IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TEXT AND IMAGE PLAGIARISM DETECTION

Mr. M. Sai Kiran[1], D. Akhil Hrishikesh[2], Naini Thanishka[3], P. Adithya Madhav[4]

[1]Assistant Professor, Department of CSE(AIML), CMR College of Engineering and Technology, Hyderabad, Telangana, India.

[2,3,4] Student, Department of CSE(AIML), CMR College of Engineering and Technology, Hyderabad, Telangana, India.

Abstract: Plagiarism in academic research is one such topic that is discussed far more now compared to earlier. The research has suffered greatly because of the assessment of the web's current status and users' capacity to perform fast, complex searches. Plagiarism detection tools focus on text and overlook graphics. Nevertheless, images are crucial for conveying large amounts of data, like those seen in papers or scientific research. Flowcharts include large amounts of data, so it's probable that because of their extensive picture library, especially when it comes to computer texts, they contain a significant amount of plagiarism. The purpose of this article is to examine the degree of plagiarism in a manuscript using artificial neural network-based flowchart images. The suggested method's average flowchart picture identification accuracy of 81.91 percent in terms of structure, nodes, and edges proved how effective it was.

Index Terms – Plagiarism, Academic research, Plagiarism detection tools, Text, Graphics, Images, Data, Flowcharts

Manuscript, Artificial neural network, Structure, Nodes, Edges.

INTRODUCTION

"The use of thoughts, ideas, words, or structures for personal gain in a setting wherein originality is expected without appropriately crediting the source" is the definition of academic plagiarism. In academic settings, plagiarism can take on a wide variety of shapes and sizes, with varying degrees of secrecy. It can manifest as straight copies (copy and paste), slightly modified forms (shake and paste), or forms that are hidden (paraphrases, translations, concept plagiarism, and even plagiarism in academic data). Students are more prone to commit easily identified copy-and-paste plagiarism, whereas researchers are more likely to use significantly modified plagiarism since they have a strong incentive to conceal their copied work from being found out. Based on studies on plagiarism detection (PD), sophisticated systems that use text retrieval to locate related documents have been developed. These methods frequently fall short of detecting covert types of academic plagiarism, but they may reliably recover documents that include copied material. Numerous tactics have been developed to improve text-matching methods and improve the identification of disguised plagiarism, as we describe in Section 2. Not many studies have been conducted on the investigation of photos to detect academic plagiarism than on the very sophisticated text-based retrieval algorithms created for PD. In this research, we explore the possibility of replacing textual similarity in plagiarism detection algorithms with visual similarity detection techniques. In our use case, "images" are defined as the visual representations of concepts, like figures that show the schematic illustrations of entities and their relationships, like component diagrams, organigrams, and flow charts, moreover data, like bar charts, scatter plots, graphs, and so forth. We also include images and renderings that are photorealistic in our definition. Words are unable to fully express the amount of information that images can in a concise manner. Photographs become a valuable characteristic to examine when assessing the semantic similarity discovered in academic texts because of these attributes. In certain instances, data plagiarism can even be identified if graphs are used to rebuild the data values. This is how the paper is organized. In Section 2, we provide a brief overview of earlier research on image-based PD and more general PD approaches. The investigation of image similarities found in texts accused of academic plagiarism then opens Section 3, providing background information for our image-based PD methodology. The rest of Section 3 details the methods we developed and subsequently integrated into a scalable and adaptable picture-based PD strategy that can target the identified categories of image similarity.

RELATED WORK

Approaches of Plagiarism Detection

The specific information retrieval (IR) issue of identifying plagiarism is to find all papers that exhibit similarities beyond an input document against a sizable collection to determine a predetermined threshold. The two steps that the majority of PD systems use are candidate recovery and comprehensive comparison. The systems frequently use effective text retrieval techniques, including n-gram fingerprinting and vector space models, for candidate retrieval. The systems commonly use exhaustive string matching for in-depth comparison. These methods can only be used to locate almost flawless copies of a text, though. Researchers have proposed other cross-lingual IR approaches in addition to monolingual text analysis tools that use syntactic and semantic properties to reveal covert academic plagiarism. Researchers shown that hybrid techniques, the enhancement of retrieval efficacy through integrated analysis of textual and other content elements for PD tasks. Alzahrani and colleagues integrated investigations of text similarities. Gipp and Meuschke gave an example of how combining the analysis of citation patterns with text similarities might improve the detection of academic plagiarism. A hybrid technique developed by Pertile et al based-on machine learning and confirmed the advantages of combining text analysis with citation analysis. Meuschke et al. have demonstrated lately that a more accurate method of detecting academic plagiarism may be achieved by studying semantic concept patterns and similarities in mathematical expressions.

Using Image Analysis to Spot Plagiarism

The study of picture similarity for Parkinson's disease (PD) has not been thoroughly examined. Horti and Horakova provide a reliable method for locating precise duplicates of images or parts that have been clipped, ignoring other picture alterations Iwanowski et al. assess the suitability of feature point techniques such as BRISK, SIFT, and SURF for extracting precise and artistically modified photo duplicates. To accomplish the same task, Srivastava et al. combine perceptual hashing with SIFT features gathered via SIFT. Methods utilizing feature points and perceptual hashing sometimes collapse when picture constituents, such as forms, are reorganized. The productivity of different feature point methods decreases when test pictures consist of several sub-images. In experiments, the top two sub-images of each compound image can be distinguished with high accuracy by the SIFT feature extractor and MSAC feature estimator, but it is unable to discover a similarity for the other sub-image pairings.

In conclusion, the study of picture resemblance to PD has not been thoroughly studied. However, Perceptual hashing and feature point techniques can be used to find visually appealing elements in a scene and gauge how similar they are.

Comparing Pictures to Identify Document Plagiarism

The study's findings are presented in this article, which documents comparisons employing framework-based image processing methods of plagiarism detection systems. Feature-point methods are the best Among all methods of image processing for figuring out picture similarity since they don't change after the picture has been altered. The study examines many combinations, including point detectors having features as well as descriptors, as possible techniques for locating similar images inside a text. These methodologies are evaluated using five widely-used techniques for processing test photographs from an academic publication database. The paper also provides an algorithmic way for determining visual similarity, which could be applied to increase the benefits of using plagiarism detection software.

Lowering Down on Computing Work for Plagiarism Identification by Limiting Retrieval Space with Citation Attributes

This paper offers a hybrid method for identifying plagiarism in academic documents that improves precision of detection for forms of disguised plagiarism by combining character-based algorithms with detection methods based on semantic word similarity, semantic argumentation structure, and citations. The plagiarism detection application available today compares text strings exclusively. These technologies identify instances of the plagiarism that aren't immediately apparent, such as concept stealing, paraphrasing, and translations. When it comes to detecting camouflaged plagiarism, methods that semantic resemblances between words and sentence structures have consistently outperformed character-based strategies in case of detection accuracy. Unfortunately, real-world detection of plagiarism scenarios cannot realistically employ these semantic techniques due to their enormous computational effort. The suggested hybrid approach makes use of citation-based approaches as initial heuristic to reduce the retrieval space while maintaining a comparatively high level of detection accuracy. There may be a s analysis based on semantics and character after this preliminary phase, that would require more computing capacity. We show that our integrated technique makes semantic plagiarism detection possible conceivable for the earliest time, even on big datasets.

METHODOLOGY

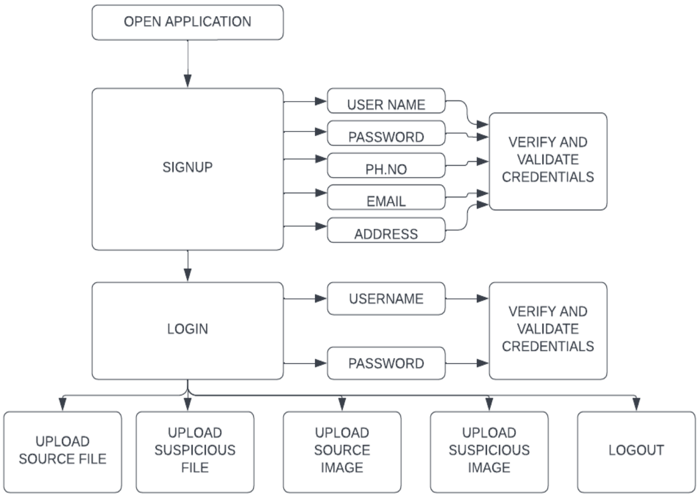


FIGURE 1: METHODOLOGY OF TEXT AND IMAGE PLAGIARISM DETECTION

New user Signup: Firstly, user will register in to Application. The user needs to register by providing the information. It is helpful to login into Application with username and password**.**

User Login: User signup process completed and now clicks on 'Login' link user is login

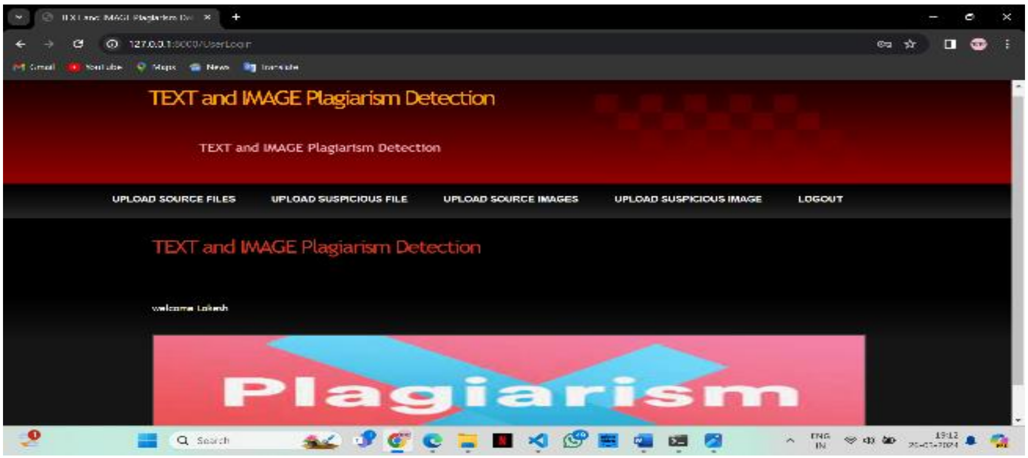
Upload Source Files: Upload Source Files link to load all files from corpus folder and all files are loaded.

Upload Suspicious Files: Upload Suspicious File module to load suspicious file

Upload Source Image: Upload Source Images module to upload all images from 'images' folder and all database images histogram will be calculated and store in array and whenever we upload new test image then both histograms will get matched.

Upload Suspicious Image: Upload Suspicious Image link to upload some image selecting and uploading '112.jpg' file and then click on 'Open' button we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected.

RESULT AND DISCUSSION



TEXT and IMAGE Plagiarism Detection

TEXT and IMAGE Plagiarism Detection

TEXT and IMAGE Plagiarism Detection

welcome Lokesh

FIGURE 2: HOME PAGE

In above screen the user can start using the application by clicking on 'Upload Source Files' link to load all files from corpus folder. Next, by selecting the suspicious file after clicking the "Upload Suspicious file" link, the user may check if any file is plagiarized. The detection of image plagiarism follows the same procedure. The testing results include an LCS score for text plagiarism identification and a histogram comparison graph with histogram matching score for image plagiarism detection.

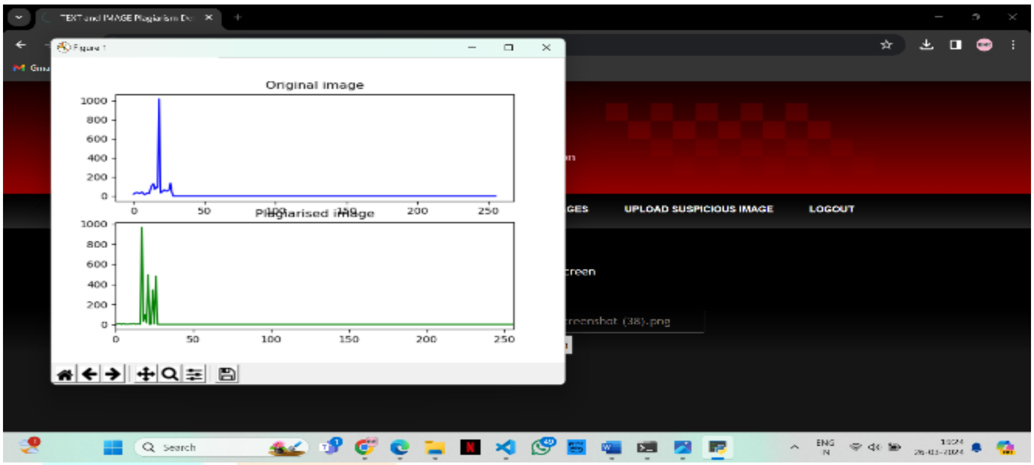


FIGURE 3

In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result

CONCLUSION

Our approach to detecting plagiarism using text and images may identify many forms of picture similarity and text file similarity that may be found in academic work. By combining methods for analyzing heterogeneous text and image features, applying methods selectively according to how well-suited they are for the input, employing a flexible procedure to find dubious text and image similarities, and making it easy to add new techniques in the future, the approach achieves adaptivity. The parameters of our approach were established by analyzing images from the Voila collection. This compilation of actual cases was created by a crowdsourcing effort that records confirmed and reported cases of plagiarism in education in the real world. Based on these instances, we created a classification system for the different types of picture similarity. Next, we introduced our flexible image-based PD approach. We improved our recognition performance by adding an extraction method (which involves perceptual hashing) for sub-images to our workflow. We created and merged two approaches that employ OCR to extract text from images and use text-based attributes for similarity assessments, since academic photographs often include textual labels. To address the issue of data reuse, we incorporated a method that can identify similar bar charts. To be able to quantify the degree of suspicion around the discovered correlations, we devised an outlier detection process. An analysis of our photodynamic process demonstrates stable performance and extends the current image-based detection approaches' detection range. We encourage other developers to build upon and alter our approach, and our code is open source.

REFERENCES

[1] Reddy, A.R., Narendar, Ch., Srinu, K., ...Reddy, R.V., Sruthi, P.” COVID 19 Patient Recognition and Prevention Utilizing Machine Learning and CNN Model Techniques”,5th IEEE International Conference on Cybernetics, Cognition and Machine Learning Applications,
ICCCMLA 2023, 2023, pp. 677–686

[2] Shirisha, N., Bhaskar, T., Kiran, A., Alankruthi, K.” Restaurant Recommender System Based on Sentiment Analysis”,2023 International Conference on Computer Communication and Informatics, ICCCI 2023, 2023

[3] Bigul, S.D., Prakash, A., Bhanu, J.S.” Futuristic evaluation of CoVID-19 spread using transfer learning: A post vaccination scenario”, IP Conference Proceedings This link is disabled., 2021, 2358, 080009

[4] Y. Ambica, Dr N. Subhash Chandra MRI brain segmentation using correlation based on adaptively regularized kernel-based fuzzy C-means clustering Int. J. Advanced Intelligence Paradigms, Vol. 19, No. 2, 2021

[5] Salha Alzahrani, Ajith Abraham, Naomie Salim, and Vasile Palade. 2011. Identifying Significant Plagiarism Cases in Scientific Publications Using Citation Evidence and Structural Information. 2011; JASIST 63(2).

[6] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Recognizing Textual Elements, Linguistic Patterns, and Detection Techniques of Plagiarism. In IEEE Trans. Syst., Man, Cybernet. C, Appl. Rev., Vol. 42.

[7] Fishman, Teddi. 2009. "When we see it, we know it"? is insufficient: in the direction of a common definition of plagiarism that goes beyond fraud, theft, and copyright. Proc. Asia-Pacific Conference on Academic Integrity.

[8] Gipp, Bela (2014). Citation-based Plagiarism Detection: Utilizing Citation Pattern Analysis to Find Cross-Language and Disguised Plagiarism. Springer.

[9] Azhar Hamdi, William Puech, Brahim Ait Es Said, and Abdellah Ait Oakman. 2012. Watermarking. Vol. 2. Intech, Chapter Perceptual Image Hashing.

[10] Petr Horti and Petra Horakova. 2015. FTIP: A tool for an image plagiarism detection. In Proc. Sopra.

[11] In 2016, Marcin Iwanowski, Arkadiusz Cacok, and Grzegorz Saras published Comparing Images for Document Plagiarism Detection. Proc. ICCVG.

[12] Trevor Darrell, Ross Kirchick, Sergio Guadarrama, Jonathan Long, Yangqin Jia, Evan Shelhamer, Jeff Donahue, and Sergey Karasev. 2014. Convolutional architecture for quick feature embedding is called Caffe. In Multimedia Proceedings.

[13] H.F. Judson. 2004. The Great Betrayal: Fraud in Science. Harcourt.

[14] Jan Kasprzak and Michal Brandeis. 2010. Improving the Reliability of the Plagiarism Detection System. In Proc. PAN WS at CLEF.

[15] IJEI 1, 1 (2005), Donald L. McCabe, Cheating Among College and University Students: A North American Perspective.

[16] Norman Meuschke and Bela Gipp. 2013. Academic plagiarism detection at the cutting edge. IJEI 9, 1 (2013).