IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Online Assignment Plagiarism Checker Using Data Mining And NLP

[1] Murtaza Sadriwala, [2] Tanvi Patil, [3]Ashita Gaikwad, [4] Shantilal Mali

[1]Student, [2]Assistant Professor, [3]Student, [4]Student

[1]Department of Information Technology,

[1]St. John College of Engineering & Management, Palghar, India

[2]Department of Information Technology,

[2]St. John College of Engineering & Management, Palghar, India

[3]Department of Information Technology,

[3]St. John College of Engineering & Management, Palghar, India

[4]Department of Information Technology,

[4]St. John College of Engineering & Management, Palghar, India

Abstract: The online assignment plagiarism checker using data mining and NLP (Natural Language Processing) is a tool designed to ensure the originality of academic work submitted by students. This system employs advanced techniques like data mining and NLP to analyze and compare text, detecting instances of plagiarism effectively.In simpler terms, data mining helps the system explore and identify patterns within the text, while NLP allows it to understand the meaning and context of the language used. By combining these technologies, the plagiarism checker can efficiently scan through vast amounts of text, highlighting similarities and potential instances of plagiarism. To put it another way, natural language processing (NLP) enables the system to comprehend the context and meaning of the language being used, while data mining assists in finding patterns within the text. Through the integration of various technologies, the plagiarism checker is able to effectively search through large volumes of text, identifying possible instances of plagiarism. Ultimately, It provides educators with a reliable means of verifying the authenticity of assignments, promoting fairness and honesty in the academic environment.

Keywords: Natural language processing, comprehend.

I. INTRODUCTION

In the world of academics, it's important for everyone to have their own thoughts and ideas in assignments.To tackle this issue and promote fair academic practices, we introduce an online assignment plagiarism checker. This tool uses smart technologies like data mining and NLP (Natural Language Processing) to make sure that the work submitted is original and truly belongs to the user.. Imagine it as a digital detective that carefully examines the words and sentences in your assignment. Data mining helps it look for patterns and connections in the text, while NLP helps it understand the meaning behind the words. So, this online plagiarism checker isn't just a tool; it's like having a trustworthy friend who watches out for you, making sure your hard work gets the credit it deserves. Let's dive into the world of this innovative technology that aims to maintain fairness and honesty in the academic journey.

II. SURVEY OF EXISTING SOLUTIONS

In the domain of plagiarism checker several noteworthy solutions have been proposed, each with its unique approach and set of features. This survey examines some of these existing systems and identifies potential areas for improvement:

[1]Online Plagiarism Detection and Result Evaluation using Data Mining and NLP.
Conclusion:
The system generates a report which shows the matching content in assignments of school and college students. Here we designed a simple method which assists us with the detection of instances of plagiarism and semantic checking.By using data mining algorithm and NLP it will provide straightforward documentation.

Research Gap:
1)The system doesn't help the user by checking grammar. 2)The system cannot completely find the mistake, but it can help up to some extent.3) It checks the plagiarism between two documents or files.

[2]Plagiarism Checker as Best Free Online Plagiarism Detection Software.
Conclusion:
Plagiarism, seven plagiarism tools that had been tested, the Plagiarism Checker is the easiest tool to learn and more efficient in searching for similar sources for textual data that can be found online.

Research Gap:
1) The Free online plagiarism checker may not be as accurate as paid versions, so there is a limited accuracy of plagiarism.
2) They often lack advanced features like deep document analysis or integration with academic databases.3) Many free tools have word limits, making them unsuitable for longer documents.

[3] Building a real-world plagiarism detection system.
Conclusion:
In this paper Docode 5, a system for plagiarism detection we have shown the problems when dealing with big document collections and when dealing with the World Wide Web. We have shown that we can take an algorithm for two-document plagiarism detection , and that we can make a system for plagiarism detection on the web that can take into account its own impact on it, maintaining the quality of the performed analyses.

Research Gap:
1) Though, as shown, the overall architecture of the system is strong enough to do a commercial deployment but most of the modules in the system could be improved to recognize structure in documents and query generators to get better
results, improving the overall performance of the system. 2) The front-end presents its results in an effective way, but this could be improved by showing the results in the documents themselves instead of in a plain text representation, thus improving the usability of the system.

[4]Plagiarism Detection through the Internet using Hybrid Artificial Neural Network and Support Vectors Machine.
Conclusion:
The general conclusion from the results is that the machine learning tested suited to solving detection of plagiarism. This is shown by the results of all methods of learning machines that do produce an average value above 90%. The experimental results show that in general there is improvement in performance in the use of hybrid machine learning methods in the case of plagiarism.

Research Gap:
1) Developing and maintaining such a system can be costly, especially when compared to simpler plagiarism detection methods.
2) Requires a substantial amount of training data to ensure accuracy, which might not always be readily available.
[5]Using Artificial Intelligence to Predict Class Loyalty and Plagiarism in Students in an Online Blended Programming Course during the COVID-19 Pandemic.
Conclusion:
We proposed a novel method combined with a convolution neural network to predict class loyalty and a novel method to discover whether students plagiarized during distance learning during the COVID-19 epidemic.We first defined some fuzzy membership functions for evaluating class loyalty and to detect instances of plagiarism.

Research Gap:
1) It does not perform a pre-test and a post-test for the students. 2) It does not Design questionnaires to reflect class loyalty. 3) The system does not warn that the students have copied the code, the students are not aware that their behavior is monitored by the teacher.

III. MATERIALS AND METHODS

The implementation of the proposed system involved the utilization of various components and technologies to ensure seamless functionality and user experience.

[1] Python:

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics developed by Guido van Rossum.Python is used for server-side web development, software development, mathematics, and system scripting, and is popular for Rapid Application Development and as a scripting or glue language to tie existing components because of its high-level, built-in data structures, dynamic typing, and dynamic binding.

[2]Django:

Django is a Python framework that makes it easier to create web sites using Python. Django takes care of the difficult stuff so that you can concentrate on building your web applications. Django emphasizes reusability of components, also referred to as DRY (Don't Repeat Yourself), and comes with ready-to-use features like login system, database connection and CRUD operations (Create Read Update Delete).

[3]Javascript:

JavaScript is a scripting or programming language that allows you to implement complex features on web pages — every time a web page does more than just sit there and display static information for you to look at — displaying timely content updates, interactive maps, animated 2D/3D graphics, scrolling video jukeboxes, etc.

[4]Html,CSS:

HTML is the standard markup language used to create web pages. It provides the structure of a webpage by using a system of tags and attributes. HTML tags are enclosed in angle brackets < >, and they define the elements of a webpage. These elements can include headings, paragraphs, images, links, forms, and more.

CSS is used for styling HTML documents. It defines how HTML elements are displayed on the screen, such as layout, colors, fonts, and sizes. By separating content from presentation, CSS allows for consistent styling across multiple web pages,

IV. METHODOLOGY

Block diagram:

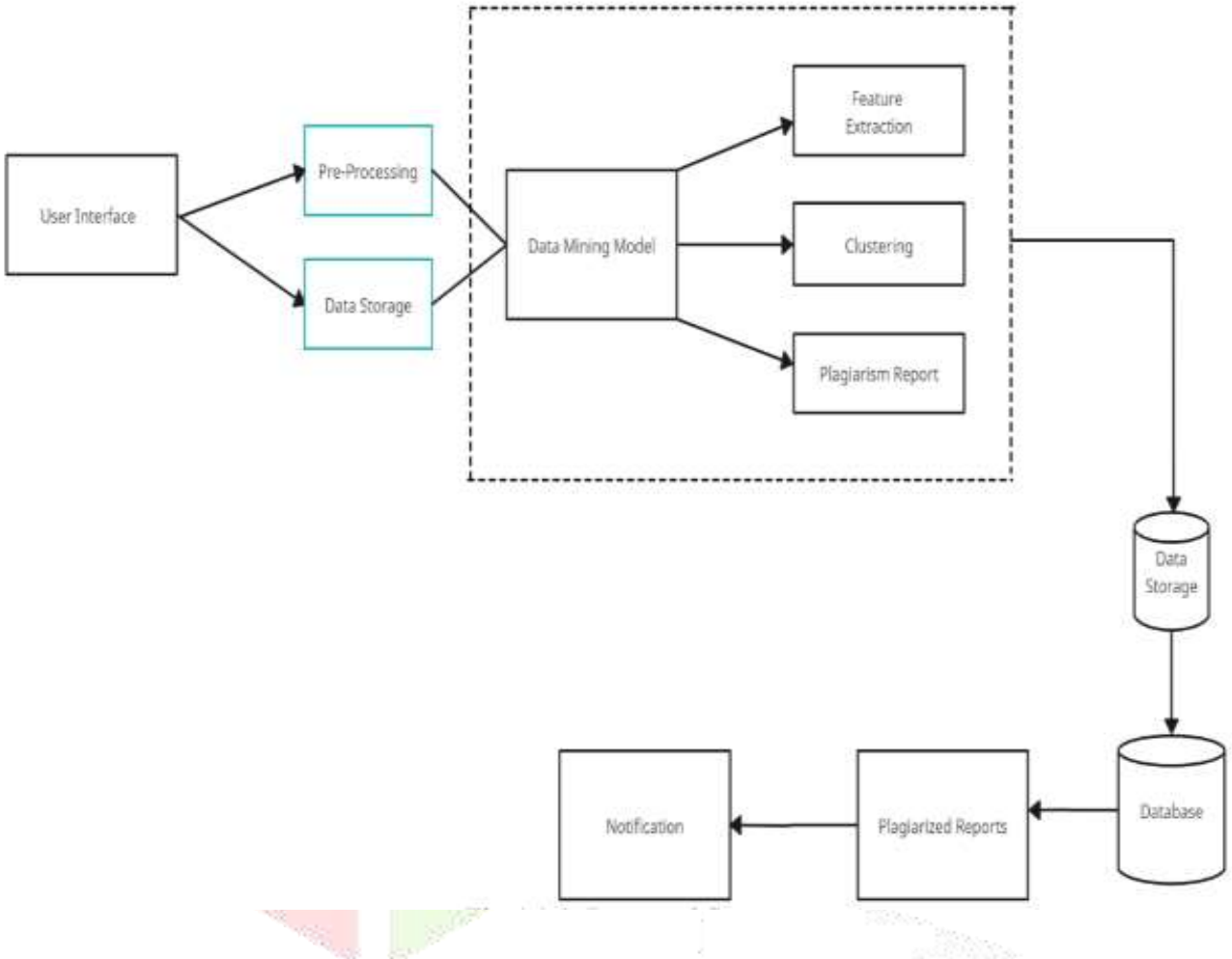

Fig 1**:** Block Diagram

Fig. 1 Illustrates the sequential steps of the methodology adopted in online assignment plagiarism checker.

1.User Interface:

Dashboard:enables users to upload files.Offers choices for customizing the settings of the plagiarism checker.

2.Prior to processing:

Upload Document:accepts a number of file formats, including Word, and text.Transforms documents into a text format that is standardized for analysis.Text Rewriting:Eliminates metadata, special characters, and unnecessary formatting.Text is normalized to provide consistent analysis.

3.DataStorage:

Original Records:keeps uploaded papers for use at a later time.Ensures the security and integrity of data.Handled Data:Keeps preprocessed documents in storage for quick and easy access.For scalability, it might make use of cloud storage or a distributed
file system.

4.Data mining Model:
The Similarity Detection Algorithm makes a comparison between the processed text and an existing document database.makes use of algorithms such as cosine similarity and file similarity.

5.Feature Extraction:
Tokenization:Divides the text into its component words or sentences.Captures semantic meaning using methods such as word embeddings.

6.Clustering:
Sorting Texts That Are Similar: Documents that are highly similar are grouped into clusters. aids in the cogent organization and presentation of results.

7.Plagiarism Report Generation:
Presentation of the Results:Gives links of the text and file results.Gives the percentage of similarity.

8.Database:
Document Metadata:Stores information about each document, upload date. Enables efficient indexing and retrieval.User Data:Stores user profiles, preferences, and history.

9.Plagiarized Reports:
Notifies users when plagiarism checks are finished.Report Archiving:Saves reports on plagiarism.

10.Logging and Monitoring:
System logs: Keep track of actions, mistakes, and user communications for auditing and troubleshooting purposes.Performance monitoring: Keeps an eye on the functionality and scalability of the system.

V. RESULTS AND DISCUSSION

1)Similarity Percentage: Discuss the overall similarity percentage of your assignment. This percentage indicates the proportion of your text that matches existing sources found by the plagiarism checker. It's essential to consider whether this similarity is due to properly cited material, common knowledge, or unintentional overlap.

2)Specific Instances: Highlight specific passages or sections flagged by the plagiarism checker. Discuss whether these instances are due to direct quotes, paraphrases, or unintentional similarities. If necessary, consider revising these sections to ensure proper citation or rephrasing.

3)Understanding Citations: Reflect on the importance of proper citation practices in academic writing. Discuss the different citation styles (e.g., APA, MLA) and how to correctly cite sources to avoid plagiarism. Emphasize the importance of giving credit to original authors and ideas.

4)Educational Value: Consider the educational value of using a plagiarism checker. Discuss how it can help students learn about proper citation practices, develop their research and writing skills, and avoid unintentional plagiarism. However, also address the limitations of plagiarism checkers and the need for critical thinking and originality in academic work.

5)Instructor Feedback: If applicable, discuss any feedback or guidance provided by your instructor based on the results of the plagiarism checker. Consider how you can use this feedback to improve your assignment and avoid plagiarism in future work.

6)Ethical Considerations: Reflect on the ethical implications of plagiarism and the importance of academic integrity. Discuss the consequences of plagiarism in academic and professional settings, as well as the importance of honesty, originality, and intellectual integrity.

By discussing these aspects, you can provide a comprehensive analysis of the results and implications of using an online assignment plagiarism checker. This discussion can help you and your peers better understand the importance of academic integrity and proper citation practices in academic writing.

VI. CONCLUSION

Plagiarism Detection:Identify instances where the assignment text closely matches or is identical to content from other sources. This includes identifying direct copying, paraphrasing, or rephrasing. Similarity Percentage:Provide a numerical value indicating the degree of similarity between the submitted assignment and existing sources. A higher percentage suggests a higher likelihood of plagiarism. Source Identification:Highlight or provide links to the specific sources or documents that match the content of the assignment. This helps instructors or evaluators to pinpoint the potential origins of plagiarism.Textual Comparison:Offer a side-by-side comparison of the submitted assignment and the identified sources, illustrating the ` and differences. This aids in manual inspection by educators. Plagiarism Report:Generate a comprehensive report summarizing the findings, including the percentage of similarity, highlight matching text, and references to potential sources. This report is often used for academic or institutional review. Algorithm Sensitivity:Allow customization of sensitivity levels to adjust the algorithm's strictness in detecting similarities. This helps in accommodating different plagiarism policies and tolerances.

VI. REFERENCES

[1] [1] Taresh Bokade, Tejas Chede, Dhanashri Kumar Prof. Rasika Shintre" Online Assignment Plagiarism Checker using Data mining and NLP"2021.

[2] "Software metrics and plagiarism detection," Cyst. Software,vol. 13.pp. 131-128, 1990.

[3] M.J Wise,"Detection of similarities in student programs:YAP'ing may be preferable to Plague'ing, ACM SIGCSE Bull.,vol.24,no.1,pp. 268-271, 1992.

[4] A. Anguita, A. Beghelli, and W. Creixell, Automatic cross-language plagiarism detection, 2011 7th International Conference
on Natural Language Processing and Knowledge Engineering, 2011.

[5] U. Bandara and G. Wijayarathna, "A Machine Learning Based Tool for Source Code Plagiarism Detection," International Journal of Machine Learning and Computing, pp. 337– 343, 2011.

[6] Sediyono, A., & Mahamud, K. (2008). Algorithm of the Longest Commonly Consecutive Word for Plagiarism Detection in Text Based Document. Digital Information Management, 253-259.

[7] Sharma, N., Bajpai, A., & M. R. (2012). Comparison the various clustering algorithms of weka tools. International Journal of Emerging Technology and Advanced Engineering , 2(5), 73-80.

[8] El-Matarawy, A., El-Ramly, M., & Bahgat, R. (2013). Plagiarism Detection using Sequential Pattern Mining. International Journal of Applied Information Systems (IJAIS) ,5.

[9] Roig, M. (2011). Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing.

[10] Atkinson, D., & Yeoh, S. (2008). Student and staff perceptions of the effectiveness of plagiarism detection software. Australasian Journal of Educational Technology, 24(2), 222-240.