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. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134107

PLAGIARISM DETECTION BASED ON MACHINE LEARNING

Snehal Golait[1], Priyanka Gupta[2], Niraj Sabre[3], Tanmay Pawar[4], Nishad Chaudhary[5], Nikhil Nirwan[6], Jivyani Bhave[7]

Professor, Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur, India[1]

Assistant professor, Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur India[2]

UG Student, Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur, India[3-7]

Abstract: Plagiarism poses a consequential challenge in academic and professional settings, requiring robust and efficient methods for detection. This study presents an innovative approach to plagiarism detection utilizing Machine Learning (ML) techniques. The proposed system leverages a diverse dataset containing both pristine and plagiarized documents, employing advanced feature extraction methods such as TF-IDF and word embeddings. The pre-processing phase involves cleaning and standardizing the text data, while feature extraction transforms documents into numerical representations felicitous for ML algorithms. Sundry ML models, including logistic regression and neural networks, are explored for their efficacy in binary relegation tasks. The system is trained on labeled datasets, distinguishing between pristine and plagiarized content. Extensive evaluations are conducted on the testing dataset, quantifying the model's precision, precision, recall, and F1-score. The study withal investigates the impact of different feature extraction techniques on the overall performance. The implementation incorporates genuine-world considerations, including the identification of variants of plagiarism, such as copy-pasting and paraphrasing. The system's adaptability to diverse domains and sources is accentuated, and scalability concerns are addressed to ascertain efficacious detection in sundry contexts.

Keywords: Paraphrase recognition, passage-level plagiarism detection, support vector machine.

I. INTRODUCTION

In the digital age, the ease of access to vast repositories of information has given rise to the pervasive issue of plagiarism, a practice that undermines the principles of academic and professional integrity. Plagiarism detection systems play a crucial role in upholding these principles by identifying instances of content misappropriation. Traditional methods, such as rule-based systems, have limitations in coping with the evolving nature of plagiarism. In response, this study introduces an innovative approach to plagiarism detection by harnessing the power of Machine Learning (ML) techniques. The ubiquity of digital content demands automated and adaptive systems capable of discerning intricate patterns and similarities within vast datasets. ML, with its ability to learn from data patterns and make predictions, offers a promising avenue for enhancing the accuracy and efficiency of plagiarism detection. This research explores the application of ML algorithms to the challenging task of distinguishing between original and plagiarized content, aiming to provide a more robust and scalable solution.

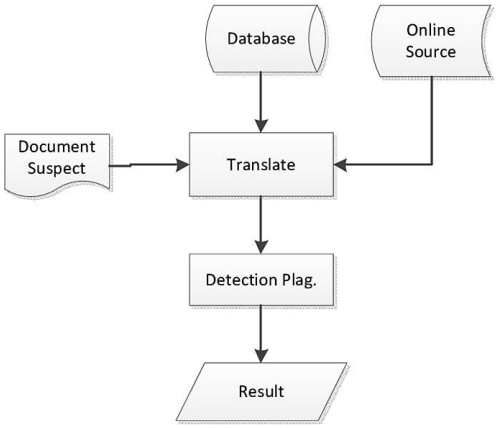


Figure 1. Block Diagram for Plagiarism Detector

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134107

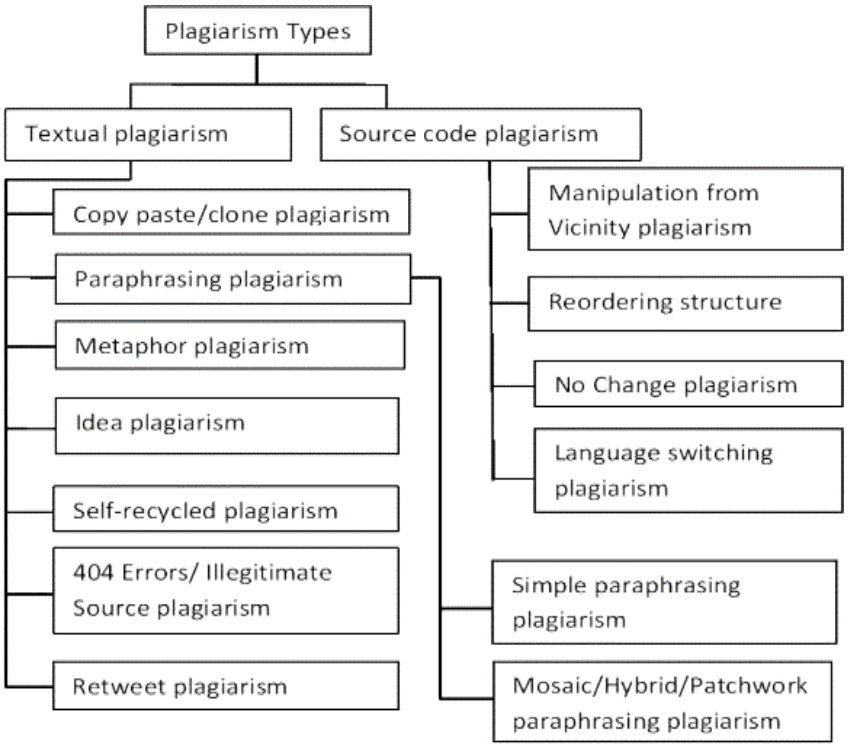
• **Motivation:** The motivation behind this study stems from the need for a dynamic and adaptive plagiarism detection system that can keep pace with the evolving tactics employed by plagiarists. Conventional methods often struggle to detect subtle forms of plagiarism, such as paraphrasing or intelligent rephrasing, necessitating a paradigm shift towards ML-driven solutions

.

• **Objective:** The primary objective of this research is to develop a plagiarism detection system that leverages ML techniques to autonomously analyze and categorize textual content. The system aims to go beyond mere rule- based comparisons and, instead, learn intricate patterns indicative of potential plagiarism, thus enhancing the accuracy and scope of detection.

• **Scope of the Study:** This study encompasses the exploration of various ML algorithms, including but not limited to logistic regression, support vector machines, and neural networks, to identify the most effective approach for plagiarism detection. The research also considers different feature extraction methods, such as TF- IDF and word embeddings, to capture the semantic nuances of text.

The remainder of this paper is organized as follows: Section 2 provides a review of related work in the field of plagiarism detection, highlighting existing methodologies and their limitations. Section 3 details the dataset used for training and testing the ML models. Section 4 delves into the methodology, outlining the steps involved in feature extraction, model training, and evaluation. Section 5, summarizes the findings, discusses implications, and proposes avenues for future research followed by the conclusion.



II. LITERATURE SURVEY

Plagiarism detectors conventionally characterize unstructured documents utilizing sundry categories of textual features such as lexical, syntactic, and semantic. The most popular lexical features are character and word n- grams, while components-of-verbalization (POS) information is utilized extensively to compute syntactic features. Semantic features depend on a thesaurus like WordNet to typify word relationships. Given a sizably voluminous document amassment, to retrieve the candidate source documents for matching against the suspicious document, traditional information retrieval techniques predicated on cosine homogeneous attribute, vector space model, and fuzzy retrieval may be utilized. Once the candidate documents are identified, they can be compared exhaustively utilizing techniques predicated on string matching, vector kindred attribute computation, syntax, and semantic, fuzzy, and structural feature-predicated methods. Semantic and fuzzy methods are more efficacious in detecting involute types of plagiarism including paraphrasing and restructuring besides the simpler copy–paste forms [1].

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134107

Clough et al. [10] have carried out some of the earliest experiments in the domain of text reuse detection and have withal constructed the METER (MEasuring TExt Reuse) corpus. The METER corpus consists of text articles amassed from the UK Press Sodality (PA) and nine British newspapers with the PA articles providing a source for the newspaper articles. The extent of text reuse has been assessed utilizing n-gram overlap, Avaricious String Tiling and sentence alignment. Clough and Stevenson have developed a corpus of plagiarized short answers [9] labeled as the Wikipedia Re-inscribe Corpus. The corpus was engendered for five questions from the Computer Science domain utilizing candidate answers engendered by participants, either independently or through the modification of the reference answer extracted from Wikipedia by sundry degrees. The kindred attribute was assessed in terms of n-gram overlap and longest mundane subsequence.

More recently the PAN-PC competitions have engendered considerable interest in this domain and have led to the development of several prosperous systems, which work on sizably voluminous-scale document amassments. Some of the approaches used include winnowing, hash function computation, finger-printing, and exact matching at sundry levels such as character-n-grams, word-n-grams, and sentences [15, 16]. Despite the immensely colossal number of plagiarism detection alternatives, the identification of paraphrased plagiarism has not been plenarily addressed [4]. As the quantity of lexical variation between the text units increases, plagiarism detection becomes tougher. Barron-Cedeno et al. have defined a typology of paraphrases comprising of 22 types predicated on the nature of changes such as morpho-lexicon-predicated, structural, semantic, and sundry. The authors of [4] have annotated a subset of the PAN-PC 2010 corpus according to their typology to engender the Paraphrasing for Plagiarism (P4P) corpus. The authors have additionally analyzed the performance of the PAN-PC 2010 competitors on the P4P corpus have optically canvassed that though the systems perform well on the entire PAN- PC 2010 corpus, they perform poorly on the P4P corpus, which involves extensive paraphrasing.

In an effort to fixate on paraphrased plagiarism, subsequent PAN competitions have introduced multiple cases of simulated plagiarism, which were engendered by workers on Amazon's Mechanical Turk by re-inditing pristine text content. While paraphrase apperception systems conventionally work on phrase-level or sentence-level inputs to determine semantic kindred attribute, plagiarism detection systems operate at the passage level [5]. Burrows et al. have adopted the crowdsourcing approach to engender paraphrased versions of text passages for constructing the Webis Crowdsourced Paraphrase Corpus (CPC). The corpus was pristinely developed as a component of the PAN 2010 competition to test the efficiency of plagiarism detection systems. The authors have withal assessed the performance of sundry paraphrase homogeneous attribute metrics for automatically filtering the engendered paraphrases. The metrics include normalized edit distance, n-gram comparison-predicated measures such as simple word n-gram overlap, BLEU metric, and longest mundane prefix n-gram overlap, besides the Sumo metric and sundry asymmetrical paraphrase detection functions proposed by Cordeiro et al. [11]. Burrows et al. have concluded that utilizing a cumulation of these metrics with a machine-learning classifier yields the best results [5].

Bar et al. have amalgamated three categories of features predicated on the content, structure, and style for quantifying text reuse [3]. Content-predicated features were engendered by comparing the text content and include string homogeneous attribute measures, acquisitive string tiling, word, and character n-gram features, Wordnet- predicated semantic homogeneous attribute measures besides latent semantic analysis and explicit semantic analysis. Structural kindred attribute was assessed in terms of word pair order, distance, as well as stop-word and POS n-grams. The stylistic homogeneous attribute was determined utilizing sentence, token length properties, function word frequencies, and lexicon richness measures such as sequential type-token ratio. The approach was tested on three different corpora, namely, Webis CPC, Wikipedia Re-indite Corpus, subset of METER corpus, and coalescing the three categories was found to yield the best results in two out of three cases. From the study of cognate work, it is visually examined that paraphrased plagiarism, though mundane, has not been addressed satisfactorily. Hence, there is a desideratum for efficient plagiarism detection approaches, which can handle paraphrased plagiarism.

III. PROPOSED SYSTEM

In this work, a machine learning-predicated paraphrase recognizer, which operates by extracting lexical, syntactic, and semantic features, has been used to detect plagiarism in text passages. The sentence-level paraphrase apperception system reported in [8] has been habituated for determining if two passages have been plagiarized.

Two different approaches have been investigated: in the first, the input source and suspicious passages have been split into sentences, and the pristine sentential paraphrase apperception system has been applied. In the second approach, the input passages have been retained as it is, and sundry features extracted from the passages have been used to judge whether the suspicious passage is a plagiarized version of the source.

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. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134107

The proposed system comprises three steps: pre-processing, detailed analysis, and post-processing. Firstly, according to basic natural language pre-processing, the suspicious and original documents are prepared into sentence and passage levels. Secondly, detailed analysis is responsible for extracting plagiarized cases by applying techniques like common n-grams, meteor scores, and intelligent deep-learning classification. Finally, post- processing is applied to find the best largest plagiarized segment by solving the overlapping issue, merging adjacent cases, and removing small cases.

A new database of lexical, syntactic, and semantic text similarity is created for the deep learning approaches, having 42 features for each similarity case. The constructed features' values are computed based on the similarity metrics of words and sentences from two benchmark datasets, that is PAN 2013 and PAN 2014. The constructed database trains the proposed system, and it is evaluated using the recall, precision, F-measure, granularity, and Plagdet measures. The performance of the proposed system based on LSTM has the first rank in PAN 2013 and PAN 2014 compared to the state-of-the-art systems.

IV. FLOW DIAGRAM

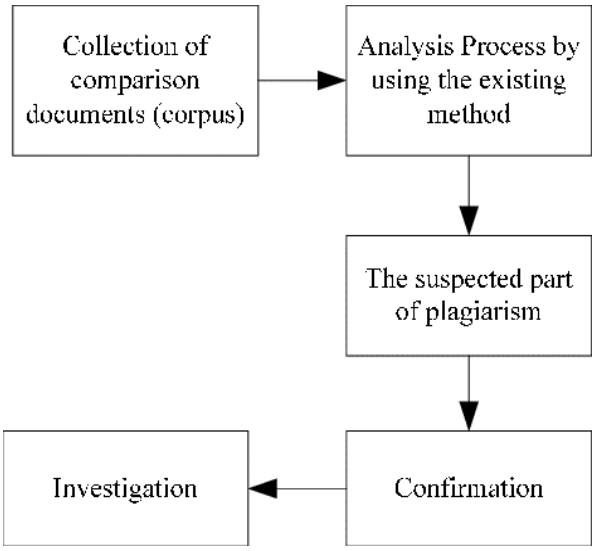


Figure 2. Flow Chart

V. PROPOSED METHODOLOGY

Building a plagiarism detector using machine learning involves several steps, from data preparation to model training and evaluation. Below are the steps to implement the project:

- **Data Collection**: Collect a dataset of documents that includes both original content and instances of plagiarism. Ensure that the dataset is diverse and representative of the type of content you expect to analyze. Each document should be labeled as either plagiarized or non-plagiarized.

- **Data Pre-processing**: Clean and pre-process the text data to make it suitable for machine learning. This may include:

a. Removing irrelevant information (e.g., formatting, metadata).
b. Tokenization: Breaking down text into individual words or tokens.
c. Removing stop words.
d. Lemmatization or stemming to reduce words to their base form.

- **Feature Extraction:** Convert the text data into numerical features that can be used as input for machine learning algorithms. Common methods include:

a. Bag-of-Words (BoW): Representing each document as a vector of word frequencies.
b. TF-IDF (Term Frequency-Inverse Document Frequency):Weighing the importance of words based on their frequency in the document and the corpus.
c. Word Embeddings: Use pre-trained word embeddings like Word2Vec, and GloVe, or train your embeddings.

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. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134107

- **Model Selection:**

a. Choose a machine learning algorithm suitable for text classification. Common choices include:
b. Logistic Regression
c. Naive Bayes
d. Support Vector Machines (SVM)
e. Neural Networks (e.g., LSTM, GRU for sequence data)

- **Model Training:** Split your dataset into training and testing sets. Train your selected model using the training set. During training, the model learns to identify patterns that distinguish between plagiarized and non-plagiarized text.

- **Model Evaluation:** Evaluate the performance of your model on the testing set using metrics such as accuracy, precision, recall, and F1 score. Make adjustments to your model or choose a different one if necessary

.

- **Hyperparameter Tuning:** Optimize the hyperparameters of your model to improve its performance. This may involve using techniques like cross-validation.

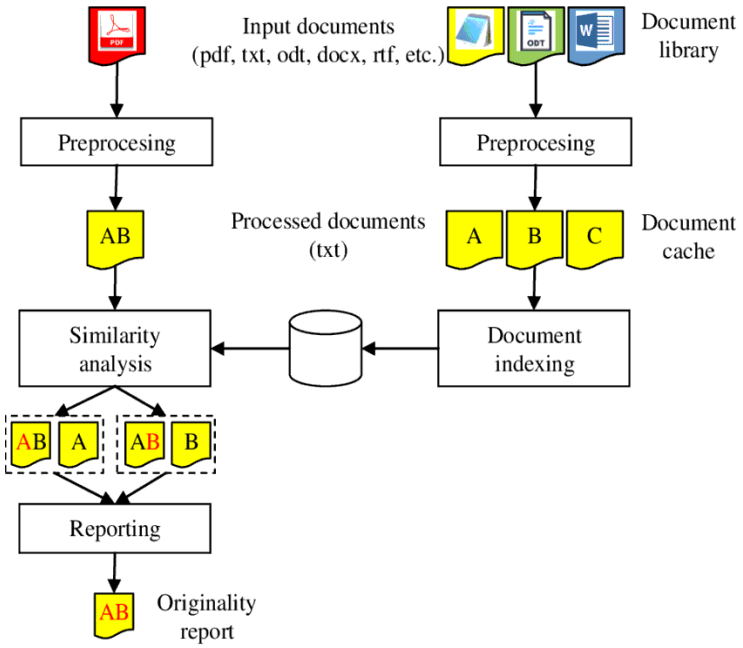
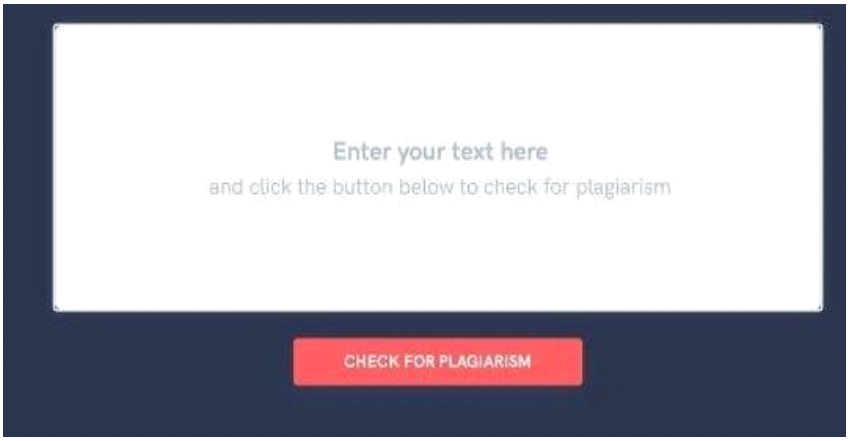


Figure 3. Working of the Modules

VI. RESULT



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134107

VII. CONCLUSION

In conclusion, the implementation of a plagiarism checker using Machine Learning (ML) represents a significant stride towards fortifying academic and professional integrity in the face of escalating challenges posed by content misappropriation. This endeavor has sought to bridge the limitations of conventional rule-based systems by harnessing the capabilities of ML algorithms, thereby offering a more adaptive and accurate solution for detecting instances of plagiarism.

REFERENCES

[1]. S. Alzahrani, N. Salim and A. Abraham, Understanding plagiarism linguistic patterns, textual features and detection methods, IEEE T. Syst. Man Cyb. 42 (2011), 133–149.
[2]. I. Androutsopoulos and P. Malakasiotis, A survey of paraphrasing and textual entailment methods, J. Artif. Intell. Res. 38 (2010), 135–187.
[3]. D. Bar, T. Zesch and I. Gurevych, Text reuse detection using a composition of text similarity measures, in: Proceedings of COLING 2012, pp. 167–184, Mumbai, December 2012.
[4]. A. Barrón-Cedeño, M. Vila, M. A. Martí and P. Rosso, Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection, Comput. Linguist. 39 (2013), 917–947.
[5]. S. Burrows, M. Potthast and B. Stein, Paraphrase acquisition via crowdsourcing and machine learning, ACM
[6]. T. Intell. Syst. Technol. 4 (2012), 1–21.
[7]. C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines, ACM T. Intell. Syst. Technol. 2 (2011), 1–27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm HYPERLINK
[8]. D. L. Chen and W. B. Dolan, Collecting highly parallel data for paraphrase evaluation, in: Proceedings of the 49th Annual Meeting of ACL, pp. 90–200, Portland, USA, June 2011.
[9]. A. Chitra and A. Rajkumar, Genetic algorithm based feature selection for paraphrase recognition, Int. J. Artif. Intell. Tool. 22 (2013), 1350007.1-17.
[10]. P. Clough and M. Stevenson, Developing a Corpus of Plagiarized Short Answers, Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis 45 (2011), 5–24.
[11]. P. Clough, R. Gaizauskas and S. L. Piao, Building and annotating a corpus for the study of journalistic text reuse, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp. 1678–1691, Spain, May 2002.
[12]. J. Cordeiro, G. Dias, and P. Brazdil, New functions for unsupervised asymmetrical paraphrase detection, J. Software 2 (2007), 12–23.
[13]. J. Jiang and D. W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of International Conference on Research in Computational Linguistics, Taiwan, pp. 19–33, September 1997.
[14]. D. Klein and C. D. Manning, Accurate unlexicalized parsing, in: Proceedings of 41st Meeting of ACL, pp. 423–430, 2003.
[15]. N. Madnani, J. Tetreault and M. Chodorow, Re-examining machine translation metrics for paraphrase identification, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.182–190, Montreal, Canada, 2012.
[16]. M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño and P. Rosso, Overview of the 1st International competition on plagiarism detection, in: Proceedings of Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), B. Stein, P. Rosso, E. Stamatatos, M. Koppel and E. Agirre, eds., pp. 1–9. Spain, September 2009.

