ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 10, Issue 2, 2022

A Framework for Detecting Data and Image Copying in Digital Content

Mr.Neralla Vivek[1]., K.Supritha[2]
1 Assistant Professor, Department of CSE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India
2, B.Tech CSE (19RG1A0590),
Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

Abstract:

Plagiarism in research is debated much more now than it was in the past. Conditions on the Web and the potential for quick, sophisticated searches are graded accordingly, and as a consequence, there have been serious harms to the study. Anti-plagiarism software only targets text and ignores graphics. On the other hand, graphics play a crucial role in conveying the vast amount of information included in scholarly articles and research. Given the vast variety of pictures, particularly the many images discovered in the computer's texts, and the fact that flowcharts are excellent for conveying a lot of information, plagiarism may be a possibility. This project's goal is to use a histogram model to analyze a paper's plagiarism rate with regard to photographs.
Keywords: KNN, Machine Learning, Copying, Text Copying, Copying Images

I. INTRODUCTION

All around the globe, the educational community often discusses the problem of plagiarism. It has to do with claiming credit for work that you have taken from someone else. In essence, it transforms the current data into a different format. "The act of using somebody else's creation or idea without permission and presenting it as one's own" is how S. Hannabuss defines plagiarism. Due to the internet's immense popularity, a vast number of papers are now publicly accessible. The internet is now a source for many kinds of data and files. Instead of creating their own original text document from scratch, people may simply get the necessary information or data from the internet and copy it. The ability to identify plagiarism has grown in importance in recent years due to the ease with which a plagiarist may locate relevant text material to copy. On the other hand, since there are so many potential online sources for data, it becomes harder and harder to detect copied content. Cases of plagiarism are often discussed in a variety of fields, including academia, media, science, politics, and even many other fields. When data collection is unavailable or not all potential copy sources are accessible, making document-to-document comparison algorithms unusable, this method of detecting plagiarism is very helpful. There are many different forms of plagiarism, including precise copying, text manipulation, integral, intrinsic, extrinsic, and literal copying. In a similar vein, there are several approaches and techniques for detecting plagiarism. Systems now in use that rely on text and picture editing methods are too imprecise for real-world use. In order to identify plagiarism between text sets and pictures, we have thus suggested a novel, simple method that is based on the texta image identification methodology via file transfer method and makes use of a machine learning approach. It compares two files to find the number of words that are similar. Next, we compute a percentage value based on the threshold value needed to identify plagiarism. The image hologram percentage aids in the detection of image plagiarism, allowing us to identify plagiarized text and image series.

II. RELATED WORK

There has been a comparison of text-based, citation-based, and shape-based plagiarism detection techniques. Text-based plagiarism detection techniques have been shown to be over 70% effective in copy-paste plagiarism comparisons, whereas citation-based techniques have shown to be ineffective in this area. Regarding the plagiarism in the translated papers, text-based techniques have been effectively less than 5%; under the citation-based approach, this percentage is over 80%.
The comparison of photos has not been carried out by the current system.

IMI Subroto, A Selamat, and Choon-Ching Ng (2009)[1] Hybrid KNN is used for Arabic Script Web Page Language recognition.One of the most important challenges in text-based language identification using the same script is how to generate trustworthy characteristics and handle the enormous number of languages spoken worldwide.

Liaqat, Ahmad Gull, and Aijaz Ahmad (2011)[2]From Back propagation to Adaptive Learning Algorithms: Advanced Supervised Learning in Multi-layer Perceptrons. Many advancements in the method for feed-forward neural network weight training have been developed since the back propagation algorithm was first presented.

Gamini Wijayrathna and Upul Bandara (2012)[3]Identifying and detecting plagiarism in source code with the use of machine learning techniques. Plagiarism using source code is a serious issue in academics right now. Programming assignments are used in academic settings to assess students enrolled in programming courses.

Ismail A great deal Ali Selamat and Ibnu Subroto (2014)[4] Hybrid Artificial Neural Network and Support Vector Machine-Based Plagiarism Detection over the Internet: The majority of plagiarism detection methods use similarity measuring techniques. In essence, two phrases that are identical each convey the same concept.

III. EXISTING SYSTEM

There has been a comparison of text-based, citation-based, and shape-based plagiarism detection techniques. Text-based plagiarism detection techniques have been shown to be over 70% effective in copy-paste plagiarism comparisons, whereas citation-based techniques have shown to be ineffective in this area. Regarding the plagiarism in the translated papers, text-based techniques have been effectively less than 5%; under the citation-based approach, this percentage is over 80%.

The comparison of photos has not been carried out by the current system.

IV. PROPOSED SYSTEM

There are two stages to the suggested system: testing and training. They are seen as using a histogram in the training phase to aid in learning, and using this network's modeling to aid in the recognition stage during the test phase. Based on the query image correlation rate with each test picture, the data analysis technique and input image similarity detection rate with images in the database pick photos with the greatest correlation. The expert is ultimately responsible for providing the final interpretation, and correlation values acquired at this point report as the tested image plagiarism.

V. SYSTEM IMPLEMENTATION

Preprocessing: Preprocessing aims to improve the picture data by reducing undesired distortions or enhancing certain key visual elements that are necessary for further processing.

• Train the Model: The module will begin to be trained using the keras functions.

• Assess the Model: Our dataset contains photos that will be used to assess the effectiveness of our model.

5.1 Formula

One machine learning method that uses the supervised learning approach is K-Nearest Neighbor. The K-NN method classifies the new instance in the category that most closely resembles the existing categories based on its assumption that the new case and the current examples are comparable. The K-NN method categorizes a new data point based on similarity after storing and comparing all of the existing data. This indicates that the K-NN algorithm can quickly classify and identify newly appearing data into a well-suited category.

Although the K-NN technique or approach is mostly utilized for classification issues, it may also be used for regression.

5.2 Architecture of the System

The conceptual model that describes a system's behavior, structure, and other aspects is called a systems architecture, or just systems architecture. A formal description and representation of a system, structured to facilitate inference about the functions and patterns of the system, is what constitutes an architectural description. System components, their outwardly evident characteristics, and the connections between them may all be included in a system design.

ISSN 2321-2152
www.ijmece.com
Vol 10, Issue 2, 2022

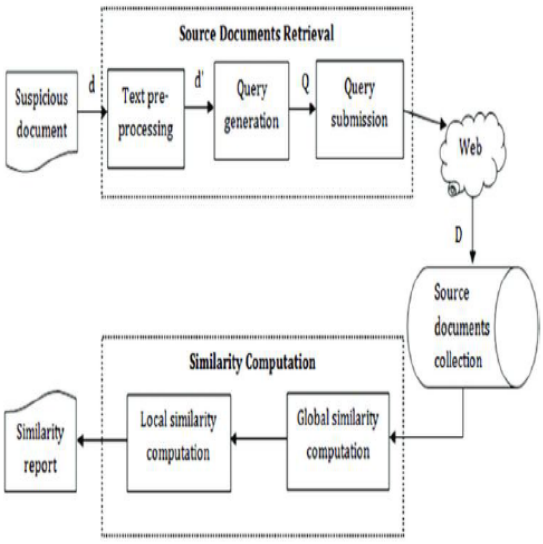

Figure 1: System Architecture

VI. RESULT

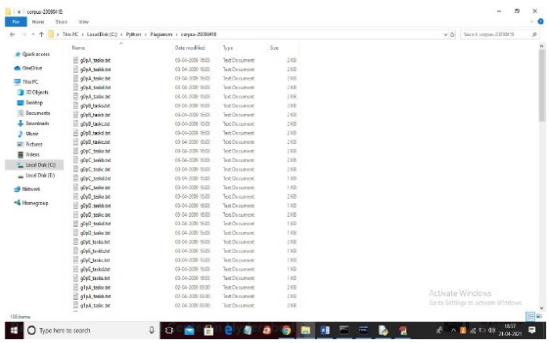

Figure 2: Data Set Input Text Files

The photos below are being used to create a histogram model; plagiarism will be identified if any questionable visual resemblance is found between this histogram and the original images. View the photos used to create the histogram model below.

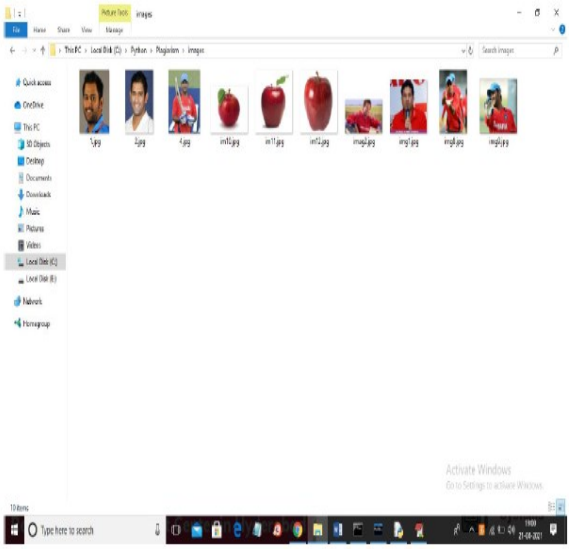

Figure 3: Input Image Files(Data Set

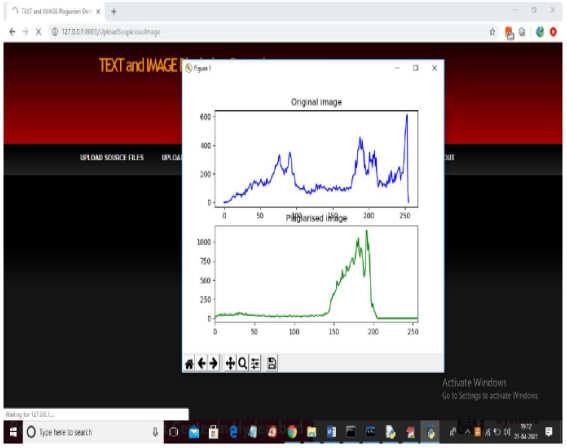

Figure 4: Graph of Output

The above screen displays the histogram that we calculated for the submitted picture and the database image. Since there is no match, plagiarism will not be identified. Close the above graph to view the result below.

ISSN 2321-2152
www.ijmece.com
Vol 10, Issue 2, 2022

Figure 5: Image plagiarism checking output

The pixel matching score in the above screen histogram is 15173 out of 40000 pixels, indicating that the picture is not plagiarized. Upload the image from the "images" folder to see the outcome.

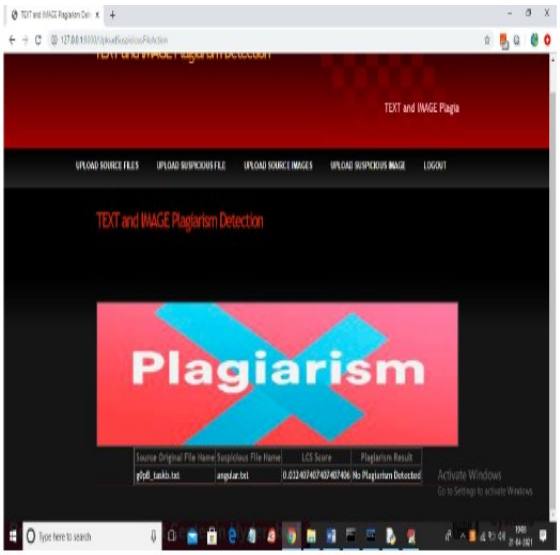

Figure 6: The text output Checking for plagiarism

The accompanying screen shows that the angular.txt file and the g)pB_taskb.txt corpus file matched very little. A similarity score of 0.03 indicates that there was no plagiarism found. You may now upload any file from the corpus to see the results.

In a similar vein, you may test the program by uploading any text file or picture.

VII. CONCLUSION

Converting preexisting knowledge into a different format is plagiarism. Nowadays, plagiarism is prevalent in almost every aspect of human endeavor due to the widespread use of the internet, prompting a great deal of focus on identifying and preventing it. Based on certain experimental findings, it seems that using hybrid machine learning techniques and methodologies to detect plagiarism often improves performance. The hybrid approach does not, however, necessarily result in more precise and superior performance. Therefore, we have developed a procedure that will increase accuracy and performance utilizing the k-NN machine learning technology. After comparing all available techniques, we can say that the k-nearest neighbor approach works very well for both pattern recognition and locating duplicate datasets (text, picture) in order to identify plagiarism. Our approach to plagiarism detection offers more speed and accuracy. To do this, we have put into practice a method that demonstrates how a text and picture collection are processed and compares a given file to related, already-existing files to identify plagiarism.

REFERENCES

[1]. 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.

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

[3]. 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.

[4]. 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.

[5]. BarrónCedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696–700, Springer.