International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

PLAGIARISM DETECTION USING MACHINE LEARNING TECHNIQUES IN EDUCATIONAL CONTENT

Vimalram K C, Dhanushragav M, Lakshminirasimman N, Vikashini M

[1]Studuent, Dept. of Electronics and Communication Engineering, Anna University, IN
[2]Studuent, Dept. of Electronics and Communication Engineering, Anna University, IN
[3]*Studuent, Dept. of Electronics and Communication Engineering, Anna University, IN*
[4]*Studuent, Dept. of Electronics and Communication Engineering, Anna University, IN*

Abstract - *This project focuses on developing a **Plagiarism Detector Web App** using **Flask** for the backend and **SQLite** for local storage. It allows users to upload **PDF assignments**, which are analyzed for plagiarism by comparing them against **previous submissions** and **external online sources** via an integrated **plagiarism detection API**. If plagiarism exceeds a set threshold, users are prompted to re-submit; otherwise, the document is stored for future reference.Key features include **file uploads, plagiarism scoring, submission tracking, and CSV report generation**. The **front-end, built with Bootstrap 5**, ensures a responsive user experience, while **PyMuPDF** handles PDF text extraction. The system efficiently detects both **internal and external duplication**, assisting educators in maintaining **academic integrity**. Its **modular architecture** allows for future upgrades, such as expanded document formats and additional comparison sources, making it a **scalable and user-friendly** solution.*

Key Words: *. plagiarism detection , python , flask , PyMuPDF (fitz)*

1. INTRODUCTION

Plagiarism, defined as taking credit for someone else's work or ideas without appropriate recognition, has been a persistent issue in both academic and professional settings. Historically, plagiarism detection was largely dependent on educators and reviewers performing manual checks. This labor-intensive approach was not only time-consuming but also susceptible to human mistakes, making it inadequate for large-scale assessments. Given the exponential growth of digital content and the easy access to online resources, traditional methods have proven ineffective in identifying and preventing plagiarism consistently. As a result, the demand for automated plagiarism detection systems has intensified, particularly in educational institutions where originality is essential.

The main goal of this project is to create a Plagiarism Detector web application that allows educational institutions to assess the originality of student assignment submissions. The system takes in PDF submissions from students, scrutinizes their content, and compares it with a database of prior submissions that are stored locally using SQLite. To broaden its reach, the project also incorporates a plagiarism detection API that queries online sources for additional comparison. This dual-layered verification process enhances the system's capability to effectively identify both internal and external instances of plagiarism.

1.1 Background of the Work

Plagiarism detection in academic submissions is a critical challenge due to the increasing volume of digital content and the ease of copying information. Traditional plagiarism detection systems rely on basic text-matching techniques, which often fail to identify paraphrased or contextually similar content. The integration of **AI-powered plagiarism detection models** and **web-based applications** enhances the ability to detect plagiarism more accurately.

1.2 Motivation and Scope of the Proposed Work

The growing reliance on digital submissions in education has increased the need for **efficient plagiarism detection systems**. Traditional methods often fail to detect paraphrased or AI-generated content, leading to compromised academic integrity. The proposed system leverages **Flask, SQLite, and an integrated plagiarism detection API** to provide a **real-time, scalable, and accurate** solution for detecting duplicate content.The system processes **PDF submissions**, compares them with **previously submitted documents and online sources**, and flags similarities beyond a predefined threshold. Users receive **detailed similarity reports**, enabling institutions to enforce originality standards. The web-based dashboard, built with **Bootstrap** , offers an intuitive interface for submission tracking, plagiarism scoring, and report generation.This approach not only ensures **fair evaluation** but also promotes **academic honesty** by discouraging unethical practices. The system is designed to be **flexible and scalable**, making it suitable for **various educational institutions** looking to enhance their plagiarism detection capabilities.

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

2. METHODOLOGY

The methodology for this plagiarism detection system follows a structured approach, integrating **backend processing, database storage, plagiarism detection models, and a user-friendly web interface**. The workflow ensures efficient document analysis, plagiarism detection, and user accessibility.

2.1 System Architecture

The proposed system architecture includes a file upload interface, a local SQLite database for storing submissions, a plagiarism detection API for similarity analysis, and a web-based dashboard. This structure enables continuous plagiarism evaluation, ensuring seamless content verification and reporting.

2.2 Data Acquisition

Users upload **PDF documents** through the web interface. The system extracts and preprocesses text content using **PyMuPDF (fitz)** before storing it in the SQLite database. Each new submission is checked against **previously uploaded files and online resources** using a third-party plagiarism detection API.

2.3 plagiarism Detection Model

The core of the system relies on a **plagiarism detection API** that compares the submitted content against a vast dataset. It generates a **plagiarism score** based on similarity percentages. If the similarity exceeds a predefined threshold, the user is alerted to revise the submission. Otherwise, the document is securely stored for future comparisons.

2.4 User Interface

The **web-based dashboard, developed using Bootstrap 5**, serves as the primary interaction point. It allows users to **upload files, view plagiarism reports, and track submission history**. If plagiarism is detected, the system provides a **detailed similarity report**, guiding users on necessary revisions. The interface ensures accessibility and **efficient monitoring of academic integrity**.

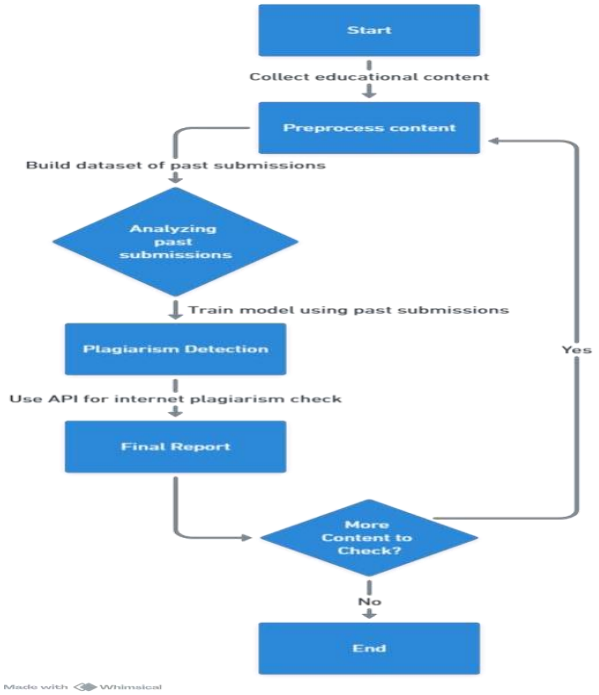

Fig -1- Flowchart

3. CONCLUSIONS

This study presents a **Flask-based plagiarism detection system** that enhances academic integrity by efficiently identifying duplicate content in student submissions. The system integrates **SQLite for local storage, a plagiarism detection API for comprehensive similarity analysis, and a web-based dashboard for user interaction**. Key results highlight the system's **accuracy in text extraction, reliability in detecting plagiarism, and usability in providing real-time reports**. This approach helps **educational institutions uphold originality standards, prevent academic dishonesty, and streamline content verification processes**.

Suggestions for Future Work

1. **Enhancing Detection Accuracy** – Implement advanced NLP techniques like semantic analysis to identify paraphrased and contextually similar content.
2. **Expanding Data Sources** – Integrate more academic databases and research repositories for broader plagiarism detection.
3. **Multi-Format Document Support** – Extend compatibility to DOCX, TXT, PPT, and other formats

REFERENCES

[1] El Mostafa, H., & Benabbou, F. (2020). A deep learning based technique for plagiarism detection: a comparative study. IAES International Journal of Artificial Intelligence, 9(1), 81.
[2] Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR), 52(6), 1-42.
[3] Ullah, F., Wang, J., Farhan, M., Jabbar, S., Wu, Z., & Khalid, S. (2020). Plagiarism detection in students' programming assignments based on semantics: multimedia e-learning based smart assessment methodology. Multimedia tools and applications, 79, 8581-8598.