PLAGIARISM DETECTION USING MACHINE LEARNING

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY:

SAHIL GULERIA (2104173)

HARNOOR SINGH (2203582)

KAUSHAL KUNDAN (2203586)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GURU NANAK DEV ENGINEERING COLLEGE, GILL PARK, GILL ROAD, LUDHIANA

Table of Contents

List of figures

Introduction

The act of plagiarism is stealing another individual's intellectual property, which comprises their literary work, creative output or ideas. The act of claiming these as ones own without due acknowledgment and consent constitutes a direct violation of the principles of academic integrity, honesty and originality. Without doubt the practice has become increasingly prevalent across myriad settings be it essay writing, research reports or artistic works a reality indicating we overlook.

A key reminder here though: while using machine learning based technologies to create content may seem like a grey area doing so without an assigned license for assessment leads it into the realm of academic dishonesty. Moreover, reusing one's own work without proper citation is still considered an act of plagiarism. In light of all this remember that exams have strict rules prohibiting instances of plagiarism which carry serious disciplinary consequences.

Learning and using the rules of good academic practice from the start of your university career is the greatest method to prevent plagiarism. Avoiding plagiarism involves using your academic skills to make your work as good as it can be, not just making sure your references are all accurate or changing enough terms so the examiner won't catch you paraphrasing.

In order to deter and fight plagiarism, educational institutions and professional communities frequently have severe regulations and sanctions in place. It is crucial to always correctly attribute and reference the sources utilized, whether through direct quotations, paraphrasing, or summarizing, as well as to adhere to the particular citation style requirements specified by the institution or publication, in order to prevent plagiarism.

To assure originality and uphold ethical standards in academic and professional work, using plagiarism detectors and developing strong research and writing practices are helpful ways.

Rationale

Plagiarism has been a problem for a long time and the problem has evolved with time. With the rise of the internet the theft of intellectual property has risen significantly and recognizing these thefts has become difficult. With this survey we have identified techniques used in detecting plagiarism. With changing times, the tools needed to detect plagiarism have to be evolved. However, to develop a tool with an ability to achieve high accuracy and greater accessibility of data has always been a demand. A comparative study on plagiarism checking tools with the technology used is presented in this paper. This study would help us determine the algorithm and methodology to proceed with the development of code to detect plagiarism

There are several benefits of using machine learning for plagiarism detection:

- **Automated Process**: Machine learning models can scan and compare massive volumes of text quickly, drastically reducing the time required for plagiarism detection.
- **Real-time Feedback**: Immediate detection and feedback help students, writers, and researchers quickly correct any issues, ensuring their work is original.
- **Continuous Improvement**: As models receive more data and feedback, they become better at detecting newer forms of plagiarism and adapting to evolving tactics.
- **Learning Aid:** Helps students understand the importance of originality and proper citation practices by providing detailed feedback.
- **Reduced Manual Effort**: Less need for manual checking and verification, saving time and resources.
- **Maintaining Standards**: Ensures that academic, professional, and creative works adhere to high standards of originality and integrity.
- **Copyright Protection:** Helps in identifying unauthorized use of copyrighted materials, ensuring compliance with copyright laws.

In summary, machine learning enhances the capabilities of plagiarism detection by making it more efficient, accurate, versatile, and scalable. It not only helps in maintaining high standards of originality and integrity but also serves as an educational tool to promote better understanding and ethical practices.

Objectives

The objectives of this project are to:

- To build a framework where the user can input their required text and get the desired result, which is getting the plagiarism percentage.
- To provide a fully automated medium to detect plagiarism from student work.
- To develop a web-based tool that is user-friendly and accessible, with an ability to achieve high accuracy and greater accessibility of data.

Literature Review

Here are summaries of some recent papers on plagiarism detection using machine learning:

1. "Plagiarism Detection in Programming Assignments using Machine Learning" By Nishesh Awale, Mitesh Pandey, Anish Dulal

   This research paper proposes a machine learning approach for detecting plagiarism in programming assignments. This calculates features like n-gram similarity, coding style similarity, and dead codes to identify similarities between source codes. Using the XGBoost model, their system predicts if pairs of source codes are plagiarized, achieving an accuracy of 94% and an F1-score of 0.905 on the test set. Their study also shows that the XGBoost model outperforms the Support Vector Machine (SVM) in this application.

2. "Online Assignment Plagiarism Checker Using Machine Learning" By Babitha V, Harshitha M, Hindumathi A, Reshma Farhin J

   This research paper discusses a plagiarism detection model designed to improve the convenience for professors in evaluating student assignments. Their approach aims to speed up and enhance the efficiency of plagiarism detection compared to traditional manual methods. The model processes the input file through stages including Tokenization, Cleaning, Stop Word Removal, and Stemming. Following these preparatory steps, the system generates a report and analyzes the degree of plagiarism in the assignments. This method reduces the effort and time required by professors, making the assessment process more efficient and less burdensome.

3. "A Deep Learning Based Technique for Plagiarism Detection: A Comparative Study" By El Mostafa Hambi, Faouzia Benabbou

   This research paper proposed a multi-layered model for plagiarism detection, which includes a preprocessing layer with word embedding, learning layers, and a detection layer. The model combines three techniques: Siamese Long Short-term Memory (SLSTM), Convolutional Neural Network (CNN), and Doc2vec, achieving an impressive accuracy score of 98.33%. By leveraging these combined approaches and utilizing internet search evidence, the model effectively detects plagiarism and determines similarity scores, showcasing significant improvements over traditional methods. Their work emphasizes the importance of an integrated approach for enhanced plagiarism detection.

Overall, these research papers illustrate significant advancements in AI and machine learning for plagiarism detection. Collectively, these studies highlight AI's potential to transform plagiarism detection, offering more precise and efficient solutions than traditional methods.

Feasibility Study

The feasibility study is a necessary part of software engineering, it is done by the engineers to understand whether the conditions are right to implement the project or not. In this assessment the practicality of a proposed project plan or methods takes place. It is done by analyzing technical, operational, and economical feasibility.

- **Technical Feasibility:** This study is carried out for understanding the technical feasibility of the project. The system has little demand for technical resources that can be fulfilled by us. The technical resources like understanding of the asp.net core language, knowledge of asp.net framework libraries, etc. along with this software like MSSQL Server for database, Visual Studio to run the code.

- **Operational Feasibility:** The project will help to users to rent car on just a click of buttons and the prices according to the given inputs which are given by the user and then user can pay and reserve the car. There are existing studies on these topics regarding the web-based car rental system. By focusing on user requirements, compatibility, operational efficiency, security, and customer satisfaction, the system can effectively meet the needs of both the car rental company and its customers. Implementing these features and considerations will ensure a robust, efficient, and user-friendly car rental system.

- **Economic Feasibility:** The project is economically feasible at the resources which are required to implement the system whether software or hardware are low in cost and their availability is there too.

Methodology

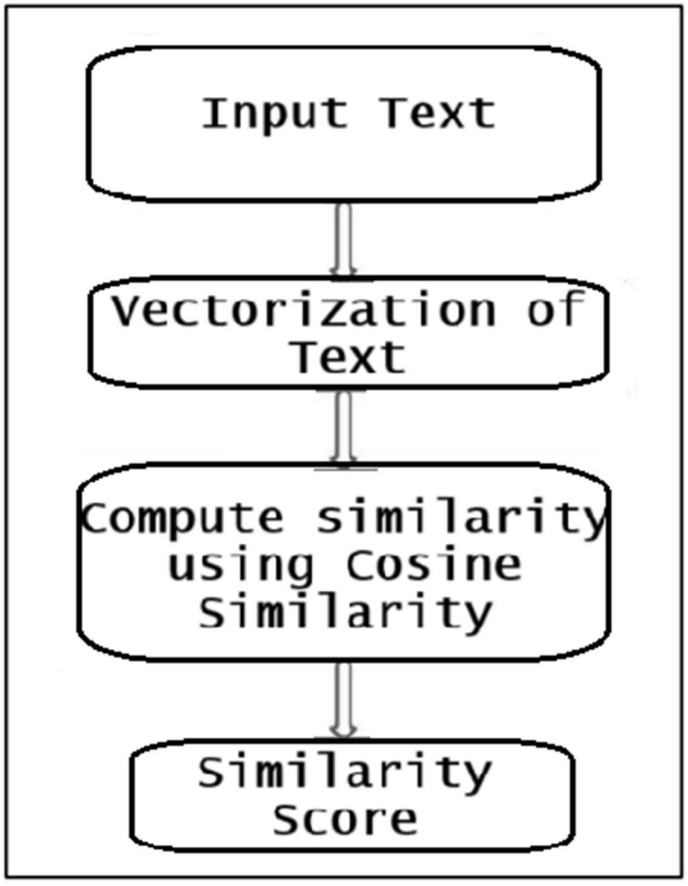

Fig 1 Flowchart of the Plagiarism Detection

Facilities required for proposed work

Hardware Required:

- **Processor:** Intel core i3 Processor
- **RAM:** Minimum 4 GB
- **Hard Drive:** Maximum 1 TB
- **Internet:** Required

Software Required:

- **Operating System:** Windows 8,10,11
- **Programming language/Framework:** Python, React
- **Libraries**: Matplotlib, Scikit-learn, Flask
- **Integrated development environment:** Visual Studio Code.
- **Web-Browser:** Microsoft Edge, Mozilla Firefox, Google Crome

Expected Outcomes

This plagiarism detector has been implemented using machine learning features like cosine similarity. The system works efficiently and detects the extent of plagiarism between the given text files. The incorporation of a user interface makes it easier for a layman to utilize the service of the system. This system can easily be used in institutions like schools and colleges to detect plagiarism in students' assignments. Operation of the system does not require any complex directions or training. It is a time efficient, easy to use, and effective plagiarism detection system

References

[1] Y. P. Ma et al., "Plagiarism Detection Based on Lexical and Syntactic Features Using Convolutional Neural Network," in Proceedings of the IEEE International Conference on Computational Science and Computational Intelligence, 2019, pp. 698-703.

[2] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer and Z. Sobhaninia, "An Improved DE Algorithm to Optimise the Learning Process of a BERT-based Plagiarism Detection Model," 2022 IEEE Congress on Evolutionary Computation (CEC), Padua, Italy, 2022, pp. 1-7

[3] El-Rashidy, Mohamed A., et al. "Reliable plagiarism detection system based on deep learning approaches." Neural Computing and Applications 34.21 (2022): 18837-18858.

[4]. Imam Much Ibnu Subroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

[5]. Upul Bandara and Gamini Wijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674- 678

[6]. Chavan, Hiten, et al. "Plagiarism detector using machine learning." International Journal of Research in Engineering, Science and Management 4.4 (2021): 152-154.

[7] Siddharth Tata, Suguri Charan Kumar and Varampati Reddy Kumar, “Extrinsic Plagiarism Detection Using Fingerprinting”, International Journal of Computer Science and Technology (IJCST) Vol. 10, Issue 4, Oct - Dec 2019.