iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** X **Month of publication:** October 2023

DOI: https://doi.org/10.22214/ijraset.2023.56180

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue X Oct 2023- Available at www.ijraset.com

A Review on Plagiarism Detection Methods in Digital Documents

Smitha R[1], Parvathy G S[2]
[1]*Dept. of Electronics and Communication Engineering, NSS Polytechnic College, Kerala, India*
[2]*Dept. of Computer Engineering, NSS Polytechnic College, Kerala, India*

Abstract: ***In the age of the Internet, it has become easy to copy someone else's work. Plagiarism refers to reusing someone else's work without acknowledging the source of the work or citing it. Because of the widespread availability of information and documents on the Internet and in online libraries, plagiarism has grown to be one of the major concerns for the educational and research sectors. It is actually a distinct offense against intellectual property. A variety of plagiarism detection software programs are currently available on the market or online for identifying certain sorts of plagiarism, but they still have some limitations. Typically, numerous documents are compared in order to find plagiarism, and a score is given depending on how similar they are to one another. In this paper, some of the plagiarism detection methods are reviewed and compared their performances.***
Keywords: Plagiarism, Semantic Role Labeling (SRL), Boyer-Moore algorithm, N-gram-based method, Graph representation, Smith-Waterman algorithm K Means Clustering

I. INTRODUCTION

The act of plagiarizing involves using another person's words, ideas, or works without giving them due credit or acknowledgment. The increasing accessibility of digital documents on the Internet has caused this behavior to become more prevalent [1]. Originality and diligence are undermined by plagiarism. In both academic and professional settings, plagiarism can have detrimental effects. There is no ethical justification for this practice, and it compromises the integrity of the work as well. It violates copyright rules. Use of someone else's work must be acknowledged correctly.

According to the Encyclopaedia Britannica, plagiarism is “the act of taking the writings of another person and passing them off as one’s own”[2]. Additionally, it can also involve direct copying of an entire work without proper attribution. Plagiarism can be classified as Global plagiarism, verbatim plagiarism, Paraphrasing plagiarism, Patchwork plagiarism, and Self-plagiarism etc[3].The term "global plagiarism" refers to the act of completely duplicating another person's work and passing it off as your own. It is considered the most serious form of plagiarism which can't occur accidentally. Global plagiarism has become more common due to the Internet and numerous websites that offer unauthorized copies of others' work. Copying word-by-bird content of another source and using it in your own work is verbatim plagiarism**.** Even if a few words are removed or synonyms are used, it still counts as verbatim plagiarism. The process of paraphrasing plagiarism involves using someone's work while making minor modifications to words or phrases [4]. Self-plagiarism occurs when someone reuses their own work. This act deceives others by portraying recycled content as fully unique and novel. Therefore, the relevance of plagiarism detection technologies is growing. In Chapter 3, the various plagiarism detection methods are reviewed. The comparison of these methodologies is presented in Chapter 4. Chapter 5 includes the conclusion of this study.

II. LITERATURE REVIEW

A plagiarism behavior detection system has been proposed in [5]. Document structure extraction and plagiarism function computation are the two primary processes in this system. Each phase of the document structure will be assessed using a recursive plagiarism evaluation method based on the Levenshtein edit distance. When a person plagiarizes, he or she performs certain common actions on the source to produce a new document from the original. Insertion, deletion, and substitution are examples of normal actions. This can be identified with the help of the Levenshtein distance. A weighted inter-chunk similarity and a structural similarity based on the Levenshtein edit distance are both considered in the recursive procedure. Based on the similarity level between the compared documents, a similarity score or value between 0 and 1 is returned by the function. In this method, unnecessary chunk comparison has been eliminated.A citation-based approach for plagiarism detection has been introduced in [6]. For plagiarism detection, this approach focuses on documents without proper citation, unlike word-by-word comparisons between documents.

The text is divided into sections, and a variety of algorithms are employed to analyze the citations, including their placement within the text. Then citations are compared to bibliographic content. Using tolerant sequence analysis methods, plagiarised content can still be discovered even if the citation order has been considerably changed.

SCAM((Standard Copy Analysis Mechanism) algorithm-based plagiarism detection methodology has been described in [7]. This system involves four main steps. At first dataset documents are indexed. Next is the pre-processing step which involves the splitting of test documents into tokens of words. To determine if the test document and the dataset document are similar, the index is called into question. Then similarity detection is performed using the SCAM algorithm. Semantic role labeling in natural language processing (NLP) defines the syntactic and semantic functions of words to decipher the meaning of the sentence [8]. Based on this a detection system has been presented in [9]. It seeks to identify conceptual similarities inside a sentence and perhaps even between sentences. Here, at first, the documents are pre-processed. The pre-processing step involves text segmentation, stop word removal, and stemming.

The sentences were then converted into arguments using SRL based on the placement of each phrase in the sentences. The following stage is argument-based label grouping, in which the arguments are organized into nodes according to their category. Then the concepts are extracted from the nodes and it is compared with the concepts extracted from another document. Based on the similarity check, a score is generated. PAN-PC-09 is the dataset used in this work.

A monolingual plagiarism detection strategy based on structural information rather than conceptual information has been proposed in [10]. Here, a group of suspicious and reference documents are considered. At first, the text data is converted to a group of n stopwords based on stopword n-grams (SWNG). Then the suspicious document is compared with the source document for any similarity i.e. this step involves identifying the common n-gram stop words between suspicious and source documents. Every discovered copied passage is given a value to show the level of plagiarism. A hybrid method for finding plagiarism is mentioned in [11]. This is based on Levenshtein distance and simplified SmithWaterman algorithm. A traditional technique for matching two strings and finding extremely similar parts between them is the Smith-Waterman algorithm.

Another hybrid strategy for plagiarism detection in academic works has been discussed in [12]. Here, mathematical equations, expressions, images, and citations of scientific papers or research works are analyzed. This method involves human intervention for final verification. This includes three steps i.e. candidate retrieval, detailed comparison, and human inspection. For basic similarity checks in mathematical expressions, the system calculates and counts the number of components in the expression. For in-depth similarity check, coverage, match depth and taxonomic distance are calculated.

By comparing the hash values produced by perpetual hashing based on the Discrete Cosine Transform, the similarity of pictures in publications is evaluated. Bibliographic Coupling (BC), Longest Common Citation Sequence (LCCS), Greedy Citation Tiling (GCT) and Citation Chunking (CC) have been measured to check citation plagiarism. The similarity between words in the text has been checked by making use of the Encoplot algorithm and the Boyer-Moore algorithm. A plagiarism detection approach based on graph representation has been implemented in [13].

The advantage of graph-based approaches is that they permit preserving the primary document's intrinsic structural details. Here, the text is split into individual sentences. Then stop-word removal process and the stemming process are done on each sentence. Graphs are created by placing the terms of every sentence into a single node, which is then connected based on the order of the sentences in the document. By separating meanings from each sentence term and putting them into this node, a topic signature node is created.

Automatic plagiarism detection using the K-means clustering method has been proposed in [14]. At first, the entire text is split into tokens. Then punctuation marks and stop words are removed.

The candidate retrieval approach uses the N-gram-based technique and the VSM approach. Both suspicious and source documents are broken into N-grams and their similarity is measured using Dice’s Coefficient. The questionable text with the highest Dice's coefficient value is chosen as the related document.

Vector Space Model (VSM) presents information in words as a vector. Cosine similarity between the original document and the suspicious documents is calculated and the document with maximum cosine similarity measure is identified. In K means clustering method, K or the number of centroids is chosen as the number of centroids.

The original text is then grouped with the matching doubtful text according to the similarity metric. As a result, each cluster is the candidate group of texts for a specific questionable document. A machine learning-based approach for plagiarism detection in multiple files has been introduced in [15]. A document is extracted and its contents are compared to those of similar ones already in existence. To detect copying, KNN is employed to categorize a text set from a specific paper and compare it to a manuscript that has already been saved in the database.

III. COMPARISON TABLE

TABLE I

No	Name	Techniques/Methods	Advantage and Disadvantage
1	Plagiarism detection through multilevel text comparison. [5]	Recursive Function based on Levenshtein distance	Advantages Low computational cost, Avoids unnecessary chunk comparison
2	Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently [6]	Method based on citation, Tolerant sequence analysis algorithms	Advantages Language independence, Immunity to paraphrasing Disadvantages Short passages with few citations cannot be detected
3	Plagiarism Detection Based on SCAM Algorithm [7]	SCAM algorithm	Advantages Constant and acceptable performance Disadvantages Several comparisons are involved in the examination of suspicious words and phrases
4	Plagiarism Detection Scheme Based on Semantic Role Labeling [6]	Semantic Role Labeling (SRL)	Advantages Better performance (accuracy, recall, precision) than Fuzzy Semantic-based String Similarity and, Longest Common Subsequence
5	Plagiarism Detection Based on Structural Information [10]	Stopword n-grams (SWNG)	Advantages Easy to understand, takes little resources, and incurs little expense in text pre-processing Disadvantages Detected cases are accurate, Many plagiarism cases are not detected
6	Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm [11]	Levenshtein distance and Smith-Waterman algorithm	Advantage Avoids globally involved string comparisons Disadvantage Smith Waterman is too time-consuming
7	HyPlag: A Hybrid Approach to Academic Plagiarism Detection [12]	Perceptual hashing,Encoplot algorithm, Boyer-Moore algorithm	Advantage Detects textual and non-textual plagiarism
8	Plagiarism Detection Using Graph-Based Representation[13]	Graph representation, concept extraction, topic signature	Advantages High efficiency, Reduction in the number of matching processes.
9	Using K-means Cluster Based Techniques in External Plagiarism Detection[14]	N-gram-based method, Vector Space Model (VSM) Method, Cluster-based method using K-means Algorithm, K-means with stemming 5. K-means with lemmatization-means with N-grams , K-means with chunking	Advantages Texts that have been cleverly modified using synonyms can be found. Disadvantage Reduced time efficiency
10	Plagiarism Detection Using Artificial Intelligence Technique In Multiple Files[15]	k-nearest neighbor,	Advantage Accurate and Fast detection

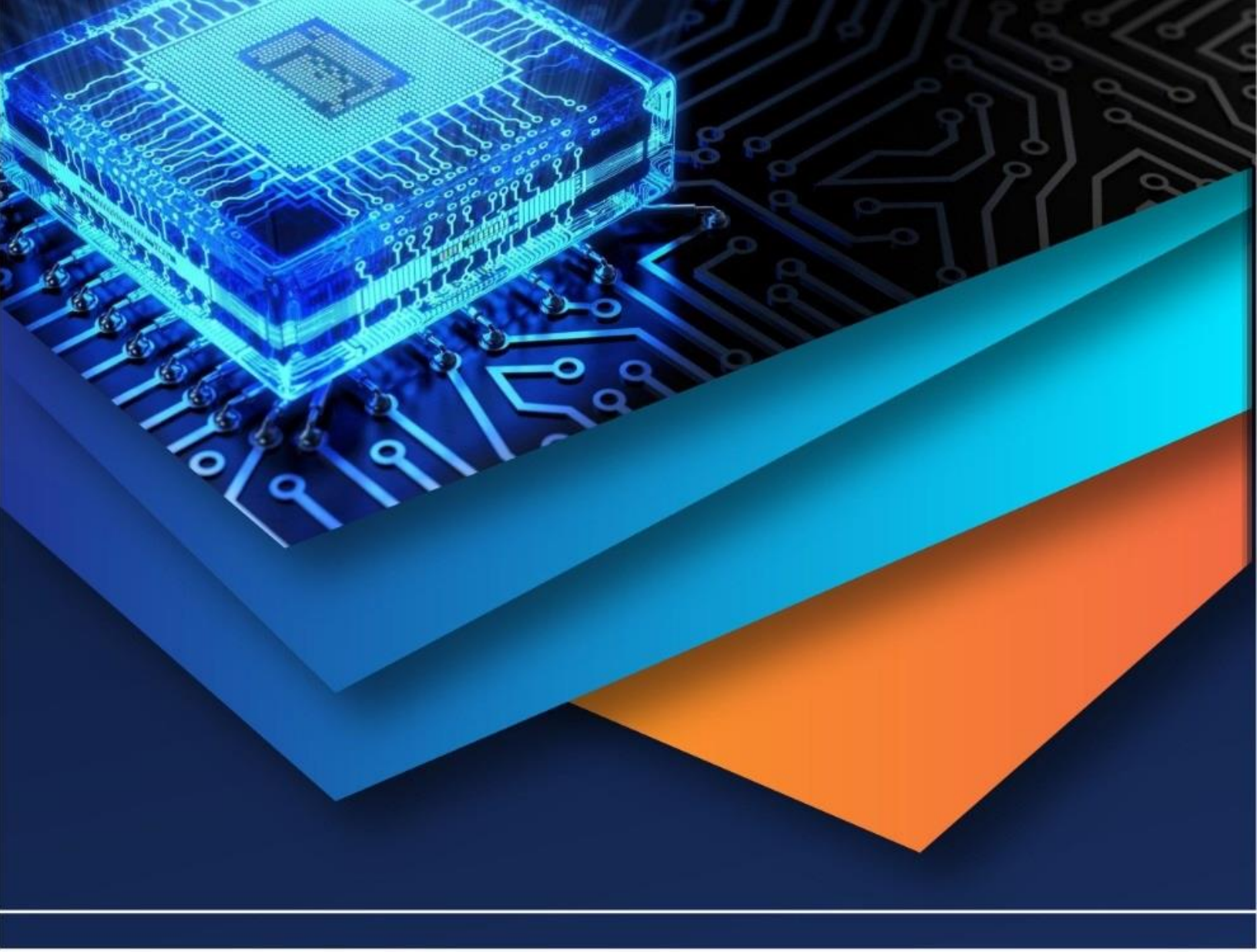
IV. CONCLUSIONS

The problem of plagiarism cannot be avoided in today's digital world. Plagiarized text or work is a revised replica of previously published data. People can easily locate the material they require and generate copies rather than producing their own content by utilizing the internet, the web, and online libraries. Because there are so many potential sources, it gets harder and harder to determine plagiarized text.

The purpose of plagiarism detection is to detect if a document contains content that has been copied from another document. It is found that the majority of detection systems rely on the concept of substring comparison. In this paper, some of the detection strategies have been compared. Every method has some sort of merits and demerits. With perfect accuracy plagiarism cannot be detected. Most of the latest detection methods use the K means clustering algorithm. In the future deep learning algorithms can be used for the same.

REFERENCES

[1] Velásquez, J. D., Covacevich, Y., Molina, F., Marrese-Taylor, E., Rodríguez, C., & Bravo-Marquez, F. (2016). DOCODE 3.0 (DOcument COpy DEtector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources. Information Fusion, 27, 64-75.
[2] Jawad F. Plagiarism and integrity in research. J Pak Med Assoc. 2013;63:1446–7.
[3] https://www.scribbr.com/plagiarism/types-of-plagiarism/
[4] https://www.grammarly.com/blog/types-of-plagiarism/
[5] Zini, M., Fabbri, M., Moneglia, M., & Panunzi, A. (2006, December). Plagiarism detection through multilevel text comparison. In 2006 Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06) (pp. 181-185). IEEE.
[6] Gipp, B., & Beel, J. (2010, June). Citation based plagiarism detection: a new approach to identify plagiarized work language independently. In Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (pp. 273-274).
[7] Anzelmi, D., Carlone, D., Rizzello, F., Thomsen, R., & Hussain, D. A. (2011, March). Plagiarism detection based on SCAM algorithm. In Proceedings of the International MultiConference on Engineers and Computer Scientists (Vol. 1, pp. 272-277).
[8] https://medium.com/@andrew_johnson_4/understanding-semantic-role-labeling-identifying-the-meaning-behind-words-in-nlp-889706cb6675
[9] Osman, A. H., Salim, N., Binwahlan, M. S., Twaha, S., Kumar, Y. J., & Abuobieda, A. (2012, March). Plagiarism detection scheme based on semantic role labeling. In 2012 international conference on Information Retrieval & Knowledge Management (pp. 30-33). IEEE.
[10] Stamatatos, E. (2011, October). Plagiarism detection based on structural information. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1221-1230).
[11] Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., & Kim, M. K. (2008, June). Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. In 2008 3rd International Conference on Innovative Computing Information and Control (pp. 569-569). IEEE.
[12] Meuschke, N., Stange, V., Schubotz, M., & Gipp, B. (2018, June). HyPlag: A hybrid approach to academic plagiarism detection. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 1321-1324).
[13] Osman, A. H., Salim, N., & Binwahlan, M. S. (2010). Plagiarism detection using graph-based representation. arXiv preprint arXiv:1004.4449.
[14] Vani, K., & Gupta, D. (2014, November). Using K-means cluster based techniques in external plagiarism detection. In 2014 international conference on contemporary computing and informatics (IC3I) (pp. 1268-1273). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY