Academic Journal Plagiarism Checker Using Tri-Gram Algorithm and KNN Algorithm

Iyanuoluwa Modupeore FATOKI
LCU/PG/002131

Being a MSc Thesis Submitted to the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria

In Partial Fulfilment of the Requirements for the Award of Master Degree (MSc) in Computer and Information Science

2022

Certification

This is to certify that this thesis report was written by Iyanuoluwa Modupeore FATOKI with matriculation number LCU/PG/002131 of the Department of Computer and Information Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria, for the award of Master Degree (MSc) in Academic Journal Plagiarism checker and that this has not been previously submitted.

_____ ----------

Mrs V. B. Oyekunle Date

Supervisor

_____ ----------

Dr. W. Sakpere Date

Head of Department

Dedication

This thesis work is dedicated to Almighty God.

Acknowledgement

I want to thank the prestigious institution Lead City University for the opportunity to learn and complete my Master's Program and to the University Library for provision of adequate materials needed.

I am grateful to the Department of Computer Science for giving me the privileged and opportunity to study and to learn, my appreciation goes to my supervisor Mrs. V.B. Oyekunle, I also like to acknowledge my lecturers, Prof. S.O. Akinola, Dr. W. Sakpere, Dr. W.S. Ajayi, Dr. A.A. Waheed, Dr. R.A. Badru for their supports and always ensuring that things are done rightly.

My appreciation goes to my family who has always supported and sponsored me through this program.

"Even though the above-mentioned institutions and persons have assisted in the process of this research work, I alone stand responsible for the errors, if any, found in the work."

Abstract

Academic journal plagiarism checker using data mining is aimed at checking for plagiarism in journals submitted by students, in which the research title has a unique name since two journals in the database are not allowed to have similar titles. The main goal of this thesis is to develop a journal plagiarism checker and provide an interface that will be used for journal submission and check for plagiarism. The objectives in achieving the set goal are: Designing of the document repository, design of the design of the plagiarism checker algorithm and the development of the whole system with HTML, CSS, PHP, JavaScript, Xamp Server. The system is targeted to helping students take their research seriously and also to reduce the rate at which students plagiarize in their research work and to also help students with the selection of research topic that has not been done by someone else before so as to prevent plagiarism. The journal plagiarism checker makes use of clustering techniques in data mining and KNN algorithm and Trigram algorithm is deployed for the implementation of the system. The system will give accurate plagiarism report and it will take less time to complete its execution. The cost of implementing the system is minimal compared to the cost of other plagiarism tools available online. The system is deployed into the Department of Computer Science in Lead City University.

Keywords: Plagiarism, KNN Algorithm, Trigram Algorithm, Journal.

Word Count: 232

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

1.1 Background to the Study

Plagiarism can be described as the act of using someone's work without permission. Plagiarism is the deliberate acquisition of someone else's work for their own benefit without the approval of the original owner[1]. In academia, if a researcher adopts someone else's work without proper reference, it may be considered a scholar called plagiarism. Researchers face many difficulties because they have to read all the journals written by each student and do not know where the plagiarism occurred in the research papers submitted by the students. The journal allows researchers to reflect, exchange ideas, and discuss issues based on specific topics. The Journal deals with specific areas where researchers are interested in gaining insights and gaining more knowledge. In computer science, journal is a general term for academic journals that focus on computers and their terms. The academic journal plagiarism checker is a system that allows researchers to submit journals online, ensuring that the submitted journals have not been plagiarized. Writing a journal involves publishing articles that have a significant impact on all areas of computer science. Journals help provide materials for training and research in relevant research areas[2].

The system incorporates a plagiarism tracker that helps prevent plagiarism of other people's articles and is implemented throughout the body of the journal. The main goal of research activities is the dissemination of research knowledge. The system provides researchers with the opportunity to retrieve information from the journal and present articles in a way that can prevent plagiarism between researchers and students. It also provides students with the

opportunity to share their views on a particular topic. Journals are important because they contribute significantly to the verification of the knowledge gained. Plagiarism can be divided into two major categories based on detection: extrinsic and intrinsic plagiarism. Extrinsic plagiarism detection refers to detecting plagiarism in a document compared to a specific reference data set, and intrinsic plagiarism detection refers to plagiarism using a variation of writing without using a reference corpus. Refers to detecting[3].

The plagiarism checker will enable uploading of journals that are not plagiarized, which will be used in the department of Computer Science Lead City University. The system will use clustering techniques from one of the techniques of Data mining.

An article on the plagiarism detection process using data mining techniques suggested a methodology that could improve the efficiency of the process. This publication uses preprocessing and clustering techniques to reduce process overhead. In addition, clusters of stolen data can be used to calculate similarity values, which can improve efficiency. There are eight processes that are loopholes, like a replica of Turnitin. Clustering techniques helps to achieve efficient results[6]. For plagiarism from paper theft detection techniques to data mining, some preliminary results suggest that using a mixed machine learning approach is generally more efficient. Hybrid systems, on the other hand, do not always give good results. As a result, we have developed a process that leverages the kNN approach to machine learning to increase efficiency. The nearest neighbor method is very useful for pattern recognition and plagiarism detection by finding the copied dataset. KNN (Knearest Neighbors Algorithm) is the method used in the work[7]. A whole new web based plagiarism checker will be created which will help prevent plagiarism.

1.2 Statement of the Problem

Plagiarism checker tools are expensive for students to afford and the checker is needed for the research work done by students so as to verify if the research work done is free from plagiarism, most of the checkers take time to complete its execution. Also not all students have access to plagiarism checker without having to go through the institution.

Academic staffs have a lots of students they supervise, and they cannot manually detect all the research work each student under their care has plagiarized, which make it easier for more students to plagiarize their work as well.

Also, there are times where the research work the researcher intends to carry out has been done before but the researcher is not having the mindset of adding something new to the research, they decide to plagiarize existing works.

An academic journal plagiarism checker would be a measure that can be put in place to provide a system where plagiarism of journals are checked and also to ensure that the journals are free from plagiarism, as well as easy accessibility to already researched works, journals and standard materials.

The plagiarism checker will allow researchers make more findings in their expected area of interest in order to gain more knowledge.

1.3 Aim and Objectives

The aim of this thesis is to develop an Academic Journal Plagiarism Checker using clustering technique which is a technique from Data Mining.

The specific objectives of the study are to:

i design the document repository which is known as the database

ii. develop the plagiarism checker system that gives precise result in less execution time

iii. evaluate and analyze the plagiarism checker system

1.4 Justification of the Study

i. Most academic staffs and students do not have access to the plagiarism checker because of its price, so this plagiarism checker can be used by students and academic staffs.

ii. Since there is a very high rate at which students and researchers plagiarize, the system will be embedded with a plagiarism tracker which will help to reduce the rate of plagiarism when researchers are working on their research work.

iii. To avoid researchers repeating the problem that has already been solved, the system will make all the documents in the database have unique titles.

1.5 Significance of the Study

This system is developed in order to give students or researchers opportunity to submit their research work and will help to reduce rate at which students can copy materials online with the aid of the plagiarism detector that will be implemented in the system.

In addition, students who are registered on the system will be able to have access to all the documents in the system and they will be able to download it for their use. The system will do the following:

i. Provide a system that is easy, secure and fast for the management of journals.

ii. Provide a system that can be used for both academic staffs and research students.

iii. Provide a system that will help to prevent plagiarism in the academic research work.

1.6 Scope of the Study

The system is going to be a web-based system and plagiarism detection function will be based on the title as well as the content of the journal submitted. The plagiarism system will run the check with the files in the database. As per IEEE standards no two papers can have the same title unless they have the same author[4].

1.7 Limitation of the Study

The thesis has limitations, but it doesn't have scientific effect on the result of the work done. The limitation of the thesis is as follows:

i. The system will check plagiarism for journals not more than 5-10 pages.

ii. The system would not accept document with the same journal title.

1.8 Operational Definition of Terms

Plagiarism: Plagiarism is the act of using or copying someone else's idea or work and pretending that you thought of it or created or taking credit for someone else's work.

Journals: Academic journals are periodicals in which academic papers related to a particular discipline are published. Scientific journals serve as a permanent and transparent forum for publication, review, and discussion of research results.

Data Mining: Data mining is the process of sorting large datasets and identifying patterns and relationships that help solve business problems through data analysis. Data mining techniques and tools enable companies to anticipate future trends and make more informed business decisions. Data mining is an important part of the overall data analysis and is one of

the core areas of data science that uses advanced analytical techniques to find useful information in datasets.

Clustering: Clustering is the process of making a group of abstract objects into classes of similar objects.

KNN Algorithm: A k-nearest-neighbor algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

Trigram Algorithm: Trigrams are a special case of the *n*-gram, where *n* is 3. They are often used in natural language processing for performing statistical analysis of texts and in cryptography for control and use of ciphers and codes.

Endnotes

[1] Bhardwaj Akanksha1, Arya Anukruti1, Vyas Tarjni1, Shivani Desai1 & Anuja Nair1 *A Survey on Plagiarism Detection.* Institute of Technology, Nirma University, 382481-Ahmedabad, India. **Advances in Computational Sciences and Technology Volume 10, Number 8** 2017. ISSN 0973-6107

[2] M. Daniels, Å. Cajander, Pears, A. & T. Clear. Engineering education research in practice: *Evolving Use of Open Ended Group Projects as a Pedagogical Strategy for Developing Skills in Global Collaboration.* **International journal of engineering education**, 2018 pp.795-806.

[3] Muna AlSallal, Rahat Iqbal, Vasile Palade, Saad Amin & Victor Chang, "*An Integrated Approach for Intrinsic Plagiarism Detection*" in , Suzhou, China: **School of Computing, Electronics and Maths, Coventry University, United Kingdom,** International Business School Suzhou, Xi'an Jiaotong Liverpool University, 2019.

[5] Abhay Nitin Pai, Chinmay & Neelmadhav Bhusari *Plagiarism Detection System,* **Department of Computer Science and Engineering Datta Meghe Institute of Engineering Technology and Research,** Wardha, Maharashtra, India February 2013, 278

[6] Mahwish Abid & Muhammad Usman, *Plagiarism Detection Process using Data Mining Techniques* Muhammad Waleed Ashraf Riphah International University Faisalabad, Pakistan. **International Journal of Engineering Studies** – Vol. 5, No. 4, 2017

[7] R. Nennuri, M. Geetha, M. Samhitha, S. Kumar & G. Roshini *Plagiarism Detection through Data Mining Techniques* Dept. of CSE, Institute of Aeronautical Engineering12345, Dundigal – **Hyderabad International Conference on Recent Trends in Computing ICRTCE**, 2021

Chapter Two

Literature Review

2.1 Introduction

Plagiarism is one of the most common problems in academia today. Plagiarism is the submission of a document that belongs to someone else, in whole or in part, without proper notice, and therefore misrepresents the author's efforts to submit. It is important to recognize that plagiarism can also occur unintentionally[1]. Scientific plagiarism is a scientific fraud that adversely affects science. Academic plagiarism is defined as the use of ideas, content, or structures without proper attribution to benefit in an environment where originality is expected. The entire plagiarism classification is broadly divided into intentional plagiarism and unintentional plagiarism, and all other types fall into these two categories. Intentionally, when the author is deliberately suffering.

Plagiarism is considered to be an academic and harmful practices. As a result, scholars are striving to understand such academic ability. One effort is to provide a clear definition of plagiaries so that confusion does not appear during police stolen. There are several sources of sources that help you understand the definition of plagiarism. For example, according to the online etymology dictionary, the word plagiarize was a word derived from the 1950s. This means THEFT[2]. Plagiar is usually defined as a discrete crime and could not give a specific credit source. But it actually asks a much more basic question for writers: "Where is my voice in this project?" As seen in this light, the strategy to help you avoid stealing it is strategic to help win as a writer[3]. As soon as your leadership question is raised according to

your relationship, "Where is my voice?" They are in the middle of the use of sources in an effective and justified way. The turder is another work, word or ideas without mission. The word "plagiarism" comes from the Latin word of "kidnapping" and is considered to be a violation of integrity in the academic world [4]. Many thinks like a stolen to copy another work that meets the original idea of others. However, how to "copy" and "Borrow" can impersonate the severity of crimes. Very new and unique ideas and products derived from existing source sources In other words, the stolen is a fraud. It contains both to steal other work and lying. However, outside of the risk of penalties, there are urgent moral and intelligent reasons to avoid plagiarism. When writing an academic audience, the author closes to a continuous conversation about the subject. When the author conflicts, this means that the author contributed to the conversation and he presented him as someone else. In addition, the action is stolen other words and ideas and clarifying her voice removing. It may be difficult to keep yourself itself to create your own contribution in the context of the class taught by experts. However, all tasks are invitations to add new things to the course triggered conversation. If you are not the owner of words or ideas, you cannot make the original contribution[5]. There are many forms of plagiarism, the three main forms of using the source language without citation, using information from the source without attribution, and paraphrasing the source in a format that is too close to the original information. It is classified into categories. This obligation applies regardless of whether the sources are public or private, and whether they are printed or available on the Internet [6].

Often**,** when students plagiarize, they do not knowingly steal someone else's work. Instead, they don't fully understand what plagiarism is, so they accidentally plagiarize it. That is the meaning of "ruthlessness". If you don't appreciate someone else's idea because you didn't

know what to do, or because you didn't know how to do it, you're faced with the same consequences as if you deliberately stole someone else's work. Therefore, it is your responsibility to understand when and how to acknowledge the contributions of others [7]. Plagiarism can be described as the following [8]:

- Give someone else's work as your own
- Copy someone else's words or ideas without approval
- Quotation marks are not enclosed in quotation marks
- Misrepresentation of citation source
- Change the word, but copy the source sentence pattern without credit
- Copy enough words and ideas from one source to make up most of the work, whether or not you grant credits (see the section on fair use rules).

However, most cases of plagiarism can be avoided by quoting the source. It is usually sufficient to prevent the plague by making sure that certain material is borrowed and providing the viewer with the information needed to find the source [9]. To understanding plagiarism, it helps to understand the process of sharing and creating ideas at the university. All knowledge is based on prior knowledge. When we read, study, experiment, and gather perspectives, we use the ideas of others. Building on their ideas and experiences, we create our own. When ideas are put on paper, the instructors wants the authors to distinguish between the buildings block ideas borrowed from other people and the author`s newly reasoned perspectives or conclusions. Providing appropriate citations will also help readers who are interested on the research find additional related material to read, in this way, the readers will be able to build on the work the author has done to find sources.

The figure below shows the types of plagiarism.

The types of plagiarism are: Direct plagiarism. If the author copies the content and uses it as their own. In other words, if the text is rearranged or rearranged, it means the same thing. Plagiarism when inadequate approved content is not quoted correctly. Mosaic plagiarism occurs when the author ignores or ignores that their work is intended to be plagiarized due to lack of knowledge or ignorance. Patchwork plagiarism. When the author copies a part of the original work and creates it by himself. Plagiarism of ideas, when the author steals the ideas of others without attribution[10].

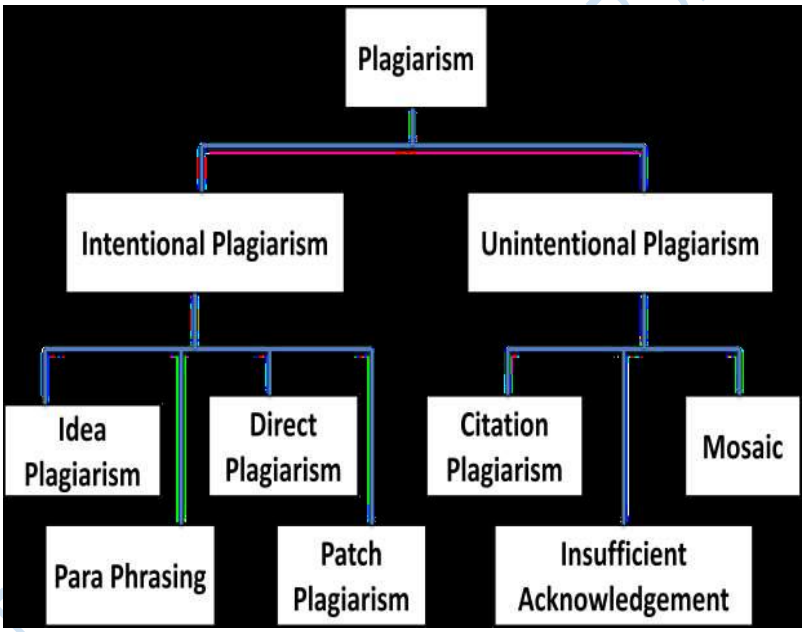


Figure 2.1: Classification of Plagiarism[10]

2.1.1 Plagiarism Types[11]

- **Word-for-Word Copy**: Also known as verbatim, it can be defined as a verbatim copy of a large passage without checking the source.

- **Mosaic Plagiarism:** Copy a portion of the entire text word by word, rather than the entire text, without accepting the borrowed idea.
- **Mock-Paraphrase:** An inline, verbatim copy of text and citations, but displayed as a paraphrase.
- **Inappropriate Paraphrase:** The language of the new text has been paraphrased, but it is still very close and imitating the original text.
- **Illegal Paraphrase:** Material paraphrased from non-attributive text.
- **Illegal Reference:** The author provides incorrect information about the reference source.
- **Intentional Plagiarism:** The author is aware of the fraud
- **Cryptomensia Plagiarism:** unconscious plagiarism
- **Other Plagiarism:** Material copied by another student with the knowledge of another student.
- **Theft:** An assignment that was unknowingly copied from another student's assignment.
- **Ghostwriter:** A task created by a third party and presented as your work.
- **Recycle:** Submitted the same assignment multiple times in different courses.
- **Internet Plagiarism:** Unlike offline sources, the author may be anonymous, so a copy from an internet source.

2.1.2 The Problem of Plagiarism

Some types of plagiarism occur when a student submits a report written by a friend. Even papers that have been revised and submitted more than once will become clearly familiar to

teachers. Students can also find essays (for free or for a fee) on many topics on the Internet. All they have to do is download their name and enter it above e.g. shayo cozy, matric No: 200039. The third form of plagiarism occurs when a student submits homework written for another class. This is easy to recognize because the topic is usually less relevant to the topic that the instructor really wants. The causes of these plagiarism cases seem to be:

1) Lack of understanding of what plagiarism means

2) Lack of prior knowledge of the process of writing an essay or research paper

3) Simple from the Internet Cut and paste

4) Poor reading and writing ability and

5) Pressure to make "perfect" paper.

Even more troublesome than plagiarism is evidence that the majority of students of this generation have not seen anything wrong with it. In a New York Times survey of plagiarism epidemics, more than 50% of respondents who said they had plagiarized did not consider it a scam[12]**.** Many of them don't have the concept of intellectual property, as young people today have almost unlimited access to music, entertainment and information at all times. Their attitude seems to be that if something is available, there is nothing wrong with using it. The search for academic literature was guided by the keywords plagiarism, students, high school, high school, research and research activities, the Internet, online, fraud, and digital. Using ERIC and Wilson Select Plus, keyword plagiarism and high school students were first used

to create a general article on student plagiarism. No mention is made of any particular use of the Internet for plagiarism. These articles do not address the causes of plagiarism and do not provide strategies to prevent plagiarism[13].

Paraphrasing is to take in the thoughts of others and put those thoughts into your own words. Paraphrasing means changing one or two words in another person's sentence, changing the structure of the sentence while preserving the original word, or changing some words to synonyms. Not. If you want to rearrange the sentence in any of these ways, it's too close to the original sentence. This is plagiarism, not paraphrase [14]. In other words, as long as the material is credited to the author and the source is credited to the last text of the sentence, it's a good way to use someone else's ideas to support your discussion. To ensure the paraphrase, the author tries to take notes about the resources used without reviewing the material and write down exactly what he is saying. This makes it easy to incorporate ideas into the author's words[15].

2.2 Plagiarism Detection

The distinct plagiarism detection strategies together with textual content primarily based totally, quotation primarily based totally and form primarily based totally are mentioned on this section. Plagiarism described because the act of taking or trying to take or to use (entire or elements) of every other person`s works, without referencing or mentioning him because the proprietor of this work[16]. There are varieties of plagiarism that happens maximum frequently:

i. Textual plagiarisms: this sort of plagiarism generally executed through college students or researchers in educational enterprises, wherein files are equal or traditional to the unique files, reports, essays medical papers and artwork design.

ii. A supply code plagiarism: additionally executed through college students in universities, wherein the scholars attempting or copying the entire or the elements of supply code written through a person else as one`s own, this varieties of plagiarism it's far tough to detect.

There are many plagiarism detection strategies together with textual primarily based totally plagiarism, quotation primarily based totally plagiarism, and form primarily based totally plagiarism for flowchart. Textual plagiarism is a sort of plagiarism that provides enjoyable effects if the plagiarized textual content is copied (reproduction & paste), with minor alterations (e.g. shake & paste) or gadget translated. However, if the textual content is paraphrased or translated through a human, the presently used techniques yield a completely bad performance. Citation primarily based totally Plagiarism Detection compares the occurrences of citations in an effort to discover similarities[17]. The maximum fundamental shape is to degree the bibliographic coupling strength. Strength of the quotation primarily based totally method lies in figuring out translation- and idea-plagiarism or disguised paraphrasing. Shape primarily based totally plagiarism for flowchart provides a way for detecting waft chart parent plagiarism primarily based totally on form primarily based totally picture processing and multimedia retrieval[18]. The technique controlled to retrieve flowcharts with ranked similarity in keeping with distinct matching sets. However, it's far not able to discover plagiarism for distinct figures and charts at the side of their contents.

Text Based Plagiarism: This sort of plagiarism makes a specialty of detecting the similarities among files through the usage of the vector area model. It can also calculate and remember the redundancy of the phrase within side the report, after which use the fingerprints for every report for matching it with fingerprints in different files and discover the similarity. This technique is appropriate for non-partial plagiarism as stated earlier than use the entire report and use vector area to healthy among the files, however if the report has been partly plagiarized it cannot attain right effects. It might also additionally consist of reproduction and paste, amendment or converting a few phrases of the unique statistics from the net, book magazine, newspaper, research, journal, private statistics or ideas[19].

The text-based plagiarism detection process involves four phases [20]:

i. Phase 1 Collection: This is the first phase of the plagiarism detection process, which requires students or researchers to upload assignments and dissertations to the web engine. The web engine acts as an interface between the student and the system.

ii. ii. Phase 2 Analysis: In this phase, each corpus or assignment submitted is performed through the similarity engine to determine which document is similar to the other documents. There are two types of similar machines, the first in-vitro machine and the second in-vitro machine. The internal engine works by returning an ordered list between similar pairs. In contrast, the in vitro engine returns the appropriate web link.

iii. Phase 3 Verification: The function of this phase is to determine if the relevant text has been stolen from other text, or if there is a high degree of similarity between the source document and another document. That is.

iv. Stage 4 Investigation: This is the final stage of the plagiarism detection process and relies on human intervention. In this step, a human expert is responsible for determining if the system is running properly and whether the results are really plagiarized or simply quoted

2.2.1 Different Methods Used for Textual Plagiarism Detection

Adoption of sentences, paragraphs, or entire sources verbatim, without quotation marks and without proper citation. The most common form of plagiarism among high school students. The Internet makes copy-and-paste very attractive because popular student plagiarism detection techniques rely on character-based methods to compare suspicious documents with the original ones. The same string can be recognized accurately or partially using the character matching approach[21]. A fingerprint method that finds string matches and plagiarism detection based on common stock of fingerprints. These methods gave good results, but failed when I paraphrased or changed some words in the suspicious text to change the plagiarized part. You can also calculate and count the verbosity of words in a document and use the fingerprints of each document to match the fingerprints of other documents to find similarities[22].

i. Grammar-based methods**:** Grammar-based methods are an important tool for detecting plagiarism. It focuses on the grammatical structure of documents, and this method uses a string-based comparison approach to detect and measure

similarities between documents. The grammar-based method is good for detecting the exact copy without modification, but for rewriting or exchanging several words with the same meaning to detect the modified copy text. Not suitable. This is considered one of the limitations of this method II.

ii. External plagiarism detection method: External plagiarism detection relies on a reference corpus consisting of documents whose passages may have been stolen. Suspicious documents are checked for plagiarism by looking for passages that overlap or nearly overlap with the passages in the documents in the reference corpus. An external plagiarism system then reports these findings to a human administrator. The administrator determines if the detected aisle has been plagiarized.

2.2.2 Plagiarism Based on Citation [CbPD]

"Citation-based plagiarism detection involves using citations and references to determine document similarity and identify plagiarism." It has long been known in academia that citations and references to scholarly publications provide valuable semantic information about the relationship between the content of a document and other works. The degree of similarity between citation patterns mainly depends on the amount of general reference (bibliographic bond strength) and the degree of similarity in the order of citations contained and the distance between each other[23].

i. Identifying citation patterns: Finding similar patterns in citations used in two academic texts is a powerful indicator of semantic text similarity and a central idea for citation-based plagiarism. The patterns are subsequences of the two texts

A and B citation tuples CA and CB, and are similar to each other because they consist of (partially) common references. The degree of similarity between patterns depends on the number of citations contained in the pattern and how similar their order and coverage are. Therefore, the literally matching subsequence of the citations of the two documents is a powerful indicator of semantic similarity[24].

ii. ii. Unknown pattern constituents: Unlike e.g. in string pattern matching the subsequences of citations to be extracted from a suspicious text and searched for within an original are initially unknown. Citations that are shared by the two documents are easily identified. However, it is unlikely that all of those shared citations represent plagiarized text passages. For instance, two documents might share 8 citations, of which 3 are contained within a plagiarized text section and 4 are distributed over the length of the text and used along with other non-shared citations without representing any form of plagiarism. Therefore, the citation order of the two documents is as follows:

Original: 1 2 3 x x x 4 x x 5 x 6 x 7 8

Plagiarism: x x 5 x x x 4 x 3 x 1 x 2 x x 7 x 8

The number 18 represents shared quotation marks and the letter x represents non-shared quotation marks. Shared Citation 13 is intended to represent the plagiarized part.

2.2.3 Citation-Based Architecture of Plagiarism Detection System

Citation-based architecture of plagiarism detection system Java's open source software system, called CitePlag, was developed for citation-based plagiarism detection. These steps are performed on a plagiarism detection system.

1. The document is analyzed and some heuristics are applied to process the citation, including its position within the document.

2. The citation is compared to the bibliography entry.

3. Document citation-based similarity is calculated. The developed prototype CbPDS consists of three main components. The developed prototype CbPDS consists of three main components.

The first is a Relational Database System (RDBS) termed CbPD database storing data to be acquired from documents as well as detection results. The second is the detection software called CbPD Detector that retrieves data from the CbPD Database, runs the different analysis algorithms to be evaluated and feeds the resulting output back to the CbPD Database. The third component, the CbPD Report Generator, creates summarized reports of detection results for individual document pairs based on adjustable filter criteria[25].

2.2.4 Methods for Detecting Plagiarism

1. Character-Based Methods: Most plagiarism detection strategies belong to this category. These strategies make the most man or woman-primarily based totally, phrase-primarily based totally, and syntax-primarily based totally functions. It makes use of those functions to discover similarity among a question record and present files. However, the similarity among

a couple of files can be expected the use of each genuine matching and approximate matching. In genuine matching, each letter in each the strings ought to be matched within side the identical order. Our survey famous that maximum detection strategies are advanced primarily based totally on n-gram or phrase n-gram primarily based totally genuine string similarity locating approach. For instance, some researchers use character 16-gram matching, while the authors of nine use phrase 8-gram matching. Similarly, a few researcher has made a powerful use of approximate string matching approach[26]. This string matching indicates diploma of similarity/dissimilarity among strings. There are numerous proximity measures to be had to aid the approximate string matching. One can use string similarity metric or vector similarity metric for the purpose.

2. Vector-Based Method: Here, lexical and syntax functions are extracted and classified as tokens in preference to strings. The similarity may be computed the use of numerous vector similarity measures like Jaccard, Dice`s, Overlap, Cosine, Euclidean and Manhattan coefficients. Our remark is Cosine coefficient and Jaccard coefficients are famous and powerful in locating similarity among vectors. Cosine coefficient in detecting partial plagiarism without sharing files content. Hence its miles beneficial to come across plagiarism in files in which submission is taken into consideration as confidential[27].

3. Syntax-Based Methods: These strategies make the most syntactical functions like a part of speech (POS) of word and phrases in distinctive statements to come across plagiarism. The factors of simple POS tag are verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions and interjections. Some authors use POS tag functions accompanied through string similarity metric to investigate and calculate similarity among texts. The authors make use of syntactical POS tag to symbolize a textual content shape as a foundation for in

addition evaluation and evaluation i.e., files containing identical POS tag functions are done for in addition evaluation and for identity of supply of a plagiarism[28].

4. Semantic-Based Methods: A sentence can be described as an ordered institution of phrases. Two sentences can be identical however the order in their phrases can be distinctive. The diploma of similarity among phrases utilized in knowledge-primarily based totally measures through thirteen is calculated the use of facts from a dictionary. This similarity among phrases is used as semantic similarity among phrases. In some other approach, the author used WordNet to calculate the semantic similarity[28]. Also, some researchers decide semantic similarity through counting the wide variety of nodes of shortest route among concepts[29].

5. Fuzzy-Based Method: In the fuzzy-based method, text similarity. B. A set of represented by values ranging from zero (completely different) to one (exact match). Here, the words in the document are represented by a set of words with similar meanings, and each word in the document is assigned a similarity, so the sentence is considered ambiguous[30]. This technique is an attractive because it can detect similarities between uncertain documents. Some research creates a correlation matrix consisting of words and their corresponding correlation coefficients. It measures the similarity between different words[31]. Then get the similarity between the sentences by calculating the correlation coefficient between the pairs of words from two different sentences in each document. The similarity between two documents or any two web documents is identified using a fuzzy IR approach[32]. Author presents a tool for this. It takes a fuzzy approach to see how similar the two Arabic statements are. To do this, they used a corpus of 4477 source statements and 303 query / suspicious statement plagiarism[33].

6. Structure-Based Method: Structure-based to find similarities between two documents, in contrast to the above method, which was developed based on the vocabulary, syntax, and semantic features of the text in the document. The method uses contextual similarity. Something like "like" Words is used throughout the document. However, in our research we can find several ways of in this category. Contextual information is typically processed using the tree-structured feature representation found in ML-SOM [34]. In a research, the author recognizes plagiarism in two steps. The first step uses a tree-structured feature representation to perform document clustering and candidate retrieval, and the second step uses ML-SOM for recognition[35].

7. Stylometry-Based Methods: These methods aim to quantify the author's writing style in order to detect plagiarism. Calculates the value of similarity between two sections or paragraphs or sentences based on the stylistic characteristics of the author. These methods are essentially cases of plagiarism. The style expression can be writer-specific or reader-specific[21]. The writing-specific style is most often, with the strength of the author's vocabulary and the complexity of document presentations. Reader-specific styles, on the other hand, deal with how readers can easily understand the text. With this approach, outlier mining helps detect plagiarism of documents.

8. Cross-Lingual Plagiarism Detection Method: Cross-lingual plagiarism detection is a difficult task. Requires in-depth knowledge of some languages. Finding a suitable similarity metric for such a procedure is also an important issue. This type of method is based on the text function between languages.

Various types of these methods include:

(1) cross-lingual syntax based methods

(2) cross-lingual dictionary based method

(3) cross-lingual dictionary based methods[23].

A detail survey on Cross-Lingual methods is done and a statistical model is used to evaluate the similarity between two documents regardless of the order in which the terms appear in suspected and original documents[35, 36].

9. Grammar Semantics Hybrid Theft Detection Methods: These methods are effective theft detection methods because they use natural language processing. They can find the copy / paste and paraphrase the plagiarism exactly. Such methods remove the limitations of semantic-based methods. Semantics-based methods usually cannot detect and determine the location of the stolen part of a document, but such grammar-based methods can effectively solve this problem [37, 18].

10. Classification and Cluster-Based Method: Grouping of 4484 documents with and without monitoring plays an important role in the information gathering process. Many research issues such as text summarization, text classification, plagiarism detection, classification and clustering help reduce search space during the information retrieval process[38, 39, 40]. This helps significantly reduce the comparison time for documents during plagiarism detection. Some methods use keywords or specific words to group similar sections of a document[41, 42].

11. Citation-Based Method: a new method for detecting citation-based plagiarism is proposed. This method is a new approach to plagiarism and detecting scientific documents that have been read but not cited[43]. The citation-based method belongs to the semantic

plagiarism detection technique because it uses the semantics contained in the citation of the document[44]. The similarity between two documents is calculated based on the similarity pattern in citation sequence [44]. Approach to identify plagiarism, most existing projects use different approaches to identify plagiarism such as exact matches, sentence-based matches, fingerprints, and substring matches. Fingerprints are a computerized technology. The fingerprint here represents a summary of the documents that are compared to identify suspicious data blocks. Substring matching matches pairs of strings, and these substrings are represented by a suffix tree. The algorithm is then applied to detect plagiarism. Another method, stylometry, is used to identify the author's attribution and capture the author's own style. Citation-based pattern analysis checks for citations and references used in text documents. Theft detection of copied data appeared in the 1970s, and common natural language processing (NLP) methods for detecting copied data are grammar-based methods, meaning-based methods, and grammar and meaning. Introduced in three different methods of hybrid method. The grammar-based method preserves the grammatical structure of the document and uses string matching techniques to calculate the similarity between the documents. The semantic procedure uses an information retrieval technique in the document and a vector space model of statistical word frequency to get the vector of the document and computes the vector of the two documents using the dot product, cousin, or other method. To do. This labeled vector is a document similarity. This technique is not effective because it does not identify the source of the stolen data. Hybrid grammatical semantic methods improve the perception of these two methods. It is important and effective to highlight or mark the stolen text in the document in parallel with the result of the similarity[45].

2.2.5 Shape-Based Plagiarism Detection for Flowchart

Most, if not all, discard the figures and charts before checking for plagiarism. Discarding the figures and charts results in look holes that people can take advantage. That means people can plagiarize figures and charts easily without the current plagiarism systems detecting it. There are very few papers which talks about flowcharts plagiarism detection. Therefore, there is a need to develop a system that will detect plagiarism in figures and charts. Flowcharts become a significant issue to explain different kinds of information based on figure types. In some documents, flowcharts are so important to illustrate a lot of details and make it easier to understand methodology of structured design is one of primary steps to build entire system and solving engineering problems that can be explained by using flowcharts and other types of figures. This system primarily focuses on flowcharts detection. The database contains flowchart images stored in a single folder. The system will retrieve and rank this database based on a given query by the user[81].

2.3 Plagiarism Detection Approaches

The extrinsic plagiarism detection approach compares suspicious documents to collections from the document that was adopted that it has accepted (referenced collection), calls all documents similar to the threshold as a potential source. In extrinsic PDS, the specified suspicious document is compared to the available reference document corpus or collection. This reference collection is either online or offline. Online source of the WWW or offline database that contains the source document. Each recognition system is intended to find intense suspicious joints of available source documents and corresponding counterparts. Suspicious document inputs are compared to available sources to determine whether they are copied or manipulated from one of these reference documents. A source corpus or database can be the entire web, some specific libraries, or some domain-specific databases. Because a

comparison database is available, it acts like a document comparison mechanism with a similarity scheme[6]. Most online plagiarism checkers work in a similar way, comparing suspicious input to documents available on the web, some databases, or a combination of both. The essential plagiarism detection approach only looks at the input document and does not compare it to the document in the reference collection. The unique recognition method uses a process known as stylometry to examine the linguistic characteristics of the text. The goal is to identify stylistic changes that the approach considers to be an indicator of potential plagiarism. Language differences can be used to enter for exogenous suppression analysis or to human beings[4].

In the intrinsic plagiarism detection approach**,** suspicious documents are compared to the reference source corpus. In contrast to extrinsic PDS, reference for unique PDS is not available. Here, suspicious documents are individually analyzed without comparing with all sources. The author's writing style, structural distribution, vocabulary, etc. are analyzed here. Thus, different stimulated features are extracted to identify these stolen cases. Extrinsic PDS, VIZ, VIZ, SYNTAX-based, semantic, based syntax-based, semantic, base, base, and cited technology, or combinations of these technologies can use various detection technologies. Current research is focusing on exogenous text stolen detection methods, methodology and condition. We also analyze the current plagiary checker limit. This study describes the steps used in exogenous plagiarism recognition first and then discuss the prior art of this domain based on the available recognition procedures and systems. Next, we will explain the PAN plagiarism contest to gain a better understanding of the various obfuscations and operations that plagiarists may impose. In the next section, we will perform an analysis of some online plagiarism tools using the obfuscated text described in PAN. Furthermore, we will point out

common problems and research gaps, and conclude the discussion based on future prospects[46].

MDR (Match Detect Reveal) is a method of first splitting a document that is checked for plagiarism into a fixed-length string by preserving the suffix tree. The string matching algorithm is used for comparison, and the longest common string can be found in the suffix tree. This allows you to get the similarity index and its position in the document. This technique is inefficient because it uses exact matches and creates an obscure plagiarized text version[46]. There are various tools that use web-based services, some of which are standalone applications. Turnitin, Article Checker, and Duplichecker are the most common examples of web-based services. While some of these tools, with the exception of Turnitin, offer free and online text-based plagiarism in limited versions, Turnitin supports both in-vivo and in-vitro detection and is not a free service. Plagiarism Checker X, Copy Catch, Plagiarism Detector, WORDCheck and CopyFind are standalone application software. There are many plagiarism detection approaches which can be used by the applications. Some uses Ngram for improving results in text base.

2.3.1 Plagiarism Detection Tool

There are many software systems that suggest that you can reliably determine if a submitted text or online document has been plagiarized. Software can only expect to compare syntax at the letter or word level and determine similarities between texts. Several experimental studies have been conducted in the field of semantic recognition. However, this seems to be successful only in areas of highly structured text, such as the Code programming language[47].

i. PlagAware: An online text plagiarism detection service that enables users to search, search, analyze, and track plagiarism on specific topics. For example, like topics, PlagAware is a search engine that is considered as follows: A strong key factor in recognizing the typical content of a given text. It uses traditional search engines to detect and scan for plagiarism and provide different types of reports to help users or document owners determine if a document has been plagiarized.

ii. PlagScan: Online software for checking text theft. Commonly used in schools, PlagScan offers different types of accounts with different features. PlagScan uses a complex algorithm to check and analyze uploaded documents to detect plagiarism based on current language research. A unique signature extracted from the document structure. This is compared to the PlagScan database and millions of online documents.

iii. CheckForPlagiarism.net: Developed by a team of professional scholars, it is one of the best online plagiarism checkers used to stop or prevent online plagiarism and minimize its impact on academic integrity. To maximize accuracy, CheckForPlagiarism.net protects documents from theft by using several methods such as document fingerprinting and document source analysis.

2.4 Data Mining

Data mining is the latest interdisciplinary field of computational science. Data mining is the process of discovering attractive information from large amounts of data stored either in data warehouses, databases, or other information repositories. This is the process of automatically finding data patterns in a huge database [48, 49]. Data mining refers to extracting or "mining" valuable information from large amounts of data [50, 51]. The extracted data can be

represented as a dataset [51, 52]. Process discovery is a learning task that helps build a process model from information system event logs[53]. Attractive insights, observable behaviors, or general information can be extracted from the database through data mining and viewed or retrieved from different angles. The knowledge found can be applied to process control, decision making, information management, and question processing. Decision makers use these methods to make clear decisions to further improve the real problems of this world. Data mining uses many data clustering techniques to track specific data patterns[51]. Figure 2.2 shows a data mining technique for better understanding. Clustering techniques are useful metalearning tools for analyzing the knowledge produced by modern applications. Clustering algorithms are used extensively not only for organizing and categorizing data but also for data modelling and data compression[52]. The purpose of the clustering is to classify the data into groups according to data similarities, traits, characteristics, and behaviors[52]. Data cluster evaluation is an essential activity for finding knowledge and for data mining. The process of clustering is achieved by unsupervised, semi supervised, or supervised manner [53]. However, there are more than 100 clustering algorithms known and selection from these algorithms for better results is more challenging.

PyClustering is an open source data mining library written in Python and C ++ that provides a variety of clustering techniques and algorithms, including bio-inspired vibration networks. PyClustering focuses primarily on cluster analysis to make it more user-friendly and understandable. Many methods and algorithms are located in the C ++ ccore: clst namespace and the Python module pyclustering.cluster. Algorithms and their availability in the PyClustering module include KMeans, KMedians, KMeans ++, XMeans. Clustering in data mining Data volumes continue to grow exponentially in various scientific and industrial

disciplines, making automated classification techniques the standard tool for dataset exploration[54]. An automatic classification technique, traditionally called clustering, helps clarify the structure of the dataset[55]. Clustering is an established method based on unsupervised data mining that handles the discovery of structures in unlabeled data collections[56]. The overall process following the development of an unsupervised learning solution can be summarized in the in Figure 2.2

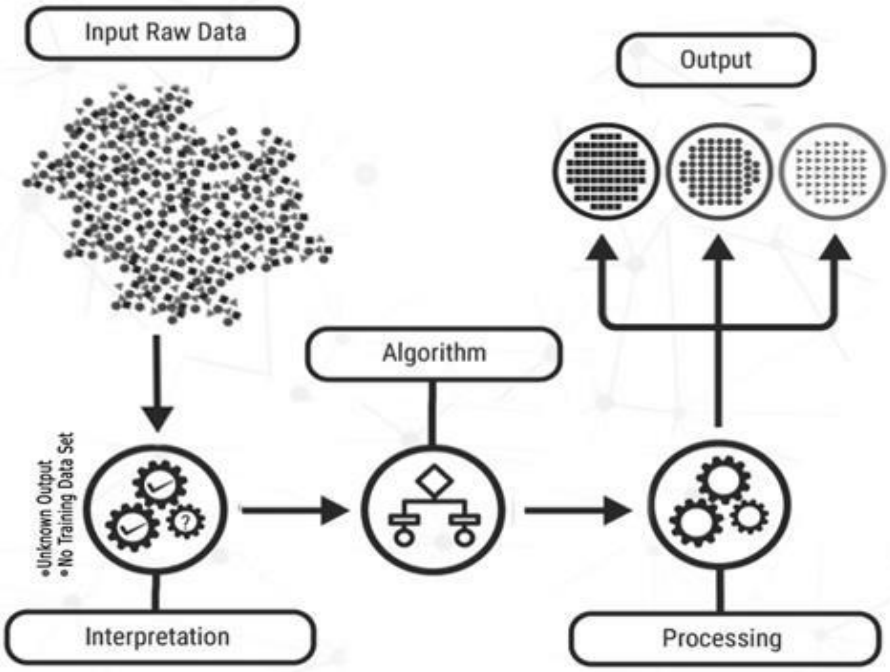


Figure 2.2: Unsupervised Learning Model[56]

The main applications of unsupervised learning are:

- Simplify datasets by aggregating variables with similar attributes.
- Detecting anomalies that do not fit any group.

- Segmenting datasets by some shared attributes.

Clustering leads to a reduction in the number of dimensions in the dataset. The goal of such a clustering algorithm is to identify different groups within dataset [57]. There are various clustering objects such as hierarchy, partition, grid-based, density-based, model-based [58]. The performance of the various methods depends on the type of data used for clustering and the amount of data available[59]. For example, document clustering has been used in various areas of text mining and information retrieval[60]. There are different metrics for quality, relative ranking, and the performance of different clustering algorithms, which can vary significantly depending on the measurements used. Two measures of cluster "goodness" or quality are used for clustering. One type of measurement allows you to compare different sets of clusters without external knowledge and is called an "internal quality measurement". Another form of measurement, called an "external quality measurement," allows you to evaluate how well clustering works by comparing the groups generated by the clustering technique with the identified classes.

Types of clustering can usually be divided into two subgroups:

Hard clustering: In hard clustering, each data point is complete or not part of the cluster. For example, each customer can be divided into one of 10 groups.

Soft Clustering: Soft clustering assigns the probability or probability of data points in a particular cluster instead of placing each data point in another cluster. For example, each customer may be in 10 clusters. Clustering method Clustering method: Since the clustering process is subjective, it is a tool that can be used to achieve many goals. Each methodology follows various rules and regulations that describe the "similarity" between data points.

Cluster analysis is not an automated task, but multiple collaborative optimization with multi-literary cooperation with repeated information investigator processes or attempts and error [61]. There are typically over 100 known clustering algorithms. However, some of these algorithms are commonly used. Some of the clustering techniques are shown in Table I below. The most well-known and most widely used partitioning methods are Kmeans [62, 63, 64]. Although there are many clustering techniques, Kmeans is an unsupervised iterative data mining approach[65]. The standard approach for all clustering techniques is to classify the cluster centers that represent each cluster. Kmeans clustering is a method of cluster analysis aimed at subdividing the observed data points into k clusters, where each observation is part of the closest mean cluster [61]. The most important advantage of the Kmeans algorithm in data mining applications is its efficiency when clustering large datasets. Kmeans and its various variants have computational time complexity proportional to the number of records, but are expected to detect inferior clusters[62].

2.4.1 Typical Clustering Methodologies

Method	Algorithm
Distance-based method	Partitioning algorithms "K-means, K-medians, K-medoids." Hierarchical algorithms, "Agglomerative, Divisive method." These algorithms are run repeatedly to find the local optimum and are very easy to understand, but lack the scalability to handle large datasets.
Grid-based method	Grid-base algorithm: Individual regions of the data space are formed into a grid-like structure. These methods use a single uniform grid to divide the entire problem area into cells. A cell uses a collection of statistical attributes from the object to represent the data object inside the cell.
Density-based method	Density-Based Spatial Clustering of Applications with Noise / DBSCAN Ordering points to identify the clustering structure OPTICS These algorithms scan the data space for areas of different data point densities within the data space. Isolate regions of different densities within the same

	cluster and assign data points within those regions.
Probabilistic and generative models	Expectation-maximization algorithm: Modeling data from a generative process. Often these models suffer from over-fitting. A prominent example of such models is the Expectation-Maximization algorithm that uses normal multivariate distributions

Table 2.1. Clustering Methodologies[62]

2.4.2 Background and discussion of clustering applications and approaches

Cluster analysis has many applications in different areas. Clustering analysis can also be used in collaborative filtering, recommender systems, customer segmentation, multimedia data analysis, biological data analysis, and social networks. Use analysis and dynamic trend detection. Table 2.2 describes some of the clustering techniques and approaches.

Clustering Techniques and Approaches with Benefits

Author	Year	Technique/ Algorithm	Approach	Outcomes
Chunhui Yuan and Haitao Yang	2019	K-Means Clustering Algorithm	Different methods applied to each dataset to determine the optimal selection of K-Value.	Concluded that these four methods (Elbow methods, silhouette coefficient, gap statistics, and canopy) satisfy the criteria for clustering small data sets. In contrast, the canopy algorithm is also the best choice for large and complex data sets[66].
Tengfei Zhang, Fumin Ma	2015	Rough k-means clustering	Improved rough k-means clustering with Gaussian function based on a weighted distance measure	An improved rough k means algorithm based on weighted distance measure with Gaussian function handles the objects which are wrongly assigned to clutters, also handles

				vulnerable sets while distributing *overlapping* objects in different clusters by rough k means with the same weighted distance for both upper and lower bounds[67].
Lior Rokach, Oded Maimon	2015	Clustering methods: Hierarchical-based, Model-based, Grid-based, Partitioning based, Density-based.	Different clustering methods/techniques are used to determine clustering efficiency in large data sets and explain how the number of clusters can be calculated.	For large dataset concluded that "K-means clustering is more efficient in terms of its time, space complexity, and its order-independent" and "Hierarchical clustering is more versatile, but it has the following disadvantages: Time complexity $O(m^2 * logm)$ and space complexity of a hierarchical agglomerative algorithm is $O(m^2)$[68].
Zengyou He,	2015	K-mean, K-	Compare different	K-Histogram is the

Xiaofei Xu, Shengchun Deng, Bin Dong	modes, K-Histogram	clustering algorithms to determine an efficient clustering algorithm for the categorical dataset.	enhanced version of K-means to categorical areas by substituting means of clusters with histograms. In general, K-Histogram is almost similar to the K-modes algorithm, but as compared to k-modes, k-histogram algorithms are more stable, and the algorithm will converge faster[69].
M.Venkat Reddy, M. Vivekananda, RUVN Satish.	2017 Divisive, and Agglomerative Hierarchical Clustering with K-means.	Discover an efficient clustering by comparing Divisive and Agglomerative Hierarchical Clustering with K-means.	To obtain high accuracy, Agglomerative Clustering with k-means will be the practical choice. Divisive clustering with K-means also works efficiently where each cluster can be

				taken fixedly[70].
Ahamed Malki, Mohamed M. Rizk, M.A. El-Shorbagy, A. A. Mousa	Al 2016	K-means, Genetic algorithm	For solving the clustering problems, introduced a hybrid approach of the Genetic algorithm with K-means.	A hybrid approach of K-means with a Genetic algorithm efficiently solves all the problems of the k-means, e.g., K-mean will produce empty clusters with initial centre vector and converge to non-optimal value, etc[71].

Table 2.2. Clustering Techniques and Approaches with Benefits[66, 67, 68, 69, 70, 71]

A. Requirement and Challenges

Despite recent efforts, the challenge of clustering with "mixed and categorical" data in the sense of big data essentially means the high computational complexity of current clustering techniques and the similarity between category objects. Remains due to the lack of certain measurements[72]. There are several things to consider in cluster analysis. Some of them are listed in Table 2.3.

2.4.2.1 Considerations for Clustering Analysis

Considerations for Clustering Analysis		
Considerations	**Options**	**Examples**
Similarity measure	Distances-based / Connectivity-based	Euclidean, road network, vector / Density, Contiguity
Partitioning criteria	Single level / Hierarchical partitioning	Often / Multi-level
Cluster space	Full space / Subspace	Low-dimensional / High-dimensional
Separation of clusters	Exclusive / Non-exclusive	Datapoint belongs to only region / Data point belongs to multiple regions

Table 2.3. Considerations of Clustering Analysis[72]

Typically, there are multiple ways to use or apply clustering analysis; some advantages and limitations of clustering techniques are mentioned in Table 2.4.

- As a stand-alone tool to get insights into data distribution.
- As a preprocessing (or intermediate) steps for other algorithms.

2.4.2.2 Clustering techniques "Advantages & Limitations"

Clustering techniques "Advantages & Limitations"

Clustering Techniques	Advantages	Limitations
Data-mining clustering algorithms	It's easy to implement. Attack on user privacy.	Do not process large amounts of data
Dimension reduction	It is very fast, reduces datasets and optimizes treatment costs.	Must be applied before the classification algorithm. Cannot provide efficient results for high dimensional datasets. Some data may be lost.

Parallel classification	It's more scalable with minimal execution time.	Challenging to implement.
MapReduce framework	Flexibility, scalability, security and authentication, batch processing, etc.	It does not do best for graphs, iterative, and incremental, multiple inputs, etc.

Table 2.4. Advantages and Limitations of Clustering Techniques[72]

A research shows that parallel classification is a better approach for big data, but it remains a major challenge due to its complex implementation. While the MapReduce framework may be suitable for implementing parallel algorithms, there is no single algorithm that can handle all big data challenges[73, 74]. In 1960, the author proposed a new "SELM" algorithm for Spark extreme learning machines based on the Spark parallel framework to speed up and improve efficiency throughout the process. SELM provides maximum speed and minimum error in all experimental results compared to Parallel Extreme Learning Machine (PELM) and Enhanced Extreme Learning Machine (ELM *). Table 2.5 shows the strengths and weaknesses of various clustering algorithms in real-world applications.

Algorithm Name	Pros	Cons	Applications in Real World
K-means	Process large amounts of data. Minimum run time. Easy to implement, etc.	Manually choose the K value. Clustering outliers. Dependent on starting point/value. Handle empty clusters, etc.	Wireless networks. System diagnostic. Search Engine. Document Analysis. Fraud detection. Call record Analysis.
Hierarchical Clustering	Do not need to specify the initial value. Easy to implement, scalable and easy to understand, etc.	Cannot process large amounts of data of different sizes. No backtrack. There is no exchange between objects.	Humans skin analysis[61] Generating a portal site. Web usage mining.

		More space and time complexity.	
Genetic Algorithm	It is easy to understand and converges on various problems. Not all problems can always provide the best results, but we can provide the best solution. Unable to search for a single point or population of points.	It is computationally intensive, for example, Time Consuming. It takes time. Data can be lost in crossover.	Engineering Designs. Robotics. Telecommunications, traffic, shipments routing. Virtual Gaming. Marketing.
DBSCAN	There is no need to define the number of clusters. Handles outliers.	Records with different densities cannot be processed. It is difficult to operate high-dimensional data.	Satellite pictures, etc.

Table 2.5. Clustering Algorithms Pros and Cons[74]

2.4.3 KNN (k-nearest neighbor algorithm) Method

KNN method (nearest neighbor algorithm). One of the simplest machine learning algorithms for pattern recognition is k-Nearest Neighbor algorithm. In some cases, the k-Nearest Neighbor (kNN) algorithm is used in most pattern recognition implementations and works fine. In the kNN algorithm, "k" is a parameter. Choosing the right k-value for the kNN algorithm requires some testing with different k-values. KNN maps the "k" nearest neighbor

in the text dataset to the neighborhood. kNN remembers all the documents in the training dataset and compares the relevant test documents with them. When you create a classification, you can assign a specific key word that is incorrectly ranked from the wrong group text or from the training package. As a result, the determination of such training sets is almost impossible. There are several disadvantages such as related documents and content needs. By using a specific text classifier, KNN can be used to solve the above drawbacks. And the training set has its own characteristics. To define and distinguish paper materials, they use taxonomic criteria and their own definition levels. The kNN algorithm has been used in various fields so far. Used for text classification due to its versatility and accuracy[75].

2.5 Review of Related Works

An automatic system is proposed to detect plagiarism. This system uses neural network techniques to create a feature based plagiarism detector and to measure the relevance of each feature in that available assessment. The paper solely focus on two different aspects namely copypaste type and paraphrasing plagiarism types only. The results were compared with commercially available online software "Article checker" [76].

Plagiarism is defined as a serious problem that violates copyrighted documents / materials. A new plagiarism detection method called SimPAD is proposed. The purpose of this method is to determine the similarity between two documents through sentence-by-sentence comparisons. Experiments have shown that SimPAD detects plagiarized documents more accurately than existing plagiarism detection approaches[77].

Focus was placed on the educational context and similar challenges faced, description on how to check the plagiarism cases. In addition they planned to build learning communities of

students, instructors, administration, faculty and staff and all collaborating and constructing strong relationships that provide the foundation for students to achieve their goals with greater success. Seamless integration with legacy and other applications was provided in some easy, modifiable, and reusable way. Learning portal may provide a support tool for these learning system. This paper gives the software to detect the plagiarism from java student assignments[78].

To plagiarize is to robe credit of another person`s work. It describes the first attempt to detect the plagiarized segments in a text employing statistical language models and perplexity. The experiments were carried out on two specialized a literary corpora. The two specialized works contained the original documents and part of speech and stemmed versions. Plagiarism was detected on these documents and the results were verified[79].

Laboratory work assignments are very important for computer science learning. A plagiarism detection tool was developed. This tool had the full toolset for helping in the management of the laboratory work assignment. Four similarity criteria was used to measure the similarities between two assignments. This paper described how the tool and the experience of using them over the last 12 years in four different programming assignment[80].

2.6 Summary of Gaps in Literature Reviewed

This section gives a summary of some research works and the limitation of the work

Plagiarism Detection through Data Mining Techniques was developed, this is a paper on plagiarism detection which makes use of data mining techniques to increase the efficiency of the plagiarism detection. Method used was k-nearest neighbor's algorithm (KNN). The limitation of the paper is that it does not identify the time taken to complete the execution[81].

AntiPlag system was developed, Plagiarism Detection on Electronic Submissions of Text Based Assignments, this paper focuses on creating an effective and fast tool for plagiarism detection for text based electronic assignments. Method used is tri-gram sequences matching technique. The limitation of the system is that the result of the system is not accurate[82].

A system to detect plagiarism for text documents on the World Wide Web is developed. This paper discusses about plagiarism detection, text mining and web mining and how to avoid the plagiarism on the web. Method used is finger printing, substring matching and pattern analysis. The limitation of the system is that it cannot detect plagiarism based on source code[83].

A Survey on Plagiarism Detection: The paper provides a brief idea about the types of plagiarism, how it is practiced and ways to detect them. The method used is clustering techniques and K means algorithm. The limitation of the paper is that it identifies that no tool can detect intelligent plagiarism specifically idea adoption[84].

Plagiarism detection for document: This paper was on plagiarism detection for document. It checks similar sounding and active passive voice detection. Method used is NET Framework and Microsoft SQL database. The limitation of the system is that it cannot detect source code based plagiarism[85].

An online system for plagiarism detection. This is a paper was on plagiarism detection which searches the Internet for evidence of plagiarism within a document. Method used was PHP scripting language in conjunction with the Google Internet search engine and various Linux applications (W3m, GNU/Linux program). The limitation of the system is that it cannot identify the performance level of the plagiarism checker[86].

Detecting the plagiarism for text documents on the World Wide Web. This paper discusses about plagiarism detection, text mining and web mining and how to avoid the plagiarism on the web. Method used is finger printing, substring matching and pattern analysis. The limitation of the system is that it takes long time to complete its execution[87].

Online assignment plagiarism checking using data mining and NLP. This paper discuss how they are going to develop a system for plagiarism detection in which whenever a student submits an assignment it detects that it is plagiarized or not by comparing with other students assignments. The method used is data mining algorithm and natural language processing. The limitation of the system is that it is limited to the institution[88].

Endnotes

1. Mike Perkins, Ulas Basar Gezgin & Jasper Roe1 *Reducing Plagiarism through Academic Misconduct Education* **International Journal for Educational Integrity** 2020 16:3

2. Durga Bhavani Dasari & Dr. Venu Gopala Rao. K. *Detecting the Plagiarism for Text Documents on the World Wide Web* **International Journal of Social Relevance & Concern Volume 2 Issue 10,** October 2014 ISSN No: 2347-9698

3. Irina Rets a & Ali Ilya b, Eliciting ELT Students' *Understanding of Plagiarism in Academic Writing* **Eurasian Journal of Applied Linguistics***, 4*(2), 193–211 EJAL Eurasian Journal of Applied Linguistics 4(2) 2018, pp. 193–211

4. Tomáš Foltýnek, Norman Meuschke & Bela Gipp *Academic Plagiarism Detection: A Systematic Literature Review* **ACM Computing Surveys Volume 52 Issue 6** January 2020 Article No.: 112 pp 1–42

5. Mahwish Abid, Muhammad Usman & Waleed Ashraf *Plagiarism Detection Process using Data Mining Techniques* **International Journal of Engineering Studies – Vol. 5,** No. 4, 2017

6. Vani K & Deepa Gupta S*tudy on Extrinsic Text Plagiarism Detection Techniques and Tools* **Journal of Engineering Science and Technology Review 9** (5) 2016 pp. 9 – 23

7. Aal M, Rizk M., El-Shorbagy M. A., & Mousa A. A., "*Hybrid Genetic Algorithm with K-Means for Clustering Problems,"* **Open J. Optim**., vol. 05, no. 02, 2016 pp. 71–83.

8. Adiningrum T. S. *Reviewing Plagiarism*: **An Input for Indonesian Higher Education. Journal of Academic Ethics***, 13*(1), 2015 pp 107-120

9. Ahmed H. O., Naomie S. & Albaraa A.," *Survey of Text Plagiarism Detection*" **International University of Africa, Faculty of Computer Studies, Khartoum, Sudan** Vol. 1, No. 1, June 2012

10. Ali H. H. & Kadhum L. E, "*K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition,*" **Int. J. Sci. Res.,** vol. 6, no. 8, 2017 pp. 1577–1584.

11. Ali, A., & Taqa, A. Y. *Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches.* **Journal of Education and Science,** 31(1), 2022, pp. 8-25.

12. Alzahrani S. M., Salim N., & Abraham A., "*Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods*," **Trans. Sys. Man Cyber Part C, vol. 42**, no. 2, pp. Mar. 2012, 133–149.

13. Alzahrani S. M., Salim N., Abraham A., *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods***, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)** 42 (2) 2012, 133-149.

14. Amidwar, S. *Plagiarism Detection Using Supervised Machine Learning Algorithm.* **Journal of Emerging Technologies and Innovative Research (JETIR),** 4(06) 2017.

15. Awale, N., Pandey, M., Dulal, A. & Timsina, B. *Plagiarism Detection in Programming Assignments Using Machine Learning.* **J Artif Intell Capsule Netw,** 2(3), 2020, pp. 177-184.

16. Bhardwaj A., Arya A., Vyas T., Shivani D. & Anuja N. *A Survey on Plagiarism Detection.* **Institute of Technology, Nirma University, 382481-Ahmedabad, India. Advances in Computational Sciences and Technology** Volume 10, Number 8, 2017, pp 2351-2365 ISSN 0973-6107

17. Duan M., Li K., Liao X, & Li K., **"***A Parallel Multi classification Algorithm for Big Data Using an Extreme Learning Machine,"* **IEEE Trans. Neural Networks Learn. Syst.,** vol. 29, no. 6, 2017 pp. 2337–2351.

18. Durga B. D. & Venu G., Rao. K. *Detecting the Plagiarism for Text Documents on the World Wide Web.* **International Journal of Social Relevance & Concern** Volume 2 Issue 10, October 2014 ISSN No: 2347-9698

19. Eman S., Al S. & Hadeel Q. G. *Plagiarism Detection Using Semantic Analysis.* **Indian Journal of Science Tech. Advances in Computational Sciences and Technology** Volume 10, Number 8, 2016, pp 2351-2365 ISSN 0973-6107

20. Eppa, A., & Murali, A. H. *Machine Learning Techniques for Multisource Plagiarism Detection.* **In 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions IEEE (CSITSS),** December 2021, pp. 1-5.

21. Faizan M., Zuhairi M., Ismail S., Sultan S., ***"****Applications of Clustering Techniques in Data Mining: A Comparative Study****"*** **(IJACSA) International Journal of Advanced Computer Science and Applications,** Vol. 11, No. 12, 2020 pp 146

22. Farhan, N. S. & Abdulmunem, M. E. *Image Plagiarism System for Forgery Detection in Maps Design*. **In 2019 2nd Scientific Conference of Computer Sciences (SCCS)**, March 2019, pp. 51-56, IEEE.

23. Gupta A., Himansshu S. & Anas A. *A Comparative Analysis of K-Means and Hierarchical Clustering*, **EPRA International Journal of Multidisciplinary research (IJMR),** Vol 7, Issue 8, August 2021, ISSN 2455-3662

24. Hossain M. Z., Akhtar M. N., Ahmad R. B., & Rahman M., *"A Dynamic K-Means Clustering for Data Mining,"* **Indones. J. Electr. Eng. Comput. Sci.,** vol. 13, no. 2, 2019 pp. 521–526.

25. Hrkút, P., Ďuračík, M., Mikušová, M., Callejas-Cuervo, M. & Zukowska, J. *Increasing K-means clustering algorithm effectivity for using in source code plagiarism detection.* **In International Conference on Smart Technologies, Systems and Applications, Springer, Cham.** December 2019, pp. 120-131.

26. Ilham, A. A., Bustamin, A., Aswad, I. & Armin, F. *Implementation of clustering and similarity analysis for detecting content similarity in student final projects*. **In IOP conference series: Materials science and engineering Vol. 875, No. 1, p. 012039. IOP Publishing** 2020, June.

27. Indirapriya P. & Ghosh D. K., "*A Survey on Different Clustering Algorithms in Data Mining Technique,"* **Int. Journal. Mod. Eng.** Res., vol. 3, no. 1, 2013 pp. 267–274, ISSN: 2249-6645

28. Jain P. K. & Pamula R., "*Two-Step Anomaly Detection Approach*." **International journal of engineering education**, 2018

29. Jazyah, Y. *Open Learning, the Issue of Plagiarism-Efficient Algorithm*. **International Journal of Computers,** 3, 2018.

30. Karthikeyan B., George D. J., Manikandan G., & Thomas T., *"A Comparative Study on K-Means Clustering and Agglomerative Hierarchical Clustering,"* **Int. J. Emerg. Trends Eng. Res.,** vol. 8, no. 5, 2020. pp. 1600–1604

31. Khanna, S., Rakesh, N., & Chaturvedi, K. NOperations on Cloud Data, *Classification and Data Redundancy.* **In Advances in Computer and Computational Sciences Springer, Singapore,** 2018, pp. 169-179.

32. Mahwish A., Muhammad U. & Muhammad W.A., *Plagiarism Detection Process using Data Mining Techniques* **Muhammad Waleed Ashraf Riphah International University Faisalabad, Pakistan. iJES –** Vol. 5, No. 4, 2017

33. Mike P., Ulas B. G. & Jasper R. *Reducing Plagiarism Through Academic Misconduct Education* **Perkins et al. International Journal for Educational Integrity** 2020 16:3

34. Muhammad F., Megat F. Z., Shahrinaz I. & Sara S. *Applications of Clustering Techniques in Data Mining: A Comparative Study.* **International Journal of Advanced Computer Science and Applications (IJACSA),** vol. 11, No 12, 2020

35. Nazir, A., Mir, R. N. & Qureshi, S. *Idea plagiarism detection with recurrent neural networks and vector space model.* **International Journal of Intelligent Computing and Cybernetics**, 2021.

36. Novak, M., Joy, M. S. & Mirza, O. M. *Improved plagiarism detection with collaboration network visualization based on source-code similarity.* **In 2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR),** May 2021, pp. 1-6. IEEE.

37. Novikov A., "*PyClustering: Data Mining Library,"* **J. Open Source Software,** vol. 4, no. 36, 2019 p. 1230.

38. Patel D., Modi R., & Sarvakar K., *"A Comparative Study of Clustering Data Mining: Techniques and Research Challenges,"* **Int. J. Latest Technol. Eng. Manag. Appl. Sci.,** vol. 3, no. 9, 2014 pp. 67–70.

39. Rashid S., Ahmed A., Barazanchi I. A, & Jaaz Z. A., *"Clustering Algorithms Subjected to K-Mean and Gaussian Mixture Model on Multidimensional Data Set,"* **Period. Eng. Nat. Sci.,** vol. 7, no. 2, 2019 pp. 448–457.

40. Reddy M., Vivekananda V., M, &. Satish R. U. V. N, *"Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,"* **Int. J. Comput. Sci. Trends Technol.,** vol. 5, no. Sep-Oct, 2017 pp. 5–11.

41. Rets I. & Ali I. 'Eliciting ELT Students' *Understanding of Plagiarism in Academic Writing* **EJAL Eurasian Journal of Applied Linguistics** 4(2) 2018, 193–211

42. Rustadi, R., Darmawan, I., Nandi Haerudin, N. H., Agus Setiawan, A. & Suharno, S. 1. *Plagiarism Checker JGSE, Groundwater exploration using integrated geophysics method.* **Journal of Groundwater Science and Engineering,** 10(1), 2022, 10-8.

43. Saeed, A. A. M. & Taqa, A. Y. *A proposed approach for plagiarism detection in Article documents*. **Sinkron: jurnal dan penelitian teknik informatika,** 7(2), 2022, 568-578.

44. Saxena A., Mukesh P., Akshansh G., Neha B., Om P.,Aruna T, Meng J., Weiping D. & Chin T. "*A Review of Clustering Techniques and Developments*," **Neurocomputing journal**, vol. 267, Issue C, 2017 pp. 664–681.

45. Seethongchuen, T. & Ratanaworabhan, P. *Improving Plagiarism Checker Throughput with Apache Storm.* **In 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology IEEE (ECTI-CON),** July 2019, pp. 393-396

46. Senosy A., Fadhil N. A. & Ahmadu M." *Shape-Based Plagiarism Detection for Flowchart Figures in Texts*" **Faculty of Computing, University Technology Malaysia, Skudai, Malaysia International Journal of Computer Science and Information Technology (IJCSIT)** Vol 6, No 1, February 2014

47. Sharma, M. K. & Sharma, S. *Plagiarism Checker and Classification of Files on Cloud Using Smart Cloud.* **In Proceedings on International Conference on Emerg** Vol. 2, October 2018, pp. 81-88.

48. Sindhu L., Bindu T., & Sumam M. I. "*Automated Plagiarism Detection System for Malayalam Text Documents*" **International Journal of Computer Applications** Volume 106 – No. 15, November 2014. 0975 – 8887

49. Tomáš F., Norman M. & Bela G. *Academic Plagiarism Detection: A Systematic Literature Review* **ACM Computing Surveys** Volume 52 Issue 6 January 2020 Article No.: 112 pp 1–42

50. Vani K & Deepa G. *Study on Extrinsic Text Plagiarism Detection Techniques and Tools* **Journal of Engineering Science and Technology** Review 9 (5) 2016 pp 9 – 23

51. Verma M., Srivastava M., Chack N., Diswar A. K., & Gupta N., "*A Comparative Study of Various Clustering Algorithms in Data Mining,"* **International Journal of Engineering Research and Applications (IJERA).** Volume 10, Number 8, 2012, pp 2351-2365

52. Widaningrum, I., Mustikasari, D., Arifin, R., & Sugianti, S. *A Review of Detection Plagiarism in Indonesian Language*. **International Journal Artificial Intelligent and Informatics,** 1(2), 2018, pp. 65-75.

53. Xu D. & Tian Y., "*A Comprehensive Survey of Clustering Algorithms,"* **Ann. Data Science journal,** vol. 2, no. 2, 2015, pp. 165–193.
54. Yuan C. & Yang H., "*Research on K-Value Selection Method of K-Means Clustering Algorithm*," **J**, vol. 2, no. 2, 2019, pp. 226–235.
55. Zerhari B., Lahcen A. A., & Mouline S., "*Big Data Clustering: Algorithms and Challenges,"* **Proc. Int. Conf. Bihree Charact.** Call. 3Vs Volume, Veloc. Var., no. May, 2015, pp. 1–7.
56. Zhang T. & Ma F., *"Improved Rough K-Means Clustering Algorithm Based on Weighted Distance Measure With Gaussian Function*," **Int. J. Comput. Math.,** vol. 94, no. 4, 2017, pp. 663–675.
57. Abhay N. P. & Chinmay N. B. *Plagiarism Detection System* **Department of Computer Science and Engineering Datta Meghe Institute of Engineering Technology and Research, Wardha, Maharashtra, International Journal of Innovations in Engineering and Technology vol 2(1)** India February 2013, 278 ISSN 2319-1058
58. Aggarwal C. C., *Data classification: Algorithms and applications*. 2014.
59. Al-suhaiqi11, M., Hazaa22, M. A. & Albared33, M. *Arabic English Cross-Lingual Plagiarism Detection Based on Key Phrases Extraction, 2 Monolingual and Machine Learning* Approach 3, 2018.
60. Alzahrani S. M., Salim N., & Abraham, A. *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods.* **IEEE transactions on systems, man, and cybernetics part c: application and reviews**, 42 (2) 2012
61. Bin D. S. D., Zengyou He & Xiaofei Xu, "K-Histrograms: *An Efficient Clustering Algorithm for Categorical Dataset*,"* no. 1, 2005, pp. 6–8.
62. Chowdhury H. A. & Bhattacharyya, D. K. *Plagiarism: Taxonomy, tools and detection techniques.* **arXiv preprint arXiv:1801.06323**, 2018.
63. Gawali, S., Thakore, D. S., Joshi, S. D. & Shinde, V. S. *A Plagiarism Detector Based on MAS Scalable Framework for Research Effort Evaluation by Unsupervised Machine Learning–Hybrid Plagiarism Model. Applied Machine Learning for Smart Data Analysis*, 41, 2019.
64. Habibah, M., Wicaksono, A., Aji, P. & Puji, M. *The Use Clustering Technique for Teaching Writing to Senior High School*, **Doctoral dissertation, Universitas Nusantara PGRI Kediri**, 2020.

65. Irsyadi, Y. *Plagiarism and Similarity Checker: Analysis of a cans waste classification system based on the CMYK color model using different metric distances on the k-means method*. 2022.

66. Lior R. O. M., "*Clustering methods*," Adv. Inf. Knowl. Process no. 9781447167341, 2015, pp. 131–167.

67. Malandrino, D., De Prisco, R., Ianulardo, M. & Zaccagnino, R. *An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering. Data Mining and Knowledge Discovery*, 2022, 1-34.

68. Mariyudi, M. *Hasil Plagiarism Checker: Jurnal Internasional Success Factors of SMEs: The Case of Indonesia*, 2019.

69. Mitra V., Wang C. J. & Banerjee S., Text Classification: *A Least Square Support Vector Machine Approach,* **Applied Soft Computing** 7 (3), 2007, 908-914.

70. Muna A., Rahat I., Vasile P., Saad A. & Victor C., "*An Integrated Approach for Intrinsic Plagiarism Detection"* in **, Suzhou, China: School of Computing, Electronics and Maths, Coventry University, United Kingdom, International Business School Suzhou, Xi'an Jiaotong Liverpool University**, 2019.

71. Nguyen T. H. T., Dinh D. T., Sriboonchitta S., & Huynh V. N., *"A Method for K-Means-Like Clustering of Categorical Data,"* J. Ambient Intell. Humaniz. Comput., no. Berkhin, 2019.

72. Rajiani, I. *Plagiarism checker: The effectiveness of acupressure intervention and birth delivery standing position to decrease the intensity of labor pain*, 2018.

73. Rajiani, I. *Plagiarism checker: The prospective consumers of the Indonesian green aviation initiative for sustainable development in air transportation,* 2018.

74. Ullah, F., Wang, J., Farhan, M., Habib, M. & Khalid, S. *Software Plagiarism Detection in Multiprogramming Languages Using Machine Learning Approach.* Concurrency and Computation: Practice and Experience, 33(4), e5000, 2021.

75. Zhang. H & Chow T. W., *A Coarse-To- Ne Framework to Efficiently Thwart Plagiarism, Pattern Recognition* 44 (2) 2011, 471-487

76. Zou D., Long W.-J. & Ling Z., *A Cluster-Based Plagiarism Detection Method*, in: **Notebook**
Papers of CLEF 2010 LABs and Workshops, September, 2010 pp. 22-23.

77. Zu S., Eissen M., Stein B. & Kulig M., *Plagiarism Detection without Reference Collections*, in: **Advances in data analysis, Springer**, 2007, pp. 359--366.

78. Ajin V. W. & Kumar L. D., "*Big Data and Clustering Algorithms*," in **International Conference on Research Advances in Integrated Navigation Systems**, RAINS 2016.

79. Alzahrani S. M. & Salim N., *The Use Of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents*, in: **Applications of Digital Information and Web Technologies**.
ICADIWT'12. **Second International Conference on the, IEEE,** pp. 539-544. 2, no. 3, 2012, pp. 1379–1384.

80. Anguita A., Beghelli A., & Creixell W., *Automatic Cross-Language Plagiarism Detection*, **7th International Conference on Natural Language Processing and Knowledge Engineering**, 2011

81. Chavan, H., Taufik, M., Kadave, R., & Chandra, N. *Plagiarism Detector Using Machine Learning.* **International Journal of Research in Engineering, Science and Management,** 4(4), 2021, pp. 152-154.

82. Dwivedi, J., & Tiwary, A. *Plagiarism Detection on Big-Data Using Modified Map-Reduced Based SCAM Algorithm*. **In 2017 International Conference on Innovative Mechanisms for Industry Applications IEEE (ICIMIA),** February, 2017, pp. 608-610.

83. Gipp B. & Beel J., *Citation Based Plagiarism Detection: A New Approach to Identify Plagiarized Work Language Independently*, in: Proceedings of the 21st **ACM conference on Hypertext and hypermedia, ACM**, 2010, pp. 273--274.

84. Gipp B. & Meuschke N., *Citation Pattern Matching Algorithms for Citation Based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence*, in: Proceedings of the 11th **ACM symposium on Document engineering**, ACM, 2011, pp. 249--258.

Chapter Three

Methodology

3.1 Introduction

This chapter describes the methodology that is be used in designing the **academic journal plagiarism checker** completely and will make the system function properly, performing the functional and nonfunctional requirements. The methods for achieving the objectives stated in this thesis will be explained below and the system architecture will be shown and explained.

3.1.1 Functional Requirements

The functional requirements describes the basic functions that the system must be able to perform. The functional requirements of the academic journal plagiarism checker are:

- User registration page
- User login page
- Plagiarism check
- Admin login page

3.1.2 Non Functional Requirements

Nonfunctional requirements are the functions that are not compulsory to be performed by the system, but if the system performs these functions, it will help to enhance the performance of the system and it will also help to make the system more secured. Some of the nonfunctional requirements includes:

- Security
- Reliability
- Usability
- User friendliness
- Fast performance

3.2 Research Approach

3.2.1 Software requirements

The following are requirements that is be needed for the journal plagiarism checker to be designed.

1. SQL: A structured query language is a standard language for storing, manipulating, and retrieving data in a database. Using SQL involves changing the database table and index structure. Add, update, and delete rows of data. Gets a subset of information from the database for transaction processing and analysis applications[1].

2. HTML: Known as the hypertext markup language, it is the standard language used to create web pages. HTML can embed programs written in scripting languages such as JavaScript and affect the behavior and content of web pages. CSS integration defines the appearance and layout of web pages[2].

3. CSS: Cascade Style Sheet (CSS) is a style sheet language used to describe the display of documents written in markup languages such as HTML. CSS, along with HTML and JavaScript, is the underlying technology of the World Wide Web. CSS is designed to separate presentation and content, such as layout, colors, and fonts. This separation gives you greater accessibility to your content, greater flexibility and control in specifying the characteristics of your presentation, and allows you to share the format across multiple web pages by specifying the relevant CSS in a separate CSS file. And can reduce the complexity and repetition of structural content[3].

4. PHP: Hypertext Preprocessor is a server scripting language and a powerful tool for creating dynamic and interactive web pages. This is a general-purpose programming language originally designed for web development[4].

5. JavaScript is a prototype based, multi paradigm, dynamic language, supporting object oriented, imperative, and declarative (e.g. functional programming) styles. JavaScript

supports event driven, functional, and imperative (including object oriented and prototype based) programming styles[5].

3.3 Methodology of the Thesis

The methodology used in this system is Object Oriented Analysis and Design. It's a structured method for analyzing, designing a system by applying the object-orientated concepts, and develop a set of graphical system models during the development life cycle of the software.

In the system, The Basic Steps for academic journal plagiarism checker using data mining includes:

i. to design the document repository y which is known as the database and it is implemented with E-R diagram

ii. to design the plagiarism checker system

iii. to carry out the implementation and analysis of the designs. Objective ii and iii which is development of the plagiarism checker and implementation of the algorithm respectively is implemented using:

HTML & CSS for design

PHP for connecting

Python plagiarism checker

For the analysis, Web Optimization is used.

iv. To evaluate the plagiarism checker system

3.3.1 Conceptual Diagram

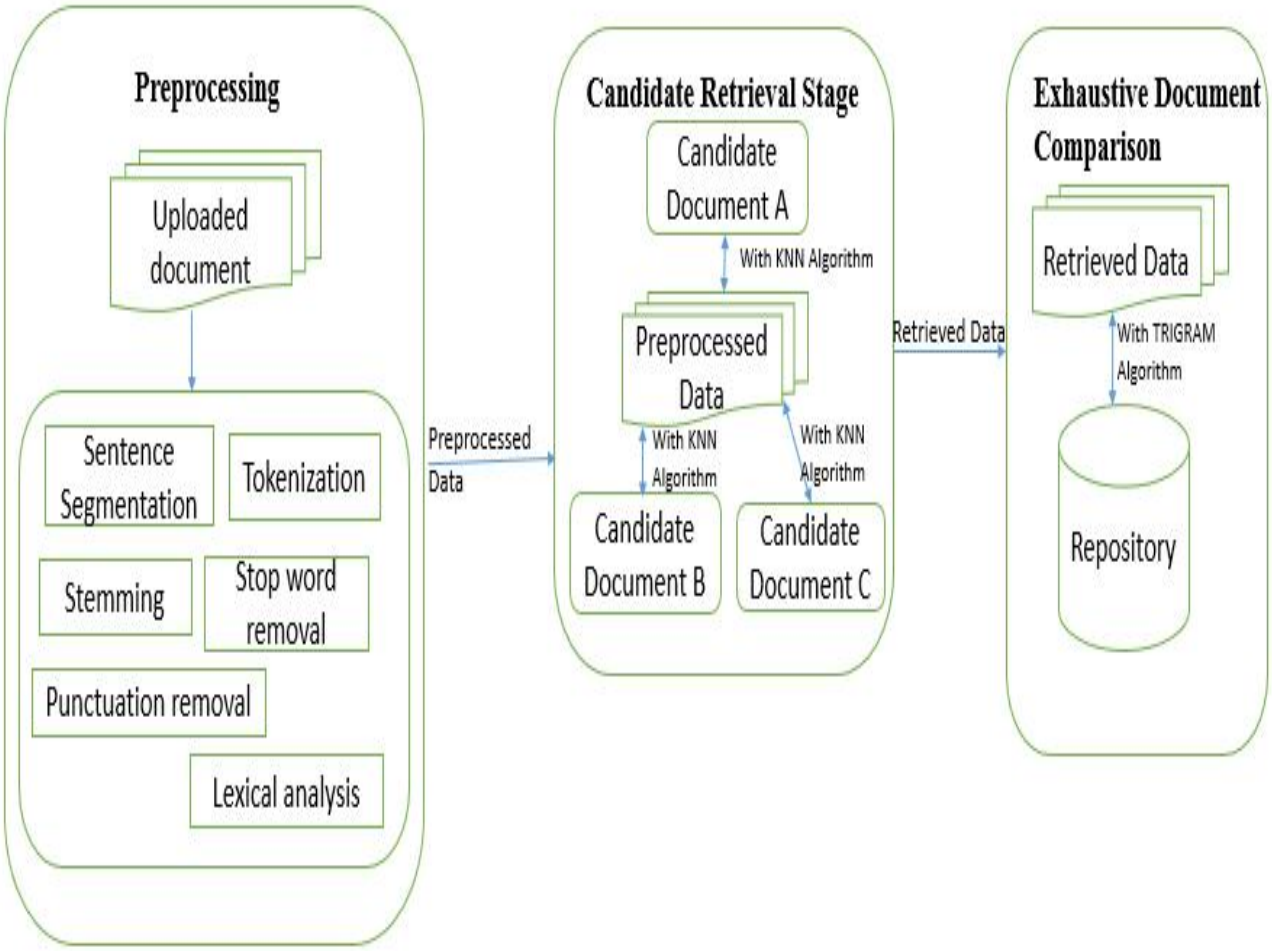
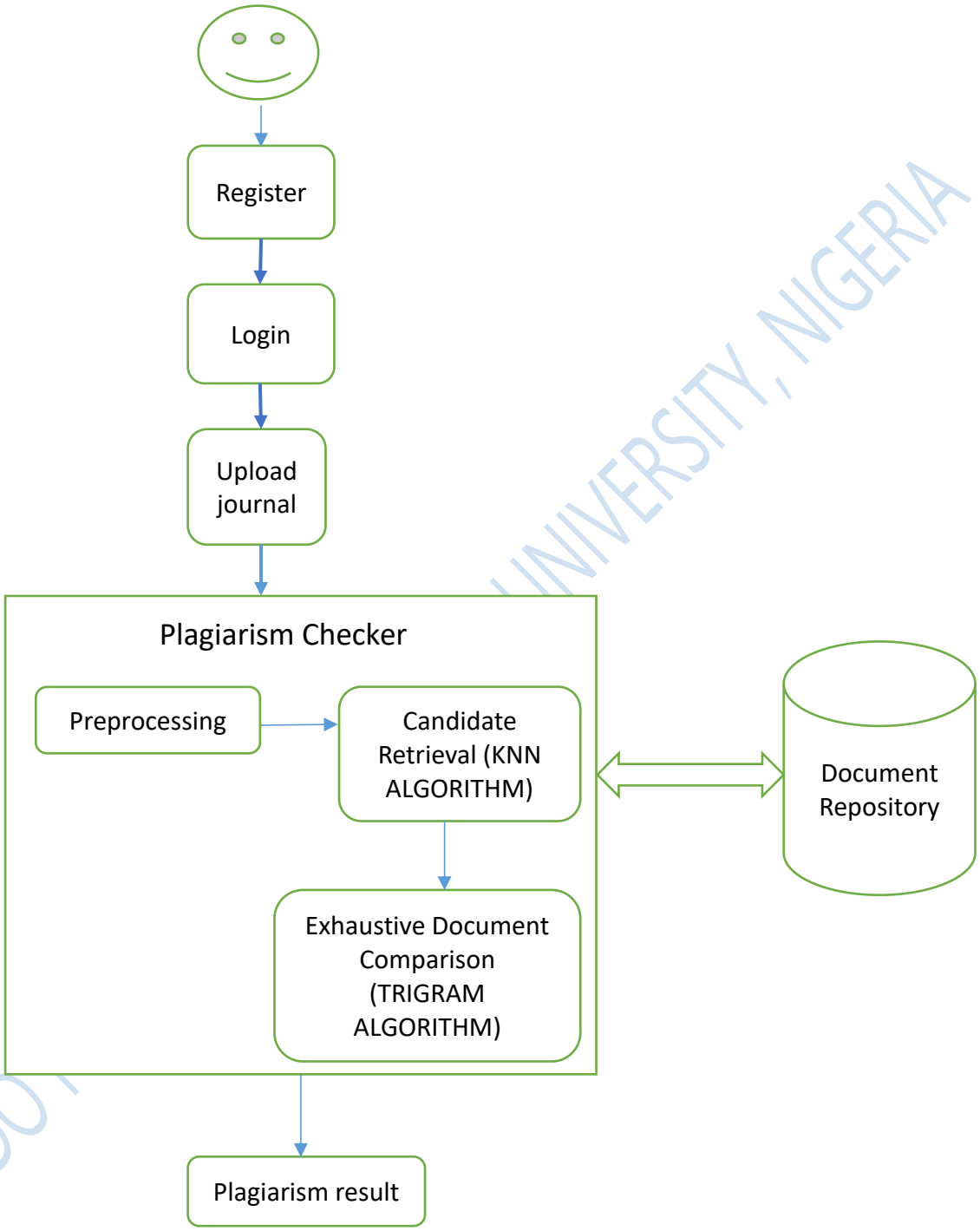


Figure 3.1 Conceptual Diagram (Researcher, Fatoki I. 2022)

3.3.2 System Architecture



The figure above is the system architecture of the academic journal plagiarism checker system. The system can allow user to upload a journal. For any user to be able to use the system, the user needs to register on the system, after the registration is complete, the user can then login into the system with the aid of the login details that has been given to the user during the registration period.

If the user wants to upload a journal, the user uploads the journal and the plagiarism checker which is implemented in the journal checks if the document uploaded by the user does not have a matching document in the repository. If the document is similar to another document in the repository, the plagiarism checker flags an error message to the user and would deny the user an access of uploading the journal. The stages involved in the plagiarism checker includes:

3.3.2.1 Pre-Processing

The document you have is preprocessed first and the irrelevant information is removed. This makes it easier to process the document. These include techniques such as sentence segmentation, tokenization, stop word removal, punctuation removal, and lowercase removal. Natural language processing (NLP) techniques, such as word stemming and lexical analysis, are also primarily used in this phase. Stemming is a heuristic process that removes affixes from words. Lemmatization uses vocabulary and morphological information to create a dictionary-based form of words. Although closely related to stemming, stemming affects only one word at a time, and lemming affects the entire text. Therefore, it is possible to distinguish words with different meanings by part of speech. The document is preprocessed based on the model or technique used. If the technique used performs a statement-based

comparison of documents, statement segmentation is performed. Here, the document is divided into sentence units based on some rough sentence boundaries or by applying some heuristics. Tokenization examines word-level documents by dividing them into tokens.

3.3.2.2Candidate Retrieval

The next important step after preprocessing is document-level plagiarism detection by retrieving near-duplicate sources. The output of the preprocessing phase is fed into the candidate retrieval phase. In real-world scenarios, the detection engine typically needs to match suspicious documents with large repositories or databases (some application-specific offline databases or the entire web). In any case, a thorough comparison of suspicious documents with all the documents in these databases can take a considerable amount of time. Therefore, to reduce this search space, document-level comparisons are performed to retrieve candidate sources for certain suspicious documents at hand. The Candidate Search task searches for globally similar source documents related to a particular suspicious document. Therefore, each suspicious document is associated with a source set called a candidate set. This process works like a search engine information search task, which retrieves documents related to a particular query.

3.3.2.3 Exhaustive Document Comparison

When candidate documents are retrieved, each suspicious document is thoroughly compared to its candidate set. Here, the suspicious plagiarism segment and its corresponding source component are identified. The detailed document comparison phase uses a variety of methods and detection techniques to compare each suspicious document with a candidate

source. The source and the suspicious component are compared using several similarity measures and the plagiarized fragment is selected.

3.3.2.4 Clustering: Clustering techniques can increase the efficiency of the recognition process. The KNN algorithm and the trigram algorithm are applied for this purpose. The KNN algorithm classifies the members of X based on their nearest neighbors, and comparison tests are performed on all data that may be called neighbors. This is a classification algorithm that shows similarity to other data. Clusters are created from similar trigrams to calculate the similarity score. Clusters assist in computation and speed up processes.

3.3.3 Algorithm for the System

Login with username and password

for *Project Upload*

Select document to upload

Check plagiarism

Classify the words into sequence of three (Trigram Algorithm)

Load the training data

Prepare data by scaling and dimensionality reduction

Find the optimal value of K (KNN Algorithm)

Predict a class for new data;

Calculate similarities X1 Xi from 1,2,3……n

Where X= new data point, X1=training data, distance as per your chosen similarity metric from the sequence

Sort the similarities in increasing order with corresponding train data

Find the most frequent class from the chosen K rows

Display plagiarism percentage

If plagiarized, return project

Else

Upload project

End if

3.3.4 E-R Diagram for Academic Journal Plagiarism Checker

According to [6] Entity Relation diagram is a pictorial representation that is used in database design which can be used to capture information of the database. The E-R diagram consists of entities, attributes of these entities and the relationships between the attributes. The database is manually loaded with journals downloaded from different scholarly websites.

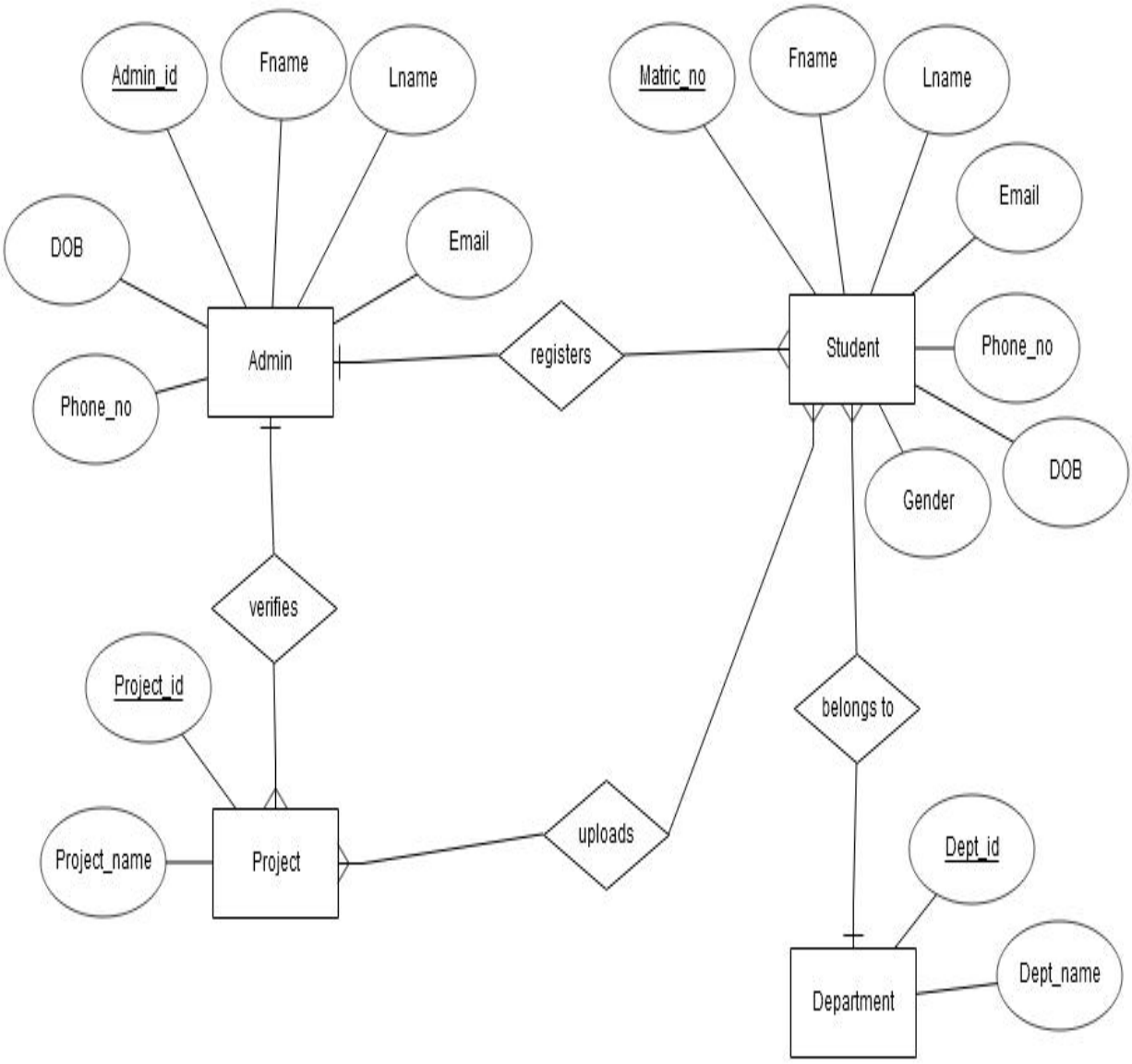


Figure 3.3 ER diagram of Academic Journal Plagiarism Checker (Researcher, Fatoki I. 2022)

The entities in the ER diagram includes Student, Admin, Department and the journal, all the entities are connected with the relationships which includes ‘registers’, ‘verifies’, belongs to’ and ‘uploads’’. The attributes of each entities are:

Students: Matric no, Fname, Lname, Email, Gender, Department, DOB, Phone_no.

Project: Project_id, project_name

Department: Department_id, Department_name

Admin: Admin_id, Fname, Lname, Email, Phone_no, Gender, DOB

Below is the detailed explanation of what the E-R diagram represents.

1. The relationship between the student and the admin is ‘register’, which shows that the admin registers the students and the cardinality between the admin to student is one to many since its only one admin that can register as many students that are in the department.
2. The relationship between the journal and the admin is ‘verify, which shows that the admin verifies the journal that is submitted by each students and the cardinality between the admin to journal is one to many since the admin can verify as many journals as possible that will be uploaded by the students.
3. The relationship between the student and the journal is ‘uploads’, which shows that the students can upload their journal and the cardinality between the student to journal is many to many because all the students in the department are all permitted to upload their journal into the repository.
4. The relationship between the student and the department is ‘belongs to’, which shows that the students belong to the department and the cardinality between the department

to student is one to many since the system is for the department of computer science and all the students in the department belongs to the department of computer science.

3.3.5 Activity Diagram for the System

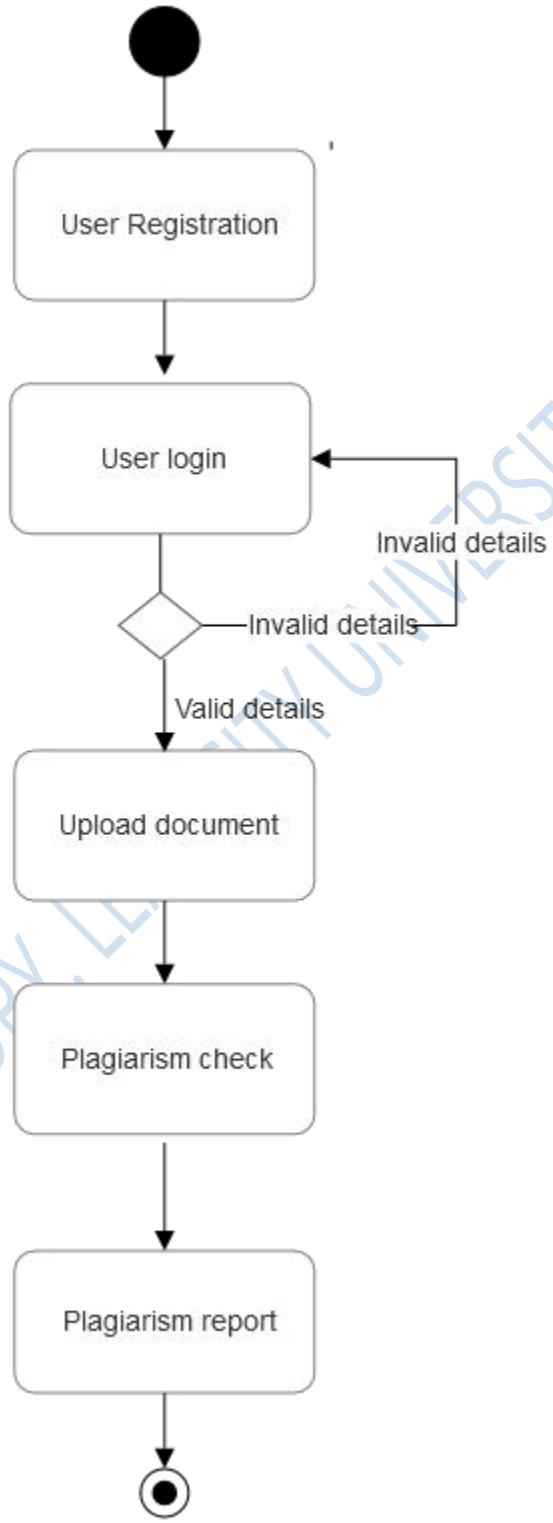


Figure 3.4 Activity Diagram for the System (Researcher, Fatoki I. 2022)

Activity Diagram

The image shows the activity diagram of the academic journal plagiarism checker. For anyone to be able to use the system, the user needs to register. The system enables the user to login into the system. The user then selects the activity that is needed to be done which is to upload journal. The figure shows that if the user selects journal upload, the system will compare the content of the document that is being uploaded to check for plagiarism. The system will check the document and will display the summary of the plagiarism that is done in the document that is to be uploaded.

3.3.6 Use Case diagram for the System

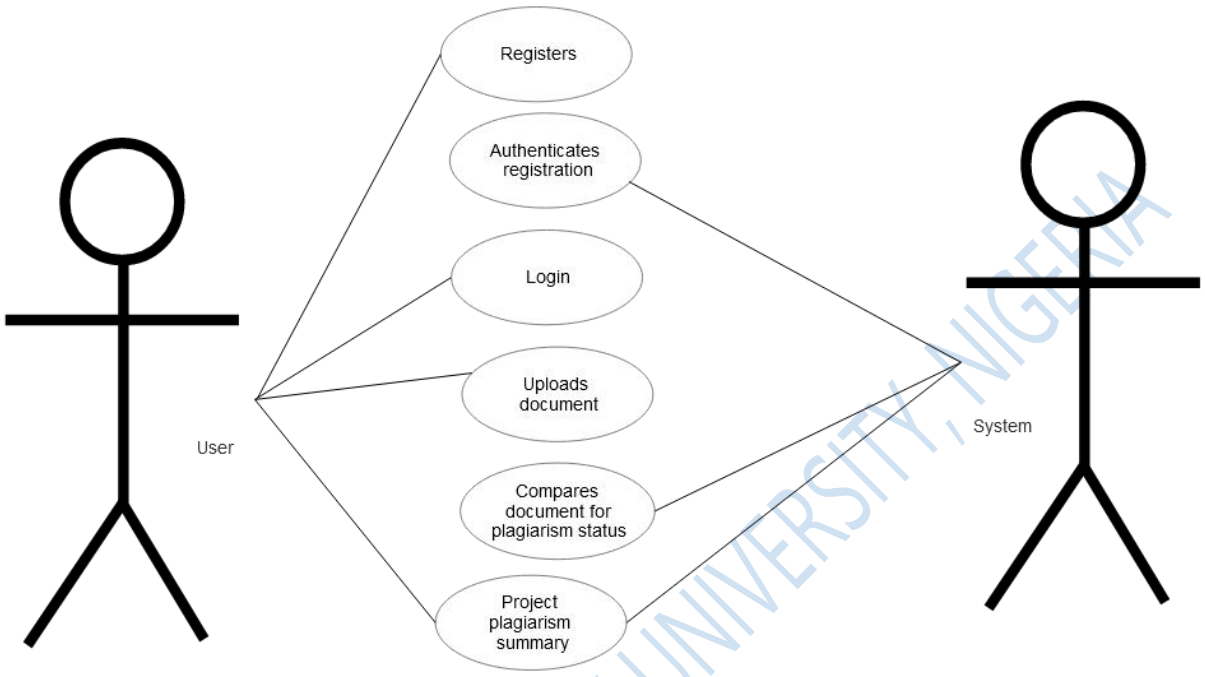


Figure 3.5 Use case diagram (Researcher, Fatoki I. 2022)

The use case diagram shows the two actors that will be included with the system, which is the user and the system administrator. The diagram shows the functions that can be performed by both the user and the system administrator.

The user can register, login, upload journal. The admin on the other end can login, view activities, register users, authenticate user registration and also validate user login, delete activities, delete user and also delete admin.

Endnotes

"Structured Query Language (SQL)". International Business Machines. October 27, 2016.

[2] "Index of elements in HTML 4". World Wide Web Consortium. December 24, 2012

[3] Bos Håkon & Wium Lie,. *Cascading Style Sheets: designing for the Web*. Harlow, England; Reading, MA. Addison Wesley Longman 1997. ISBN 0-201-41998-X.

[4] Lerdorf, Rasmus. "*PHP on Hormones – history of PHP* given at the MySQL **Conference in Santa Clara, California**". The Conversations Network. 2007-04-26

[5] "JavaScript". Collins English Dictionary – Complete & Unabridged 2012 Digital Edition. William Collins Sons & Co. 2012

[6] Chen, Peter. *"The Entity-Relationship Model* - Toward a Unified View of Data". ACM Transactions on Database Systems. **1** (1): 9. March 2016

Chapter Four

Implementation and Evaluation

4.1 Introduction

This chapter shows the result to the academic journal plagiarism checker system that has been created. This chapter displays the results of the areas the system is working on.

4.1 Login Page

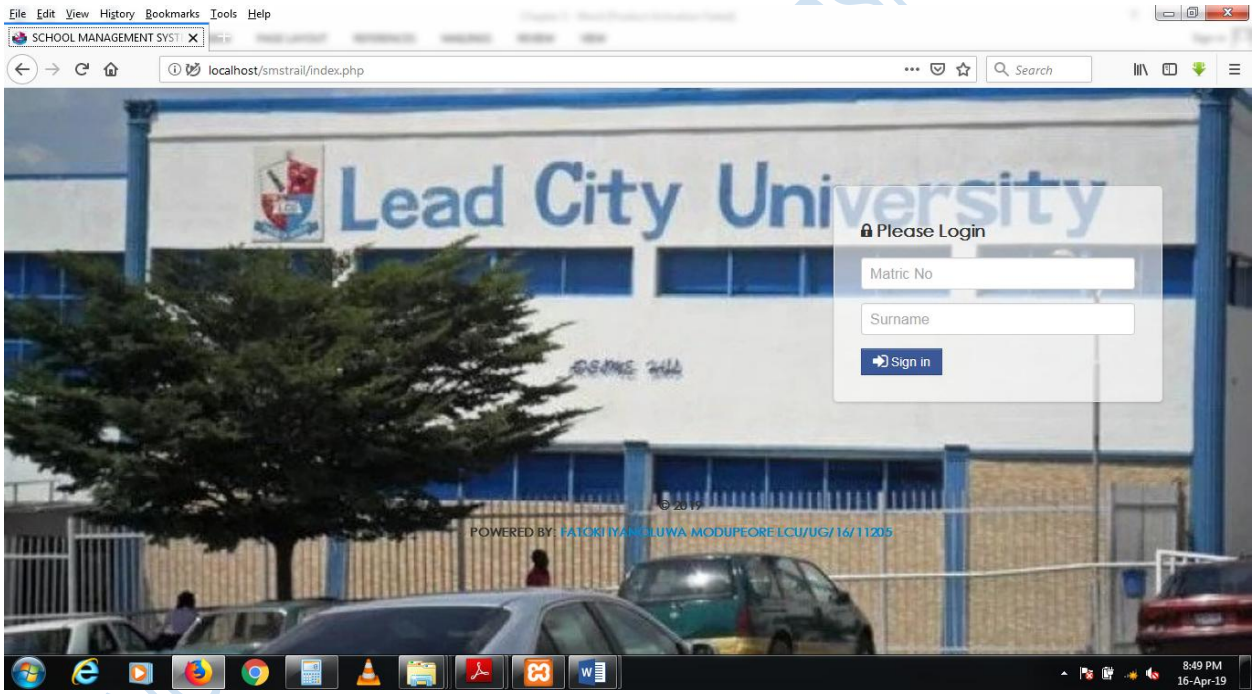


Figure 4.1 Login page (Researcher, Fatoki I. 2022)

The image in figure 4.1 is the home page of the plagiarism system. This page allows the students or the admin to sign in to the system. The students' login to the system with their matric number as the username and their surname as the password. The students cannot register themselves to use the system. The admin is the one who registers the students as a

user of the system. This will help to make the system more secure and prevent unauthorized user from the system. The admin sign in into the system with 'admin' as username and 'admin' as password. The page also allows the student or the admin to change their password.

4.2 Home Page

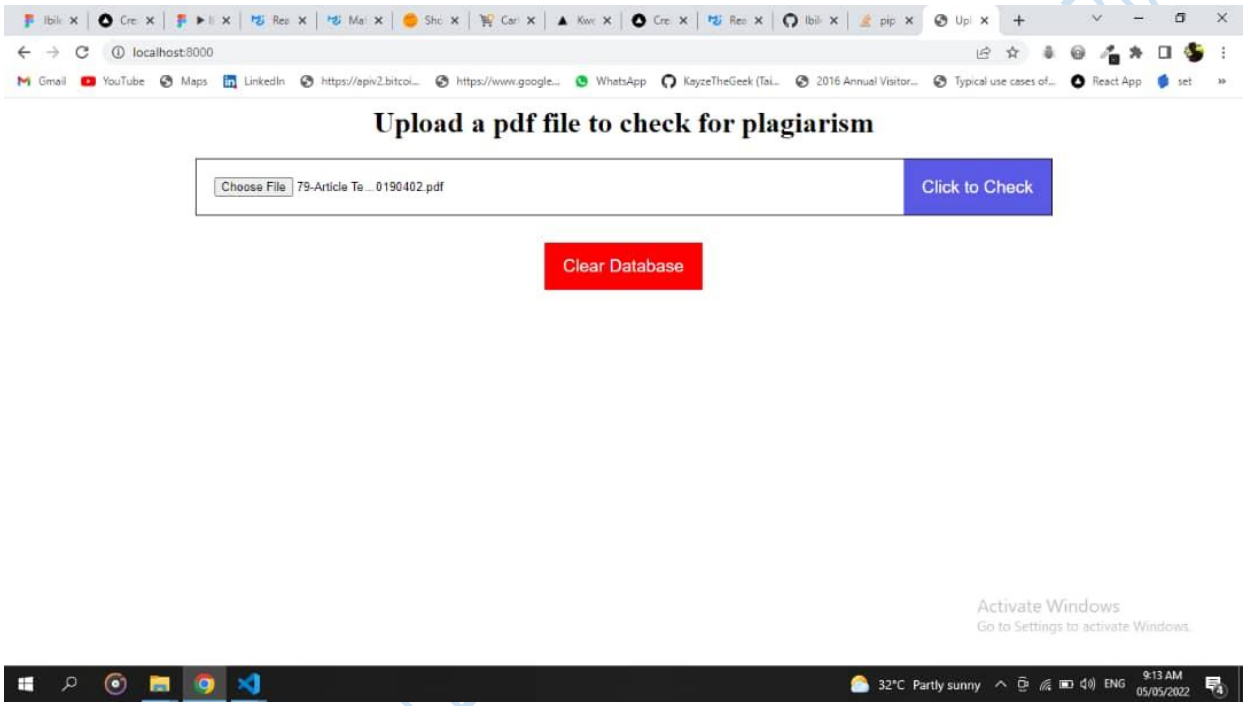


Figure 4.2 User's Homepage (Researcher, Fatoki I. 2022)

The snapshot in figure 4.2 is the homepage of the plagiarism system. The homepage displays the activities that can be performed by the users on the system. The users can upload journals to the system. The database can be cleared by the admin.

4.3 Plagiarism Result Page

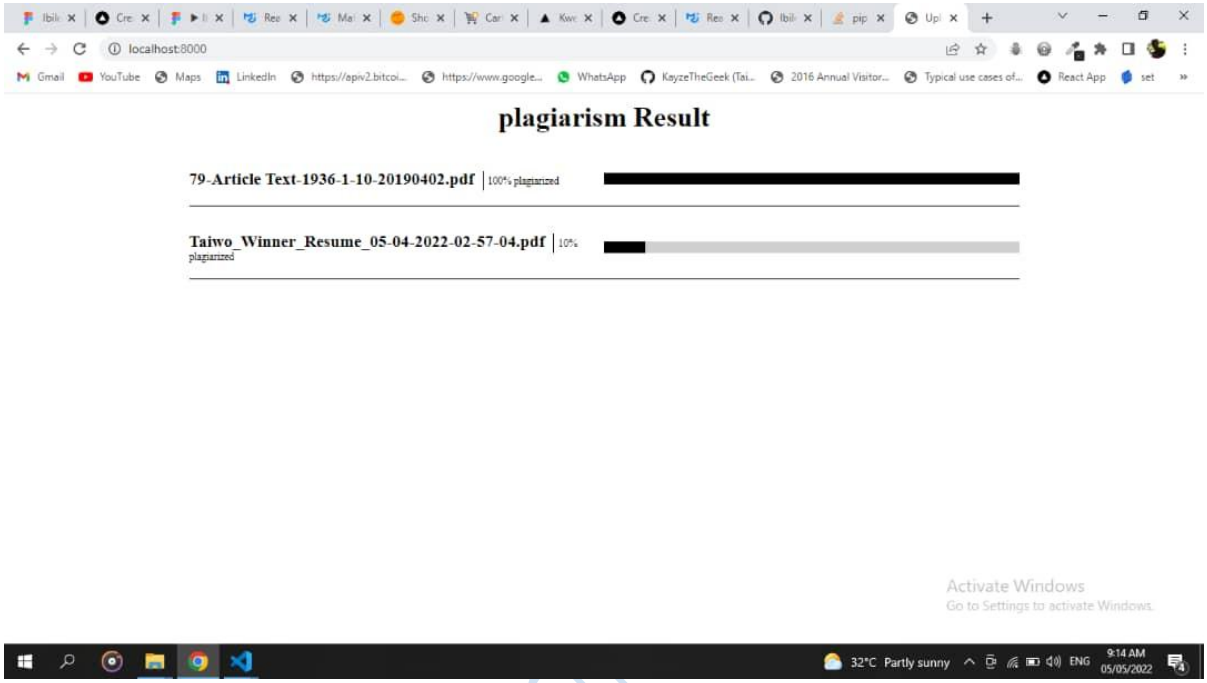


Figure 4.3 Plagiarism result page (Researcher, Fatoki I. 2022)

The snapshot in figure 4.3 shows the efficiency of the plagiarism report when a journal is uploaded in the system. The system determines the result based on the documents available in the database.

4.4 Result Page

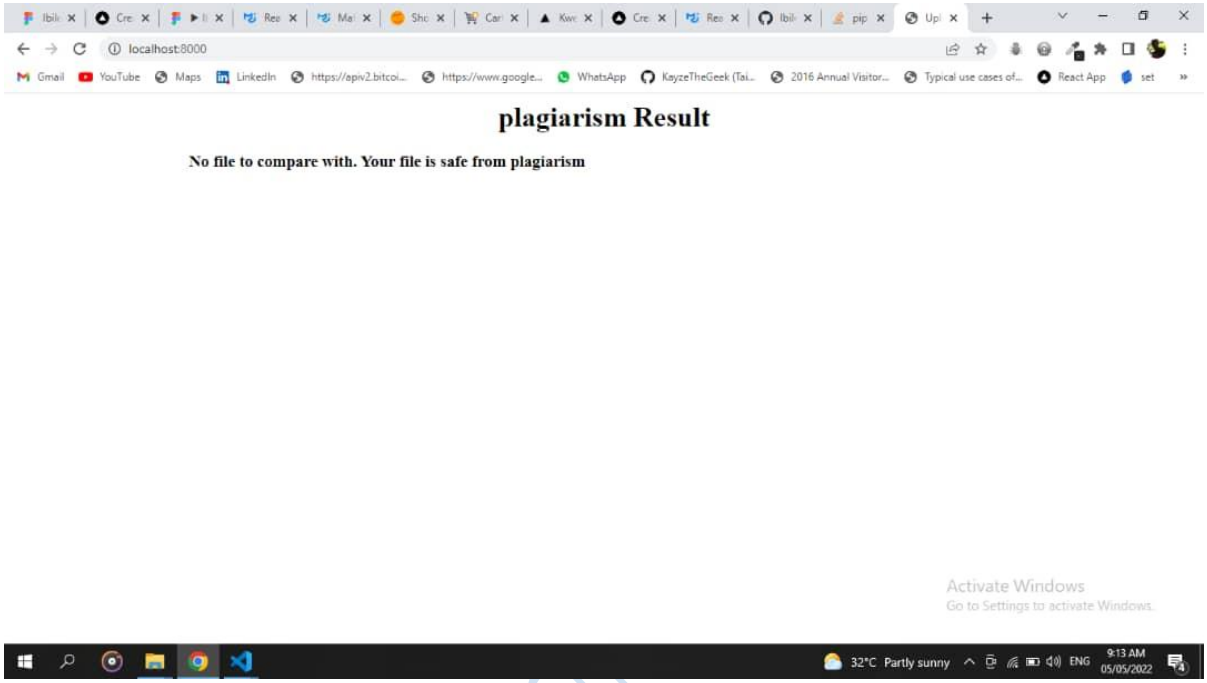


Figure 4.4 Result page (Researcher, Fatoki I. 2022)

The snapshot in figure 4.4 shows the plagiarism result when the database is cleared and there is no document to compare the journal that is to be checked with the system.

4.5 Chapter Summary: System Evaluation and Performance Metrics

Identifying the key components or parameters from our thesis. This includes us checking through previous developed plagiarism software. This is subjected to the following parameters:

- Response time
- Efficiency
- Accuracy

"Anti-plagiarist" is the software used to perform comparison test for the Plagiarism checker system developed and it makes use of offline database. Anti-plagiarist is a system that checks for similarity between two comparable documents or files.

Anti-plagiarist system uses 23.94secs to check for plagiarism of 100 words while our developed Plagiarism checker system uses 18seconds to check for plagiarism of 100 words.

Plagiarism Checker uses KNN Algorithm which will help the system to select words from the neighboring data from the offline database and the comparison test will take place. The KNN algorithm helped to increase the efficiency of the result provided by the developed system.

Plagiarism Checker uses Tri-gram algorithm which helps to select a sequence of three words, this will help to increase the accuracy of the plagiarism system and also to aid the KNN algorithm used in the system.

4.5.1 Plagiarism Tools and Features

Plagiarism Tool	Features
TRI-KNN Plagiarism Checker	• Uses 18secs for execution of 100 words, and 178secs for 1000 words • Uses KNN algorithm to increase the efficiency and the efficiency is rated 80% • Uses Trigram algorithm to increase its accuracy and the efficiency is rated 90.5% • It uses an offline database

Anti-Plagiarist	• Uses 23.94secs for execution of 100 words and 220secs for 1000 words. • It uses an offline database • It checks for similarities between two comparable documents. • Its Accuracy is rated 50% and the efficiency is rated 75 % as it uses KNN algorithm

Table 4.1 Plagiarism Tool and Features

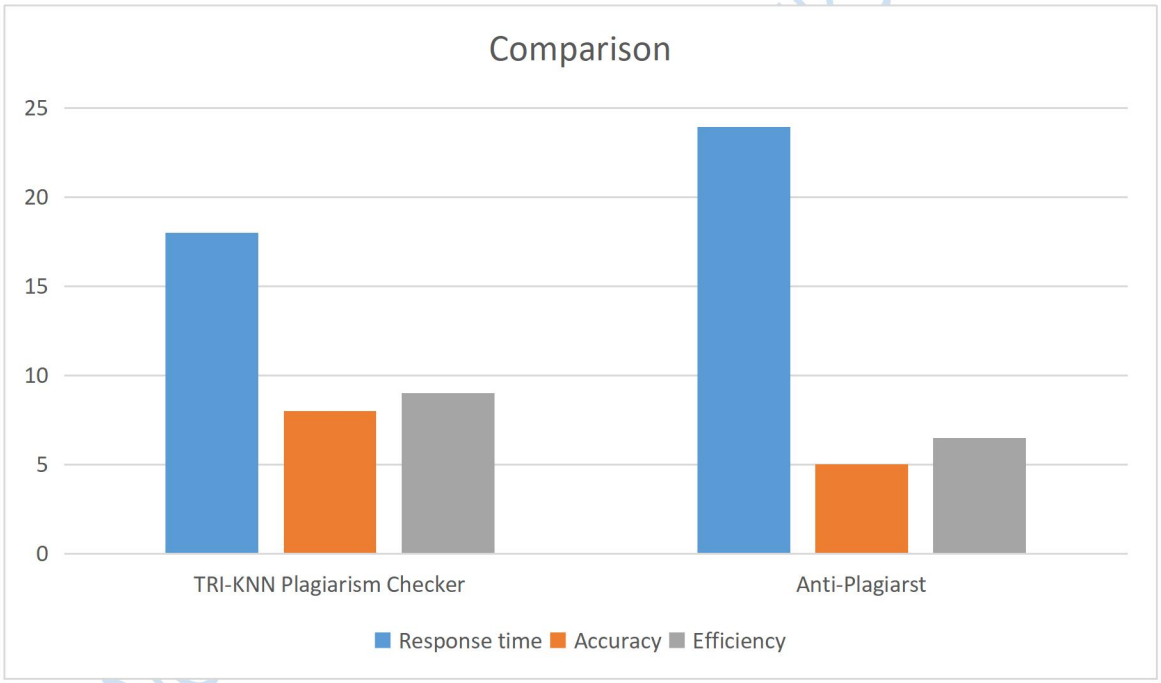


Figure 4.5 Comparison Chart (Researcher, Fatoki I. 2022)

The system makes use of offline database, and is freely available, each sentence in the journal uploaded is given as a query and the results is retrieved if the similarity within the document is detected. With respect to KNN algorithm and trigram algorithm used in the

development of the system, 90% uniqueness of the system is shown and 10% of the system deficiency is shown. This shows the accuracy of the system.

The scalability is rated at 90% as the database is populated with thousands of journals and the system still performs excellently well. The performance metrics used to measure the efficiency of this system is accuracy and scalability.

Chapter Five

Conclusion

5.1 Summary

Journal is an academic research work that is submitted by students in an academic environment. The academic journal plagiarism checker system is a system that enables students to upload documents. The system provides a platform for administrator to clear the database at any time. The academic journal plagiarism checker is developed in other to facilitate its users to sharing resources. The system is of great benefit to the department, both lecturers and students.

5.1.1 Conclusion

In conclusion, this thesis is developed with the aim of helping students have access to plagiarism checker system. The system is an effective, efficient, simple and fast tool for plagiarism detection on text based electronic journals. The system will help to prevent repetition of research work that has already been done and it will make students develop new research problems based on the limitation of the existing research work that they can work on during their research work. This thesis used a methodology that uses data mining techniques. The efficiency of the process is improved. Preprocessing and clustering techniques is used to reduce process overhead. The plagiarism checker system is implemented using KNN algorithm and trigram algorithm.

5.2 Contribution to Knowledge

In this research explore the problem of text plagiarism and the possibility of its detection by the use of computer algorithms. In view of this, techniques and approaches to detect digital automated plagiarism detection have been introduced. One of the first problems the systems face is the collection of possible sources to compare the suspected documents with. This represent an entire problem in itself, and it is common that the ideal and real sources are not always available, limiting the potential of algorithms that compute similarity document-to-document. It was recently tested and studies utilizing different writing style markers are being introduced. In this research study a self-based information algorithm, whose basic idea is the use of a function to quantify the writing style based solely on the use of words

5.3 Suggestions for Further Studies

The study aimed at disclosing researcher's perceptions and their competencies in identification of plagiarism, which is to check for authenticity of the research done; however, researcher's actual behavior and how well they understand their research of plagiarism in their submitted research might also be explored in the forthcoming studies. Another research direction might be a comparative exploration of supervisors' and students' perceptions in respect to the research titles. A study that compares the views and student's research work majoring in different departments in Lead City University. Lastly, a plagiarism checker system that would check for plagiarism in larger documents within the offline database to reach more generalizable findings can be developed.

Bibliography

Journals

Aal M, Rizk M., El-Shorbagy M. A., & Mousa A. A., "*Hybrid Genetic Algorithm with K-Means for Clustering Problems,"* **Open J. Optim**., vol. 05, no. 02, 2016 pp. 71–83.

Adiningrum T. S. *Reviewing Plagiarism*: **An Input for Indonesian Higher Education. Journal of Academic Ethics***, 13*(1), 2015 pp 107-120

Ali H. H. & Kadhum L. E, "*K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition,*" **Int. J. Sci. Res.,** vol. 6, no. 8, 2017 pp. 1577–1584.

Ali, A., & Taqa, A. Y. *Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches.* **Journal of Education and Science,** 31(1), 2022, pp. 8-25.

Amidwar, S. *Plagiarism Detection Using Supervised Machine Learning Algorithm.* **Journal of Emerging Technologies and Innovative Research (JETIR),** 4(06) 2017.

Awale, N., Pandey, M., Dulal, A. & Timsina, B. *Plagiarism Detection in Programming Assignments Using Machine Learning.* **J Artif Intell Capsule Netw,** 2(3), 2020, pp. 177-184.

Bhardwaj A., Arya A., Vyas T., Shivani D. & Anuja N. *A Survey on Plagiarism Detection.* **Institute of Technology, Nirma University, 382481-Ahmedabad, India. Advances in Computational Sciences and Technology** Volume 10, Number 8, 2017, pp 2351-2365 ISSN 0973-6107

Duan M., Li K., Liao X, & Li K., **"***A Parallel Multi classification Algorithm for Big Data Using an Extreme Learning Machine,"* **IEEE Trans. Neural Networks Learn. Syst.,** vol. 29, no. 6, 2017 pp. 2337–2351.

Durga B. D. & Venu G., Rao. K. *Detecting the Plagiarism for Text Documents on the World Wide Web.* **International Journal of Social Relevance & Concern** Volume 2 Issue 10, October 2014 ISSN No: 2347-9698

Eman S., Al S. & Hadeel Q. G. *Plagiarism Detection Using Semantic Analysis.* **Indian Journal of Science Tech. Advances in Computational Sciences and Technology** Volume 10, Number 8, 2016, pp 2351-2365 ISSN 0973-6107

Eppa, A., & Murali, A. H. *Machine Learning Techniques for Multisource Plagiarism Detection.* **In 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions IEEE (CSITSS),** December 2021, pp. 1-5.

Faizan M., Zuhairi M., Ismail S., Sultan S., ***"****Applications of Clustering Techniques in Data Mining: A Comparative Study****"*** **(IJACSA) International Journal of Advanced Computer Science and Applications,** Vol. 11, No. 12, 2020 pp 146

Farhan, N. S. & Abdulmunem, M. E. *Image Plagiarism System for Forgery Detection in Maps Design*. **In 2019 2nd Scientific Conference of Computer Sciences (SCCS)**, March 2019, pp. 51-56, IEEE.

Gupta A., Himansshu S. & Anas A. *A Comparative Analysis of K-Means and Hierarchical Clustering*, **EPRA International Journal of Multidisciplinary research (IJMR),** Vol 7, Issue 8, August 2021, ISSN 2455-3662

Hossain M. Z., Akhtar M. N., Ahmad R. B., & Rahman M., *"A Dynamic K-Means Clustering for Data Mining,"* **Indones. J. Electr. Eng. Comput. Sci.,** vol. 13, no. 2, 2019 pp. 521–526.

Hrkút, P., Ďuračík, M., Mikušová, M., Callejas-Cuervo, M. & Zukowska, J. *Increasing K-means clustering algorithm effectivity for using in source code plagiarism detection.* **In International Conference on Smart Technologies, Systems and Applications, Springer, Cham.** December 2019, pp. 120-131.

Ilham, A. A., Bustamin, A., Aswad, I. & Armin, F. *Implementation of clustering and similarity analysis for detecting content similarity in student final projects*. **In IOP conference series: Materials science and engineering Vol. 875, No. 1, p. 012039. IOP Publishing** 2020, June.

Jain P. K. & Pamula R., "*Two-Step Anomaly Detection Approach*." **International journal of engineering education**, 2018

Jazyah, Y. *Open Learning, the Issue of Plagiarism-Efficient Algorithm*. **International Journal of Computers,** 3, 2018.

Karthikeyan B., George D. J., Manikandan G., & Thomas T., *"A Comparative Study on K-Means Clustering and Agglomerative Hierarchical Clustering,"* **Int. J. Emerg. Trends Eng. Res.,** vol. 8, no. 5, 2020. pp. 1600–1604

Khanna, S., Rakesh, N., & Chaturvedi, K. NOperations on Cloud Data, *Classification and Data Redundancy.* **In Advances in Computer and Computational Sciences Springer, Singapore,** 2018, pp. 169-179.

Mahwish A., Muhammad U. & Muhammad W.A., *Plagiarism Detection Process using Data Mining Techniques* **Muhammad Waleed Ashraf Riphah International University Faisalabad, Pakistan. iJES –** Vol. 5, No. 4, 2017

Mike P., Ulas B. G. & Jasper R. *Reducing Plagiarism Through Academic Misconduct Education* **Perkins et al. International Journal for Educational Integrity** 2020 16:3

Muhammad F., Megat F. Z., Shahrinaz I. & Sara S. *Applications of Clustering Techniques in Data Mining: A Comparative Study.* **International Journal of Advanced Computer Science and Applications (IJACSA),** vol. 11, No 12, 2020

Nazir, A., Mir, R. N. & Qureshi, S. *Idea plagiarism detection with recurrent neural networks and vector space model.* **International Journal of Intelligent Computing and Cybernetics**, 2021.

Novak, M., Joy, M. S. & Mirza, O. M. *Improved plagiarism detection with collaboration network visualization based on source-code similarity.* **In 2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR),** May 2021, pp. 1-6. IEEE.

Novikov A., "*PyClustering: Data Mining Library,"* **J. Open Source Software,** vol. 4, no. 36, 2019 p. 1230.

Rashid S., Ahmed A., Barazanchi I. A, & Jaaz Z. A., *"Clustering Algorithms Subjected to K-Mean and Gaussian Mixture Model on Multidimensional Data Set,"* **Period. Eng. Nat. Sci.,** vol. 7, no. 2, 2019 pp. 448–457.

Reddy M., Vivekananda V., M, &. Satish R. U. V. N, *"Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,"* **Int. J. Comput. Sci. Trends Technol.,** vol. 5, no. Sep-Oct, 2017 pp. 5–11.

Rets I. & Ali I. 'Eliciting ELT Students' *Understanding of Plagiarism in Academic Writing* **EJAL Eurasian Journal of Applied Linguistics** 4(2) 2018, 193–211

Rustadi, R., Darmawan, I., Nandi Haerudin, N. H., Agus Setiawan, A. & Suharno, S. 1. *Plagiarism Checker JGSE, Groundwater exploration using integrated geophysics method.* **Journal of Groundwater Science and Engineering,** 10(1), 2022, 10-8.

Saeed, A. A. M. & Taqa, A. Y. *A proposed approach for plagiarism detection in Article documents*. **Sinkron: jurnal dan penelitian teknik informatika,** 7(2), 2022, 568-578.

Saxena A., Mukesh P., Akshansh G., Neha B., Om P.,Aruna T, Meng J., Weiping D. & Chin T. "*A Review of Clustering Techniques and Developments*," **Neurocomputing journal**, vol. 267, Issue C, 2017 pp. 664–681.

Seethongchuen, T. & Ratanaworabhan, P. *Improving Plagiarism Checker Throughput with Apache Storm.* **In 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology IEEE (ECTI-CON),** July 2019, pp. 393-396

Sharma, M. K. & Sharma, S. *Plagiarism Checker and Classification of Files on Cloud Using Smart Cloud.* **In Proceedings on International Conference on Emerg** Vol. 2, October 2018, pp. 81-88.

Tomáš F., Norman M. & Bela G. *Academic Plagiarism Detection: A Systematic Literature Review* **ACM Computing Surveys** Volume 52 Issue 6 January 2020 Article No.: 112 pp 1–42

Vani K & Deepa G. *Study on Extrinsic Text Plagiarism Detection Techniques and Tools* **Journal of Engineering Science and Technology** Review 9 (5) 2016 pp 9 – 23

Widaningrum, I., Mustikasari, D., Arifin, R., & Sugianti, S. *A Review of Detection Plagiarism in Indonesian Language*. **International Journal Artificial Intelligent and Informatics,** 1(2), 2018, pp. 65-75.

Xu D. & Tian Y., "*A Comprehensive Survey of Clustering Algorithms,"* **Ann. Data Science journal,** vol. 2, no. 2, 2015, pp. 165–193.

Yuan C. & Yang H., "*Research on K-Value Selection Method of K-Means Clustering Algorithm*," **J**, vol. 2, no. 2, 2019, pp. 226–235.

Zerhari B., Lahcen A. A., & Mouline S., "*Big Data Clustering: Algorithms and Challenges,"* **Proc. Int. Conf. Bihree Charact.** Call. 3Vs Volume, Veloc. Var., no. May, 2015, pp. 1–7.

Zhang T. & Ma F., *"Improved Rough K-Means Clustering Algorithm Based on Weighted Distance Measure With Gaussian Function*," **Int. J. Comput. Math.,** vol. 94, no. 4, 2017, pp. 663–675.

Theses/Dissertations

Abhay N. P. & Chinmay N. B. *Plagiarism Detection System* **Department of Computer Science and Engineering Datta Meghe Institute of Engineering Technology and Research, Wardha, Maharashtra, International Journal of Innovations in Engineering and Technology vol 2(1)** India February 2013, 278 ISSN 2319-1058

Al-suhaiqi11, M., Hazaa22, M. A. & Albared33, M. *Arabic English Cross-Lingual Plagiarism Detection Based on Key Phrases Extraction, 2 Monolingual and Machine Learning* Approach 3, 2018.

Chowdhury H. A. & Bhattacharyya, D. K. *Plagiarism: Taxonomy, tools and detection techniques.* **arXiv preprint arXiv:1801.06323**, 2018.

Gawali, S., Thakore, D. S., Joshi, S. D. & Shinde, V. S. *A Plagiarism Detector Based on MAS Scalable Framework for Research Effort Evaluation by Unsupervised Machine Learning–Hybrid Plagiarism Model. Applied Machine Learning for Smart Data Analysis*, 41, 2019.

Habibah, M., Wicaksono, A., Aji, P. & Puji, M. *The Use Clustering Technique for Teaching Writing to Senior High School*, **Doctoral dissertation, Universitas Nusantara PGRI Kediri**, 2020.

Irsyadi, Y. *Plagiarism and Similarity Checker: Analysis of a cans waste classification system based on the CMYK color model using different metric distances on the k-means method*. 2022.

Lior R. O. M., "*Clustering methods*," Adv. Inf. Knowl. Process no. 9781447167341, 2015, pp. 131–167.

Malandrino, D., De Prisco, R., Ianulardo, M. & Zaccagnino, R. *An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering. Data Mining and Knowledge Discovery*, 2022, 1-34.

Mariyudi, M. *Hasil Plagiarism Checker: Jurnal Internasional Success Factors of SMEs: The Case of Indonesia*, 2019.

Muna A., Rahat I., Vasile P., Saad A. & Victor C., "*An Integrated Approach for Intrinsic Plagiarism Detection"* in **, Suzhou, China: School of Computing, Electronics and Maths, Coventry University, United Kingdom, International Business School Suzhou, Xi'an Jiaotong Liverpool University**, 2019.

Nguyen T. H. T., Dinh D. T., Sriboonchitta S., & Huynh V. N., *"A Method for K-Means-Like Clustering of Categorical Data,"* J. Ambient Intell. Humaniz. Comput., no. Berkhin, 2019.

Rajiani, I. *Plagiarism checker: The effectiveness of acupressure intervention and birth delivery standing position to decrease the intensity of labor pain*, 2018.

Rajiani, I. *Plagiarism checker: The prospective consumers of the Indonesian green aviation initiative for sustainable development in air transportation,* 2018.

Ullah, F., Wang, J., Farhan, M., Habib, M. & Khalid, S. *Software Plagiarism Detection in Multiprogramming Languages Using Machine Learning Approach.* Concurrency and Computation: Practice and Experience, 33(4), e5000, 2021.

Zou D., Long W.-J. & Ling Z., *A Cluster-Based Plagiarism Detection Method*, in: **Notebook Papers of CLEF 2010 LABs and Workshops**, September, 2010 pp. 22-23.

Conference Proceedings

Ajin V. W. & Kumar L. D., "*Big Data and Clustering Algorithms*," in **International Conference on Research Advances in Integrated Navigation Systems**, RAINS 2016.

Alzahrani S. M. & Salim N., *The Use Of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents*, in: **Applications of Digital Information and Web Technologies**.
ICADIWT'12. **Second International Conference on the, IEEE,** pp. 539-544. 2, no. 3, 2012, pp. 1379–1384.

Chavan, H., Taufik, M., Kadave, R., & Chandra, N. *Plagiarism Detector Using Machine Learning.* **International Journal of Research in Engineering, Science and Management,** 4(4), 2021, pp. 152-154.

Dwivedi, J., & Tiwary, A. *Plagiarism Detection on Big-Data Using Modified Map-Reduced Based SCAM Algorithm*. **In 2017 International Conference on Innovative Mechanisms for Industry Applications IEEE (ICIMIA),** February, 2017, pp. 608-610.

Gipp B. & Beel J., *Citation Based Plagiarism Detection: A New Approach to Identify Plagiarized Work Language Independently*, in: Proceedings of the 21st **ACM conference on Hypertext and hypermedia, ACM**, 2010, pp. 273--274.

Gipp B. & Meuschke N., *Citation Pattern Matching Algorithms for Citation Based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence*, in: Proceedings of the 11th **ACM symposium on Document engineering**, ACM, 2011, pp. 249--258.

Moncada, I. R. & Rodriguez, V. F. *Natural Language Processing for Plagiarism Detection–A Survey*. **In INTED2022 Proceedings IATED,** 2022, pp. 9624-9629.

Puflović, D., Gligorijević, M. F. & Stoimenov, L. CSPlag: *A Source Code Plagiarism Detection Using Syntax Trees and Intermediate Language*. **In Proceedings of the 52nd International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST 2017),** 2017, pp. 102-105.

Rajashekar N., Geetha M., Samhitha M., Sandeep S, & Roshini G *"Plagiarism Detection through Data Mining Techniques"* **International Conference on Recent Trends in Computing (ICRTCE-2021) Journal of Physics**: Conference Series **1979** 2021

Rajashekar N., Geetha M., Yadav, Samhitha M. S., Sandeep K. & Roshini G. *Plagiarism Detection through Data Mining Techniques* Dept. of CSE, Institute of Aeronautical

Engineering12345, Dundigal – **Hyderabad International Conference on Recent Trends in Computing** ICRTCE-2021

Ratna, A. A. P., Ekadiyanto, F. A., Ibrahim, I., Husna, D. & Rahimullah, F. *Investigating Parallelization of Cross-language Plagiarism Detection System Using the Winnowing Algorithm in Cloud Based Implementation.* **In 2019 IEEE 10th International Conference on Awareness Science and Technology IEEE (iCAST),** October 2019, pp. 1-7.

Singh, M., & Gupta, V. *Review of Extrinsic Plagiarism Detection Techniques and Their Efficiency Comparison.* **In International Conference on Advanced Network Technologies and Intelligent Computing, Springer, Cham**, December 2021, pp. 609-624.

Su Z., Ahn B.-R., Eom K.-Y., Kang M.-K. & Kim J.-P., *Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm*, in: **Innovative Computing Information and Control, 2008. ICICIC' 08. 3rd International Conference on, IEEE**, 2008, pp. 569-569.

Telhadi M. & Al-Tobi A., *Duplicate Detection in Documents and Webpages Using Improved Longest Common Subsequence and Documents Syntactical Structures*, in: **Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on, IEEE**, 2009, pp. 679--684.

Appendix

```
 <div class="row-fluid">
                    <!-- block -->
 <div class="block">
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Register New Student</i></div>
</div>
<div class="block-content collapse in">
                        <div class="span12">
                                                                <!--------------------form--------------------->
                                                                    <form method="post">
                <div class="control-group">
                                <div class="controls">
                                  <p>
                                    <input class="input focused" name="RegNo" id="focusedInput" type="text" placeholder = " Matric No  E.g. LCU/2016/23456">
                                  </p>
                                </div>
                                </div>
                                                <div class="control-group">
                            <p> <div class="controls">
                              <p>
                                <input class="input focused" name="FirstName" id="focusedInput" type="text" placeholder = "FirstName E.g. John" required>
                              </p>
                            </div>
```

```
                    </div>

                                                        <div class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="LastName"
id="focusedInput" type="text" placeholder = "Last Name E.g. Kamau" required>
                          </p>
                        </div>
                        </div> </p>

                                                                <div
class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="Othername"
id="focusedInput" type="text" placeholder = "Other Name E.g. Muthare">
                          </p>
</div>
                      </div>

                                                        <div class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="DateOfBirth"
id="focusedInput" type="date" placeholder = " yearOfBirth E.g. 1999" >
                          </p>
                        </div>
                        </div>
```

```
                                                    <div class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="LastName"
id="focusedInput" type="text" placeholder = "Last Name E.g. Kamau" required>
                          </p>
                        </div>
                        </div> </p>


                                                                    <div
class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="Othername"
id="focusedInput" type="text" placeholder = "Other Name E.g. Muthare">
                          </p>
</div>
                      </div>
                            <input class="input focused" name="Mobile" id="focusedInput"
type="number" placeholder = "contact e.g. 07100 000 000" >
                          </p>
                        </div>
                        </div>


                                                    <div class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="PostAddress"
id="focusedInput" type="text" placeholder = "PostAddress E.g. 59 Kiambuu" >
                          </p>
<?php include('session.php'); ?>
```

```
<body>

        <?php include('navbar.php'); ?>

    <div class="container-fluid">

      <div class="row-fluid">

                    <?php include('activity_log_slidebar.php'); ?>

        <div class="span9" id="content">

          <div class="row-fluid">

            <!-- block -->

                    <div class="empty">

          <div class="alert alert-success alert-dismissable">

            <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>

             <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checbox
if you want to delete?

          </div>

        </div>



                    <?php

            $count_log=mysql_query("select * from activity_log");
```

```
$count = mysql_num_rows($count_log);

?>

<div id="block_bg" class="block">

<div class="navbar navbar-inner block-header">

<div class="muted pull-left"><i class="icon-user"></i> System User Activity Log</div>

<div class="muted pull-right">

Number of System user Activity Log: <span class="badge badge-info"><?php  echo $count; ?></span>

</div>

</div>

<div class="block-content collapse in">

<div class="span12">

<form action="delete_log.php" method="post">

<table cellpadding="0" cellspacing="0" border="0" class="table" id="example">

<a data-placement="right" title="Click to Delete checked item" data-toggle="modal" href="#delete_log" id="delete"  class="btn btn-danger" name=""><i class="icon-trash icon-large"> Delete</i></a>
```

<script type="text/javascript">

$(document).ready(function(){

$('#delete').tooltip('show');

$('#delete').tooltip('hide');

});

</script>

<?php include('modal_delete.php'); ?>

<thead>

<tr>

<th>Check</th>

<th>Date</th>

<th>System User</th>

```
<th>Action</th>

</tr>

</thead>

<tbody>

<?php

</div>

</div>

<div
class="control-group">

<div class="controls">

<input class="input focused" name="YearAdmitted" id="focusedInput"
type="text" placeholder = " YearAdmitted E.g. 1999" required>

</div>

</div>

<div class="control-group">

<div class="controls">

<p>

<input class="input focused" name="YearOut"
id="focusedInput" type="text" placeholder = "YearOut  E.g. 1999" required>

</p>

</div>

</div>
```

```
                                                                                <div
class="control-group">
                            <div class="controls">
 <button name="save" class="btn btn-info" id="save" data-placement="right" title="Click to
Save"><i class="icon-plus-sign icon-large"> Save</i></button>

        <script type="text/javascript">
                                     $(document).ready(function(){
                                     $('#save').tooltip('show');
                                     $('#save').tooltip('hide');
                                     });
                                     </script>
                            </div>
                           </div>
                      </form>

                                                                 </div>
                    </div>
                  </div>
                  <!-- /block -->
                </div>
<?php
if (isset($_POST['save'])){
$RegNo = $_POST['RegNo'];
$FirstName = $_POST['FirstName'];
$LastName = $_POST['LastName'];
$Othername = $_POST['Othername'];
$DateOfBirth = $_POST['DateOfBirth'];
$Mobile = $_POST['Mobile'];
$postalCode= $_POST['postalCode'];
```

```
$YearAdmitted= $_POST['YearAdmitted'];

$YearOut = $_POST['YearOut'];

$PostAddress= $_POST['PostAddress'];



$query = @mysql_query("select * from student where  RegNO = '$RegNO'  ")or
die(mysql_error());

$count = mysql_num_rows($query);


if ($count > 0){ ?>

<script>

alert('Matric No Already Registered');

</script>

<?php

}else{

mysql_query("insert into student
(RegNo,FirstName,LastName,Othername,DateOfBirth,YearAdmitted,YearOut,Mobile,PAdd
ress)

values('$RegNo','$FirstName','$LastName','$Othername','$DateOfBirth','$YearAdmitted','$Y
earOut','$Mobile','$PostAddress')")or die(mysql_error());


mysql_query("insert into activity_log (date,username,action)
values(NOW(),'$admin_username','Added Student $RegNo')")or die(mysql_error());

?>

<script>

window.location = "add_student.php";

$.jGrowl("Student Successfully added", { header: 'Staff add' });

</script>

<?php

}

}
```

```
?>

 <div class="row-fluid">
                    <!-- block -->
 <div class="block">
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Insert New
Result</i></div>
</div>
<div class="block-content collapse in">
                            <div class="span12">

                                                                                <!--------------------form-----------
-------->
                                                                 <form class="form-inline"
method="POST" enctype="multipart/form-data">

                                                                        <input class="form-control"
type="file" name="upload"/>

                                                                                                      <div
class="control-group">
                                    <div class="controls">
 <button name="save" class="btn btn-info" id="save" data-placement="right" title="Click to
Save"><i class="icon-plus-sign icon-large"> Upload Project </i></button>

        <script type="text/javascript">
                                            $(document).ready(function(){
                                            $('#save').tooltip('show');
                                            $('#save').tooltip('hide');
                                            });
```

```
                                    </script>
                                </div>
                            </div>
                        </form>


                                                                    </div>
                    </div>
                </div>
                <!-- /block -->
            </div>
<?php
if(ISSET($_POST['save'])){
   if($_FILES['upload']['name'] != "") {
     $file = $_FILES['upload'];
     $file_name = $file['name'];
     $file_temp = $file['tmp_name'];
     $name = explode('.', $file_name);
     $path = "files/".$file_name;
$query = @mysql_query("select * from file where name = '$file_name'")or
die(mysql_error());
$count = mysql_num_rows($query);


if (file_exists($path)) { ?>
<script>
alert('Project Already Exist');
$count = 0;
</script>
<?php
}else{
```

```
$score_query = mysql_query("INSERT INTO `file` VALUES('', '$name[0]', '$path')") or
die(mysqli_error());


    move_uploaded_file($file_temp, $path);


?>
<script>
window.location = "add_project.php";
$.jGrowl("Project Successfully added", { header: 'Project added' });
</script>
<?php
  <div class="row-fluid">
                <!-- block -->
 <div class="block">
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Insert New
Result</i></div>
</div>
<div class="block-content collapse in">
                    <div class="span12">
}
}
}
?>
<?php include('header_dashboard.php'); ?>

<?php include('session.php'); ?>

   <body>

            <?php include('navbar_staff.php'); ?>
```

```
<div class="container-fluid">
    <div class="row-fluid">
                        <?php include('staff_sidebar.php'); ?>
                        <div class="span3" id="adduser">
                        <?php include('addproject.php'); ?>

                        </div>
        <div class="span6" id="">
            <div class="row-fluid">
              <!-- block -->

                        <div class="empty">
            <div class="alert alert-success alert-dismissable">
                <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>
                 <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checkbox
if you want to delete?
            </div>
        </div>
```

```
                                <?php
$score_query = mysql_query("select * from file")or die(mysql_error());
                $count = mysql_num_rows($score_query);
            ?>
                <div id="block_bg" class="block">
                  <div class="navbar navbar-inner block-header">
                    <div class="muted pull-left"><i class="icon-user"></i>  Project
Upload(s) List</div>
                                                        <div class="muted pull-right">
                                                        Number of Project File Uploaded:
<span class="badge badge-info"><?php  echo $count; ?></span>
                                                    </div>
                  </div>
                  <div class="block-content collapse in">
                    <div class="span12">
                                                        <table cellpadding="0"
cellspacing="0" border="0" class="table" id="example">
                                                        <?php
```

```
                    $score_query = mysql_query("select * from file")or
die(mysql_error());
                while($row = mysql_fetch_array($score_query)){
                    extract($row);
                ?>
                <tr>
                        <?php
                            $name = explode('/', $row['file']);
                        ?>
                        <td><?php echo $row['name']?></td>
                        <td><a href="download.php?file=<?php echo
$name[1]?>" class="btn btn-primary"><span class="glyphicon glyphicon-
download"></span> Download</a></td>
                        </tr>
                <?php
                    }
                ?>
                                    </tbody>
                                </table>
```

```
                                                                    </form>

                        </div>

                      </div>

                    </div>

                    <!-- /block -->

                  </div>

                </div>

              </div>

                    <?php include('footer.php'); ?>

          </div>

                    <?php include('script.php'); ?>

    </body>

</html>

<!DOCTYPE html>

<html lang="en" dir="ltr">

  <head>

    <meta charset="utf-8">

    <title>INSERT PAGE</title>
```

```
  <link rel="stylesheet" href="css/insert.css" />
 </head>
 <body>


 <?php
   $servername = "localhost";
   $username = "root";
   $password = "Sp09iner001";
   $dbname = "db_course";


   // Create connection
   $conn = new mysqli($servername, $username, $password, $dbname);
   // Check connection
   if ($conn->connect_error) {
      die("Connection failed: " . $conn->connect_error);
   }
#    else {
#      echo "Connected Successfully";
```

```
#   }
    $CourseCode = mysqli_real_escape_string($conn, $_POST['CourseCode']);
    $CourseTitle = mysqli_real_escape_string($conn, $_POST['CourseTitle']);
    $sql = "INSERT INTO course (CourseCode, CourseTitle) VALUES ('$CourseCode',
'$CourseTitle')";
    if ($conn->query($sql) === TRUE) {
      echo "New course added successfully";
    } else {
    echo "Error: " . $sql . "<br>" . $conn->error;
    }

  ?>
<center>
  <div class="maindiv">
  <!--HTML form -->
    <div class="form_div">
      <div class="title"><h2>Enter the Course details below.</h2></div>
```

```
    <form action="insert.php" method="post">    <!-- method can be set POST for hiding
values in URL-->
        <h2>Course Details</h2>
        <label>Course Title:</label>
        <br />
        <input class="input" type="text" name="CourseTitle" value="" required = "required"/>
        <br />

        <label>Course Code:</label>
        <br />
        <input class="input" type="text" name="CourseCode" value="" required = "required"/>
        <br />

        <input class="submit" type="submit" name="submit" value="Insert Course" />
    </form>
  </div>
  </div>
</center>
```

```
 </body>
</html>
<?php include('header_dashboard.php'); ?>
<?php include('session.php'); ?>
   <body>
                <?php include('navbar_staff.php'); ?>
      <div class="container-fluid">
         <div class="row-fluid">
                                <?php include('staff_sidebar.php'); ?>

                                <div class="span9" id="content">
            <div class="row-fluid">

                                        <script type="text/javascript">
                        $(document).ready(function(){
                        $('#add').tooltip('show');
                        $('#add').tooltip('hide');
                        });
```

```
 <div class="row-fluid">
                <!-- block -->
 <div class="block">
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Register New Student</i></div>
</div>
<div class="block-content collapse in">
                    <div class="span12">
                                                        <!--------------------form--------------------->
                                                        <form method="post">
            <div class="control-group">
                        <div class="controls">
                          <p>
                            <input class="input focused" name="RegNo" id="focusedInput" type="text" placeholder = " Matric No  E.g. LCU/2016/23456">
                          </p>
                      </div>
                      </div>
                                <div class="control-group">
                    <p> <div class="controls">
                      <p>
                        <input class="input focused" name="FirstName" id="focusedInput" type="text" placeholder = "FirstName E.g. John" required>
                      </p>
                    </div>
                    </div>

                                                        <div class="control-group">
                        <div class="controls">
```

```
                    <p>
                      <input class="input focused" name="LastName"
id="focusedInput" type="text" placeholder = "Last Name E.g. Kamau" required>
                    </p>
                  </div>
                  </div> </p>


                                                                    <div
class="control-group">
                    <div class="controls">
                      <p>
                        <input class="input focused" name="Othername"
id="focusedInput" type="text" placeholder = "Other Name E.g. Muthare">
                      </p>
</div>
                  </div>


                                                          <div class="control-group">
                    <div class="controls">
                      <p>
                        <input class="input focused" name="DateOfBirth"
id="focusedInput" type="date" placeholder = " yearOfBirth E.g. 1999" >
                      </p>
                  </div>
                  </div>


                                                          <div class="control-group">
                    <div class="controls">
                      <p>
```

```
                                <input class="input focused" name="LastName"
id="focusedInput" type="text" placeholder = "Last Name E.g. Kamau" required>
                                </p>
                            </div>
                            </div> </p>


                                                                    <div
class="control-group">
                                <div class="controls">
                                 <p>
                                  <input class="input focused" name="Othername"
id="focusedInput" type="text" placeholder = "Other Name E.g. Muthare">
                                 </p>
</div>
                                </div>
                                  <input class="input focused" name="Mobile" id="focusedInput"
type="number" placeholder = "contact e.g. 07100 000 000" >
                                 </p>
                            </div>
                            </div>


                                                                    <div class="control-group">
                                <div class="controls">
                                 <p>
                                  <input class="input focused" name="PostAddress"
id="focusedInput" type="text" placeholder = "PostAddress E.g. 59 Kiambuu" >
                                 </p>
<?php include('session.php'); ?>

   <body>

                <?php include('navbar.php'); ?>
```

```
<div class="container-fluid">
    <div class="row-fluid">
                        <?php include('activity_log_slidebar.php'); ?>
        <div class="span9" id="content">
            <div class="row-fluid">
              <!-- block -->
                        <div class="empty">
            <div class="alert alert-success alert-dismissable">
              <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>
               <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checbox
if you want to delete?
            </div>
        </div>

                        <?php
            $count_log=mysql_query("select * from activity_log");
            $count = mysql_num_rows($count_log);
        ?>
```

```
<div id="block_bg" class="block">
  <div class="navbar navbar-inner block-header">
    <div class="muted pull-left"><i class="icon-user"></i> System User Activity Log</div>
                                                    <div class="muted pull-right">
                                                    Number of System user Activity Log: <span class="badge badge-info"><?php  echo $count; ?></span>
                                                    </div>
  </div>
  <div class="block-content collapse in">
    <div class="span12">
                                                    <form action="delete_log.php" method="post">
                                                        <table cellpadding="0" cellspacing="0" border="0" class="table" id="example">
                                                    <a data-placement="right" title="Click to Delete checked item" data-toggle="modal" href="#delete_log" id="delete"  class="btn btn-danger" name=""><i class="icon-trash icon-large"> Delete</i></a>
                                                        <script type="text/javascript">
```

```
$(document).ready(function(){

$('#delete').tooltip('show');
                                                    $('#delete').tooltip('hide');
                                                    });
                                                    </script>
                                                    <?php
include('modal_delete.php'); ?>
                                                        <thead>
                                                            <tr>
                            <th>Check</th>
        <th>Date</th>
        <th>System User</th>
        <th>Action</th>
        </tr>
```

```
                                                                    </thead>

                                                                    <tbody>

                                    <?php

                                </div>
                                </div>


                                                                    <div
class="control-group">
                                        <div class="controls">
                                <input class="input focused" name="YearAdmitted" id="focusedInput"
type="text" placeholder = " YearAdmitted E.g. 1999" required>
                                        </div>
                                    </div>
                                                                        <div class="control-group">
                                        <div class="controls">
                                          <p>
                                            <input class="input focused" name="YearOut"
id="focusedInput" type="text" placeholder = "YearOut  E.g. 1999" required>
                                        </p>
                                    </div>
                                </div>




                                                                    <div
class="control-group">
                                        <div class="controls">
 <button name="save" class="btn btn-info" id="save" data-placement="right" title="Click to
Save"><i class="icon-plus-sign icon-large"> Save</i></button>
```

```
<script type="text/javascript">
                    $(document).ready(function(){
                    $('#save').tooltip('show');
                    $('#save').tooltip('hide');
                    });
                    </script>
                </div>
            </div>
        </form>

                                    </div>
      </div>
    </div>
    <!-- /block -->
  </div>
<?php
if (isset($_POST['save'])){
$RegNo = $_POST['RegNo'];
$FirstName = $_POST['FirstName'];
$LastName = $_POST['LastName'];
$Othername = $_POST['Othername'];
$DateOfBirth = $_POST['DateOfBirth'];
$Mobile = $_POST['Mobile'];
$postalCode= $_POST['postalCode'];
$YearAdmitted= $_POST['YearAdmitted'];
$YearOut = $_POST['YearOut'];
$PostAddress= $_POST['PostAddress'];
```

```
$query = @mysql_query("select * from student where  RegNO = '$RegNO'  ")or
die(mysql_error());

$count = mysql_num_rows($query);


if ($count > 0){ ?>

<script>

alert('Matric No Already Registered');

</script>

<?php

}else{

mysql_query("insert into student
(RegNo,FirstName,LastName,Othername,DateOfBirth,YearAdmitted,YearOut,Mobile,PAdd
ress)

values('$RegNo','$FirstName','$LastName','$Othername','$DateOfBirth','$YearAdmitted','$Y
earOut','$Mobile','$PostAddress')")or die(mysql_error());


mysql_query("insert into activity_log (date,username,action)
values(NOW(),'$admin_username','Added Student $RegNo')")or die(mysql_error());

?>

<script>

window.location = "add_student.php";

$.jGrowl("Student Successfully added", { header: 'Staff add' });

</script>

<?php

}

}
?>


 <div class="row-fluid">

                <!-- block -->

<div class="block">
```

```
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Insert New Result</i></div>
</div>
<div class="block-content collapse in">
                        <div class="span12">

                                                                <!--------------------form----------
-------->
                                                        <form class="form-inline"
method="POST" enctype="multipart/form-data">

                                                                <input class="form-control"
type="file" name="upload"/>

                                                                        <div
class="control-group">
                                <div class="controls">
 <button name="save" class="btn btn-info" id="save" data-placement="right" title="Click to Save"><i class="icon-plus-sign icon-large"> Upload Project </i></button>

        <script type="text/javascript">
                                        $(document).ready(function(){
                                        $('#save').tooltip('show');
                                        $('#save').tooltip('hide');
                                        });
                                        </script>
                                </div>
                               </div>
                        </form>
```

```
                                                    </div>
                    </div>
                  </div>
                  <!-- /block -->
                </div>
<?php
if(ISSET($_POST['save'])){
   if($_FILES['upload']['name'] != "") {
     $file = $_FILES['upload'];
     $file_name = $file['name'];
     $file_temp = $file['tmp_name'];
     $name = explode('.', $file_name);
     $path = "files/".$file_name;
$query = @mysql_query("select * from file where name = '$file_name'")or
die(mysql_error());
$count = mysql_num_rows($query);

if (file_exists($path)) { ?>
<script>
alert('Project Already Exist');
$count = 0;
</script>
<?php
}else{
$score_query = mysql_query("INSERT INTO `file` VALUES('', '$name[0]', '$path')") or
die(mysqli_error());

     move_uploaded_file($file_temp, $path);

?>
```

```
<script>
window.location = "add_project.php";
$.jGrowl("Project Successfully added", { header: 'Project added' });
</script>
<?php
  <div class="row-fluid">
                    <!-- block -->
 <div class="block">
 <div class="navbar navbar-inner block-header">
<div class="muted pull-left"><i class="icon-plus-sign icon-large"> Insert New Result</i></div>
</div>
<div class="block-content collapse in">
                        <div class="span12">
}
}
}
?>
<?php include('header_dashboard.php'); ?>

<?php include('session.php'); ?>

   <body>

              <?php include('navbar_staff.php'); ?>

      <div class="container-fluid">

          <div class="row-fluid">

                                  <?php include('staff_sidebar.php'); ?>
```

```
                    <div class="span3" id="adduser">

                    <?php include('addproject.php'); ?>


                    </div>

            <div class="span6" id="">

                <div class="row-fluid">

                  <!-- block -->


                    <div class="empty">

                <div class="alert alert-success alert-dismissable">

                  <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>

                   <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checkbox
if you want to delete?

                </div>

              </div>

                    <?php

$score_query = mysql_query("select * from file")or die(mysql_error());

                   $count = mysql_num_rows($score_query);
```

```
?>

        <div id="block_bg" class="block">

          <div class="navbar navbar-inner block-header">

            <div class="muted pull-left"><i class="icon-user"></i>  Project
Upload(s) List</div>

                                                    <div class="muted pull-right">

                                                    Number of Project File Uploaded:
<span class="badge badge-info"><?php  echo $count; ?></span>

                                                    </div>

          </div>

          <div class="block-content collapse in">

            <div class="span12">

                                                    <table cellpadding="0"
cellspacing="0" border="0" class="table" id="example">

                                                    <?php

                                    $score_query = mysql_query("select * from file")or
die(mysql_error());

          while($row = mysql_fetch_array($score_query)){

                    extract($row);
```

```
?>
<tr>
<?php
$name = explode('/', $row['file']);
?>
<td><?php echo $row['name']?></td>
<td><a href="download.php?file=<?php echo $name[1]?>" class="btn btn-primary"><span class="glyphicon glyphicon-download"></span> Download</a></td>
</tr>
<?php
}
?>
</tbody>
</table>
</form>
</div>
</div>
</div>
```

```
                <!-- /block -->
              </div>
            </div>
          </div>
                <?php include('footer.php'); ?>
        </div>
                <?php include('script.php'); ?>
    </body>
</html>
<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head>
    <meta charset="utf-8">
    <title>INSERT PAGE</title>
    <link rel="stylesheet" href="css/insert.css" />
  </head>
  <body>
```

```
<?php
   $servername = "localhost";
   $username = "root";
   $password = "Sp09iner001";
   $dbname = "db_course";

   // Create connection
   $conn = new mysqli($servername, $username, $password, $dbname);
   // Check connection
   if ($conn->connect_error) {
      die("Connection failed: " . $conn->connect_error);
   }
#   else {
#      echo "Connected Successfully";
#   }
   $CourseCode = mysqli_real_escape_string($conn, $_POST['CourseCode']);
   $CourseTitle = mysqli_real_escape_string($conn, $_POST['CourseTitle']);
```

```
    $sql = "INSERT INTO course (CourseCode, CourseTitle) VALUES ('$CourseCode',
'$CourseTitle')";
    if ($conn->query($sql) === TRUE) {
      echo "New course added successfully";
    } else {
    echo "Error: " . $sql . "<br>" . $conn->error;
    }


  ?>
<center>
  <div class="maindiv">
  <!--HTML form -->
    <div class="form_div">
      <div class="title"><h2>Enter the Course details below.</h2></div>

     <form action="insert.php" method="post">    <!-- method can be set POST for hiding
values in URL-->
        <h2>Course Details</h2>
        <label>Course Title:</label>
```

```
        <br />
        <input class="input" type="text" name="CourseTitle" value="" required = "required"/>
        <br />

        <label>Course Code:</label>
        <br />
        <input class="input" type="text" name="CourseCode" value="" required = "required"/>
        <br />

        <input class="submit" type="submit" name="submit" value="Insert Course" />
      </form>
    </div>
    </div>
</center>
  </body>
</html>

                    </script>
```

<image src="images/sclogo.png" width="45%" height="45%"/>

<?php

$course_query = mysql_query("select * from business where id = '$session_id'")or die(mysql_error());

$count = mysql_num_rows($course_query);

?>

 My Business s

<h4 id="sc">score

Date:

<?php

$date = new DateTime();

echo $date->format('l, F jS, Y');

```
?></div>

                    </h4>



<div class="container-fluid">

 <div class="row-fluid">

   <div class="empty">

          <div class="pull-right">



              <script type="text/javascript">

               $(document).ready(function(){

               $('#print').tooltip('show');

               $('#print').tooltip('hide');

               });

              </script>

    </div>

  </div>
```

```
    </div>
</div>


<div class="block-content collapse in">
    <div class="span12">
        <form action="" method="post">
        <table cellpadding="0" cellspacing="0" border="0" class="table" id="example">
                <thead>
                    <tr>
                <th>Business Name</th>
                                    <th>Business Category  </th>
                                        <th>Office Address   </th>
            <th>Office Number  </th>
                <th>Subscribe Rank  </th>
```

```
                </tr>

            </thead>

<tbody>

<!----------------------------------Content---------------------------------->

<?php

            $course_query = mysql_query("select * from business where id =
'$session_id'")or die(mysql_error());

            while($row = mysql_fetch_array($course_query)){

            $bname = $row['bname'];

            $cat = $row['cat'];

            $address = $row['address'];

            $office = $row['office'];

        $rank = $row['rank'];



            ?>



            <tr>
```

```
                    <td><?php echo $row['bname']; ?></td>

                        <td><?php echo $row['cat']; ?></td>

                        <td><?php echo $row['address']; ?></td>

                        <td><?php echo $row['office']; ?></td>

                        <td><?php echo $row['rank']; ?></td>

            </tr>

            <?php } ?>


</tbody>

</table>

</form>




</div>

</div>

</div>

</div>

</div>
```

```
</div>

<?php include('footer.php'); ?>

</div>

<?php include('script.php'); ?>

 </body>

</html>

<?php

	include('admin/lib/dbcon.php');

			dbcon();

			session_start();

			$username = $_POST['username'];

			$password = $_POST['password'];


			/*..... admin .....*/

				$query = "SELECT * FROM admin WHERE username='$username'
AND password='$password'";

				$result = mysql_query($query)or die(mysql_error());

				$row = mysql_fetch_array($result);
```

```
$num_row = mysql_num_rows($result);

/*..... Staff .....*/

$query_staff = mysql_query("SELECT * FROM staff WHERE
username='$username' AND password='$password'")or die(mysql_error());

$num_row_staff = mysql_num_rows($query_staff);

$row_staff = mysql_fetch_array($query_staff);

if( $num_row > 0 ) {

$_SESSION['id']=$row['admin_id'];

echo 'true_admin';

mysql_query("insert into user_log
(username,login_date,admin_id)values('$username',NOW(),".$row['admin_id'].")")or
die(mysql_error());

}else if ($num_row_staff > 0){

$_SESSION['staff']=$row_staff['staff_id'];

echo 'true';
```

```
                mysql_query("insert into user_log
(username,login_date,staff_id)values('$username',NOW(),".$row_staff['staff_id'].")")or
die(mysql_error());


                }else{
                        echo 'false';
                }


                ?>
<?php include('header.php'); ?>
<?php include('session.php'); ?>
  <body>
                <?php include('navbar.php'); ?>
    <div class="container-fluid">
      <div class="row-fluid">
                        <?php include('activity_log_slidebar.php'); ?>
        <div class="span9" id="content">
          <div class="row-fluid">
```

```
                <!-- block -->

                                <div class="empty">

            <div class="alert alert-success alert-dismissable">

                <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>

                 <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checbox
if you want to delete?

            </div>

          </div>

<?php include('session.php'); ?>

  <body>

                <?php include('navbar.php'); ?>

     <div class="container-fluid">

       <div class="row-fluid">

                                <?php include('activity_log_slidebar.php'); ?>

         <div class="span9" id="content">

           <div class="row-fluid">

             <!-- block -->

                                <div class="empty">
```

<button type="button" class="close" data-dismiss="alert" aria-hidden="true">×</button>

 Note!: Select the checbox if you want to delete?

<?php

$count_log=mysql_query("select * from activity_log");

$count = mysql_num_rows($count_log);

?>

 System User Activity Log

Number of System user Activity Log: <?php echo $count; ?>

```
                                                    </div>
                    </div>
                    <div class="block-content collapse in">
                        <div class="span12">
                                                    <form action="delete_log.php"
method="post">
                                                        <table cellpadding="0"
cellspacing="0" border="0" class="table" id="example">
                                                    <a data-placement="right" title="Click
to Delete checked item" data-toggle="modal" href="#delete_log" id="delete"  class="btn btn-
danger" name=""><i class="icon-trash icon-large"> Delete</i></a>
                                                        <script
type="text/javascript">

$(document).ready(function(){

$('#delete').tooltip('show');
                                                        $('#delete').tooltip('hide');
                                                        });
                                                    </script>
```

```
                                                    <?php
include('modal_delete.php'); ?>
                                                        <thead>
                                                            <tr>
                                <th>Check</th>
        <th>Date</th>
        <th>System User</th>
        <th>Action</th>
        </tr>
                                                        </thead>
                                                        <tbody>
                                <?php
                                <?php
```

```
$count_log=mysql_query("select * from activity_log");

$count = mysql_num_rows($count_log);

?>

<div id="block_bg" class="block">

<div class="navbar navbar-inner block-header">

<div class="muted pull-left"><i class="icon-user"></i> System User Activity Log</div>

<div class="muted pull-right">

Number of System user Activity Log: <span class="badge badge-info"><?php  echo $count; ?></span>

</div>

</div>

<div class="block-content collapse in">

<div class="span12">

<form action="delete_log.php" method="post">

<table cellpadding="0" cellspacing="0" border="0" class="table" id="example">
```

```
                                        <a data-placement="right" title="Click
to Delete checked item" data-toggle="modal" href="#delete_log" id="delete"  class="btn btn-
danger" name=""><i class="icon-trash icon-large"> Delete</i></a>

                                                                <script
type="text/javascript">
$(document).ready(function(){
$('#delete').tooltip('show');
$('#delete').tooltip('hide');
                                                                 });
                                                                </script>
                                                                <?php
include('modal_delete.php'); ?>
                                                                        <thead>
                                                                            <tr>
                                        <th>Check</th>

        <th>Date</th>
        <th>System User</th>
        <th>Action</th>
```

```
</tr>
                                                    </thead>
                                                    <tbody>
                                <?php
                                                    $query =
mysql_query("select * from  activity_log
                                                    LEFT JOIN
admin ON activity_log.username = admin.username
                                                    order by date
DESC")or die(mysql_error());
                                                    while($row =
mysql_fetch_array($query)){
                                                    $id =
$row['activity_log_id'];
                                                    $username =
$row['username'];
                                                                ?>
                        <tr>
                                                <td width="70">
```

<input

id="optionsCheckbox" class="uniform_on" name="selector[]" type="checkbox"

value="<?php echo $id; ?>">

</td>

<td>

<?php echo

$row['date']; ?></td>

<td><img id="avatar1" src="<?php echo

$row['adminthumbnails']; ?>">

<?php echo

$row['username']; ?></td>

<td>

<?php echo

$row['action']; ?></td>

</tr>

<?php } ?>

</tbody>

</table>

</form>

```
                    </div>
                  </div>
                </div>
                <!-- /block -->
              </div>
            </div>
          </div>
                <?php include('footer.php'); ?>
      </div>
                <?php include('script.php'); ?>
  </body>
```

Activity Log

Number of System user Activity Log: <?php echo $count; ?>

```
            <div class="block-content collapse in">
              <div class="span12">
```

```
Activity Log</div>
                                                    <div class="muted pull-right">
                                                    Number of System user Activity
Log: <span class="badge badge-info"><?php  echo $count; ?></span>
                                                    </div> </div>
                    <div class="block-content collapse in">
                        <div class="span12">
                    <!-- block -->
                                <div class="empty">
                <div class="alert alert-success alert-dismissable">
                    <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>
                     <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checbox
if you want to delete?
                </div>
            </div>
<?php include('session.php'); ?>
   <body>
                    <?php include('navbar.php'); ?>
```

```
<div class="container-fluid">
    <div class="row-fluid">
                        <?php include('activity_log_slidebar.php'); ?>
        <div class="span9" id="content">
            <div class="row-fluid">
              <!-- block -->
                        <div class="empty">
            <div class="alert alert-success alert-dismissable">
              <button type="button" class="close" data-dismiss="alert" aria-
hidden="true">&times;</button>
               <i class="icon-info-sign"></i>  <strong>Note!:</strong> Select the checbox
if you want to delete?
            </div>
        </div>

                        <?php
            $count_log=mysql_query("select * from activity_log");
            $count = mysql_num_rows($count_log);
        ?>
```

```
<div id="block_bg" class="block">
  <div class="navbar navbar-inner block-header">
    <div class="muted pull-left"><i class="icon-user"></i> System User Activity Log</div>
                                                    <div class="muted pull-right">
                                                    Number of System user Activity Log: <span class="badge badge-info"><?php  echo $count; ?></span>
                                                    </div>
  </div>
  <div class="block-content collapse in">
    <div class="span12">
                                                    <form action="delete_log.php" method="post">
                                                        <table cellpadding="0" cellspacing="0" border="0" class="table" id="example">
                                                    <a data-placement="right" title="Click to Delete checked item" data-toggle="modal" href="#delete_log" id="delete"  class="btn btn-danger" name=""><i class="icon-trash icon-large"> Delete</i></a>
                                                        <script type="text/javascript">
```

```
$(document).ready(function(){

$('#delete').tooltip('show');
                                                    $('#delete').tooltip('hide');
                                                    });
                                                    </script>
                                                    <?php
include('modal_delete.php'); ?>
                                                        <thead>
                                                            <tr>
                            <th>Check</th>
        <th>Date</th>
        <th>System User</th>
        <th>Action</th>
        </tr>
```

```
                                                    </thead>

                                                    <tbody>

<?php
```


Biodata

Personal Data:

Surname: FATOKI

Other Names: IYANUOLWA M.

Date of Birth: 25th June, 1997

Gender: Female

Local Government Area: Lagelu

State of Origin: Oyo State

Nationality: Nigerian

Marital Status: Single

Religion: Christianity

Next of Kin: Fatoki, Bolaji O

Contact:

Residential Address: House 40, Road 3, Christ Lodge Villa, Ojoyeye, Olodo, Ibadan.

Postal Address: Same as above.

Email Address: iyanuoluwafatoki@gmail.com

Mobile No: 07066752276

Educational Background:

Institutions attended with Dates and Qualification:

Lead City University, Toll Gate, Ibadan, Oyo State: 2020 till date (MSc. Software Engineering)

Lead City University, Toll Gate, Ibadan, Oyo State: 2016-2019 (BSc. Computer and Information Sciences)

Highland College of Technology, Samonda, Ibadan, Oyo State: 2013-2015 (OND. Computer Networking and System Security)

Oyo State School of Science, Pade, Oyo state, Nigeria: 2010-2013 (WAEC O'Level, NECO O'Level)

Treasureland Junior Academy: 2002-2007 (First School Leaving Certificate)

Awards and Fellowship

Best student in professional certification passed in Computer Science department, 2019, Lead City University.

Work Experience with Dates

Information Technology Unit (ITU), University College Hospital, Ibadan, Oyo state. SIWES Trainee (3 Months) 2014

Information Technology Unit (ITU), University College Hospital, Ibadan, Oyo state. SIWES Trainee (6 Months) 2018

Detip Educational INC, Sango Ota, Ogun State. NYSC 2020

Lead City University, Ibadan. Oyo State. 2021 till date

Courses Taught

- CSC 111: Introduction to computer science
- CSE 112: Introduction to problem solving
- SEN 111: Introduction to computer science II
- CYB 111: Fundamentals of cyber security
- IFT 111: Introduction to information systems

- OIM 215: Desktop publishing and web design
- ISM 412: Digital rendering
- BED 402: ICT application

Membership of Academic Professional Bodies

Associate member: Institute of personality development and customer relationship management ID-CRM

Academic Papers (Unpublished These):

Design and construction of 2.5kva uninterrupted power supply. ND. Thesis, 2015.

Project submission management system. B.Sc. Thesis, 2019.

Published Journal Articles

Ayemowa Matthew, Dr. W. Ajayi, Emmanuel Adediran, **Iyanuoluwa Fatoki** and Alonge Opeyemi "Testing of embedded system with a slight look at mobile devices" American Journal of Computer Science and Information Technology. Volume 9, Number 9, Number 9.9885, October 2021.

Iyanuoluwa Fatoki, Kikelomo Okesola, Oluwaseyi Afe, Victoria Oyekunle and Akintayo Ayoade "Keyword based academic project management system" International Conference on Applied ICT (ICAICT). Volume 2, ISBN: 978-978-977-446-3. October 2019.

Skills

- Computer networking and maintenance
- System analysis and design
- Computer hardware maintenance
- Software engineering

Referees

1. **Mrs. Kikelomo Okesola**

   Department of Computer science

   Lead City University, Ibadan, Oyo State.

   08081451461

2. **Rev. Adetule**

   ITU Department

   College of Medicine, UCH, Oyo State.

   07033981596

3. **Dr. A.A. Waheed**

   Department of Computer science

   Lead City University, Ibadan, Oyo State.

   07031199441

…… ……..

Signature Date

University Compliance Form

This is to certify that this thesis by Fatoki Iyanuoluwa Modupeore with Matriculation Number LCU/PG/002131 in the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan is in full compliance with the approval of the University's format and style.

…… ……..

Signature Date