[1] **Dr. Palvadi Srinivas Kumar**

[2] **Dr. Krishna Prasad**

A Comprehensive Survey of Advanced Image Processing and OCR Techniques for Enhanced Image Plagiarism Detection

Abstract: - In today's digital world, sharing files, documents, and presentations online has become routine for both work and study purposes. However, this surge in online sharing has brought forth a significant challenge: plagiarism. Plagiarism will happen whenever the copied content in various sources by not giving proper credit. This paper delves into the realm of plagiarism detection, particularly focusing on images. We discuss various tools and techniques designed to detect plagiarism and compare their effectiveness based on factors like accuracy. Finally, we analyze the work of different authors and share common findings. Our research employs a thorough review process to ensure the accuracy of our conclusions. By emphasizing the use of AI-powered tools, our agenda is to promote sharing of original dataover the online domain.

Keywords: Plagiarism detection, Image plagiarism, Online data sharing, Document sharing, Plagiarism checking software, AI-powered tools, Digital collaboration, Academic integrity, Content attribution, Comparative analysis.

Introduction

Plagiarism, is a way of submitting someone else's task as their own work without giving the proper credit to the concerned individual who should get the credit. Ethical challenges in research and scholarly endeavors. It encompasses various forms, from directly copying text to using images and data without acknowledgment, undermining the integrity of the research process and leading to severe consequences such as reputational damage and legal repercussions. To uphold academic integrity, it is imperative to cite all sources properly and give credit where it is due.

Moreover, the complexity of plagiarism extends to different manifestations, including copy-and-paste, paraphrasing, self-plagiarism, patchwork, mosaic, and accidental plagiarism. Each form presents unique challenges and implications, emphasizing the need for robust detection mechanisms to combat this pervasive issue effectively.

In response to the evolving landscape of plagiarism, exploring diverse tools and techniques for enhanced image plagiarism detection has become indispensable. Artificial intelligence (AI) emerges as a pivotal ally in this endeavor, leveraging advanced algorithms to scrutinize images comprehensively and identify instances of plagiarism with precision. These AI-powered tools offer comprehensive solutions for detecting copied content, thereby upholding the integrity of academic and professional research practices.

What is Image Processing?

Image processing is the technique of altering or enhancing digital images through mathematical operations. It involves tasks like improving image quality, extracting useful information, and recognizing patterns. Image processing finds applications in various fields such as medicine, surveillance, and entertainment.

Identifying plagiarism in Images

Identifying plagiarism in images involves comparing visual content to determine if it has been copied or altered without proper authorization. This can be achieved using image processing techniques such as undo image search, where an image is compared against a database of existing images to find similar or identical matches. Additionally, analyzing metadata, watermarks, and image features can help detect alterations or unauthorized use of images.

Various image processing algorithms were there for identifying the plagiarism in images. Some of the algorithms like Histogram Comparison, Feature Matching, Template Matching, Deep Learning, Hashing Algorithms, Watermark Detection, Edge Detection, Edge Analysis, Optical Character Recognition (OCR) etc.,

Over the set of algorithms were available in image processing choosing Optical Character Recognition (OCR) for getting text in images can be advantageous for several reasons like Text Extraction, Accuracy, Versatility, Automation, Integration, Cross-Verification, Integration with Existing Systems, Customization and Tuning, Metadata Extraction, Accessibility and Usability, Scalability, Language Support, Character Recognition

[1] *Post-Doctoral Research Fellow, Institute of Computer and Information Sciences, Srinivas University, Mangalore, Karnataka, INDIA. srinivaskumarpalvadi@gmail.com

[2]Professor & HOD (Cyber Security and Cyber Forensics), Institute of Computer and Information Sciences, Srinivas University, Mangalore, Karnataka, INDIA. krishnaprasadkcci@srinivasuniversity.edu.in

Accuracy, Document Layout Preservation, Support for Various Image Sources, Batch Processing Capabilities, Integration with Text-Based Plagiarism Detection, Continual Improvement through Feedback, Selective Text Extraction, Error Handling and Correction, Support for Handwritten Text Recognition, Batch Metadata Extraction, Integration with Document Management Systems, Customizable Workflows and Pipelines, Compliance with Industry Standards, Text Verification and Comparison, Language Identification, Contextual Analysis, Metadata Enrichment, Integration with Text Mining Tools, Collaborative Annotation and Review, Feedback Mechanisms for Improvement, Adaptive Learning and Training, Detection of Text in Complex Environments, Multimodal Content Analysis, Scalability and Performance Optimization, Support for Regulatory Compliance, Integration with Learning Management Systems (LMS), Cross-Domain Plagiarism Detection, Continuous Monitoring and Alerting, Support for Various Document Types, Quality Control and Validation, Adaptability to Domain-Specific Vocabulary, Support for Non-Latin Scripts and Languages, Collaborative Research and Development, Interoperability with External Tools and Services, User-Friendly Interfaces and Workflows and many more….

Literature Survey

Several mechanisms came into the existence in knowing the plagiarism over textual data and images. Regarding detection of plagiarism over textual data, the longest common subsequence (LCS) algorithm was the most commonly used technique.

LCS algorithm [1] categorise the uniqueness among two various data by sorting out the same type of information in both files. Moreover there are many challenges in this technique like less elasticity, more complexity, cant able to handle properly as well as proper grammar checking.

For avoiding these pitfalls [2] different versions in LCS algorithm are termed for least common subsequence (LCS) algorithm, that findout the common amont the two text files by the help of the least common subsequence (LCS) which were present in both files.

The LCS algorithm has several benefits like [3] such more scalable, reduce computing power as well as checking the grammar.

In the realm of image plagiarism detection [4] numerous methodologies have been proposed, including the scale-invariant feature transform (SIFT) algorithm. This algorithm aims to analyze and extract the primary elements of two images to ascertain their similarity. However, the SIFT algorithm is hindered by various constraints, such as susceptibility to image noise, variation, and distortion.

In response to these limitations, a recent advancement known as the five-modulus method has emerged. This technique gauges the likeness between two images by segmenting [5] each image into non-overlapping blocks and computing the modulus of the sum of pixels within each block. The five-modulus method offers several advantages over the SIFT algorithm [6] notably improved resilience to image noise, variation, and distortion.

This paper provides a concise overview of classification methods based on language within documents. It distinguishes between mono-lingual as well as cross-lingual classifications. Mono-lingual plagiarism identification focuses on identifying along extracting text over documents as well as detecting plagiarism within the same language, such as English-English plagiarism [7]. Cross-lingual else multi-lingual plagiarism detection extends this by identifying and extracting text from documents, detecting plagiarism across different languages, for instance, English-Arabic plagiarism.

Shape-Based Plagiarism Detection over flowcharts in textual data begins with preprocessing, which involves delineating boundaries, identifying edges, measuring distances, and storing figures in a database after removing accompanying text. During the training phase, the system preprocesses a sample figure to construct a query vector, which is then compared with figure-documents in the database [8]. Subsequently, during testing, the system takes test figures as input and compares them with those stored in the database to ascertain the number of figures replicated from the original paper.

This paper presents a comparison between two sets, A and B, comprising RGB images of the same dimensions. The objective is to identify identical pixels in the same positions between A and B, outlining the steps and algorithms involved [9]. We define C as an image matrix resulting from subtracting the color matrix of B from that of A, denoted as C = A - B. whenever the following pixle in A & B share the unique color, RGB significance the next pixels in C must be zero, indicating black. As a result, when transferring pixels from images A to B, the following pixel in c must display as black. Let's denote H as a collection containing black pixels obtained by picture C. Therefore the copied pixels from images A to B must include within set H. In comparison process, if images A and B share the same background color, that particular section with the matching background color is isolated into set H. Subsequently, any repetitive elements are filtered out, considering that the background color is typically monochromatic.

The exploration of Histogram in this study delves into its diverse applications, elucidating how histograms can unveil various properties of images, facilitate image enhancement, and aid in detecting exposure saturation, brightness levels, gaps, among others. Additionally, histograms play a pivotal role in thresholding [10]. Furthermore, this paper delves into Histogram stretching, a technique aimed at enhancing image contrast.

Histogram sliding, another technique discussed, offers insights into image intensity and brightness variations. Moreover, Histogram Equalization is explored, which standardizes all pixels of an image to a uniform distribution, resulting in a flattened histogram.

The Flowchart Plagiarism identification technique employs an area detection mechanism for identify instances of plagiarism. Flowchart images serve like inputs for the device and undergo preprocessing, wherein edges are detected using the 'Canny edge detection' algorithm. Subsequently, every shape within an image and bordersof the image are identified, and the Euclidean distance from the centroid to the boundary is computed, generating a graph [11]. This produced graph was then cross checks by graph of the actual image. Outcome is like indicating that the image exhibits signs of plagiarism or not. However, it's important to note a limitation of this approach: it is solely applicable to flowchart images.

The paper "Edge Detection Methods" provides an in-depth exploration of edge detection, a crucial technique for feature extraction, particularly in image analysis. The author categorizes various edge detection mechanisms, including Sobel, Prewitt's, Roberts Cross, Laplacian of Gaussian, and Canny methods [12]. In the experiments detailed within the paper, sample images are first converted to grayscale, serving as the basis for analysis across all techniques. The author conducts a thorough comparison of these techniques, evaluating their effectiveness in detecting edges within the images. Ultimately, the paper concludes that the Canny edge detection method outperforms the others, emerging as the most suitable choice for edge detection tasks compared to Sobel, Prewitt's, Roberts Cross, and Laplacian of Gaussian techniques.

Content-Based Image Retrieval (CBIR) serves as feature extraction method, utilizing various attributes for image such as shape, color, and texture in defining its representation along indexing within a database [13]. CBIR involves the process of retrieving images based on their visual content, focusing on similarities in shape, color, and texture. This method encompasses both color and shape feature extraction techniques.Color feature extraction involves the utilization of the HSV color space for comparison, enabling the identification of similarities in color distribution among images. On the other hand, shape feature extraction plays pivotal role for discerning original shapes along their representations within the images.

The exploration of plagiarism detection in images highlights advancements in technology for identifying image-based plagiarism, particularly in areas like diagrams and flowcharts where research has been less extensive. Th article [14] concentrates on Content-Based Image Retrieval (CBIR) for extracting features like color, shape, and texture from images.The pre-processing phase comprises three key steps they are Grayscaling, Thresholding and Boundary Detection and Cropping

In their work titled "Arabic Script Web Page Language Detection Using Hybrid KNN" the authors[15]address a critical aspect of text-based language identification: generating dependable features and managing the extensive variety of languages worldwide.

In their publication titled "Advanced Supervised Learning in Multi-layer Perceptrons - From Backpropagation to Adaptive Learning Algorithms" (Ahmad Gull Liaqat & Aijaz Ahmad, 2011) [16], the authors discuss the evolution over supervised learning techniques within multi-layer perceptrons, specifically focusing over advancements fromoriginal backpropagation algorithm to adaptive learning algorithms. They highlight the considerable progress made in refining an mechanism in training weights over feed-forward neural networks since the introduction of the backpropagation algorithm.

In their paper titled "Detection and Identification of Source Code Plagiarism Using Machine Learning Approach" the authors [17] delve into the pressing issue of source code plagiarism within academic settings. They emphasize the significance of this problem, particularly in academia, where programming assignments serve as crucial evaluation tools in programming courses. The paper explores how machine learning techniques were leveraged for identifyinga part of source code plagiarism, offering a potential solution to address this pervasive issue.

In their paper titled "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vector Machine" the authors [18] explore a novel approach to detecting plagiarism via the internet, employing a hybrid model combining Artificial Neural Networks and Support Vector Machines. They note that conventional plagiarism detection methods rely heavily on uniqueness identification procedures, where the set of unique data typically conveys identical ideas.

Chaudhari, et al.[19] have created a technique for retrieving images. In this study, they created an image extraction mechanism that takes the image as a input in the form of the query and extracts their properties, such as color and shape. They use the Color Coherence Vector for color, and mass, centroid, and dispersion for shape.

Ait-Aoudia, et al. [20] created an image retrieval system and published a paper about it. Query image was initially provided like the input. The paper outlines the calculation of various features essential for image analysis, encompassing color, shape, and points of interest. In extracting color features, the authors utilized the HSV color space. Texture features were derived from measures such as contrast, energy, and entropy. Additionally, point of interest features were extracted using the Harris detector. Variations were made as well as the outcome is then shown.

Joshi, M., & Khanna, K. [21]The paper 'Edge Detection Methods' tells regarding the edge detection which is a prominent technique for feature extraction method which helps in identifying lines over images.

Bhattacharjee, et al. [22] here the author designed the plagiarism identification mechanismforequations.Initiallyby taking each and every line from the word file. The author focused on "=" character in each line. When a "=" symbol appears, it indicates that there are equations present. The equation line image is then converted to characters using character recognition, and the resulting characters are then compared to equations in the database before the results are shown.

Ghassan MahmoudhusienAmerandDr.Ahmed Mohamed Abushaala [23]introduced a type in feature extraction technique called Content-Based Image Retrieval (CBIR) makes use of a variety of an image's contents, including shape, colour, and texture, for the representation and indexing of the image.

Sakhare, et al. [24] They have created a picture retrieval system, focusing on the qualities of color and texture. They employ RGB color

The author [25] presentedA technique for detecting plagiarism in flowchart figures relies over shape as well as various kinds of multimedia data grabbing. This procedure retrieves flowcharts and ranks their uniqueness based on diverse matching sets, facilitating the identification of plagiarized figures.

The author [26] Perceptual hash functions, combined with rotation checks, are employed to detect instances of image plagiarism.

The author [27] discusses the significant problem over plagiarism as well as evaluates effectiveness for specific plagiarism identification software.

The author [28] provides an overview of various methods currently available for detecting web-based plagiarism.

The author [29] utilizes loop structures as a main agenda in identifying directional strokes within a learning mechanism. Although thisprocedure successfully recognized 97.6% for provided datasets, its practicality is limited as it necessitates user input for drawing loop structures. Flowcharts serve to convey information about the workflow process to readers, thereby enabling the textual data having within charts to utilize for adescriptive categories.

The author [30] introduced the technique to bring information from flowchart using image processing as well as neural network dependent optical character recognition. This technique gives the general information as well as contours for the device features data. Remarkably, this device achieved a recall rate of 98.0%.

The author [31] adopted a novel approach to semantically categorize flow chats with the help of a grammatical technique. It emphasizes that, in addition to the data havingwithin flowchart, shapes utilized with each node are equally significant as they can differentiate between various processes in the workflow. It asserts that a flowchart can be considered subset over images and therefore can be categorized with various image features. Moreover, in order to design the flowchart plagiarism detection system the device must have capable of recognizing the features, enabling it to address the queries posed by humans effectively.

The author [32] examined techniques for representing and describing figures depending overthe shape. This mechanism relies over features within every document, such as color, shape, and texture accurately represent along describe the figures contained therein.

The author [33] The approach to detecting text plagiarism involves utilizing semantic role labeling along sentence ranking techniques in order to expedite andifferentiating process and reduce delays.

Several mechanisms were brought up into the existence for identifying the plagiarism in images [34] regarding the concept of text generation over image uniqueness identification, longest common subsequence (LCS) [35] technique is most commonly used.

The LCS technique which identifies uniqueness among different text data by finding out the uniqueness among them. Moreover LCS having the limitations such as scalability, less user friendly, low computational power, can't handle grammar paraphrasing in text [36].For overcoming these problems, various techniques regarding The LCS (Least Common Subsequence) algorithm is a well-explored method for measuring [37] the similarity between two texts by identifying the least common subsequence shared by both texts. It offers several advantages, including higher scalability, lower computational complexity, and improved handling of synonymy and paraphrasing[38].

In the realm of image plagiarism detection, various techniques have been proposed, such as the scale-invariant feature transform (SIFT) algorithm. [39] This algorithm aims to analyze and extract the major elements of two photos to determine their similarity. [40] However, the SIFT algorithm is known to have limitations, including sensitivity to image noise, variation, and distortion. [41]

Solutions brought to overcome the plagiarism

Year	Approach	Obtained Result
2016	A proposed method involved two stages: retrieving candidates and assessing pairwise document similarity. This method was devised around a keyword-focused	In this work they got the effectiveness of the work[24]

	technique. Given data within the text, the primary goal was on segmenting statements into fragments or chunks to identify underlying similarities.	
2017	A system named Decode 5 was developed to identify plagiarism by analyzing data on the World Wide Web as well as user-defined data. Furthermore, its implementation was carried out as a Decision Support System (DSS).	The conducted work involved load testing on the deployed system, showcasing the advantages of distributed deployment, and demonstrating satisfactory results from both scientific and business viewpoints. This led to the conclusion that algorithms initially designed for small-scale plagiarism detection could be adapted for utilization in a commercial-level platform. [25]
2017	A method for detecting plagiarism was proposed utilizing the weighted local maximum value of Longest Common Subsequence (LCS) with a distributed format. The dataset by a plagiarism finding contest wasto evaluate the suggested approach with additional basic methods based on LCS.	The extensive experiments showcased the effectiveness of the suggested approach in applications that demand stringent plagiarism identification. [26]
2017	An external system for detecting plagiarism (EPDS) integrates semantic and syntactic data alongside Semantic Role Labeling (SRL) methodology.	The proposed approach demonstrated efficacy in detecting multiple types of plagiarism. Results from the experiment also indicated that this method could enhance performance. [27]
2017	The plagiarism detection method employed SCAM (Standard Copy Analysis Mechanism), which utilizes a relative scale to identify overlap by comparing the number of shared words between the test file and the registered document.	The proposed detection method compared documents using natural language processing. For controlling lot volumes over information efficiently, the study implemented a Map-Reduce-based SCAM method with Hadoop for plagiarism detection. While the typical SCAM algorithm is suitable for processing small amounts of data, this adaptation was specifically designed for handling larger datasets. [28]
2017	The method's implementation saw a notable performance boost, along with the integration of a plagiarism detection technique tailored for "copy and paste" plagiarism types, leveraging approximate string matching. This enhancement facilitated the skipping of the majority of calculations by utilizing two different types of output estimates for plagiarism detection. Although there was slight accuracy trade-off, it was deemed acceptable.	Based on research observations, the enhancement could potentially reduce processing time by half, albeit with 6.4% reduction in accuracy compared to the algorithm's standard deployment. [29]
2017	The study presented and evaluated an application designed to validate similarities among documents. It utilized word similarity percentages to gauge the degree of resemblance between texts, enabling the identification of plagiarism in written work. The program incorporated a web-based k-gram and winnowing method for this purpose.	This effort assessed accuracy was done by comparing of its output with output generated by a human evaluator. Discrepancies between the system's results and human evaluations were noted based on the number of pages. Additionally, the study mentioned the processing time required by the application. [30]
2018	A plagiarism detection solution tailored for Urdu text documents was developed, capable of identifying various categories of plagiarism. The paper utilized several procedures including cosine similarity, Generalized Jaccard similarityand the Waterman algorithm to achieve its objectives. Additionallymachine learning classifiers, Naive Bayes and Support Vector Machine (SVM), were employed. The study defined two types of classification: binary and multi-classification.	According toresearch observations, the suggested DLDM approach outperformed existing methods for both binary and multi-class classification. Furthermore, the cosine and Waterman algorithms showed better performance compared to the G-Jaccard algorithm. In pursuit of even better outcomes, researchers plan to integrate information on syntactic and semantic similarities in the future. [31]
2018	Two systems were constructed using MCANN and BP neural networks to identify plagiarism in Nepali language literature. These frameworks underwent testing on two separate datasets, and the findings were	After conducting a detailed comparison of the papers line-by-line and paragraph-by-paragraph, it was found that the mean accuracy of BP and MCANN was in the range of 98.657 and 99.864,

	thoroughly evaluated and critiqued. It was observed that MCANN converged more rapidly compared to the conventional BP algorithm.	respectively. In contrast to BP, MCANN proved to be effective in detecting plagiarism in documents written in Nepali. [32]
2018	The study introduced two-step method for detecting plagiarism: text alignment and candidate retrieval. Initially, candidate documents were extracted using k means mechanism later creating a vector representation at record level with a Convolutional Neural Network (CNN). Subsequently, features were retrieved at the sentence level using another CNN to align the text.	The corpora created for both the AAI competition and PAN2015 competition were evaluated. The precision as well as recall was 0.84 along 0.80, respectively, for the second corpus, it was 0.83 as well as 0.82 [33]
2018	In this study, a tool was developed to assist instructors in detecting the similarity of student assignments. The Rocchio approach was utilized to identify text similarity.	The accuracy of the classification reached 74.6%. The Rocchio approach demonstrated its capability in accurately classifying comparable documents. [34]
2019	A fusion of CKR method was proposed for plagiarism identification. The outcomes of conceptual analysis were integrated with BoW (bag-of-words) systems using a dynamic interpolation factor.	The outcomes illustrated that the suggested method outperformed existing techniques in accuracy and they have proved for development the tool over cross language mechanism [35]
2019	A DNN-based word embedding framework was constructed using a Sinhala text corpus. This was achieved with the assistance of the UCSC Sinhala News corpus and the word2vec technique.	The developed framework was implemented and evaluated using a new dataset, revealing an impressive 97% accuracy in identifying plagiarism. [36]
2019	This work integrated several heuristics, including string compression as well as detection probability, with the basic detection approach for improving accuracy of identifying copied sections and reduce computation times.	This approach was straightforward to use and set up, requiring only one parameter to determine plagiarism. The final contribution of the work showcased, through real data, the effectiveness for the suggested strategy, especially highlighting the added heuristics [37].
2019	This technique for the device efficiency of a winnowing method used via different support mechanism for Plagiarism detection.	After performing all the surveys, the total values were found that 1.07 to 3.52. [38]
2019	A plagiarism corpus for Thai was presented to evaluate and compare all algorithms for plagiarism detection.	They demonstrated that the suspicious documents in the corpus were manually crafted using various techniques, making them more realistic and challenging.[39]
2019	Data matchingmechanism was the "character-by-character" identification technique. The approach could also define hashing blocks of the characters later use n grams for comparing the hash value blocks. Moreover while utilizing n-gram, the document needs to follow all pre-processing steps.	This technique employing a character checking methodology establishes the same level forplagiarismdetecting with N-Gram result by the help of Dice's Likeness Coefficient. For the text extracting mechanism tokenization system brought into existence as well as streaming was designed.Subsequently, the text is randomized using the Rabin-Karp technique. [40]
2021	A method was developed for identifying text and cross-language plagiarism in both English and Albanian. By implementing this approach to monitor student work, it was anticipated that this paper would enhance standards and accountability in educational settings and universities. The system was constructed using Python and PHP and was web-based.	Multiprocessing was employed in this work for enhancing the system's performance. The results of the trials demonstrated that this approach was effective in detecting both textual and cross-text plagiarism. [41]
2021	A computerized method was developed capable of swiftly identifying plagiarism-causing similarities in scientific or writing articles. The VSM technique by the help of TF along IDF methods was utilized for this purpose.	The findings revealed that using these methods together resulted in a similarity value of 1 (100%). [42]
2021	An attempt was made to develop a system or application to identify similarities between Indonesian documents utilizing Rabin-Karp algorithm, stemming methodology, as well as cosine uniqueness approach as	This algorithm swiftly identified document similarities, particularly in documents, with minimal time.[43]

	a distance-based similarity measure.	
2022	This paper proposed an approach to automatically identify various forms of plagiarism from two languages. Using the Doc2Vec model, that identifies semantic among documents as well as other data, this strategy was built on sentence modeling to detect copied passages from documents. For Arabic and English, respectively, this work utilized AraPlagDet and the PAN corpus for detecting plagiarism.	The PAN alongAraPlagDet corpora provided record of questionable file was were manually as well as intentionally copied with their origins.[44]
2022	Deep learning system in statement text plagiarism detection was developed, which utilized then/w for Siamese LSTM as well as word-dependingembeddings.This suggested approach employed Word2Vec and Glove approaches to create the network's input. Then, a combined with Manhattan length along cosine uniqueness metrics was used to calculate the percentage of plagiarism at network's two results.	Extensive experiments over PAN-PC-11 along Webis-CPC-11 dataset demonstrated an merging of Word2Vec architecture yielded mostexactness in identifying scores. By utilizing Word2Vec embeddings, the approach achieved F1-measures of 0.81, 0.910 recall as well as 0.92 accuracy for the PAN-PC-11 corpus as well as F1-measures 0.79, 0.852 recall, and 0.902 precision for the Webis-CPC-11 corpora. [45]

Plagiarism Detection Tools

As of now, there are numerous plagiarism detection tools as live which offerstheir own characters as well as working style. While it's challenging to provide an exhaustive list, I can certainly highlight some of the best and most widely recognized plagiarism detection tools were Turnitin,Grammarly, Copyscape, Plagscan, Quetext, Unicheck, Plagiarism Checker X, iThenticate, DupliChecker, Plagscan etc., the mentioned tools defines the best and most widely used features were available for identifying copied data in research papers and other types of content. Depending on our specific needs and preferences, we will find that one of these tools better suits your requirements than others.

Plagiarism Detection Procedure with combination of Text and Images in the Document

Plagiarism detection processes typically focus on analyzing text, but when dealing with documents that contain a combination of text and images, the process becomes slightly more complex. Here's how the process can be adapted:

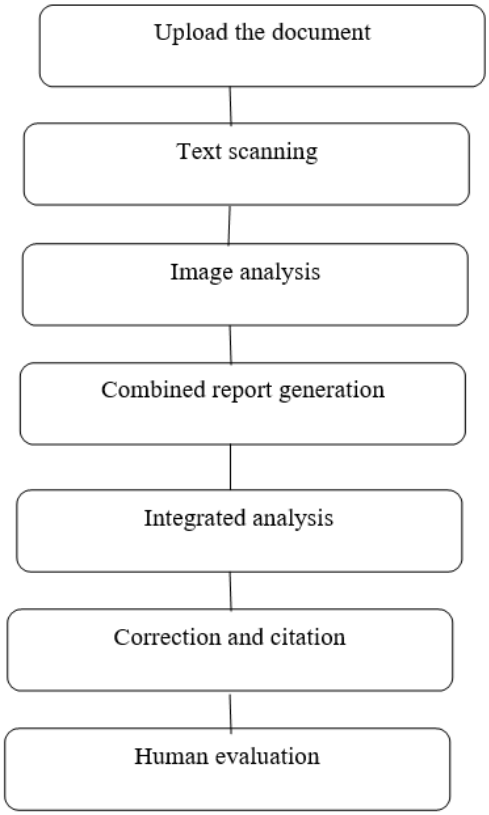

Figure 1: Workflow of the Plagiarism detection which is merged with text and images

Combining text scanning and image analysis enhances plagiarism detection processes, enabling them to effectively assess documents containing both text and images. This integrated approach offers a more comprehensive evaluation of originality.

Common Findings and Scope of Research

After reviewing different mechanisms or techniques in the domain of plagiarism identification, several research gaps have been identified:

- Previous methods for checking plagiarism often overlook referenced content. It is crucial to consider referenced material as potentially plagiarized due to its influence on the model.
- Machine learning (ML) models can be leveraged to enhance accuracy in plagiarism detection.
- There is a need for a proper algorithm to retrieve documents from the internet. Currently, the title of the paper is commonly used as a query for document retrieval. However, this approach may fail to retrieve the correct document if the paper's title is not accurately provided, leading to incorrect similarity scores.
- It is important to generate PDF or HTML reports after plagiarism checking, and the text highlighting method should be implemented effectively to improve results.

Conclusion

After examining various plagiarism detection methodologies, it's clear that despite the diversity of techniques available, there remain notable research gaps and challenges. Some methods struggle with detecting plagiarism when the text has been paraphrased or reworded, while others face issues with handling diverse document formats or languages. Additionally, ethical and legal considerations, such as safeguarding user privacy, are paramount in implementing plagiarism detection systems.

A review of papers from 2016 to 2022 suggests several areas for enhancement. Firstly, developing an intra-corpus productive system alongside a robust AI model could bolster counterfeit recognition. Incorporating a comprehensive local database within the system could further enhance its efficacy, mirroring the approach of many commercial software solutions that utilize local corpora to expedite searches and reduce processing time.

Furthermore, there's a pressing need for improved plagiarism detection tools for images. Leveraging machine learning and image processing techniques could aid in the development of such tools. Additionally, the lack of web deployment for existing methods limits accessibility. Creating a web interface for these tools would improve user experience and accessibility, fostering wider adoption and utilization.

REFERENCES

[1] U. Garg and V. Goyal, "Maulik: A Plagiarism Detection Tool for Hindi Documents," Indian Journal of Science and Technology, vol. 9, no. 12, pp. 1-11, 2016.

[2] N. Ehsan and A. Shakery, "Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information," Information Processing & Management, November 2016.

[3] G. P. V. and J. D. Velásquez, "Docode 5: Building a real-world plagiarism detection system," Engineering Applications of Artificial Intelligence, vol. 32, no. 4, pp. 1703-1712, September 2017.

[4] K. Baba, T. Nakatoh and T. Minami, "Plagiarism detection using document similarity based on distributed representation," Procedia Computer Science, vol. 10, pp. 89798-89822, 2017.

[5] A. Abdi, S. M. Shamsuddin and R. M. Aliguliyev, "A linguistic treatment for automatic external plagiarism detection," Knowledge-Based Systems, vol. 19, no. 3, pp. 1817-1826, 1 November 2017.

[6] J. Dwivedi and A. Tiwary, "Plagiarism detection on big data using modified map-reduced based SCAM algorithm," 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, pp. 608-610.

[7] K. Baba, "Fast plagiarism detection based on simple document similarity," 2017 Twelfth International Conference on Digital Information Management (ICDIM), 2017, pp. 54-58.

[8] R. Sutoyo et al., "Detecting documents plagiarism using winnowing algorithm and k-gram method," 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 2017, pp. 67-72.

[9] W. Ali, T. Ahmed, Z. Rehman, A. U. Rehman and M. Slaman, "Detection of Plagiarism in Urdu Text Documents," 2018 14th International Conference on Emerging Technologies (ICET), 2018, pp. 1-6.

[10] R. K. Bachchan and A. K. Timalsina, "Plagiarism Detection Framework Using Monte Carlo Based Artificial Neural Network for Nepali Language," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), 2018, pp. 122-127.

[11] S. Lazemi, H. Ebrahimpour-Komleh and N. Noroozi, "Persian Plagiarism Detection Using CNNs," 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE), 2018, pp. 171-175.

[12] D. Soyusiawaty, A. H. S. Jones and P. Widiandana, "Similarity Detection of Student Assignments Using Rocchio Method," 2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA), 2018, pp. 1-4.

[13] N. Meuschke, V. Stange, M. Schubotz, M. Kramer and B. Gipp, "Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations," 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, pp. 120-129.

[14] M. Roostaee, M. H. Sadreddini and S. M. Fakhrahmad, "An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes," Information Processing & Management, vol. 70, no. 1, pp. 248-260, 1 November 2019.

[15] T. KasthuriArachchi and E. Y. A. Charles, "Deep Learning Approach to Detect Plagiarism in Sinhala Text," 2019 14th Conference on Industrial and Information Systems (ICIIS), 2019, pp. 314-319.

[16] D. Sakamoto and K. Tsuda, "A Detection Method for Plagiarism Reports of Students," Procedia Computer Science, 14 October 2019.

[17] S. Thaiprayoon, P. Palingoon and K. Trakultaweekoon, "Design and Development of a Plagiarism Corpus in Thai for Plagiarism Detection," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-5.

[18] W. G. S. Parwita, I. G. A. A. D. Indradewi and I. N. S. W. Wijaya, "String Matching based Plagiarism Detection for Document in Bahasa Indonesia," 2019 5th International Conference on New Media Studies (CONMEDIA), 2019, pp. 54-58.

[19] L. Shkurti, J. Ajdari, F. Kabashi and V. Fusa, "PlagAL: Plagiarism detection system for Albanian texts," 2021 10th Mediterranean Conference on Embedded Computing (MECO), 2021, pp. 1-5.

[20] R. Wahyudi, M. Zarlis, S. Efendi and T. F. Abidin, "Determination of Sentence Similarity Level Using Vector Space Model (VSM) and Word Relationship Weighting for Plagiarism Detection for Indonesian Documents," 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), 2021, pp. 142-153.

[21] A. D. Hartanto, Y. Pristyanto and A. Saputra, "Document Similarity Detection using Rabin-Karp and Cosine Similarity Algorithms," 2021 International Conference on Computer Science and Engineering (IC2SE), 2021, pp. 1-6.

[22] I. Setha and H. Aliane, "Enhancing automatic plagiarism detection: Using Doc2vec model," 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), 2022, pp. 1-5.

[23] A. A. M. Saeed and A. Y. Taqa, "Textual Plagiarism Detection Using Embedding Models and Siamese LSTM," 2022 International Conference for Natural and Applied Sciences (ICNAS), 2022, pp. 95-100.

[24] K. Chandra Sekhar and K. Santhosh Kumar, “Undergraduate Student’s Campus Placement Determination Using Logistic Regression Analysis for Predicted Probabilities on Uncertain Dataset,” Int. J. Intell. Syst. Appl. Eng., vol. 10, no. 2s, pp. 14–20, 2022.

[25] K. Chandra sekhar and k. Santhosh kumar, “an ensemble approach to feature detection and performance to predict placements in technical institutions,” vol. 20, no. 11, pp. 7949–7961, 2022, doi: 10.14704/nq.2022.20.11.nq66789.

[26] K. Chandra Sekhar and K. Santhosh Kumar, “Data Preprocessing and Visualizations Using Machine Learning for Student Placement Prediction,” Proc. Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2022, pp. 386–391, 2022, doi: 10.1109/ICTACS56270.2022.9988247.

[27] Boro, A., & Patil, N. (2016). A Survey on Plagiarism Detection Techniques. International Journal of Computer Science and Information Technologies, 7(2), 857-861.

[28] Eskandari, M., Hedayati, A., &Razzazi, F. (2019). Plagiarism detection techniques in academic writings: A systematic literature review. Education and Information Technologies, 24(2), 1533-1561.

[29] Gogate, M., & Patil, A. (2017). Image plagiarism detection using modified SIFT algorithm. International Journal of Computer Applications, 174(37), 1-4.

[30] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, ”Understand- ing Plagiarism Linguistic Patterns, Textual Features, and Detection Methods”, IEEE, Vol:42, Issue:2, PP:133-149,2012.

[31] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim,” Shape-Based Plagiarism Detection for Flowchart Figures in Texts”, International Journal of Computer Science Information Tech- nology , Vol:6, No:1,2014.

[32] Wang Wen, Wang Yanb and Li Bingbing , ”Research on Plagiarism IdentificationofDigitalImages”,IEEE,2010.

[33] Harpreet Kaur and Neelofar Sohi, ”A Study for Applications of Histogram in Image Enhancement”, The International Journal of Engineering and Science (IJES), Vol:6, Issue:6, PP:59-63,2017.

[34] Jithin S Kuruvila, Midhun Lal V L, Rejin Roy, Tomin Baby, Sangeetha Jamal and Sherly K K, ”Flowchart Plagiarism Detection System: An Image Processing Approach”, 7th International Conference on Advances in Computing Communications,2017.

[35] Joshi, M., & Khanna, K. A Similarity Measure Analysis Based Improved Approach For Plagiarism Detection.

[36] Ghassan MahmoudhusienAmerand Dr. Ahmed Mohamed Abushaala, ”Edge Detection Methods”, IEEE,2015.

[37] Reshma Chaudhari and A.M Patil, ”Content Based Image Retrieval Using Color and Shape Features”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol:1, Issue:5,2012.

[38] A Selamat, IMI Subroto and Choon-Ching Ng, "Arabic Script Web Page Language Identification Using HybridKNN Method," International Journal of Computational Intelligence and Applications, 2009, pp. 315-343.

[39] Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.

[40] Upul Bandara and Gamini Wijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674- 678.

[41] Imam Much IbnuSubroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

[42] R. Chaudhari, A. M. Patil, Content Based Image Retrieval Using Color and Shape Features, “International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering”, Vol.1, Issue 5, November 2012.

[43] Samy Ait-Aoudia, RamdaneMahiou, BillelBenzaid, “Yet Another Content Based Image Retrieval system,” In IEEE.14th International Conference on Information Visualisation (IV), 2010 (pp. 570-575). 1550- 6037/10 $26.00 © 2010 IEEE DOI 10.1109/IV.2010.83

[44] Joshi, M., & Khanna, K. A Similarity Measure Analysis Based Improved Approach For Plagiarism Detection.

[45] Debotosh Bhattacharjee and Sandipan Dutta, “Plagiarism Detection by Identifying the Equations”, ELSEVIER, International Conference on Computational Intelligence: Modelling Techniques and Applications (CIMTA) 2013.