ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

Plagiarism detection using natural language processing

Rayana Vikas Goud[1] Dr. Rambabu[2]

[1] PG Student from the Dept of Computer Science & Engineering, Mallareddy University, Hyderabad

[2] Professor from the Dept of Computer Science & Engineering, Mallareddy University, Hyderabad

[1] rayanavikas1@gmail.com [2] drrambabu@mallareddyuniversity.ac.in

ABSTRACT:

Today, there is a greater emphasis on discussing plagiarism in research compared to the past. The availability of advanced web technologies and the ability to do complicated and efficient searches within a short timeframe have led to substantial negative consequences for research. Plagiarism detection tools only target text and do not consider graphics. Furthermore, photographs have a vital role in conveying a significant amount of information within an article or scientific study. Given the wide range of graphics, especially those found in computer texts, and the substantial information flowcharts hold, they could potentially serve as a vehicle for plagiarism. The objective of this research is to analyze the level of plagiarism in a document, specifically in relation to picture plagiarism, using the histogram model.

INTRODUCTION:

The global educational community often debates the topic of plagiarism. Plagiarism refers to the act of appropriating someone else's work and presenting it as one's own. Essentially, it transforms the current information into a new format. According to S. Hannabuss, plagiarism is defined as the unauthorized use of someone else's product or idea while falsely presenting it as one's own. Today, because of the widespread popularity of the internet, there is a vast abundance of publicly accessible materials. Currently, the internet serves as a repository for many sorts of files and data. Individuals may easily obtain the necessary information or data from the internet and replicate it, rather than composing their own textual content using their own cognitive abilities. The importance of detecting plagiarism has increased in recent years due to the ease with which plagiarists can access and replicate suitable textual content. However, it is becoming more challenging to accurately

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

detect copied data because of the vast number of potential data sources available on the internet. Instances of plagiarism are a common subject in several fields, such as academia, media, scientific research, politics, and other industries. This method of plagiarism detection is particularly valuable in situations where there is a lack of data gathering or when not all potential sources of copied material are accessible, making document-to-document comparison algorithms impractical. Plagiarism encompasses several forms, including literal copying, picture plagiarism, integral plagiarism, intrinsic plagiarism, extrinsic plagiarism, exact copy plagiarism, and text manipulation. There are several strategies and procedures available to identify plagiarism. Currently, text and picture manipulation algorithms lack the necessary accuracy for practical use. Thus, we have introduced a novel and straightforward way that utilizes the text-image identification methodology via file transfer. This method employs a machine learning algorithm to accurately identify instances of plagiarism between text sets and photographs. The program compares two files and determines the number of shared words between them. It then calculates a percentage value based on a specified threshold for detecting plagiarism. The hologram percentage for images assists in identifying image plagiarism, allowing us to obtain the plagiarized text and image series.

Existing System:

The current technique may be enough for identifying picture plagiarism when the source and suspected image have not undergone significant rotation. However, if there are rotational changes, the current methodology will be ineffective. The suggested technique would guarantee the detection of plagiarism even in the case of picture rotation, whether it is due to intentional manipulation or an assault including rotational changes. Furthermore, the current method lacks efficiency in accurately identifying plagiarism across various picture formats. The suggested system will guarantee the use of adaptive threshold values. The technique minimizes the time required for photo matching by gradually shrinking the search area with each refining iteration.

Disadvantages of Existing System:

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

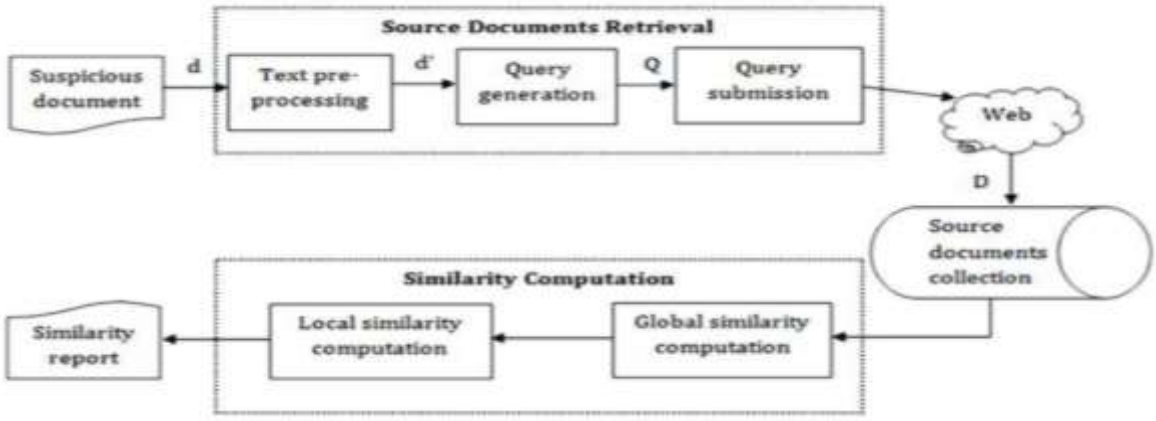
Plagiarism detection often examines the textual material using several reliable platforms to find instances of duplicated or closely similar text. However, these systems are generally not designed to detect plagiarism in pictures or files, since their primary focus is on text.

Proposed System:

The proposed plagiarism detection system would accept the user's input in the form of suspected plagiarized photos. We would create the picture's Phash value using the corpus technique. We will now examine the supplied picture for plagiarism by comparing it to the photographs in the local database. We save images in a database along with their corresponding phash values. To identify instances of plagiarism, the plagiarism detection engine will use a sequence of procedures. This would include calculating the hamming distance between the input picture's phash values and the photos in the database. The detection engine will display its findings upon completion. The corpus method may also be used to identify text files.

Advantages of Proposed System:

In this context, it is worth noting that there is no standardized approach to evaluating plagiarism detection algorithms. Custom-made datasets and a variety of performance metrics typically assess these algorithms. 1 This scenario makes the available studies almost impossible to compare.



System Architecture

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

LITERATURE SURVEY:

Ayoub Ali M. Saeed, Alaa Yaseen Taqa[1] states that"According to the scientific institutes, Plagiarism is defined as claiming someone else's ideas or efforts as one's own without citing the sources. Systems of plagiarism detection typically use a text similarity algorithm in a text document to look for common sentences between source and suspicious documents, either by directly matching the sentences or by embedding the sentences into a vector using TFIDF-like or other methods, and then calculating the distance or the similarity between the source and suspect sentence vectors. The cosine similarity method is one of the methods for determining that distance. To cluster the documents and choose only related documents for detection, an unsupervised Machine learning technique such as K-means could be utilized. In this paper, a plagiarism detecting application was created and tested on many text document types, including doc, docx, and pdf of research papers that collected from the web to build the source corpus. To calculate the level of similarity between the suspicious article and the corpus of source articles, the TFIDF text encoding approach is used with NLP ,Kmeans clustering and cosine similarity algorithms. The proposed application were carried out with five different documents and result in different ratios of plagiarism , the first document has 0.27 ratio, second document has 0.15 ratio, third document has 0.19 ratio while document 4 has 0.42 ratio and finally document 5 has 0.37 ratio of plagiarism. The generated detailed plagiarism ratio report presents the percentage of plagiarism in the suspicious article document. Depending on the threshold value, the application will decide if the suspicious document is acceptable or not".

Zhizhi Wang, Chaoji Zuo, Dong Den[2] states that "In this paper, we study the near-duplicate text alignment search problem, which, given a collection of source (data) documents and a suspicious (query) document, finds all the near-duplicate passage pairs between the suspicious document and every source document. It finds applications in plagiarism detection. Specifically, the first two steps in plagiarism detection are source retrieval and text alignment. Source retrieval finds candidate source documents in a corpus that share content with the suspicious document while text alignment finds all the similar passage

pairs between the suspicious document and every candidate source document. This problem is computation-intensive, especially for long documents. This is because there are O(n 2m2) passage pairs between a single source document with n words and a suspicious document with m words, not to mention the large number of source documents in a corpus. Due to the high computation cost, existing solutions primarily rely on heuristic rules, such as the "seeding-extension-filtering" pipeline, and involves many hard-to-tune hyper-parameters. To address these issues, a recent work Allign leverages the min-wise hash sketch for the text alignment problem. However, Allign only works for two documents and leaves the source retrieval problem unattended. In this paper, we propose to leverage the bottom-k sketch (a.k.a., conditional random sampling) to estimate the similarity of two passages. We observe that many nearby passages in a document would share the same bottom-k sketch. Thus we propose to group all the passages in a document by their sketches. We prove that all the O(n 2) passages can be partitioned into O(nk) groups in a document with n words and develop an algorithm to generate these groups in O(n logn + nk) time. Then, to address the source retrieval problem, we only need to find groups of passages with "similar" bottom-k sketches. Every passage pair in two groups with "similar" sketches are near-duplicates. Experimental results on real-world datasets show that our techniques are highly efficient".

Ravi Sankar Landu [3] states that "In this age of Internet and Cloud based applications, image compression has become all the more important, especially in the areas of Telemedicine, Satellite data etc. The images need to be compressed with high computation efficiency and regenerated with minimal losses. There are many techniques related to Image compression, mainly classified into Lossless and Lossy compression techniques. Many computer programs are developed for implementing these compression techniques. However, Deep Learning has been used for image compression since 1980s and has been adopted into most of the Artificial Intelligence (AI) platforms. This paper gives brief insights into some Deep Learning techniques, some AI platforms and different functions/methods available in those platforms."

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

Venkataravana Nayak K, J S Arunalatha, etal [4] states that "This paper proposes an Image Retrieval model using Multiple Feature Sets and Artificial Neural Network (IR-MFS-ANN), where the multiple features Histogram of oriented Gradient (HoG), Overlapping Local Binary Pattern (OLBP), Color and Statistical features are considered. Visual information is one of the most important data in the field of social networking, medicine, military, and these areas contain an enormous volume of semi-organized and organized heterogeneous information related to explicit subjects. However, retrieval and usage of suitable information from the comprehensive information archives are important to meet the content extraction and retrieval challenges. To improve retrieval performance, the image representation, modeling, scalable algorithm that permit accessing large archives are integrated into the retrieval framework. The proposed model utilizes ANN to find the distance between feature vectors. The proposed algorithm is tested and analyzed with various retrieval techniques and it is found that ANN-based image retrieval outperforms the state-of-the-art techniques [1–5]. The proposed method results in accuracy of 94%, 92%, 95%, and 94% for Wang, Cifar-10, Oxford Flower, and ImageNet standard databases respectively"

Hossam Elzayady, Mohamed S. Mohamed, etal [5] states that "Recent studies show that social media has become an integral part of everyone's daily routine. People often use it to convey their ideas, opinions, and critiques. Consequently, the increasing use of social media has motivated malicious users to misuse online social media anonymity. Thus, these users can exploit this advantage and engage in socially unacceptable behavior. The use of inappropriate language on social media is one of the greatest societal dangers that exist today. Therefore, there is a need to monitor and evaluate social media postings using automated methods and techniques. The majority of studies that deal with offensive language classification in texts have used English datasets. However, the enhancement of offensive language detection in Arabic has gotten less consideration. The Arabic language has different rules and structures. This article provides a thorough review of research studies that have made use of artificial intelligence (AI) for the

identification of Arabic offensive language in various contexts."

Jiabao Liu, Yang Fu, Bin Wang,[6] states that "The automatic identification technology of goods in intelligent three-dimensional warehouse has the problems of high cost of RFID electronic tag and low recognition rate of goods image recognition technology. This paper presents an automatic cargo image recognition method based on the combination of feature matching and template matching. Firstly, the minimum container rectangle (MBR) in the image is extracted as the template for image matching. Secondly, the scaleinvariant feature transform (SIFT) algorithm is used for rough matching of cargo images. Finally, the deformable multiple similarity measure (DDIS) method was used to match accurately. Experimental results show that the proposed method has high recognition accuracy and relatively fast recognition speed, and has certain practical value".

Rakesh Ranjan, Dr. Vinay Avasth [7] states that "In image processing, edge detection is a critical issue. Edge detection is a key approach for evaluating the edge of various objects in a digital image. These edges are found using the gradients, which are present in the image. The intensity and value of pixels determine the gradients. In digital images, edge detection lowers the quantity of data and filters out irrelevant data while maintaining the image's key structural features. In this paper, a new edge detection approach based on a fuzzy rule-based system is proposed. In digital image processing, the proposed method typically depends on fuzzy logic systems. The main goal of this system is to show how fuzzy logic may be used in image processing. This paper provides a fuzzy logic-based edge detection technique that uses a sharpening Gabor filter to regulate edge quality and a Gaussian filter to reduce noise caused by sharpening. This is determined by utilizing applications such as "Peak Signal to Noise Ratio (PSNR) F-Measure, and Hausdorff distance (HOD) to prove that fuzzy logic outperforms the proposed system. The findings for edge detection approaches are included in high quality. The proposed approach outperforms most commonly used traditional edge detection methods. The proposed method also reduces the number of noisy features and may be used for a wide range of images".

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 3, 2024

Farah Khaled, Mohammed Sabbih H. Al-Tamimi [8] statesthat“Plagiarism Detection Systems play an important role in revealing instances of a plagiarism act, especially in the educational sector with scientific documents and papers. The idea of plagiarism is that when any content is copied without permission or citation from the author. To detect such activities, it is necessary to have extensive information about plagiarism forms and classes. Thanks to the developed tools and methods it is possible to reveal many types of plagiarism. The development of the Information and Communication Technologies (ICT) and the availability of the online scientific documents lead to the ease of access to these documents. With the availability of many software text editors, plagiarism detectionsbecomes a critical issue. A large number of scientific papers have already investigated in plagiarism detection, and common types of plagiarism detection datasets are being used for recognition systems, WordNet and PAN Datasets have been used since 2009. The researchers have defined the operation of verbatim plagiarism detection as a simple type of copy and paste. Then they have shed the lights on intelligent plagiarism where this process became more difficult to reveal because it may include manipulation of original text, adoption of other researchers' ideas, and translation to other languages, which will be more challenging to handle. Other researchers have expressed that the ways of plagiarism may overshadow the scientific text by replacing, removing, or inserting words, along with shuffling or modifying the original papers. This paper gives an overall definition of plagiarism and works through different papers for the most known types of plagiarism methods and tools”.

Kaustubh Gayadhankar, Rishi Patel, Hrithik Lodha, Mr. Swapnil Shinde[9] states that “In Today's date plagiarism is a very important aspect because content originality is the client's prior requirement. Many people on the internet use others' images and get publicity while the owner of the image or data won’t get anything out of it. Many users copy the data or image features from the other users and modify it a little bit or create an artificial replica of it. With sufficient computational power and volume of data, the GAN models are capable enough to produce fake images that look very much similar to the real images. These kinds of images are generally not detected by modern

plagiarism systems. GAN stands for generative adversarial network. It has two neural networks working inside. The first one is the generator which generates a random image and the second one is the discriminator which identifies whether the image being generated is a real or a fake image. In this paper, we have proposed a system that has been trained on both fake images (GAN Generated images) and real images and will help us in flagging whether the image is plagiarised or a real image".

S Sowmya Mithra, K P Supreeth [10] states that "plagiarism is a big problem in academics, researches and it can be a big problem in every department of education sector. Students plagiarize in different areas like homework, assignments, projects, essays etc. Collecting information from multiple sources is considered as a process of learning but this learning experience is diminished when students plagiarize by copying assignments and getting credits for work they have not done. In this project we have developed a system that helps in detecting plagiarism of images in which whenever a student submits an assignment, it detects whether it is plagiarized or not by comparing with other student assignments. For this we are using Reverse Image Processing to get proposed output."

MODULES:

1.New user Signup

Firstly, user will register into Application. It helpful to login into Application with username and password.

2.Login

User will login into Application through username and password**.**

3.Upload Source File

Folder is created into Upload Source Files' link to load all files from corpus folder.

4.Upload Suspicious files

To load suspicious file and get result, user will upload file to upload suspicious files the result is execute. LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u you may enter any text file and get result.

5. Upload Source Image

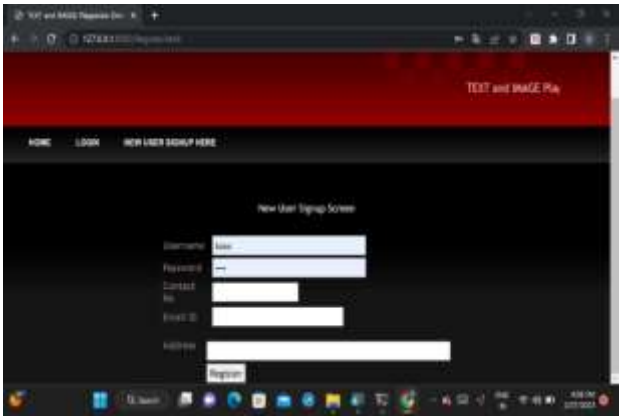
In this module from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched.

6. Upload Suspicious Image

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. histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarized and now upload image from “images” folder and see result. We can both original and uploaded image histogram is matching 100% so plagiarism is detected and now get below result. Histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result

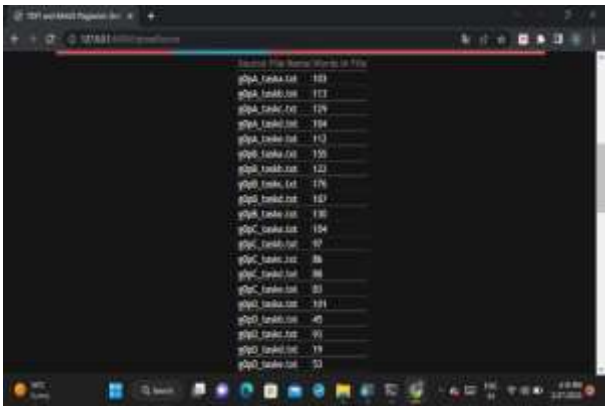
Results:

Screenshot of Home page

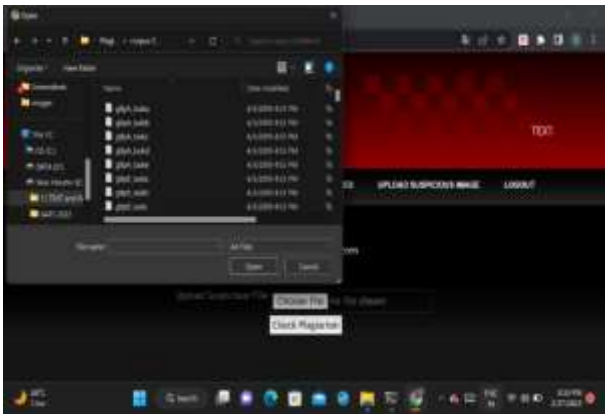


Screenshot of Registration Page

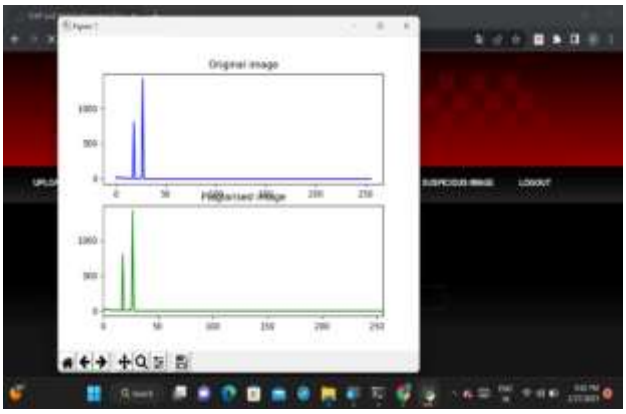
Screenshot of Login page



Screenshot of Uploaded Source Text Files



Screenshot of Uploading Suspicious Text Files



Screenshot of Histograms of Original Image & Plagiarized Image

Screenshot of Plagiarism Detection in Suspected Image File

Conclusion:

The Corpus is the first standardized collection of texts specifically designed for assessing the effectiveness of automated plagiarism detection systems. The inaugural International Competition on Plagiarism Detection effectively utilized the corpus. This study asserts that the corpus and performance measurements will serve as a potent tool for evaluating future research on plagiarism detection. Presently, a more refined iteration of the corpus is under construction.

REFERENCES:

[1] Ayoub Ali M. Saeed, Alaa Yaseen Taqa, "A proposed approach for plagiarism detection in Article documents",Sinkron: Journal dan Penelitian Teknik Informatika, ISSN: 2541-2019, 2022, pp: 568-578.

[2] Zhizhi Wang, Chaoji Zuo, Dong Den, "TxtAlign: Efficient Near-Duplicate Text Alignment Search via Bottom-k Sketches for Plagiarism Detection", SIGMOD ’22, June 12–17, 2022, Philadelphia, PA, USA, pp:1146-1159.

[3] Ravi Sankar, “Image compression using AI:Brief insights into deep learning techniques and frameworks”, International Journal of Engineering, Science, Technology and Innovation (IJESTI), ISSN: 2582-9734, 2022, pp: 1-6.

[4] Venkataravana Nayak K, J S Aruna Latha, G U Vasanthakumar, K R Venugopal, "Soft Computing based Artificial Neural Network and Multiple Feature set Intelligence

system for Image Retrieval", Research square, 2022, pp: 1-34.

[5] Hossam Elzayady, Mohamed S. Mohamed, Khaled M. Badran, Gouda I. Salam, “Detecting Arabic textual threats in social media using artificial intelligence: An overview”, Indonesian Journal of Electrical Engineering and Computer Science Vol. 25, No. 3, March 2022, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v25.i3.pp1712-1722

[6] Jiabao Liu, Yang Fu, Bin Wang, "Resolution system based on Image processing Technology Applied to Logistics system", International Conference on Information Technology, Education and Development,2021, pp: 509-513.

[7] Rakesh Ranjan, Dr. Vinay Avasthi, "Enhanced Edge Detection Technique in Digital Images Using Optimized Fuzzy Operation", Webolog, ISSN: 1735-188X, 2021, pp: 5402-5416.

[8] Farah Khaled, Mohammed Sabbih H. Al-Tamimi, "Plagiarism Detection Methods and Tools: An Overview", Iraqi Journal of Science, ISSN: 0067-2904, 2021, Vol. 62, No. 8, pp: 2771-2783.

[9] Kaustubh Gayadhankar,Rishi Patel HrithikLodha, Mr.Swapnil Shinde, "Image plagiarism detection using GAN - (Generative Adversarial Network)", ITM Web of Conference, 2021, pp:1-8.

[10] S Sowmya Mithra, K P Supreeth, "Online Plagiarism detection for images",Journal of Emerging Technologies and Innovative Research (JETIR), ISSN-2349-5162, 2021, pp: d391-d395.

PG Student from the Dept of Computer Science & Engineering, Mallareddy University, Hyderabad

Dr. Rambabu, Professor from the dept of Computer Science & Engineering, Mallareddy university, Hyderabad