Automated Plagiarism Detection in Moodle

Ravi Kishore Kodali, Tanvi Shekhar and Lakshmi Boppana
Department of Electronics and Communication Engineering
National Institute of Technology, Warangal
Warangal, Telangana, India
ravikkodali@gmail.com

***Abstract*—The digital revolution has made access to information very easy. The onset of the COVID-19 pandemic also called for further digitization. Every organization; be it an office, an educational institute or a government entity, was forced to shift to an all virtual mode of operation. This led to the conduct of online examinations with very little time for formulating an anti-cheating examination pattern. Audio and video proctoring tools are considered helpful but are very expensive and do not provide a method to detect plagiarism in the handwritten text. This is a serious problem for academic enterprises and institutes where there is a need for plagiarism detection in the submitted assignments, answer-scripts against the information available on the Internet as well as against other submissions. This paper presents a plagiarism detection system for handwritten text in English. The proposed system uses authentication tools/services, cloud storage, and optical character recognition (OCR) services to automate the process of checking plagiarism between two handwritten documents, as well as plagiarism with respect to all information available online.**

Keywords - Plagiarism Detection, Optical Character Recognition, Artificial Intelligence

I. Introduction

Plagiarism is the act of using the ideas, words, or works of another person without giving them credit or citing the source. This is done with the intention of gaining some sort of benefit or recognition, not necessarily for financial gain. This type of behavior is considered academic misconduct [1].

Textual plagiarism is a threat to the fundamentals of the education system as it challenges the legitimacy and honesty of students' work and might lead to unfair grading. Students may be able to complete courses without achieving the intended results and gain credits based on other work; this in the long run will affect their learning curve and hinder their growth and performance as students and members of the industry in the upcoming future. Using automated plagiarism detection, it is possible to identify the original source.

The onset of the COVID-19 pandemic has intensified the need for a feasible plagiarism detection system, as it becomes difficult to detect plagiarism with limited human resources. Automated plagiarism detection becomes essential, considering various data sources facilitated by various search engines [2].

The post-COVID-19 era brought about an urgent necessity to accelerate and improve existing plagiarism detection methods. But coherent systems that allow for this automation are not in place. There are standalone handwritten text detection, i.e. OCR software, services, and products such as AWS Textract, Tesseract OCR, etc. And many authentication tools have also been developed, but none of them provides the facility to scan handwritten text followed by checking its authenticity. Learning Management Systems (LMSs) like Moodle can use compatible authentication tools/software/service APIs and plugins to help automate the plagiarism detection system.

II. Literature Survey

Plagiarism in handwritten documents, computer codes / programs, etc. is prevalent in the student community. This must be prevented because it affects the learning of the students along with their ability to assess themselves. Fortunately, plagiarism detection is a well-researched field. Various techniques and methods have been developed to detect plagiarism in digital text.

Five different software tools used for textual plagiarism were compared based on their features and performance, and it was concluded that no software tool was 100% accurate [3].

A system called CHECK was used to identify similarity in domain-specific documents [4]. The approach based on natural language processing (NLP) commonly used among free online tools was compared with the proposed method [5].

A MatchDetectReveal system was proposed that uses an exact string matching algorithm and was claimed to be more accurate [6]. The use of few online textual matching tools, such as Duplicheck, PlagScan, and Plagiarisma, using their respective URLs was considered [7].

The most common technique used for plagiarism detection is known as similarity technique. But this had been replaced by more refined modern techniques using content comparison. String tiling techniques [8] and tree parsing techniques [9] have been used extensively. Another technique, the Fast Plagiarism Detection System (FPDS), is also being used on the basis of its performance [10]. This method incorporated a special indexed data structure while storing file collections. Variable renaming is removed in the tokenization technique [11].

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

III. Proposed System

In this work, Moodle LMS has been used as the starting point of the automation process. Answer scripts, assignments, and other submissions from students are stored here. From here, the data will be moved to AWS S3 to help with the automation process. Now, AWS Textract has been used to detect handwritten text and convert it to digital text. The digital text is then moved to Moodle LMS after which the plagiarism is detected using the Turnitin Plugin. Digital text is also passed through the custom Python code to verify similarity between text files (.txt)

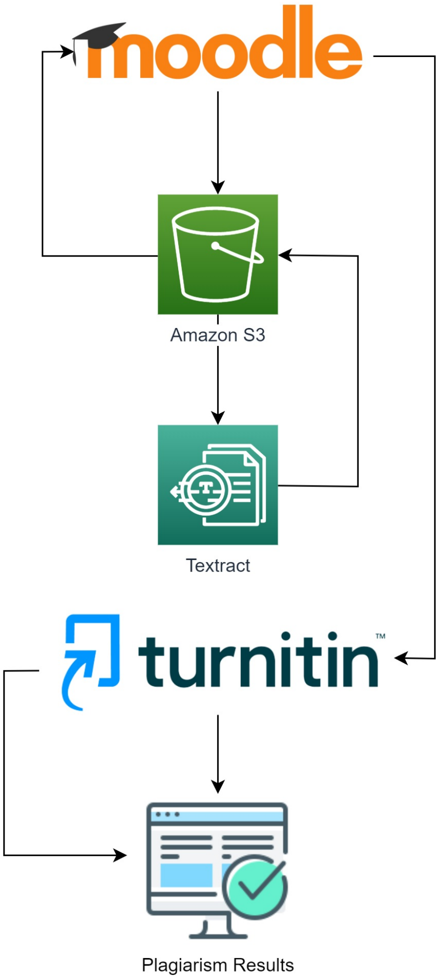


Fig. 1. Proposed System Flow Diagram

IV. Implementation

As in the flow diagram shown in Fig.1, the following services, software, and products have been used in the implementation of the proposed system.

A. Moodle LMS

Moodle LMS is an open source LMS that can be customized for any course and customized according to the requirements of individuals, organizations, or educational companies and institutions. It greatly helps educators and others conceptualize the various courses, course structures, and curriculum. This helps facilitate and organize online interaction between students and educators. Moodle LMS also has various integration options, as shown in Fig. 2, it can be well integrated with AWS to enable the various services provided like S3 for storage, AWS Textract for handwritten text detection, etc. Many tools, services, and products like Turnitin also provide compatible plugin options with Moodle LMS.

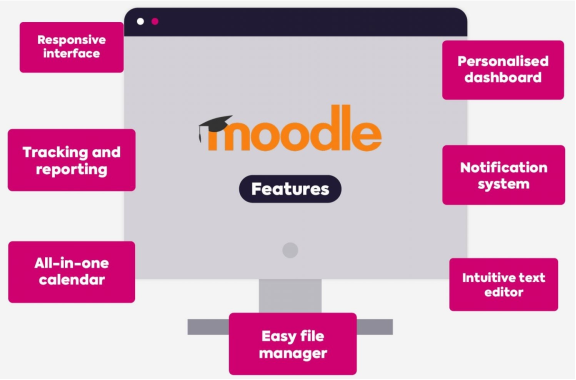


Fig. 2. Moodle LMS

B. Amazon Web Services

1) AWS simple storage service (S3): Amazon S3 uses object storage service. Among the storage types supported by Amazon S3 is quite inexpensive and may be accessed from anywhere and highly scalable, available, secure and reliable. This is the most suitable storage type for storing large files such as educational resources and scanned answer scripts. A cloud architecture involving S3 is shown in Fig. 3.

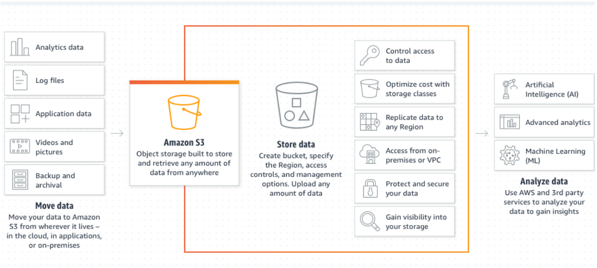

Analytics data
Log files
Application data
Videos and pictures
Backup and archival
Move data
Move your data to Amazon S3 from wherever it lives – in the cloud, in applications, or on-premises
Amazon S3
Object storage built to store and retrieve any amount of data from anywhere
Store data
Create bucket, specify the Region, access controls, and management options. Upload any amount of data
Control access to data
Optimize cost with storage classes
Replicate data to any Region
Access from on-premises or VPC
Protect and secure your data
Gain visibility into your storage
Artificial Intelligence (AI)
Advanced analytics
Machine Learning (ML)
Analyze data
Use AWS and 3rd party services to analyze your data to gain insights

Fig. 3. A Cloud Architecture using S3

2) AWS Textract: Amazon Textract is an AI service that allows one to automatically extract text, handwriting, and data from scanned documents.

Being a commercial product, no specifics on the models used in Textract can be obtained. Textract automatically identifies the layout of a document. The location and geometry (bounding box) for lines and words in a document are returned as a block object when the Textract operation is invoked. A list of block objects is generated along with the relationships and context among the detected text.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

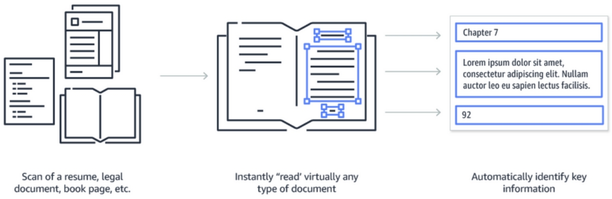

Chapter 7
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam auctor leo eu sapien lectus facilisis.
92
Scan of a resume, legal document, book page, etc.
Instantly "read" virtually any type of document
Automatically identify key information

Fig. 4. Flow Diagram of AWS Textract OCR

C. iThenticate Turnitin

ithenticate Turnitin, a web-based plagiarism tool widely used among the academic community, uses its database collection comprising billions of web pages which include millions of published scholarly articles facilitated by Crossref Similarity Check. It obtains a similarity score of the submission and shows the original submission highlighting the set of matches along with the matched sources.

Turnitin can be integrated with other products such as Moodle LMS. This enables the work to be sent through Turnitin's repository automatically as and when a submission is made, without the need to log in separately and upload the submission. The Similarity Report shows the exact portions of the copied content and the same may be used during the assessment. The information flow between the Turnitin database hosted Moodle instance on AWS is shown in Fig. 5.

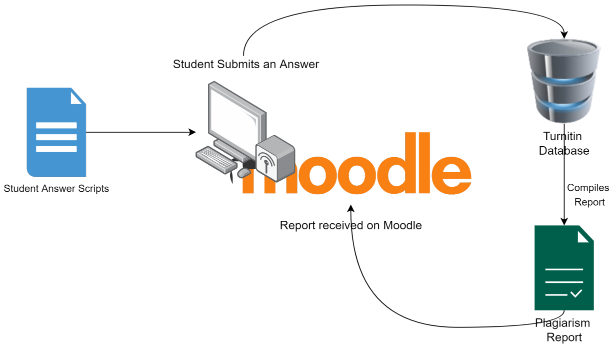


Fig. 5. Turnitin integration with Moodle LMS

V. Other OCR Tools - Tesseract OCR

Optical Character Recognition (OCR) software tools use AI and ML to detect and recognize text characters and extract text from non-text documents such as images and pdf files. The extracted text may be edited later. Tesseract OCR is an open source HP OCR engine.

Tesseract OCR uses a two-step recognition process after converting the original image to its binary version and detection phases as shown in Fig. 6.

VI. Comparison of Textract and Tesseract OCR

Textract can be compared to Tesseract OCR to better analyze and understand it. The long-short-term memory (LSTM) network is a recurrent neural network (RNN) and is widely used in natural language processing (NLP).

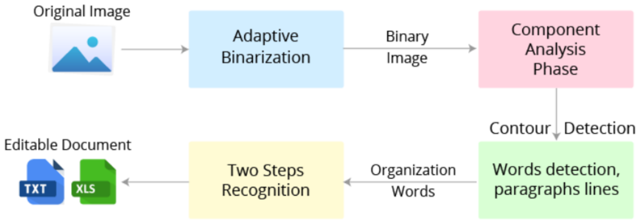
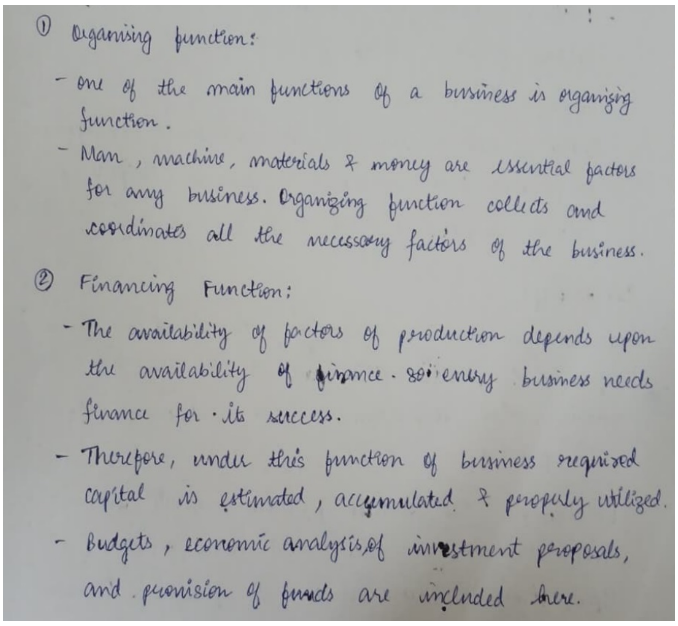


Fig. 6. Flow Diagram of Tesseract OCR

Tesseract OCR extensively uses the long-short-term memory (LSTM) network, a recurrent neural network (RNN), and is widely used in Natural Language Processing (NLP). Tesseract OCR model can also be trained to recognize other languages, as it is completely open source and Tesseract fails in the case $< key - value >$ and table extraction is involved.

Textract OCR also uses deep learning, and it is a pre-trained AI service, which parses and extracts the data within a document. When in doubt, the human-in-the-loop option may be chosen. Textract performs fairly well in key-value pair analysis and in achieving accuracy, but it is limited to only a few languages and document formats as of now. The drawback of Tesseract OCR with respect to recognizing handwritten text is that, for any handwritten text to be detected, the Tesseract engine first has to be trained to recognize that particular handwriting. Therefore, it cannot be used for the system proposed in this work.



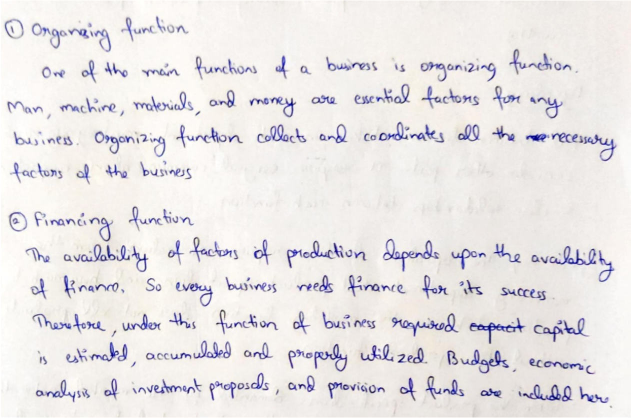
① Organising function:

- one of the main functions of a business is organising function.
- Man, machine, materials & money are essential factors for any business. Organizing function collects and coordinates all the necessary factors of the business.

② Financing Function:

- The availability of factors of production depends upon the availability of finance. So every business needs finance for its success.
- Therefore, under this function of business required capital is estimated, accumulated & properly utilized.
- Budgets, economic analysis of investment proposals, and provision of funds are included here.

Fig. 7. Handwritten Text Sample A



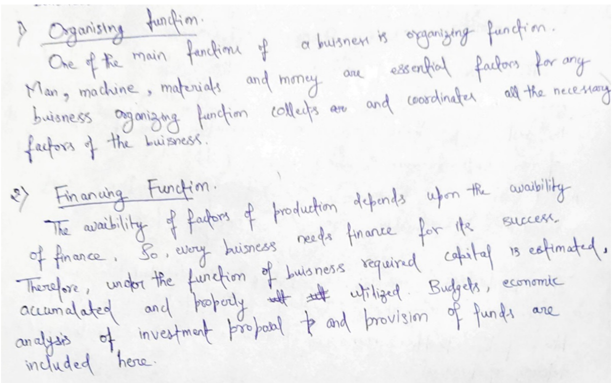
① Organzing function
One of tho main functions of a business is organizing function.
Man, machine, materials, and money are essential factors for any
business. Organizing function collects and coordinates all the ~~ne~~ necessary
factors of the business

② Financing function
The availability of factors of production depends upon the availability
of finance. So every business needs finance for its success.
Therefore, under this function of business required ~~capacit~~ capital
is estimated, accumulated and properly utilized. Budgets, economic
analysis of investment proposals, and provision of funds are included here.

Fig. 8. Handwritten Text Sample B

VII. Dataset

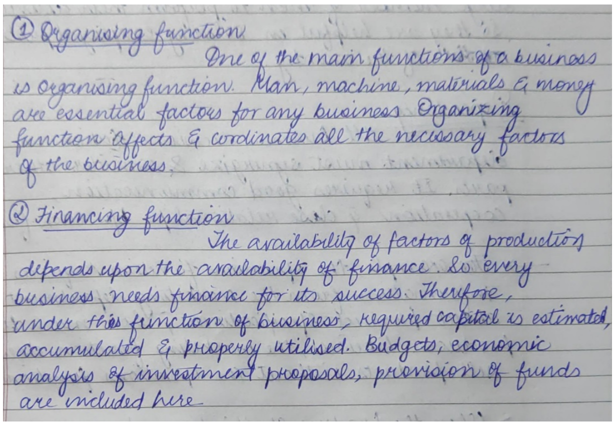
Three samples of the similar text are taken in three different hand-writings. These particular samples have been taken to help accommodate various handwriting styles, letter size, neatness, legibility etc. while testing.



1) Organising function.
One of the main functions of a buisness is organizing function.
Man, machine, materials and money are essential factors for any
buisness organizing function collects ~~are~~ and coordinates all the necessary
factors of the buisness.

2) Financing Function.
The availibility of factors of production depends upon the availibility
of finance. So, every buisness needs finance for its success.
Therefore, under the function of buisness required capital is estimated,
accumalated and poperly ~~ut~~ ~~ut~~ utilized. Budgets, economic
analysis of investment proposal ~~p~~ and provision of funds are
included here.

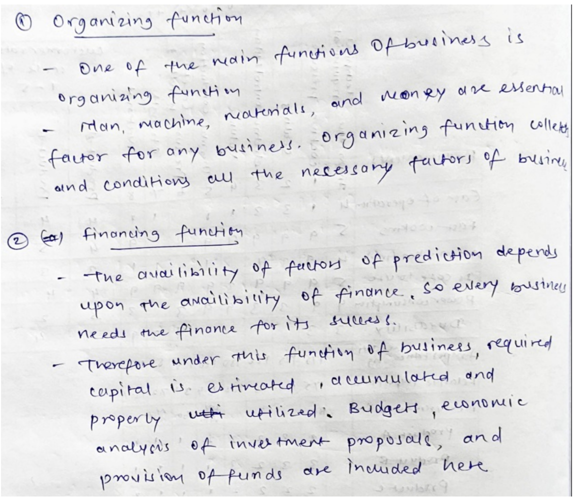
Fig. 9. Handwritten Text Sample C



① Organising function
One of the main functions of a business
is organising function. Man, machine, materials & money
are essential factors for any business Organizing
function affects & cordinates all the necessary factors
of the business.

② Financing function
The availability of factors of production
depends upon the availability of finance. So every
business needs finance for its success. Therfore,
under this function of business, required capital is estimated,
accumulated & properly utilised. Budgets, economic
analysis of investment proposals, provision of funds
are included here

Fig. 10. Handwritten Text Sample D



① Organizing function
- One of the main functions Of business is
organizing function
- Man, machine, materials, and money are essential
factor for any business. organizing function collects
and conditions all the necessary factors of busines

② ~~(a)~~ financing function
- The availibility of factors of prediction depends
upon the availibility of finance. So every business
needs the finance for its success.
- Therefore under this function of business, required
capital is estimated, accumulated and
properly ~~uti~~ utilized. Budgets, economic
analysis of investment proposals, and
provision of funds are included here

Fig. 11. Handwritten Text Sample E

VIII. AWS Textract Results

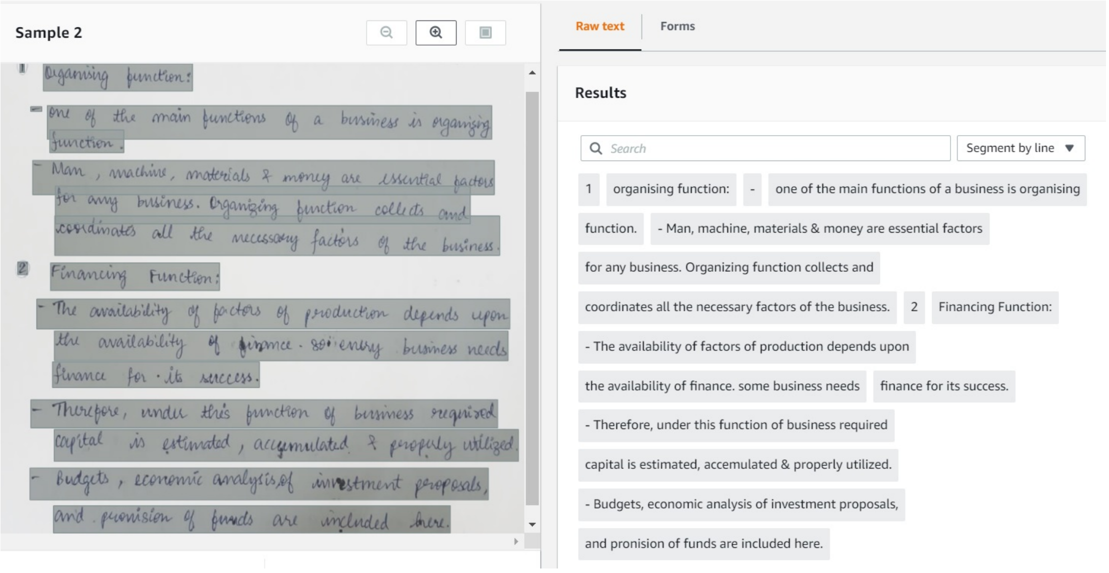
The results from the AWS Textract analysis of samples A, B, C, D, and E are as shown in Figs. 11, 12, 13, 14 and 15. The handwritten text detected was segmented line by line, that is, the text was detected line by line based on how it was written in the uploaded sample.

IX. Analysis of the Results

The samples A, B and D in Figs. 11, 12 and 14 have been detected accurately, as the bounding boxes were able to precisely enclose the handwritten text line by line. However, the bounding boxes in samples C and E as shown in Figs. 13 and 15 are not precise, as the handwritten text is not properly aligned. This leads to incorrect detection of the text, resulting in jumbled words and sentences. This demonstrates that illegible handwriting can be a limitation for the AWS Textract OCR, as misalignment of the text can cause confusion in the overlapping of the boundaries of the bounding boxes, leading to incorrect output.

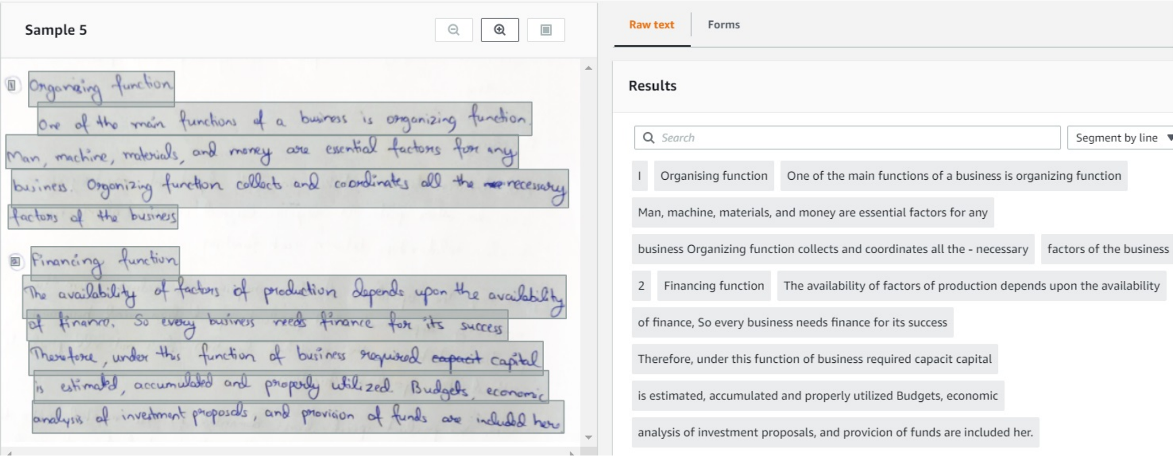
X. Conclusions

This paper proposes an automated system for detecting plagiarism in handwritten documents, such as answer-scripts. It utilizes Moodle LMS to upload the documents, and then uses AWS Textract, a commercially available OCR tool, to convert the handwritten text into digital ASCII characters. The digital text is then sent back to Moodle LMS via AWS S3, where the Turnitin plugin checks for plagiarism on the Internet and its database. Although the system is automated, there are certain limitations in terms of the accuracy of the OCR tool and the efficiency of the plagiarism checker. As seen in Figs. 15 and 17, AWS Textract fails to detect handwritten text with high precision. The detected words are correct, but the sentences are jumbled due to misalignment in the handwritten text, which causes the bounding boxes to overlap. To improve the accuracy of the proposed model, advancement of OCR



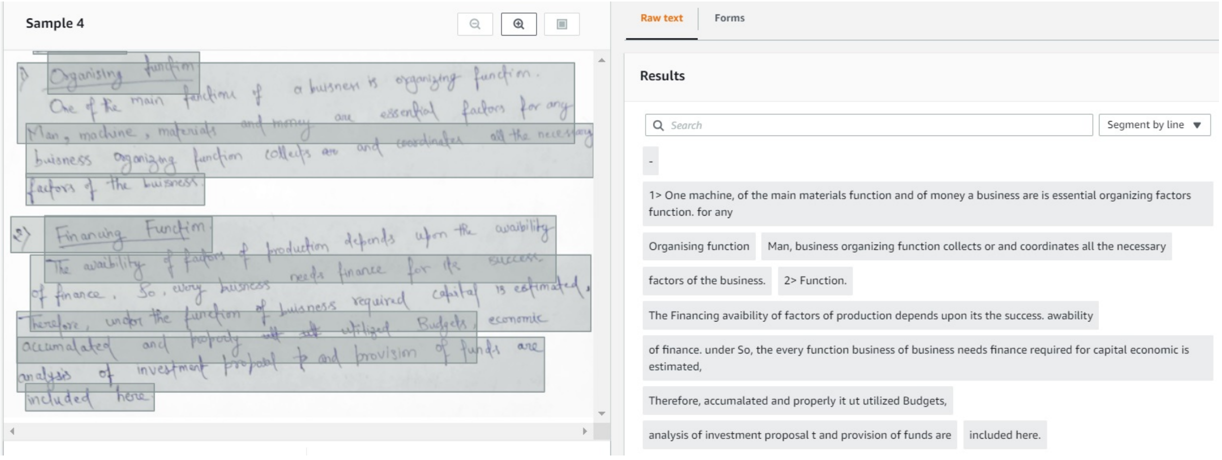
1
organising function:
-
one of the main functions of a business is organising
function.
- Man, machine, materials & money are essential factors
for any business. Organizing function collects and
coordinates all the necessary factors of the business.
2
Financing Function:
- The availability of factors of production depends upon
the availability of finance. some business needs
finance for its success.
- Therefore, under this function of business required
capital is estimated, accemulated & properly utilized.
- Budgets, economic analysis of investment proposals,
and pronision of funds are included here.

Fig. 12. Handwritten Text as detected from Sample A



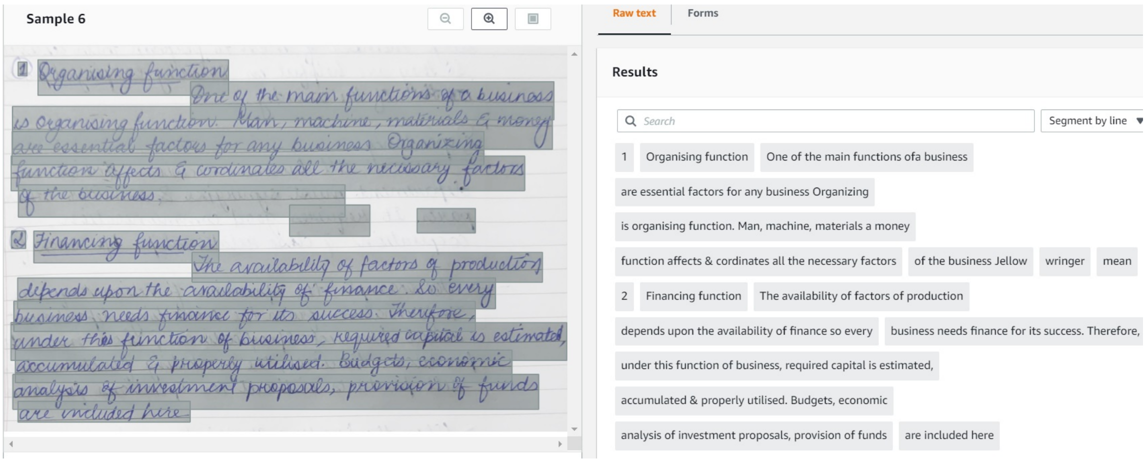
I
Organising function
One of the main functions of a business is organizing function
Man, machine, materials, and money are essential factors for any
business Organizing function collects and coordinates all the - necessary
factors of the business
2
Financing function
The availability of factors of production depends upon the availability
of finance, So every business needs finance for its success
Therefore, under this function of business required capacit capital
is estimated, accumulated and properly utilized Budgets, economic
analysis of investment proposals, and provicion of funds are included her.

Fig. 13. Handwritten Text as detected from Sample B



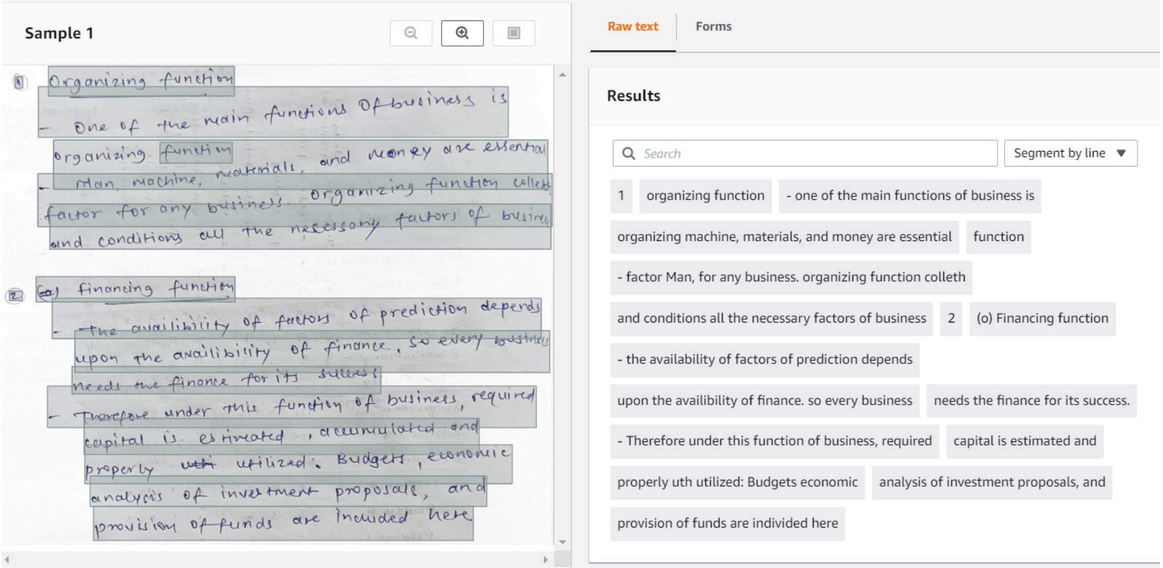
-
1> One machine, of the main materials function and of money a business are is essential organizing factors function. for any
Organising function
Man, business organizing function collects or and coordinates all the necessary
factors of the business.
2> Function.
The Financing avaibility of factors of production depends upon its the success. awability
of finance. under So, the every function business of business needs finance required for capital economic is estimated,
Therefore, accumalated and properly it ut utilized Budgets,
analysis of investment proposal t and provision of funds are
included here.

Fig. 14. Handwritten Text as detected from Sample C



1 | Organising function | One of the main functions ofa business
are essential factors for any business Organizing
is organising function. Man, machine, materials a money
function affects & cordinates all the necessary factors | of the business Jellow | wringer | mean
2 | Financing function | The availability of factors of production
depends upon the availability of finance so every | business needs finance for its success. Therefore,
under this function of business, required capital is estimated,
accumulated & properly utilised. Budgets, economic
analysis of investment proposals, provision of funds | are included here

Fig. 15. Handwritten Text as detected from Sample D



1 | organizing function | - one of the main functions of business is
organizing machine, materials, and money are essential | function
- factor Man, for any business. organizing function colleth
and conditions all the necessary factors of business | 2 | (o) Financing function
- the availability of factors of prediction depends
upon the availibility of finance. so every business | needs the finance for its success.
- Therefore under this function of business, required | capital is estimated and
properly uth utilized: Budgets economic | analysis of investment proposals, and
provision of funds are individed here

Fig. 16. Handwritten Text as detected from Sample E

tools and plagiarism checkers is necessary, along with their integration with other tools, software and services.

REFERENCES

[1] Foltýnek, Dita Dlabolová, Alla Anohina-Naumeca, Salim Raz, Jlius Kravjar, Laima Kamzola, Jean Guerrero-Dib, zgür elik7 and Debora Weber-Wulff, ”Testing of support tools for plagiarism detection”, International Journal of Educational Technology in Higher Education (2020)

[2] Prasanth.S, Rajshree.R, Saravana Balaji.B, A Survey on Plagiarism Detection, International Journal of Computer Applications (0975 – 8887), Volume 86 – No 19, January 2014

[3] A. M. E. T. Ali, H. M. D. Abdulla, V. Snasel, ”Overview and Comparison of Plagiarism Detection Tools,” In DATESO (pp. 161-172).

[4] A. Si, H. V. Leong, and R. W. Lau, ”Check: a document plagiarism detection system,” ACM symposium on Applied computing 1997 (pp. 70-77).

[5] R. Tripathi, P. Tiwari, and K. Nithyanandam, ”Avoiding plagiarism in research through free online plagiarism tools”, in Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on (pp. 275-280), IEEE.

[6] K. Monostori, A. B. Zaslavsky, and H. W. Schmidt, ”MatchDetectReveal: finding overlapping and similar digital documents,” InIRMA Conference May 2020 (pp. 955-957)

[7] Shivani, Vishal Goyal, ”A Novel Approach for Plagiarism Detection in English Text,” International Journal of Computer Applications (0975 – 8887) Volume 154 – No.2, November 2016

[8] M. J. Wise, ’YAP3: improved detection of similarities in computer program and other texts,’ Proc. of SIGCSE’96 Technical Symposium, 2006

[9] L. Prechelt, G. Malpohl, and M. Philippsen, “Finding plagiarisms among a set of programs with JPlag”, Journal of Universal Computer Science, 2008.

[10] D. Gitchell and N. Tran, “Sim: a utility for detecting similarity in computer programs”, the 30th SIGCSE Technical Symposium on Computer Science Education, 2006

[11] M. Mozgovoy, K. Fredriksson, and D. White, ’Fast plagiarism detection system’, Lecture Notes in Computer Science, 2005.