IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Plagiarism Checking with extracted text from image using NLP: A Comprehensive Overview

Daksha D, Bhuvana K, N N Siri, Sanjana M

Abstract

Plagiarism, the unauthorized use or reproduction of someone else's work, poses a significant challenge in academic and professional settings. This research explores an innovative approach to plagiarism detection by integrating Natural Language Processing (NLP) techniques with text extraction from images. The proposed system leverages advanced NLP algorithms to analyze and compare textual content extracted from images, aiming to identify instances of potential plagiarism. The study employs a comprehensive set of experiments to evaluate the effectiveness and accuracy of the proposed system, showcasing its potential as a valuable tool for enhancing plagiarism detection in diverse domains. The outcomes of this research contribute to the ongoing efforts in developing more robust and inclusive strategies for ensuring academic and professional integrity. There are a lot of potential users who want to extract the text from images, archiving documents etc. For this reason, user need an Optical Character Recognition (OCR). It aims at detecting textual regions from the document and separating it from the graphics portion. Getting information directly from applications forms and it saves a lot of time

Introduction

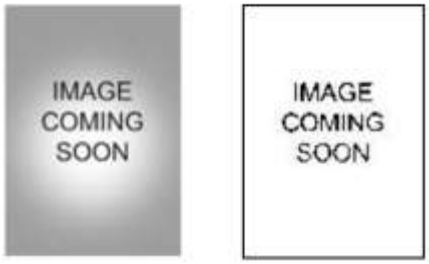

Fig 1: IMAGE COMING SOON will be extracted from the above image using OCR Technique.

Fig : Extracted text from grey Scale image

Plagiarism is defined as to take or theft some work and present it has one's own work. This grammar and plagiarism checker system is used to analyze the plagiarism data. Plagiarism affects the education quality of the students and thereby reduce the economic status of the country. Plagiarism is done by paraphrased works and the similarities between keywords and verbatim overlaps, change of sentences from one form to other form, which could be identified using Word Net etc. This plagiarism detector measures the similar text that matches and detects plagiarism. Many methods are employed in detecting plagiarism. Usually plagiarism detection is done using text mining method. NLP is a branch of artificial intelligence that focuses on the interaction between computers and human language. Its primary goal is to enable machines to understand, interpret, and generate human-like text. In the context of plagiarism checking, NLP plays a pivotal role by leveraging its capabilities to analyze and compare textual content efficiently. The traditional methods of plagiarism detection, often reliant on exact matching or simple keyword analysis. NLP however ,excels in understanding the semantics and context of language, allowing for a more sophisticated approach to identifying instances of plagiarism. One of the challenges in plagiarism detection arises when dealing with content embedded within images. Whether it's due to the limitations of existing text recognition

technologies or intentional efforts to obfuscate plagiarism, extracting text from images has proven to be a formidable task. This process, often referred to as Optical Character Recognition (OCR), allows NLP algorithms to transform image-based text into a format that can be analyzed and compared more effectively. The extracted text from images can then be subjected to a range of NLP techniques for plagiarism checking. OCR is a technology that enables you to convert different types of documents such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data(5).Text Extraction involves a computer system designed to translate captured or scanned documents into Machine editable text. OCR began as a field of research in artificial intelligence and computational Vision. The process by which Image Text converted into plain text that computer can recognize its ASCII character is Text Extraction. The users can scan a document and have the text of that document in .txt or .doc

Literature survey

Optical Character Recognition from Text Image Several approaches for text detection in images and videos have been proposed in the past. Based on the methods being used to localize text regions, these approaches can be categorized into two main classes: connected component based methods and texture based methods. Optical character recognition (OCR) is the conversion of scanned images of printed, handwritten or typewritten text into machine-encoded text. This technology allows to automatically recognizing characters through an optical[7]. Text Detection and Recognition in Imagery

Early text detection and recognition research was a natural extension of document analysis and recognition research, moving from scanned page images to camera captured imagery, focusing on basic preprocessing, detection and OCR technology. Efficiently extracts and interprets text from images, enhancing data accessibility and automating information retrieval. But the disadvantage in this paper is the complexity of environments, flexible image acquisition styles and variation of text contents pose various challenges are Fonts, Uneven lighting, Scene complexity[4]. Automated Text Extraction from Images using OCR System

Text extraction can be achieved by applying text detection that identifies image parts containing text, text localization finds the exact position of the text, text segmentation separates the text from its background and binarization process converts the coloured images into binary. An OCR system, accordingly may face challenges in accurately recognizing handwritten or complex fonts, limiting its effectiveness in certain contexts[11]. Online assignment Plagiarism Checker Using Machine Learning It employs advanced algorithms to analyze and compare student submissions with a diverse dataset. By training on labeled data, the system identifies patterns and anomalies, offering a highly accurate and adaptable solution for detecting plagiarism. This technology enhances academic integrity by providing educators with a robust tool to ensure the originality of student work. Include occasional false negatives, where subtle instances of plagiarism go undetected due to evolving tactics. Overreliance on training data quality may introduce biases, impacting the system's fairness. Privacy concerns arise as the technology involves analyzing student submissions, necessitating careful handling of sensitive information and clear communication with users[1][7]. Online Assignment Plagiarism Checking integrates Data Mining and NLP to ensure academic integrity. By analyzing the similarity between student submissions and a reference corpus, the system employs advanced algorithms to detect potential plagiarism. Through a user-friendly interface, it generates detailed reports for instructors, facilitating efficient review and maintenance of academic honesty. It include heightened accuracy in identifying plagiarism through sophisticated similarity analysis. By leveraging NLP, the system comprehensively understands linguistic nuances, enhancing its ability to detect subtle instances of academic dishonesty .Privacy concerns may arise as the technology involves scanning and analyzing student submissions, necessitating careful consideration of data protection measures[12]. Online Plagiarism Detection for images utilizes advanced algorithms to compare and analyze visual content, ensuring the originality of images submitted online. Through image recognition and similarity metrics, the system identifies instances of image plagiarism, providing a robust solution for maintaining integrity in visual-based assignments and content creation. This technology provides a crucial tool for educators, content creators, and image-based platforms, enhancing the ability to maintain originality and intellectual property rights. Additionally, it contributes to a fair and ethical online environment by preventing the unauthorized use of images .Involve challenges in discerning context and intent, leading to potential false positives when images are similar but used legitimately[10]. Online Plagiarism Detection for images utilizes advanced algorithms to compare and analyze visual content, ensuring the originality of images submitted online. Through image recognition and similarity metrics, the system identifies instances of image plagiarism, providing a robust solution for maintaining integrity in visual-based assignments and content creation .Include its effectiveness in identifying visual content similarities, offering a

comprehensive approach to ensure the authenticity of graphical materials. This technology provides a crucial tool for educators, content creators, and image-based platforms, enhancing the ability to maintain originality and intellectual property rights. The effectiveness of the system may be influenced by the quality and diversity of the reference image database, impacting its ability to identify nuanced instances of plagiarism accurately[3][4].

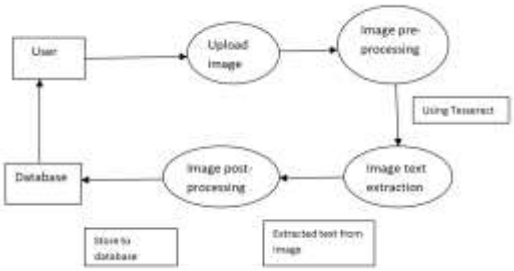

Fig 2 : System Design.

Proposed Work

In this chapter, we will discuss the different design constraints and available options for implementing the various stages of plagiarism checking with extracted text from image using NLP. The focus of this chapter is on the design steps involved in the text extraction implementation .The system we are discussing is designed to identify the plagiarism using Natural language processing and Optical character recognition algorithms. The process of predicting plagiarism involves several steps. Firstly, relevant images is collected, including diagram type image formats. In the realm of plagiarism checking, the integration of Natural Language Processing (NLP) to analyze and compare text content has become paramount. The proposed system aims to elevate the efficacy of plagiarism detection by incorporating NLP techniques, particularly in the context of extracting textual information from images. Leveraging advanced algorithms, the system will employ Optical Character Recognition (OCR) to convert image-based text into machine readable format. Subsequently, NLP will be harnessed to analyze and comprehend the semantic meaning of the extracted text. This innovative approach not only broadens the scope of plagiarism detection to include content embedded in images but also enhances the accuracy of comparisons by delving into the intricacies of language semantics. The system's architecture will encompass a multi-layered design, comprising an OCR module for image text extraction and an NLP module for semantic analysis. The OCR component will utilize cutting-edge image processing techniques to convert text from images into a structured format. The NLP module, empowered by pre-trained models, will then scrutinize the linguistic nuances, syntactic structures, and contextual semantics of the extracted text. The integration of these modules will furnish a comprehensive plagiarism checking system capable of detecting similarities not only within textual content but also across diverse media, thereby providing a robust defense against intellectual property infringement in the digital landscape.

In the devised model as shown the idea proposed is to take in a number of images of documents like identity proofs of individuals and classify them into classes, such aspassport and license .Once the images are classified, they are subjected to the text extraction module. The text data are extracted from the classified images. The extracted credentials from the images are then stored in the database. Text Extraction Text extraction is implemented using the Tesseract OCR package which contains an optical character recognition (OCR) engine - libtesseract and a command line program –Tesseract. Tesseract includes a new neural net called Long Short-Term Memory (LSTM) based OCR engine, which focuses on line recognition and also recognizes character patterns. The LSTM network is the units of Recurrent Neural Networks. The Python-Tesseract is an optical character recognition (OCR) tool in python used for text extraction .The data flow of this project is users upload the image then image scaling or gray scale conversion takes place and store the record. once it is stored so we can view, delete, update, and record the text. And it is stored in the database

Methods

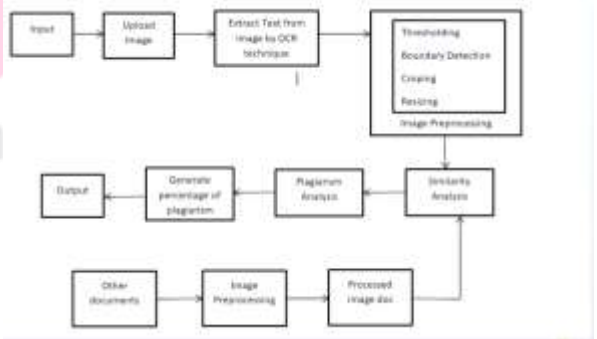

Fig 3 : Method for Text Extraction

Multiple packages such as 4 dimensional cube have their own dictionary for this process that give due results. NLP is a process of manipulating or understanding the text or speech by any software or machine. An analogy is that humans interact and understand each other views and respond with the appropriate answer .In NLP, the interaction, understanding, and response are made by a computer instead of a human. Natural Language toolkit(NLTK) The natural language toolkit is a suite of libraries and programs for symbolic and statistical natural language processing(NLP) for English written in Python programming language. NLTK is a leading Platform for building python programs to work with human language data. The system goes through multiple stages before the final value can be predicted accurately. These stages are pre-processing, text extraction, text recognition and plagiarism analysis of the algorithm. Once the desired accuracy is obtained, we can integrate the system with an application for real-world use .We will design a system to detect plagiarism in academic assignments in the suggested system, which will help to prevent students from copying other students' assignments and will improve the quality of education and also will help to improve personal skills of student and student can also check the grammar from the assignment.

The plagiarism detector in this system compares comparable texts and detects plagiarism. As well semantically checking will be also done with respect to assignment. For detecting the plagiarism, we will use data mining algorithm and natural language processing. Step 1 : Giving the input The input can be given as any kind of document by entering the location of the file where all the documents have been be stored all the assignments submitted by the students should be stored in that location. Once we enter the file location it will take all the document and process to the next step. Step 2 : Conventing to text document Once all the files are got the step is to convert those files to text format. Whatever the file can be it can be document, word, pdf (Portable Document Format) it will be converted to text format. We are converting to text files because it will easy to do pre-processing in the text files rather than in other files. Step 3: Pre-Processing. The text files will go through the pre-processing where those following process will happen and it will be ready to find the similarity in the between the files. Preprocessing consists of cleaning, stopword, stemming. Step 4: Tokenization The process of tokenizing or breaking a string of text into a list of tokens is known as tokenization. Tokens can be thought of as components; for example, a word in a sentence is a token, and a sentence is a token in a paragraph. Step 5: Cleaning The practise of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are numerous ways for data to be duplicated or mislabeled when merging multiple data sources. Even if the data is right, outcomes and algorithms are untrustworthy if the data is erroneous. Because the methods differ from dataset to dataset, there is no definite way to prescribe the exact phases in the data cleaning process. Step 6: Stop WordA stop word is a widely used word (such as "the," "a," "an," or "in") that a search engine has been configured to disregard while indexing and retrieving entries as the result of a search query. Step 7: SteammingThe process of developing morphological variants of a root/base word is known as stemming. Stemming algorithms or stemmers are terms used to describe stemming programmes. The phrases "chocolates," "chocolatey," and "Choco" are reduced to the root word "chocolate," and "retrieval," "retrieved," and "retrieves" are reduced to the stem "retrieve."Step 8 : Similarity Analysis After removing stop words and stemming. the phrases, the algorithm developed to discover similarity among text documents determines content similarity among specified texts. Various measures, such as cosine similarity, are used to quantify the similarity between document samples (assignments).Step 9 :Plagiarism Analysis Plagiarism is the act of claiming to be the author of material that someone else actually wrote. This definition relates to text documents, which is also the focus of this project. Clearly, a question of central importance is to what extent such and similar tasks can be automated. This will show the similar lines between two documents among the submitted documents.

OCR Algorithm

STEP 1: The input text image is converted into binary image.

STEP 2: The binary image is complimented so that the letters constitute by binary 1 (one) and background constitute by binary 0 (zero).

STEP 3: All text lines are separated from the binary image. This is done by finding the sum of all values in a row. When the sum is 0, a new line is identified and separation is done. The sum of all rows in between two lines should be zero.

STEP 4: For each line, the characters are to be extracted. This is done by finding the sum of all pixels value in a column. When sum is zero, a new character is identified and separation is done.

STEP 5: Total 12 features value f1 to f12 are extracted for each character.

STEP 6: The features value are matched with the trained features set to recognize the exact character.

Conclusion

People easily plagiarizes images because many plagiarism detectors are based on text plagiarism detection, they simply discard images from the file. There is very less research in this area. In this project, a system will be developed to retrieve images having certain characteristics based on the given input image.

Some research has been performed on flowchart image but we are planning to develop a plagiarism system which will be applicable for all kind of images. For attaining accurate results, features such as color and grey scaling will be used. After removing stop words and stemming the phrases, the algorithm developed to discover similarity among text documents determines content similarity among specified texts. Various measures, such as cosine similarity, are used to quantify the similarity between document samples (assignments).Plagiarism is the act of claiming to be the author of material that someone else actually wrote. This definition relates to text documents, which is also the focus of this paper. Clearly, a question of central importance is to what extent such and similar tasks can be automated. This will show the similar lines between two documents among the submitted documents.

Results

The output of this paper is by implementing OCR technique we draw the text from the image and display the extracted text. By constructing the master document in text format to verify the submitted document using NLP. Implementing the NLP to check the plagiarism percentage with extracted document and show the percentage of plagiarism found in new document comparing with the master document using NLP.

Output

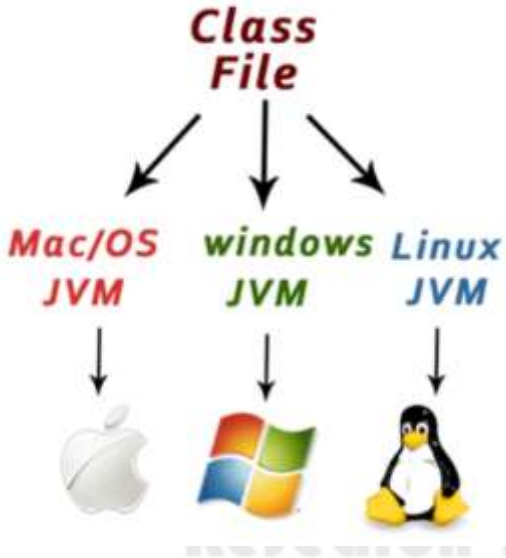

Fig : Original Image

Fig : Extracted text from image

References

1. “Online Assignment Plagiarism Checker Using Machine Learning”,Babitha, Harshitha M,Hindumathi A,Reshma Farhin J,ISSN (O) 2278-1021, ISSN (P) 2319-5940,Issue 4, April 2022.

2. “Extracting text from image document and displaying its related information”, K.N. Natei journal of Engineering Research and Application (ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May2018.

3 .J. Pradeep, E. Srinivasan and S. Himavathi, “Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network”, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

4. “Text Recognition using image processing”, International journal of Advanced Research in Computer Science by Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar (Vol-8,No.5,May-June 2017).

5. Deepa, R., & Lalwani, K. N. (2019), “Image Classification and Text Extraction using Machine Learning”, 2019 3rd International Conference on Electronics,Communication and Aerospace Technology(ICECA).doi:10.1109/iceca.2019.8821936. w

6. K.N. Natei, J. Viradiya, S. Sasikumar. “Extracting Text from Image Document and Displaying Its Related Information”, K.N. Natei Journal of Engineering Research and Application, ISSN : 2248- 9622, Vol. 8, Issue5 (Part -V) May 2018, pp 27-33.

7. “Optical Character Recognition from Text Image” Ranjan Jana,Amrita Roy Chowdhury,Mazharul Islam ISSN: 2319–8656 ,2020.

8. K. Gaurav and Bhatia P. K., “Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition”, 2nd International Conference on Emerging.Trends in Engineering & Management, ICETEM, 2013.

9. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, ‘‘Evaluating the visualization of what a deep neural network has learned,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

10. J. Schneider, A. Bernstein, J. V. Brocke, K. Damevski, and D. C. Shepherd, ‘‘Detecting plagiarism based on the creation process,’’ IEEE Trans. Learn. Technol., vol. no. 3, pp. 348–361, Jul. 2018.

11 . Chandni Kaundilya , Diksha Chawla, Yatin Chopra Automated Text Extraction from Images using OCR System. Studies in Fuzziness and Soft Computing 352, DOI 10.1007/978-3-319-50252-6_2(2017).

12. “Online Assignment Plagiarism Checking integrates Data Mining and NLP” aresh Bokade, Tejas Chede, Dhanashri Kuwar, Prof. Rasika Shintre ISO 9001:2008 Certified Journal(2021).

International Research Journal
IJNRD
Research Through Innovation