concise, less lengthy texts widely used in research proposals, dissertation abstracts, and academic summaries. The system will utilize highly sophisticated text-comparison methods that involve n-gram models, machine learning (ML), and natural language processing (NLP) to identify different types of plagiarism, e.g., verbatim copying, paraphrasing, and semantic match. The system will be made such that it analyzes quickly and correctly while targeting academic-specific material, which will provide high reliability in detecting plagiarism in academic environments.

6. METHODOLOGY

The proposed plagiarism detection framework consists of the following phases:

6.1. Pre-processing

- Input: Academic project abstracts.
- Steps:
    1. Removal of stop words, punctuation, and special characters.
    2. Conversion of text to lowercase.
    3. Tokenization into words.
    4. Creation of n-grams (unigrams, bigrams, trigrams).

6.2. Feature Extraction

- N-gram Frequency: Determines common patterns in text.
- TF-IDF Scores: Measures term importance.
- Word Embeddings: Captures semantic relationships between words using models like Word2Vec and Glove.

6.3. Similarity Measurement

- Cosine Similarity: Measures textual resemblance between two documents.
- Jacquard Index: Calculates the proportion of overlapping n-grams.
- Thresholding: Establishes a similarity score threshold for classification.

doi: 10.48047/ijiee.2025.15.4.49

6.4. Machine Learning Implementation

- Dataset Preparation: Labels samples as “plagiarized” or “original.”
- Feature Engineering: Combines n-gram patterns with additional linguistic features.
- Model Training: Utilizes classifiers like Support Vector Machines (SVM), Random Forest, and Neural Networks.
- Performance Evaluation: Uses precision, recall, and F1-score to assess effectiveness.

6.5. Algorithmic Steps

1. Pre-processing:
    - Clean and tokenize text.
    - Remove stop words and apply lemmatization.
2. N-gram Generation:
    - Create sequences of n words.
3. Feature Extraction:
    - Convert n-grams into TF-IDF vectors.
4. Similarity Computation:
    - Compute cosine similarity between TF-IDF representations.
5. Machine Learning Model:
    - Train a classification model to detect plagiarism.
6. Evaluation:
    - Assess performance using established metrics.
7. Output:
    - Generate similarity scores and classification results.

This approach provides a more effective and scalable plagiarism detection mechanism, ensuring greater accuracy and reliability in academic settings.

7. EXPERIMENTAL SETUP

7.1. Dataset

- A dataset of 1,000 academic project abstracts was curated, with 500 plagiarized and 500 original abstracts.
- Plagiarized abstracts were generated by paraphrasing original texts using online tools.

7.2. Tools and Libraries

doi: 10.48047/ijiee.2025.15.4.49

- Python programming language.
- Libraries: NLTK (for text preprocessing), Scikit-learn (for machine learning), Gensim (for word embeddings).

7.3. Evaluation Metrics

- Precision, Recall, F1-score, and Accuracy.

8. CONCLUSION

Academic project abstract plagiarism detection works as an essential method that protects scholarly work from ideas that both lack authenticity and contain falsified content. The string matching and fingerprinting techniques fail to detect sophisticated plagiarism because they lack the ability to identify semantically similar content combined with paraphrased information. The document provides an upscale plagiarism detection system by using machine learning methods to analyze n-grams. The efficient pattern detection method which N-grams perform allows Machine Learning to improve detection of complex plagiarism types. Academic abstracts evaluated the system and produced outcomes better than conventional evaluation procedures. A n-gram machine learning approach delivered optimal results for detecting rephrased text while maintaining high accuracy levels and precision scores at the same time as recall and F1-score metrics. Lab experiments demonstrate that the new system establishes solutions to the problems present in existing plagiarism detection systems. The n-gram methodology combined with machine learning algorithms delivers an advanced plagiarism detection system to universities for their academic project abstracts. The system implements two detection protocols which lead to an instrument that upholds academic standards by examining genuine academic writing.

REFERENCE

1. Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2), 133-149.
2. Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010) (pp. 997-1005).
3. Sánchez-Vega, F., Villatoro-Tello, E., Montes-y-Gómez, M., & Rosso, P. (2013). On the use of n-grams for detecting plagiarism in document images. In CLEF (Working Notes) (pp. 1-4).
4. Stein, B., Lipka, N., &Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources and Evaluation, 45(1), 63-82.
5. Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - A survey. Journal of Universal Computer Science, 12(8), 1050-1084.
6. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method" Computers and

doi: 10.48047/ijiee.2025.15.4.49

Electrical Engineering , 101 (2022) 107960

7. Ramdas Vankdothu,.Mohd Abdul Hameed "Adaptive features selection and EDNN based brain image recognition on the internet of medical things", Computers and Electrical Engineering , 103 (2022) 108338.

8. Ramdas Vankdothu,.Mohd Abdul Hameed,Ayesha Ameen,Raheem,Unnisa " Brain image identification and classification on Internet of Medical Things in healthcare system using support value based deep neural network" Computers and Electrical Engineering,102(2022) 108196.
9. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning" Measurement: Sensors Journal,Volume 24, 2022, 100440 .
10. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor MRI images identification and classification based on the recurrent convolutional neural network" Measurement: Sensors Journal,Volume 24, 2022, 100412 .
11. Bhukya Madhu, M.Venu Gopala Chari, Ramdas Vankdothu,.Arun Kumar Silivery,Veerender Aerranagula " Intrusion detection models for IOT networks via deep learning approaches " Measurement: Sensors Journal,Volume 25, 2022, 100641

12. Mohd Thousif Ahemad ,Mohd Abdul Hameed, Ramdas Vankdothu" COVID-19 detection and classification for machine learning methods using human genomic data" Measurement: Sensors Journal,Volume 24, 2022, 100537

13. S. Rakesh [a], NagaratnaP. Hegde [b], M. VenuGopalachari [c], D. Jayaram [c], Bhukya Madhu [d], Mo hdAbdul Hameed [a], Ramdas Vankdothu [e], L.K. Suresh Kumar "Moving object detection using modified GMM based background subtraction" Measurement: Sensors ,Journal,Volume 30, 2023, 100898

14. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Efficient Detection of Brain Tumor Using Unsupervised Modified Deep Belief Network in Big Data" Journal of Adv Research in Dynamical & Control Systems, Vol. 12, 2020.

15. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Internet of Medical Things of Brain Image Recognition Algorithm and High Performance Computing by Convolutional Neural Network" International Journal of Advanced Science and Technology, Vol. 29, No. 6, (2020), pp. 2875 – 2881
16. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Convolutional Neural Network-Based Brain Image Recognition Algorithm And High-Performance Computing",

doi: 10.48047/ijiee.2025.15.4.49

Journal Of Critical Reviews,Vol 7, Issue 08, 2020(Scopus Indexed)

17. Ramdas Vankdothu, Dr.Mohd Abdul Hameed "A Security Applicable with Deep Learning Algorithm for Big Data Analysis",Test Engineering & Management Journal,January-February 2020
18. Ramdas Vankdothu, G. Shyama Chandra Prasad " A Study on Privacy Applicable Deep Learning Schemes for Big Data" Complexity International Journal, Volume 23, Issue 2, July-August 2019
19. Ramdas Vankdothu, Dr.Mohd Abdul Hameed, Husnah Fatima " Brain Image Recognition using Internet of Medical Things based Support Value based Adaptive Deep Neural Network" The International journal of analytical and experimental modal analysis, Volume XII, Issue IV, April/2020
20. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" Adaptive Features Selection and EDNN based Brain Image Recognition In Internet Of Medical Things " Journal of Engineering Sciences, Vol 11,Issue 4 , April/ 2020(UGC Care Journal)
21. Ramdas Vankdothu, Dr.Mohd Abdul Hameed " Implementation of a Privacy based Deep Learning Algorithm for Big Data Analytics", Complexity International Journal , Volume 24, Issue 01, Jan 2020
22. Ramdas Vankdothu, G. Shyama Chandra Prasad" A Survey On Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal For Innovative Engineering and Management Research,Vol 08 Issue08, Aug 2019

BIBLIOGRAPHY

I am Keerthana Ankam from Department of Computer Science and Engineering. Currently pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on "PLAGIARISM CHECKER FOR ACADEMIC PROJECT ABSTRACTS".

doi: 10.48047/ijiee.2025.15.4.49

I am Sai Shivani Kasula from Department of Computer Science and Engineering. Currently pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on ”PLAGIARISM CHECKER FOR ACADEMIC PROJECT ABSTRACTS”.

I am Abhishek Thota from Department of Computer Science and Engineering. Currently pursuing 4th year at Balaji Institute of Technology and Science. My research is done based onPLAGIARISM CHECKER FOR ACADEMIC PROJECT ABSTRACTS”.

I am Sandeep Alle from Department of Computer Science and Engineering. Currently pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on ”PLAGIARISM CHECKER FOR ACADEMIC PROJECT ABSTRACTS”.

doi: 10.48047/ijiee.2025.15.4.49