Open Access
Online ISSN: 2992-5487
Print ISSN: 2992-6068

NIJEP
Publications

NEWPORT INTERNATIONAL JOURNAL OF ENGINEERING AND PHYSICAL SCIENCES (NIJEP)

Volume 4 Issue 2 2024

https://doi.org/10.59298/NIJEP/2024/421018.2.1

Development of an Effective Hybrid Text Plagiarism Detection System using Machine Learning Techniques

[1]Adejumo S.O., [2]Alade S.M., [3]Akawuku G.I. and [4]Eze H.E.

[1,2,3,4] Department of Computer Science, Nnamdi Azikiwe University, Awka
Email: so.adejumo@unizik.edu.ng;sm.alade@unizik.edu.eg;gi.akawuku@unizik.edu.ng

ABSTRACT

In recent times, there has been a great spread of plagiarism as a result to the advancement on internet technology, which has brought about large volume of information to be share on the internet in several fields and discipline. The design and implementation of a Plagiarism Detection System for Nigerian Public University' is a detailed approach to dealing with academic plagiarism at the university. The study examined the existing systems relating to the plagiarism detection model as well as system, designed the model for the plagiarism system, implement the proposed system, and evaluate the developed system on various performance metrics. The system used a Hybrid Methodology by combining text-based, semantic-based, and machine learning-based techniques to analyze academic submissions. The proposed system has been shown to be efficient and effective, with an average accuracy of 80%, precision score of 0.90, recall score of 0.80, F-measure of 0.82, granularity of 0.93(≈1.000) and a plagdet percent or score of 0.8856 in detecting plagiarized documents. The study highlights the risks associated with academic plagiarism, the development of a database of research papers, and the evaluation of machine-learning techniques for plagiarism detection. The report emphasizes the importance of developing effective plagiarism detection systems to promote academic integrity and originality in academic institutions. This project is a significant contribution to the computer science field and has the potential to positively impact academic integrity in Nigerian public University and beyond.

Keywords: Hybrid Text, Plagiarism, Detection, Machine Learning and Techniques

INTRODUCTION

The advancement in information Technology, particularly the advent of internet and its wide coverage has brought about a rising issue not for the academics alone but also for all educational, training, and research organization [1]. In fact, the internet has made it easy to access a large amount of unstructured data, particularly text-based information [2]. However, this has led to an increase in the practice of copying and reusing content without proper permission, which is a significant downside to this easy access to information. It is particularly problematic in the education sector where students' research projects are graded based on their resourcefulness, ingenuity, and diligence, and it can be frustrating for anyone who has had their work exploited without proper credit given [3]. This practice is known as plagiarism and poses a threat to academic integrity and authenticity. Plagiarism detection involves using reliable sources or technology to identify copied work. Before the internet, teachers used several methods to detect plagiarism, such as manually comparing papers, searching through the library, double-checking sources, and monitoring group projects. Experienced instructors could detect inconsistencies in a student's work that could indicate potential plagiarism. Sharing information about plagiarism through networks was also a common practice [4]. Like a double-edged sword, the internet has also made plagiarism detection easier. Prior to the advent of the internet, it might have been tedious and difficult to detect plagiarism, but this is no longer the case. Plagiarism detection online software such as iThenticate, Turnitin, Grammarly have made it easier to detect plagiarism. Turnitin for example gives an overall similarity index score of any write-up to previous write-ups online, and the user (Journal, researcher, university) decides what percentage to use as a cut-off to determine an acceptable level of similarity score. It has been observed that in this era of intense plagiarism, many institutions of learning do not have a functional automatic plagiarism detection system in place [5]. This is a serious issue as plagiarism is one of the main problems faced by academic institutions, and a reliable automatic plagiarism detection system can help prevent academic dishonesty. Hence this paper aims to address this issue by proposing the design and implementation of a web-based plagiarism detection system (PDS) using advanced automatic plagiarism detection techniques to aid in curbing plagiarism among students, researchers, academics etc. in the university.

Intellectual honesty is a sacred principle that underlies academia such that one is obliged to acknowledge the originators of ideas, words and data which form the basis for one's work. Research has demonstrated that using plagiarism detector to assess the contents of academic works among students, researchers and scholars improves the provision of high-quality research output, fund and grants. Moreso, a variety of manual techniques, including manual perusal of academic works, including projects, thesis and dissertations, have been employed to evaluate the textual information content that students have supplied [6]. In fact, in this era of intense plagiarism where higher institution of education, research institutes and others do not have a functional automatic plagiarism detection system in place, intellectual integrity is compromised. Nevertheless, the existing approaches, such as text matching as well as other techniques like linguistic analysis are ineffective and, hence a difficult issue in the education arena (HEI). This is because the seriousness of plagiarism is not relative to its quantity but to the extent of its consequences. Some of the potential consequences of plagiarism include the fact that it infringes upon the creator's personality rights and can have devastating effects, erodes the connection to historical information and knowledge by refusing readers access to the original sources or individuals, aids loss of confidence in the abilities of students or researchers and can make them become addicted to cheating. Worse still, the act of plagiarism results in sloppy research, disregard for ethics, and overall damages to the reputation of academic system [7]. Therefore, the need for computational technique that will analyze and check for similarities of the contents submitted by university students for evaluation and approval of their creative output is therefore required, hence this study. There have been various approaches that have been employed for solving academic plagiarism of different types. These include the methodologies such as extraction of sentence components [8], string matching [9], structured based analysis, citation-based analysis, vector state model and others. For example, [10] addressed the problem of plagiarism of text in academic institution using string matching technique, particularly for English language texts. Similarly, the study conducted by [11] using string matching produced an appreciable level of accuracy. However, this research makes use of hybrid model that is a mix of text based, semantic based and machine learning techniques to achieve a high level of accuracy and robustness [12]. The main aim of this study is to design and implement a plagiarism detection system (PDS) with a view to curbing fraud and illegal use of contents among researchers, academics and higher institutes of education using advanced automatic plagiarism detection techniques, to examine (i.e., study and analyze) the existing research relating to the development and performance of plagiarism detection model or system, design the model for the plagiarism system, implement the proposed system, and evaluate the developed system on various performance metrics.

Literature Review

The issue of plagiarism has become widespread worldwide, spanning various areas including music, arts and crafts, entertainment, businesses, academics, etc. Plagiarism is the act of using someone else's words, ideas, creations, or work without proper acknowledgement, permission, or attribution and presenting it as one's own. Plagiarism is considered unethical and a breach of academic, professional, and creative integrity. It can have serious consequences, including damage to one's reputation, academic penalties, legal actions, and professional repercussions [13]. As the spread of plagiarism has increased so has how they can be detected or discovered continue to grow and become more sophisticated [3]. Traditionally, plagiarism detection would have to be done using manual methods such as comparing suspected plagiarized text against the source document, citation analysis, checking inconsistencies in writing style in sections of a text etc. The advancement in technology brought about automatic plagiarism detection techniques which allow plagiarism detection to be done automatically, much faster, and more reliably using plagiarism detection techniques such as cosine similarity, jacquard similarity, co-efficient measure, logistic regression, naïve Bayes, support vector machine, decision trees, KNN and many more [5]. These different techniques vary based on their different approaches to plagiarism detection for instance, some of the techniques require to be trained while some do not, some techniques are simple and efficient while others are more computationally expensive, some have a higher level of accuracy while others do not, and some are suitable for detecting more advanced plagiarism such as paraphrasing, rephrasing etc.

Forms of Plagiarism

Generally, plagiarism comes in two (2) broad forms namely textual and source code plagiarism [8]. The textual form is the commonest form of plagiarism in research and scientific discipline, where the whole textual information content is copied without giving credit to the original owner of the idea. This is further divided into several subclasses [9]. According to individual researchers' intuition, practical application and exposure, the types of textual form of plagiarism can be categorized into several sub-classes.

Textual Plagiarism

[10] highlights some of the most common forms of textual plagiarism which include:

- *Copy and paste or Complete plagiarism*
  It is a class of plagiarism which involves copying the original text or idea as if it were your work without referring to the original paper. In other words, this overt type of plagiarism occurs when a writer submits someone else's work in their own name. This involves paying somebody to write a paper for you, then handing that paper in with your name on it, is an act of complete plagiarism as is stealing or "borrowing" someone's work and submitting it as your own [4].

- *Direct plagiarism*
  Direct plagiarism is similar to complete plagiarism in that it, too, is the overt passing-off of another writer's words as your own. The difference between the two is how much of the paper is plagiarized. With complete plagiarism, it is the entire paper. With direct plagiarism, specific sections or paragraphs are included without crediting (or even acknowledging) the author [6].
- *Paraphrasing plagiarism*
  This form of plagiarism can be of two types: simple and mosaic plagiarism. The simple paraphrasing plagiarism is what happens when a writer reuses another's work and changes a few words or phrases. It is a common type of plagiarism that many students do but do not even realize it is a form of plagiarism. But if you are presenting someone else's original idea in your writing without crediting them, even if you are presenting it in your own words, it is plagiarism. However, the other type is known as patchwork or mosaic plagiarism. It is a type which refers to instances where plagiarized work is interwoven with the writer's original work. This kind of plagiarism can be subtle and easy to miss, and it may happen in conjunction with direct plagiarism. A typical example of patchwork plagiarism is taking a clause from a source and embedding it in a sentence of your own without referring to the original paper [7].
- *Self-plagiarism*
  This is also referred to as recycled plagiarism. It is a form of plagiarism where the author uses or reuse content from their past work for a new publication. It is important to know that you can indeed plagiarize your own work. It may seem like your original thoughts are yours to use however you please, but that's not entirely true. If you wrote an essay on a particular subject in the past and are now writing a research paper on the same topic, reusing content from your previous work would be considered an act of self-plagiarism. It is important to be aware of this fact and avoid such practices. To avoid any accusations of plagiarism, always ensure that you cite your sources properly. Remember, using the same sources is acceptable as long as you cite them correctly [5].
- *Accidental plagiarism*
  Accidental plagiarism is a frequent form of plagiarism that arises when a writer recklessly uses another individual's work without giving proper credit. This commonly occurs when a writer neglects to cite their sources, cites them inaccurately, or omits the use of quotation marks around cited material. It is crucial to avoid this type of plagiarism at all costs, as it can have severe consequences [7].
- *Idea Plagiarism*
  This is a form of plagiarism where the entire solution and ideas are stolen from others, claiming that it is an original research paper [6].

Source-Code Plagiarism

This is the other form of plagiarism which occur in the educational arena, where lines of programming code of specific program written by the source is reused or modified by another person either partially or completely [7]. It has 5 sub-categories including,

- Manipulation plagiarism: This is the form where the source code is modified by other developers by either deleting or inserting sub-codes or strings to an original one without acknowledgment.
- Reordering structure plagiarism: this is the type in which the language rules or sentence structure of the source code is modified by functions or statements recording without referring to the original paper.
- No-change plagiarism is another form of plagiarism where the developers or programmer do not change anything in the program code but add/remove spaces or comments as it was their work.
- Language switching plagiarism: This is a form where the source code language is rewritten in other programming languages and declared as original code. However, textual and source code plagiarism can be detected by humans or automated detection methods.

In addition, [8, 9] classified plagiarism into two main categories: Monolingual and Cross-lingual. Monolingual plagiarism is one which is about the same language working with most detectors, whereas Cross-lingual plagiarism works with diverse languages. However, the main forms of plagiarism can be identified by humans or automatic means. Before the internet, professors and teachers used various methods to check for plagiarism in assignment, essay, mini-project and academic writings. A few common approaches include manually comparing papers to detect/identify any similarities or differences with other sources, personally searching through the library to find appropriate sources and then comparing them to their students' work, double-checking sources and comparing them to citations to detect any discrepancies or missing information, experienced instructors can detect inconsistencies in a student's work that may indicate potential plagiarism, such as sudden improvements in writing or the use of advanced vocabulary, Oral exams and presentations help teachers evaluate a student's understanding of the subject and identify inconsistencies between their verbal and written responses, monitoring group projects for signs of plagiarism, such as irregularities in style, tone, or content, while evaluating student contributions and interactions, sharing information about plagiarism through their networks to identify patterns [6]. Today, there are so many freely available online contents. As a result, manual identification has grown more and more difficult and now a critical issue to our institution of higher learning [8]. Therefore, the exploration for plagiarism detection for checking our assignment, essay, mini-project and academic writings has been an issue in the academic arena. But the development of technology and data mining techniques has made plagiarism checking easy. Like a double-edged sword, the internet has also made plagiarism detection easier. Prior to the advent of the internet, it might have been tedious and difficult to detect plagiarism, but this is no longer the case. Plagiarism detection online software such as iThenticate, Turnitin, Grammarly, Plagiarism Checker, Duplicheck [1,2,3,4,5,] etc., have made it easier to detect plagiarism. Turnitin, for example, gives an overall similarity index score of any write-up to previous write-ups online, and the user (Journal, researcher, university) decides what percentage to use as a cut-off to determine an acceptable level of similarity score. Nevertheless, even when the similarity score is lower than the specified cut-off, the author may still be considered to have plagiarized if he/she had copied a statement(s) verbatim from other sources without the use of quotation marks ("") or italics on the copied texts before referencing. However, plagiarism can be detected by humans or by automated detection methods. The plagiarism detection method can be internal or external detection. The intrinsic detection is one in which the plagiarized passages is located within a document without access to the potential original text [7]. Whereas, the external detection is a detection method, where a suspected document is examined against the source document text which may be online or offline [8].

Plagiarism comparative study

Table 1: Plagiarism comparative study

S/N	Plagiarism Detection Technique	Description	Examples	Limitations
1	Text/Lexical Based Comparison	Compares the text of a suspected plagiarized document to the text of known sources.	Word matching N-gram matching Single matching	Can be fooled by synonym substitution and paraphrasing. May not be able to detect plagiarism if the text has been heavily rewritten or paraphrased.
2	Semantic Based Comparison	Compares the meaning of a suspected plagiarized document to the meaning of known sources.	Latent semantic indexing (LSI) Topic modeling Natural language processing (NLP)	More robust to synonym substitution and paraphrasing than text-based comparison techniques. May not be able to detect plagiarism if the meaning of the text has been significantly changed.
3	Citation Based Analysis	Compares the citations in a suspected plagiarized document to the citations in known sources.	Citation frequency analysis Citation source analysis Citation pattern	Can be fooled by self-citation and misattribution. May not be able to detect plagiarism if the citations

			analysis	have been removed or changed.
4	Vector Space Model	Compares the vectors of two documents to determine how similar they are.	Term frequency-inverse document frequency (TF-IDF) Latent semantic indexing (LSI)	More robust to synonym substitution and paraphrasing than other techniques. May not be able to detect plagiarism if the text has been significantly changed.
5	Structural Based Analysis	Compares the structure of a suspected plagiarized document to the structure of known sources.	Section heading analysis Table of contents analysis Reference list analysis	Can be fooled by the use of different formatting styles or the removal of section headings. May not be able to detect plagiarism if the document has been heavily rewritten or paraphrased.
6	Machine Learning Based	Uses machine learning algorithms to identify patterns in suspected plagiarized documents that are indicative of plagiarism.	Support vector Machines (SVMs) Decision trees Random forests	More robust to different types of plagiarism than other techniques. Require a large dataset of known plagiarized and non-plagiarized documents to train the machine learning model.

- https://www.ithenticate.com
- https://www.turnitin.com
- http://www.grammarly.com/
- http://www.dustball.com/cs/plagiarism.checker
- http://www.duplicheck.com/

Related Work

Academic plagiarism research has gained significant attention in recent years, focusing on various stakeholders such as teachers, students, decision-makers, and institutions. The issue of plagiarism is primarily used to improve teaching, management, and evaluation. [7, 8, 9] proposed a method for detecting plagiarism in text documents using a vector space model to capture semantic similarity. The methodology was broken down into four stages: preprocessing, seeding, extension, and filtering. [10], presented a context skip n-grams and context n-grams model, which is dependent on context to address plagiarism. The model uses context 3-grams to compare suspected documents with the source document, skips context n-grams, joins adjacent detections simultaneously, and eliminates sentences shorter than 190 characters. However, this approach's drawback is trying to combine different n-gram characteristics to boost performance overall. [11] proposed a system for detecting plagiarism based on context 5-gram comparisons between suspicious and source documents. The system employed Euclidean distance clustering to determine the extent of seeds detected while passes that are shorter than 190 characters are eliminated. However, the system demonstrated that it was unable to identify some forms of summary plagiarism. [13] proposed a set-based algorithm to detect obscured plagiarism. The algorithm involved text preprocessing, tokenization, whitespace collapse, chunking documents with a semi-fixed window size of 250 characters, and using Dice's coefficient to compare every suspicious chunk with all of the source chunks. However, the algorithm fails when it comes to extremely obscured plagiarism. [7] applied data mining methods to detect similarities in titles, abstracts, or topics of students' final scientific papers to prevent plagiarism. Overall, academic plagiarism research aims to improve teaching, management, and evaluation in academic settings.

Methodology

The proposed plagiarism detection system for the higher education institute (HEI) aims to detect the extent to which the text, ideas and contents are copied with the highest accuracy score [4]. The proposed system is constructed based on three main phases: preprocessing, analysis and post-processing as depicted in Fig.1. Similarly, the proposed system is premised on the development of a database to learn the machine learning and deep learning models for the identification of the plagiarized text. Furthermore, the model was constructed by taking into consideration the similarity features that show the variety of linguistic properties like the lexical, syntactic and semantic components of the document text [6].

Preprocessing

This is an important role in data mining. This is because the fundamental goal of data preparation stage is to make original dataset suitable for data mining techniques and algorithms (ML/DL) to be implemented [7]. Owing to the massive volume of dataset available in educational data repositories and institution learning portals and platforms, there are challenges facing the database that affect the quality of data. To improve the quality of data available for the model, it is essential to carry out data cleaning to handle missing and inconsistent data. This part discusses the data cleaning method applied to remove noise and handle missing values [9]. This is the preprocessing task that involves converting the original or raw data collected into an appropriate format. This phase involves sentence segmentation, tokenization, case conversion, removal of stop words and special characters and lemmatization. The preprocessing stage was accomplished on the passage and sentence-based data by applying a segmentation method to split the document into meaningful sentences for data cleaning. The stop words, punctuation, hashtags, hyperlinks, username, numbers, special characters and irrelevant symbols like $, @, &, as well as non-relevant POS tags are removed to prepare the data for further analysis. Furthermore, characters are changed to lower case using the "Swapcase" function in Python Language before the tokenization of the document text. To make the data cleaning process simpler, a data cleaning pipeline (function) was created to carry out the data cleaning task to prepare the dataset for tokenization. During the process of tokenization, each sentence was converted into small units called token. Moreover, the process of lemmatization is used to produce and derive the primary/basic form of the word through its context using LSA, where morphological analysis of the words is performed according to parts of speech. Additionally, at the paragraph level, the document text was converted into meaningful passages using segmentation method to combine the adjacent/neighbouring or close sentences till the length of the passage extended to at most 255 characters or words. Thereafter, each passage or paragraph was split into trigram and then cleaned [8].

Detailed analysis

In the analysis stage, the focus is geared towards detecting/discovering the copied or lifted instances between the documents, which is dependent on three (3) operational steps: paragraph, sentence and intelligent classification [6]. The first step is applied to understand or make the set of suspected and source passages having the highest probability of plagiarism, which cut down/ reduces or lowers/brings down the search area for plagiarized instances from the whole document. Likewise, the next step is centered on n-gram to detect the plagiarized/ copied cases. While the third step in the operational step involved in detecting plagiarism in the document text is built up using deep learning techniques/approach. At the paragraph stage, the first step is applied to understand or make or get the set of suspected and source passages having the highest probability of plagiarism. This is done by inspecting each suspected and source document with the highest likelihood score of plagiarism corresponding to the n-gram approach where n = 3 (that is, maximum trigram). Then, if plagiarism is applicable, it retrieves the source passage's prior and subsequent passages [7]. The duplicate pairs are eliminated once the passage pairs have been extracted. For every instance of document text, the text document plagiarism is a part or section of the suspected and source document, similar in context and meaning. This phase is conducted using artificial intelligent techniques to accomplish the objective of the detection system. This phase is determined by the steps namely passage, sentence and intelligent classification. Here at the paragraph-level of the analysis stage, the focus is geared towards comparing each paragraph of the suspected document with the paragraphs of the source document to obtain the pairs or set of suspected and source passage having the highest likelihood of plagiarism by analysing every suspected and source passage corresponding to common trigram based on n-gram approach. At the paragraph level, the preprocessing steps earlier discussed are applied to these paragraphs [9]. As a result, every or each suspected sentence is compared with all the sentences of the source document. Afterwards, the source sentence will be extracted if it contains the maximum value of the shared trigram compared with it as described in Algorithm 2. Here, two conditions are employed such that if the value obtained is greater or equal to the limit or threshold, it is believed to be plagiarized case or instance, else, if the score is below the threshold, then it indicates a non-plagiarized instance. However, if neither of the conditions stated are satisfied, the instance (pair of sentences) is classified will be analysed using machine/deep algorithm, particularly the random forest. At this stage of the detailed analysis, the intelligent classification step comes into play by taking into account the relevant features that have the capability to classify the suspected cases and differentiate or distinguish or make a distinction between the differences in the text similarities with high accuracy. Here, a Random Forest classifier (RF) algorithm is used to detect plagiarism [10]. This is because, it is a powerful machine learning algorithm used for classification tasks. It belongs to the ensemble learning family, which means that it combines multiple decision trees (DT) to make predictions. The classifier is trained on the features extracted from the text documents in the training set, where the database of the RF algorithm was designed containing the similarity features that can be used to detect the various types of document text plagiarism. The RF classifier identifies patterns in the data, understands the relationship between them, and learns to differentiate between a plagiarized text and an original one. Moreover, we must first identify and extract relevant characteristics or attributes [11]. This involves extracting features such as the number of shared trigrams, the Jaccard and containment similarity and also the similarity of the documents' word vectors. Consequently, similarity features are calculated and recorded combining with the class label for each of the extracted case of the first stage. The computed sentence similarity features were based on similarity feature [8] like the hybrid syntactic and semantic similarity features [4], cosine measure [7] and containment similarity [9]. Similarly, a WordNet [8] database was designed to offer or present similarity in the meaning between the words, containing other synonyms and concepts for each word. The computed word and sentence similarity criteria are described as follow:

Jaccard Similarity: This is a common proximity measurement used to compute the similarity between two objects, such as two text documents. It can be used to find the similarity between two asymmetric binary vectors or to find the similarity between two sets. It measures the similarity between two sets of data to see which members are shared and distinct. The Jaccard similarity is calculated by dividing the number of observations in both sets by the number of observations in either set. In other words, the Jaccard similarity can be computed as the size of the intersection divided by the size of the union of two sets. This can be written in set notation using intersection ($A\cap B$) and unions ($A\cup B$) of two sets as expressed mathematically in (1). where $|A\cap B|$ gives the number of members shared between both sets and $|A\cup B|$ gives the total number of members in both sets (shared and un-shared). The Jaccard Similarity will be 0 if the two sets do not share any values and 1 if the two sets are identical. The set may contain either numerical values or strings. Additionally, this function can be used to find the dissimilarity between two sets by calculating $d(A, B) = 1–J(A, B)$. **Containment similarity:** It is a technique used in plagiarism detection to compare the fingerprints of passages and determine the presence of plagiarism. It is a type of similarity measure that focuses on the overlap in word usage between two documents, which can help identify cases of cut-and-paste as well as paraphrased levels of plagiarism [6]. In the context of plagiarism detection, containment similarity can be represented by the following formula expressed in (2). Where A represents the fingerprint of a suspect passage and B represents the fingerprint of a source passage. If a suspect passage is plagiarised from a small portion of a source passage, containment similarity will yield a high similarity, while Jaccard similarity (another similarity measure) will yield a smaller similarity [9].

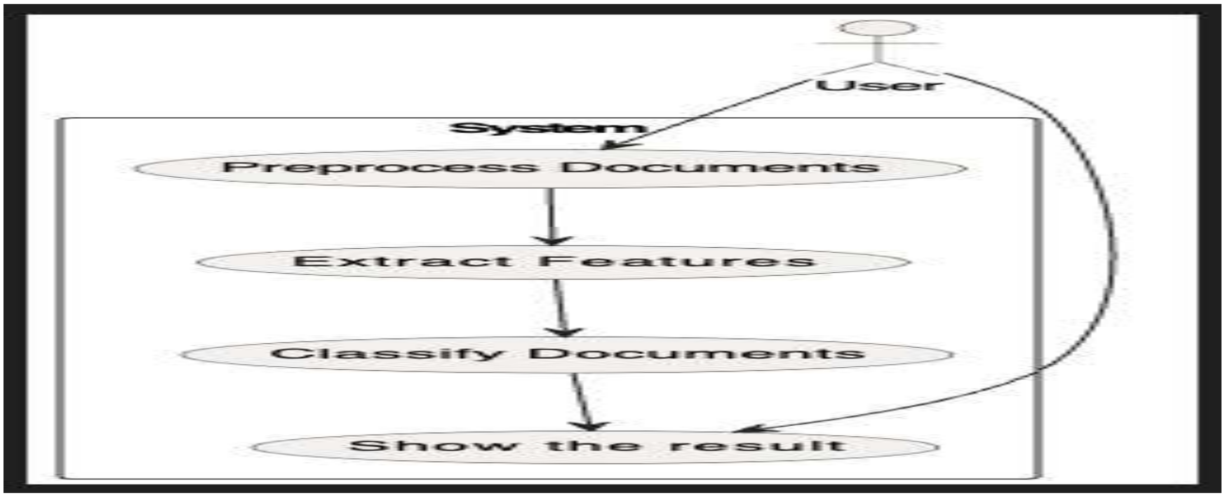
System Design

The system architecture was designed to provide a prototype that describes the structure and views of the plagiarism detection system as depicted in Figure 3 and 4. The requirements for the development of the detection system was defined using Use case as depicted in Table 2. The System requirement specifications (SRS) were designed using Unified Modelling Language (UML) tools, including Use case, Class diagram, and Sequence diagram illustrating the design of the proposed plagiarism detection system as shown in Figures 1, 2, and 3 respectively [9]. The use case diagram defines the system, its actors (Users: Teachers, Lecturers, librarian, students), and the roles the actors perform such as preprocessing, extracting features, classifying document and showing result of plagiarism. The class diagram illustrated in Figure 2 presents the various object and classes with its attributes and methods. The classes shown in the diagram are Document, User and their many subclasses. Similarly, Figure 3 depicts system dynamics by presenting the participating objects (classes, components, etc.) in the interaction and the sequence of information that is transferred between the modules or classes.

Table 2: Sample Table of Use Case for the Proposed Plagiarism Detection System

S/N	Use case Name	Description	Role
1	Login	The user logs in to the plagiarism system using a username and password to gain authorized access.	Users: Teachers, Lecturers, Librarian, Students
2	Preprocess Document	The user is allowed to carry out the initial preprocessing of the document to be checked.	Users: Teachers, Lecturers, Librarian, Students
3	Extract features	The user is permitted to extract relevant features to determine the plagiarism form	Users: Teachers, Lecturers, Librarian, Students
4	Classify	The user is allowed to classify the document as “plagiarized” or “not plagiarized”	Users: Teachers, Lecturers, Librarian, Students

The proposed system architecture designed in Figure 4 was implemented using a Python Programming Language with a lightweight Web Framework called Flask. Flask framework uses a Model View Concurrency Controller (MVCC) framework to provide useful tools and features for creating web applications in Python, while React, a JavaScript library, was used to create fast user interfaces for the web application [9].



REFERENCES

1. Arabi H, Akbari M. Improving plagiarism detection in text document using hybrid weighted similarity. Expert Systems with Applications. 2022 Nov 30;207:118034.
2.Ahuja L, Gupta V, Kumar R. A new hybrid technique for detection of plagiarism from text documents. Arabian Journal for Science and Engineering. 2020 Dec;45:9939-52.
3. Fuad AJ, Wicaksono AK, Aqib MA, Khoiruddin MA, Fajar AS, Mustamir K. AI Hybrid Based Plagiarism Detection System Creation. In2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) 2024 May 14 (pp. 1495-1500). IEEE.
4. Meuschke N, Stange V, Schubotz M, Gipp B. HyPlag: a hybrid approach to academic plagiarism detection. InThe 41st international ACM SIGIR conference on research & development in information retrieval 2018 Jun 27 (pp. 1321-1324).
5. El-Rashidy MA, Mohamed RG, El-Fishawy NA, Shouman MA. Reliable plagiarism detection system based on deep learning approaches. Neural Computing and Applications. 2022 Nov;34(21):18837-58.
6. Foltýnek T, Meuschke N, Gipp B. Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR). 2019 Oct 16;52(6):1-42.
7. Alzahrani SM, 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). 2011 May 12;42(2):133-49.
8.Ilyas M, Malik N, Bilal A, Razzaq S, Maqbool F, Abbas Q. Plagiarism detection using natural language processing techniques. Technical Journal. 2021 Apr 10;26(01):90-101.
9. Xiong J, Yang J, Yan L, Awais M, Khan AA, Alizadehsani R, Acharya UR. Efficient reinforcement learning-based method for plagiarism detection boosted by a population-based algorithm for pretraining weights. Expert Systems with Applications. 2024 Mar 15;238:122088.
10. Manzoor MF, Farooq MS, Haseeb M, Farooq U, Khalid S, Abid A. Exploring the Landscape of Intrinsic Plagiarism Detection: Benchmarks, Techniques, Evolution, and Challenges. IEEE Access. 2023 Dec 1;11:140519-45.
11. Chang CY, Lee SJ, Wu CH, Liu CF, Liu CK. Using word semantic concepts for plagiarism detection in text documents. Information Retrieval Journal. 2021 Oct;24:298-321.
12. 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. 2021 Feb 25;33(4):e5000.

13. Pertile SD, Moreira VP, Rosso P. Comparing and combining C ontent-and C itation-based approaches for plagiarism detection. Journal of the Association for Information Science and Technology. 2016 Oct;67(10):2511-26.

CITE AS: Adejumo S.O., Alade S.M., Akawuku G.I. and Eze H.E. (2024). Development of an Effective Hybrid Text Plagiarism Detection System using Machine Learning Techniques. NEWPORT INTERNATIONAL JOURNAL OF ENGINEERING AND PHYSICAL SCIENCES, 4(2):10-18. https://doi.org/10.59298/NIJEP/2024/421018.2.1