ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

TEXT AND IMAGE PLAGIARISM DETECTION

Mr.P.Sreekanth Reddy [1], B. Chandresh [2], D.Sadaf [3], P.Naveen kumar reddy [4]

[1]Assistant professor, [2][3][4] Student.

[1] sreekanth.cse@svrec.ac.in, [2] bonthalachandresh133@gmail.com ,[3] dudekulasadaf@gmail.com,

[4] naveenkumarreddypalle781@gmail.com .

Department of CSE, SVR Engineering College, Ayyalur Metta, Nandyal(DIST)

Andhra Pradesh, India

Abstract: Plagiarism is when someone takes another author's works, thoughts, ideas, etc. without proper referencing and claim it as his/her own works. Plagiarism detection is the process to find the plagiarism within a work or documents. With the advance of modern technology, it makes it easier for people to search for information and plagiarize the work of others. Although the effort and ideas for an image-based plagiarism detection has been increasing over the years, flaws are still presence in the current systems. This paper proposes a new system that can cover those flaws. It consists three stages: the pre-processing, feature extraction and comparison stage. The results showed in an ascending order of similarity index and true and false.

Index terms *– Plagiarism, suspicious files & images, documents.*

1. INTRODUCTION

Plagiarism, the unauthorized use or reproduction of another's work, poses a significant challenge in both textual and visual domains.

Today, much more than in the past are discussed of plagiarism in the research. Conditions of the Web, Possibility of complex and smart searches in a short time, is rated to this, and as a result has arrived significant damages to the research. Tools designed to deal with plagiarism act on the text and ignore images.

The project addresses the issue of plagiarism by utilizing two distinct methodologies.

For textual content, a corpus is established as a reference to detect similarities in suspicious files.

Images are examined through histogram matching, comparing uploaded images to a database of reference images.

The project leverages these techniques to detect potential plagiarism occurrences.

The issue of plagiarism is often discussed in the educational community across the world. It relates to the act of taking another person's work and passing it as your

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

own. Basically it converts the existing information in a modified format.

Today, much more than in the past are discussed of plagiarism in the research. Conditions of the Web and Possibility of complex and smart searches in a short time, it is rated to this, and as a result has arrived significant damages to the research.

Tools designed to deal with plagiarism act on the text and ignore images. On the other hand, an inseparable part of information transfer images that transfer includes large volume of information in an article or scientific research. Because of the images include a very wide range and especially found large amounts of images in the computer's texts, and as respects, flowcharts are carrying a lot of information, could be one of the options of plagiarism.

The objective of the project is to develop a comprehensive system that detects plagiarism in both textual content and images.

By utilizing a predefined text corpus and image dataset, the system aims to identify similarities between uploaded suspicious files and the existing corpus.

Textual plagiarism is determined by calculating the Longest Common Subsequence (LCS) score between files.

For images, histogram-based comparisons are performed, analyzing pixel distribution.

This project addresses the critical issue of academic and visual content integrity by providing a reliable tool for detecting unauthorized use of textual and image-based resources, thereby promoting originality and authentic content creation.

2. LITERATURE SURVEY

Plagiarism, the act of using someone else's work without proper attribution, is a significant issue across various domains, including academia, journalism, and the arts. With the advent of digital technologies, plagiarism detection has evolved beyond textual content to encompass multimedia elements such as images. This literature survey provides an overview of recent research efforts in the field of image plagiarism detection, focusing on methodologies, challenges, and advancements.

Image plagiarism detection has garnered increasing attention due to the proliferation of digital content and the ease of sharing images online. Researchers have proposed various techniques and algorithms to address this challenge effectively. Amirul S. Bin Ibrahin et al. [11] presented a method for plagiarism detection of images, utilizing advanced image processing algorithms. Their approach aims to identify similarities between images by analyzing their visual features, such as color histograms and texture patterns.

Similarly, Simon Sepulveda et al. [12] developed a plagiarism detection engine specifically tailored for images in Docode, leveraging deep learning techniques. Their model employs convolutional neural networks (CNNs) to extract discriminative features from images and compares them to detect instances of plagiarism accurately.

In a comprehensive survey by Hermann Maurer, Frank Kappe, and Bilal Zaka [13], the authors provide an extensive overview of plagiarism detection techniques across various types of content, including text, code, and multimedia. While focusing primarily on textual plagiarism, the survey highlights the importance of addressing image plagiarism and calls for further research in this area.

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

Advancements in technology have also influenced plagiarism prevention and detection mechanisms. James Douglas Beasley et al. [14] discuss the impact of technology on plagiarism prevention, emphasizing the role of research process automation in mitigating plagiarism risks. Their approach advocates for proactive measures to deter plagiarism through systematic documentation and automated checks.

Moreover, Akshay S, Chaitanya B N, and Rishabh Kumar [15] proposed a novel approach for image plagiarism detection using compressed images. By analyzing the compression artifacts introduced during image compression, their method can identify similarities between images even when modifications have been made to evade detection.

Despite these advancements, image plagiarism detection still faces several challenges. One of the primary challenges is the vast amount of digital content available, making it difficult to perform efficient and accurate comparisons. Additionally, the proliferation of image editing tools enables plagiarists to manipulate images easily, further complicating detection efforts.

Furthermore, ensuring the scalability and computational efficiency of plagiarism detection algorithms remains a pressing concern, particularly when dealing with large datasets. Balancing detection accuracy with computational resources is crucial for deploying effective plagiarism detection systems in real-world scenarios.

Another challenge lies in addressing cross-modal plagiarism, where plagiarists may translate textual content into images to evade detection. Developing techniques capable of detecting such instances of plagiarism requires interdisciplinary approaches combining natural language processing (NLP) and computer vision.

In conclusion, image plagiarism detection is an evolving field with significant implications for maintaining academic integrity and intellectual property rights. While existing research has made notable strides in developing detection techniques, further efforts are needed to overcome remaining challenges and ensure the effectiveness of plagiarism detection systems across diverse multimedia content types.

3. METHODOLOGY

i) Proposed Work:

The proposed system integrates text and image analysis techniques to detect potential plagiarism occurrences. Users can upload files and images, which are then compared against a text corpus to assess their originality. If any suspicious file data exhibits similarity to the corpus, plagiarism is flagged. This approach not only enhances content authenticity but also fosters a more accountable and trustworthy information environment. The system aims to overcome existing limitations by incorporating advanced methodologies in text and image analysis. By optimizing for scalability and enhancing cross-modal capabilities, it provides a comprehensive solution for detecting plagiarism across diverse content types. Through these innovations, the proposed system aims to offer a robust and efficient tool for maintaining academic integrity and intellectual property rights in the digital age.

ii) System Architecture:

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

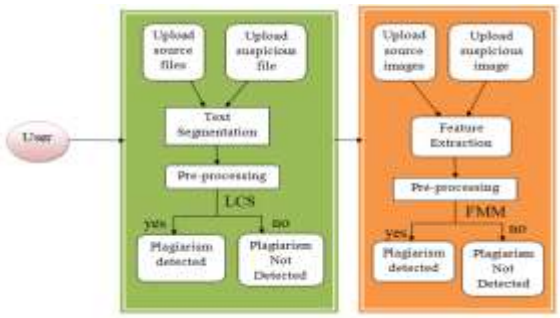


Fig 1 Proposed Architecture

iii) User signup:

The New User Signup module facilitates the registration process for individuals intending to utilize the plagiarism detection system. Users input essential details such as username, password, and contact information to create their account. Upon successful registration, they gain access to the system's functionalities and can subsequently log in to utilize its features. This module serves as the initial step for new users to engage with the system, ensuring a seamless onboarding experience and enabling them to leverage the system's capabilities effectively.

iv) User login:

The Login module offers an authentication mechanism for users to access the plagiarism detection system securely. Users input their username and password, and upon validation of their credentials, gain entry to the system. This module serves as a gatekeeper, ensuring that only authorized users with valid login credentials can access the system's functionalities. By providing a robust authentication process, the Login module enhances system security and safeguards sensitive user data. It offers users a seamless and reliable means of accessing the plagiarism detection system while maintaining the integrity and confidentiality of their accounts.

v) Upload source file:

The Upload Source File module enables users to submit text files containing original content they wish to check for plagiarism. Upon uploading, the system processes the source text, comparing it with existing files to detect any similarities indicative of plagiarism. This module plays a pivotal role in text plagiarism detection, empowering users to ensure the originality and integrity of their content. By facilitating the comparison process, it aids in identifying potential instances of plagiarism, thereby promoting academic integrity and intellectual property rights preservation. Users benefit from a streamlined process for verifying the authenticity of their text content within the plagiarism detection system.

vi) Upload suspicious file:

The Upload Suspicious File module functions similarly to the Upload Source File module but is intended for detecting potential plagiarism within submitted text files. Users upload documents suspected of containing plagiarized or copied content. The system then compares the suspicious file against its database, searching for resemblances to existing texts that may indicate plagiarism. This module serves as a critical tool in text plagiarism detection, empowering users to identify and address instances of academic dishonesty or intellectual property infringement. By facilitating the comparison process, it assists in maintaining the integrity and originality of textual content within the plagiarism detection system.

vii) Upload source image:

The Upload Source Images module allows users to submit images considered as original or authenticated visuals for image plagiarism detection. Upon upload, the system analyzes the distinctive features and patterns within these images. It then compares them with other images to identify any instances of similarity or exact matches that

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

may indicate unauthorized usage without proper attribution. This module plays a crucial role in safeguarding intellectual property rights and promoting content authenticity within the realm of image plagiarism detection. By facilitating the identification of potential image plagiarism, it contributes to maintaining integrity and accountability in visual content creation and distribution.

viii) Upload suspicious image:

The Upload Suspicious Image module serves as a counterpart to the Upload Source Images module, specifically designed for identifying potentially plagiarized images. Users submit images they suspect may have been copied from other sources. The system then conducts a search for similarities between the uploaded suspicious images and those stored within its database. By analyzing visual features and patterns, this module assists in detecting instances of image plagiarism, thereby safeguarding against unauthorized usage and promoting content authenticity. It provides users with a vital tool for ensuring the integrity and originality of visual content within the plagiarism detection system.

4. EXPERIMENTAL RESULTS

To run project install python 3.7 and then install DJANGO server and deploy code on that server and run from browser to get below screen

In above screen click on 'New User Signup Here' link to get below screen

In above screen user signup details entered and then click on 'Register' button to get below screen

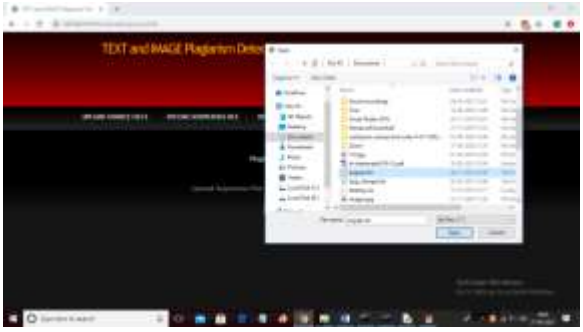
In above screen user signup process completed and now click on 'Login' link to get below screen

In above screen user is login and then click on button to get below screen

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

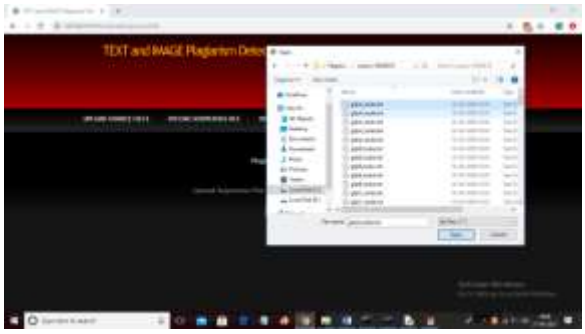
In above screen click on 'Upload Source Files' link to load all files from corpus folder

In above screen all files are loaded now click on 'Upload Suspicious File' button to load suspicious file and get result



In above screen I am selecting and uploading 'angular.txt' file and then click on 'Open' button to get below result and then click on 'Check Plagiarism' button to get result

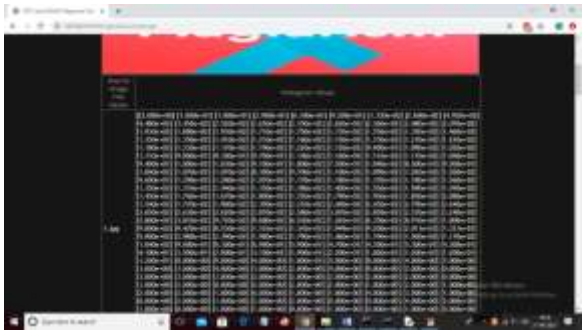
In above screen angular.txt file matched very little with g)pB_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result



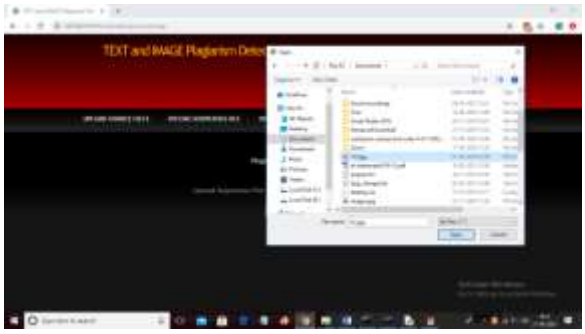
In above screen I am selecting and uploading first file and then click on button to get below result

In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on 'Upload Source Images' link to upload all images from 'images' folder

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024



In above screen from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched and now click on 'Upload Suspicious Image' link to upload some image

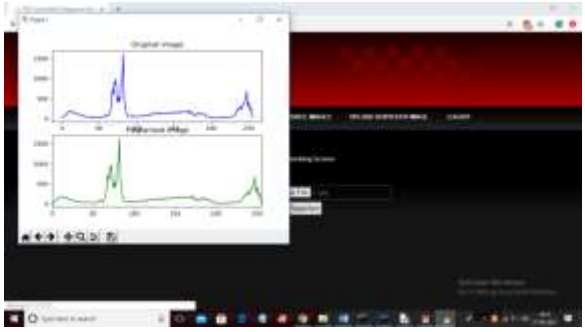


In above screen I am selecting and uploading '112.jpg' file and then click on 'Open' button to get below result

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

In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from "images" folder and see result

In above screen I am selecting and uploading '2.jpg' file from "images" database folder and below is the result



In above screen we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now close above graph to get below result

ISSN: 2278-4632
Vol-14, Issue-4, April: 2024

In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

Similarly u can upload other files and images to detect plagiarism or not.

5. CONCLUSION

Plagiarism involves converting the existing information in modified format. Today it is found in almost all fields of human activities because use of internet is high, so a lot of attention is given to identify and detect plagiarism. The project aims to identify instances of plagiarism using two distinct methods: text corpus comparison and image histogram matching. By analyzing a predefined text corpus, the system assesses the similarity of uploaded text files and reports potential plagiarism based on comparison scores. Additionally, the project employs histogram-based image analysis to detect similarities between uploaded images and a reference set of images. The system efficiently aids in detecting both textual and visual plagiarism, contributing to the preservation of academic integrity and content originality. Its integration of text and image analysis provides a comprehensive solution for plagiarism detection across different media types.

6. FUTURE SCOPE

In the future, the proposed project can be enhanced by incorporating real-time detection capabilities, enabling immediate identification of plagiarism instances as content is uploaded. Integration with machine learning models can further refine detection accuracy and adaptability to evolving plagiarism techniques. Additionally, leveraging blockchain technology for security can provide a tamper-proof and transparent framework for storing plagiarism detection records, enhancing data integrity and trust. These advancements would elevate the project's effectiveness in combating plagiarism and ensuring the integrity of digital content in diverse contexts.

REFERENCES

[1]G.MohamedKandahar,"100SocialMed iaStatisticsYoumustknow,"[online]Avail ableat:HTTP://blog. statutorily/social- media-statistics-2018-for business /[Accessed02Mar2019].

[2]Damian Radcliffe , Amanda Lam, "Social Media in middlemost,"[online] Available:https://www.researchgate.net/publication/323185146_Social_Media_in_the_Middl e_East_The_Story_of_2017[Accessed06 Feb2019].

[3]GM_BLOGGER,"Saudi Arabia Social Media Statistics, "G MI_blogger.[online]Availableat:HTTP://WWW. Global media insightful/ blog/Saudi-Arabia-social- media-statistics/[Accessed04May2019].

[4]Kit Smith,"49Incredible Instagram Statistics ,".Brand watch. [online] Available at :HTTP://www.brandwatch.com/blog/Instagram-stats/[Accessed10May2019].

[5]Selling Stock.(2014).Selling Stock.[online]Available at:HTTP://WWW. selling-stockroom /Article/18-billion-

images-uploaded –to - the-web-every-d [Accessed12Feb2019].

[6]Li,W.,Prada,S.,Fowler,J.E.,&Bruce,L.M.(2012).Locality-preserving dimensionality reduction and classification for hyperspectralimageanalysis.IEEETran sactionson Geoscience and RemoteSensing ,50(4),1185–1198.

[7]A. Krizhevsky,I. Sutskever,&Hinton, (2012).Image net classification with deep convolutional neural networks.In Advances in Neural Information Processing Systems,1097–1105.

[8]K. Ravi,(2018).Detecting fake images with Machine Learning.Harkuch Journal

[9]L.Zheng,Y.Yang,J.Zhang,Q.Cui,X.Zhang,Z.Li,etal.(2018).TI-CNN: Convolutional Neural Networks for Fake News Detection.

[10] R.Raturi,(2018).Machine Learning Implementation for Identifying Fake Accounts in Social Network.International Journal of Pureand Applied Mathematics,118(20),4785- 4797.

[11] Amirul S. Bin Ibrahin; Othman O. Khalifa; Diaa Eldein M. Ahmed , et. al., "Plagiarism Detection of Images" published in ieee open Access, available at https://ieeexplore.ieee.org/document/9250940.

[12] Simon Sepulveda; Gaspar Pizarro V.; Juan D. Velasquez, et. al., "A Plagiarism Detection Engine for Images in Docode" published in ieee open Access, available at https://ieeexplore.ieee.org/document/8609605.

[13] Hermann Maurer, Frank Kappe, Bilal Zaka , et. al., "Plagiarism - A Survey" published in jucs_org open Access, available at https://www.jucs.org/jucs_12_8/plagiarism_a_survey/jucs_12_08_1050_1084.

[14] James Douglas Beasley , et. al., "The Impact of Technology on Plagiarism Prevention and Detection: Research Process Automation, a New Approach for Prevention" published in docplayer open Access, available at https://docplayer.net/11862090-The-impact-of-technology-on-plagiarism-prevention-and-detection-research-process-automation-a-new-approach-for-prevention.html.

[15] Akshay S, Chaitanya B N, Rishabh Kumar, et. al., "Image Plagiarism Detection using Compressed Images" published in research gate open Access, available at https://www.researchgate.net/publication/334226542.