Journal of Nonlinear Analysis and Optimization
Vol. 15, Issue. 1, No.6 : 2024
ISSN : **1906-9685**

Journal of Nonlinear Analysis and Optimization : Theory & Applications

Editors-in-Chief :

PLAGIARISM SYSTEM FOR INDIAN LANGUAGES USING PYTHON

Rahul Rai Department of Computer Science and Engineering Shri Shankaracharya Institute of Professional Management and Technology Raipur 492001, India
Ashish Pandey Department of Computer Science and Engineering Shri Shankaracharya Institute of Professional Management and Technology Raipur 492001, India
Muskan Dewangan Department of Computer Science and Engineering Shri Shankaracharya Institute of Professional Management and Technology Raipur 492001, India

Abstract —
Plagiarism is inextricably intertwined with intellectual property rights and copyright laws, both of which were created to preserve the concept's ownership. When tested with sample Bengali text, most of the available tools for detecting plagiarism failed to recognize the Bengali text, and those that did support Bengali text, as well as other texts like Telugu, etc., did a simple string comparison to detect suspected copy-paste plagiarism, ignoring other forms of plagiarism such as word switching, synonym replacement, sentence switching, and so on. Plagiarism is a major issue in the research field. It is a primary topic of academic researchers. Students often duplicate text or programming assignments from one another to make their work easier. Copying decreases the time and effort required to think about and formulate program logic, as well as the coding and debugging processes. Plagiarism, on the other hand, significantly increases the work of assessors. In this study, a two-phase design is developed that incorporates approaches from both categories to include their benefits while overcoming their limitations when used separately. Multi-language Plagiarism and content-based Plagiarism are the techniques.

Keywords — Plagiarism, word switching, synonym replacement, content-based plagiarism, multi-language plagiarism

INTRODUCTION

Plagiarism, defined as the act of passing off someone else's original words and ideas as one's own, is considered a moral as well as a legal offense. Plagiarism has a long history, since the term is derived from the Latin terms "plagiaries," which meaning abductor, and "plagiare," which means to steal (1). Plagiarism may be divided into two types: literal and intellectual. In the former, the plagiarist simply copies and pastes content from the internet. Intelligent plagiarism, on the other hand, attempts to deceive consumers by employing paraphrasing skills to make it appear as their own (6).

The "Plagiarism system for Indian languages using python" project aims to address the challenges in the plagiarism detection system in India. It addresses the need for a specialized system that can effectively detect plagiarism in a context where multiple languages and dialects are prevalent. The project recognizes the need for a specialized system that can effectively handle content in various Indian languages and writing styles.

One of the primary challenges is adapting the system to diverse educational standards within India. The project seeks to align with various academic norms and evaluation methods, ensuring that the system provides accurate assessments of originality according to the specific requirements of Indian educational institutions. This adaptability is crucial for widespread acceptance and utilization across different levels of education and professional sectors.

To create a comprehensive database for comparison, the project will integrate the system with local repositories, online journals, and educational databases. This integration ensures that the system has access to a wide range of Indian content sources, enhancing its capability to identify instances of plagiarism effectively. Furthermore, the system will address algorithmic complexity by incorporating features that go beyond simple text matching, including the detection of paraphrased content and translational similarities.

User-friendliness is a key consideration in the project's design. The system will feature an intuitive interface, making it accessible to educators, students, and professionals with varying levels of technical expertise. This user-friendly approach is essential for the successful adoption of the plagiarism detection system in diverse educational and professional settings.

Legal and ethical considerations are integral to the project's development. The system will adhere to Indian copyright laws and ethical guidelines for plagiarism detection, ensuring that it provides valuable insights without compromising privacy or violating intellectual property rights. This commitment to legal and ethical standards is vital for the system to gain trust and acceptance within the academic and professional communities.

In summary, the "Plagiarism system for Indian languages using python" project aims to create a sophisticated and culturally sensitive tool that addresses the unique challenges posed by academic dishonesty in India. By combining advanced algorithms, linguistic diversity adaptation, and a user-friendly interface, the project seeks to contribute significantly to the promotion of academic integrity and originality in the Indian educational landscape.

LITERATURE REVIEW

Plagiarism is discussed in this work. What does plagiarism entail? These publications' primary aspects are how to identify plagiarism, how to prevent it, and the distinctions between plagiarism and copyright infringement, as well as which citations to avoid. This article discusses efficient tools for detecting plagiarism and provides a foundation for determining how much work is authentic and how much is fake (1). This study described a python-based plagiarism detecting program or tool. This program is used to discover similarities in text in a variety of disciplines, including student assignments, instructor study materials, and blogger content, in order to minimize plagiarism. The availability of many materials on the internet allows students to effortlessly copy-paste, resulting in grades without knowledge background. Plagiarism is defined as "taking someone else's idea or work and passing it off as your own," which is obviously a poor behavior. This study also discussed another idea known as paraphrase. Reading the accessible information or content of others and writing it in your own knowledge is what paraphrasing is all about. (2). A research project is a unique and methodical inquiry conducted to find new facts and information about a topic. This paper discusses plagiarism. Students, Research Scholars, and Teachers can verify research work using software such as Turnitin, Ithenticate, Plagiarism Checker, Viper, Duplichecker, Copyleaks, Paperrater, Plagium,Plagiarisma, Plagscan, and others. The current study illustrated many phases of plagiarism detection as well as how to avoid duplication in research output (3). Plagiarism is becoming a more serious problem in academics. It is exacerbated by the ease of access to and cutting and pasting from a diverse variety of resources available on the internet. It is academic theft since the offender 'taken' the work of others and presented the stolen material as his or her own. It pertains to a person's integrity and honesty. It stifles creativity and originality and undermines education's mission. Plagiarism is a common and rising issue in the academic process. Traditional human detection of plagiarism is a difficult, inaccurate, and time-consuming method since it is difficult for anybody to verify with current data. The primary (4). This paper discusses plagiarism, Plagiarism Checker, and Plagiarism Detection Tools from software such as Turnitin, Ithenticate, Plagiarism Checker, Viper, Dupli checker, Copy leakage, Paperrater, Plagium, Plagiarism, Plagscan, and others. The current study made use of several plagiarism detection technologies and checkers (5). Plagiarism is the illegal act of exploiting someone else's work entirely or partially as one's own in any discipline, including art, poetry, literature, cinema, research, and other creative kinds of study. It has become a significant crime in academic and research disciplines, and the availability of a wide range of resources on the internet has exacerbated the matter. As a result, automated detection of plagiarism in text is required. This document provides an overview of several plagiarism detection strategies used for various languages (6).

PLAGIARISM DETECTION

Detecting and preventing plagiarism has become an issue in colleges and institutions. This is because most professionals and students cheat when they are given an assignment or project. This is why there are so many materials available on the internet. It is incredibly simple for anyone to use a search engine to find any article on any topic they want to copy from without examining who produced it. As a result, having plagiarism detection software to remove or limit cheating or duplicating documents has become a must or must for the entire academic platform.

There are a few types of plagiarism that are easily detectable; all we need to do is copy and paste our content into the text area provided or upload the document to any plagiarism checking websites available on the internet, and it will show the percentage of plagiarism within 4-5 seconds, or it may not be checked at all.

Plagiarism is a practice that is not limited to students. However, in their desire to be renowned, some staff members directly duplicate or change ideas from other resources in order to write papers.

There are two forms of plagiarism that have been seen to occur more frequently:

1. Textual plagiarism: Textual plagiarism happens when someone replicates someone else's content without providing credit to the creator. This is the form of plagiarism that is commonly committed by academic researchers or students, where the contents are identical to the original contents, such as essays, reports, articles, research papers, and so on.
2. Source code plagiarism Source code plagiarism, also known as programming plagiarism, is commonly committed by college or high school students. It is described as the act or tendency of utilizing, converting, modifying, or copying the entirety or a portion of the source code produced by someone else and incorporating it into your program without the owner's consent.

I. DETECTION TECHNIQUES

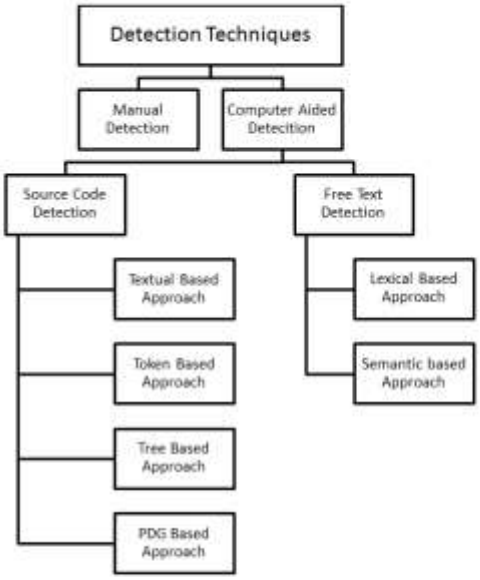


Figure 1: plagiarism detection technique

Following are the detection techniques:

a) Manual detection: entails human evaluation of papers in order to discover and assess instances of plagiarism. This approach is time-consuming since it depends on human skill to discover similarities in information.
b) Computer-Aided Detection (CAD): CAD is the use of software tools to help in recognizing probable plagiarism. Algorithms are often used in these programs to evaluate textual material and identify probable matches or similarities.

a. Source Code Detection: Source code detection is concerned with detecting plagiarism in programming code. It compares code structures using algorithms to discover instances of code repetition or resemblance.

i. Textual Approach: A textual approach compares the textual content of texts to detect commonalities. This method frequently employs techniques such as string matching and similarity scoring.

ii. Token-Based Approach: Token-based techniques divide text into smaller components (tokens) like words or sentences. This approach may find similarities and differences across documents by comparing token sequences.

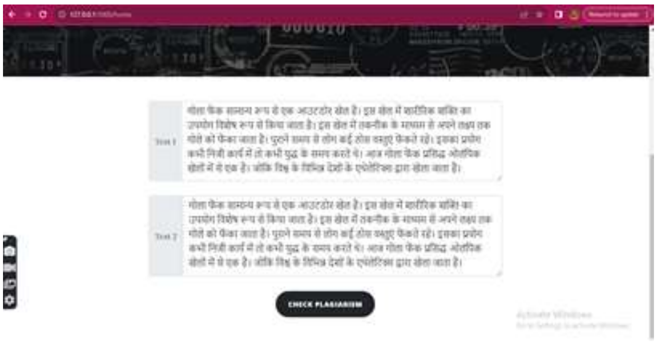
iii. PDG-Based Approach: In source code plagiarism detection, Program Dependency Graph (PDG)-based techniques are used. They depict program structures as graphs, allowing program dependencies and structures to be compared.

b. Free content Detection: Detecting plagiarism in free-form content is what free text detection is all about. This method is frequently used to find similarities in essays, articles, and other written content.

i. Lexical-Based Approach: Lexical-based techniques compare document lexical components such as words and punctuation. This approach may be used to find textual similarities while ignoring structural differences.

ii. Semantic-Based Approach: Semantic-based techniques examine the significance of material rather than just surface-level similarities. These strategies use natural language processing and semantic analysis to detect plagiarism based on context.

II. SNAPSHOTS



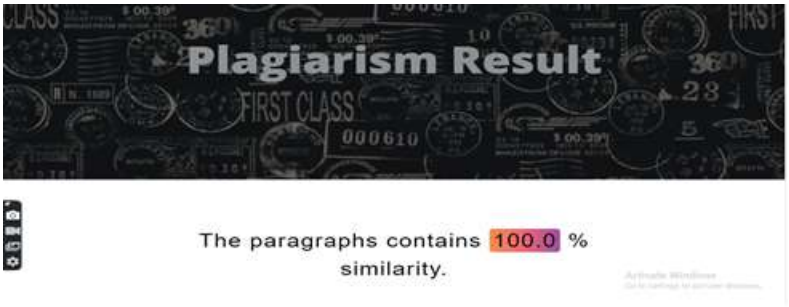
Text 1

गोला फेंक सामान्य रूप से एक आउटडोर खेल है। इस खेल में शारीरिक शक्ति का उपयोग विशेष रूप से किया जाता है। इस खेल में तकनीक के माध्यम से अपने लक्ष्य तक गोले को फेंका जाता है। पुराने समय से लोग कई ठोस वस्तुएं फेंकते रहे। इसका प्रयोग कभी निजी कार्य में तो कभी युद्ध के समय करते थे। आज गोला फेंक प्रसिद्ध ओलंपिक खेलों में से एक है। जोकि विश्व के विभिन्न देशों के एथलीटों द्वारा खेला जाता है।

Text 2

गोला फेंक सामान्य रूप से एक आउटडोर खेल है। इस खेल में शारीरिक शक्ति का उपयोग विशेष रूप से किया जाता है। इस खेल में तकनीक के माध्यम से अपने लक्ष्य तक गोले को फेंका जाता है। पुराने समय से लोग कई ठोस वस्तुएं फेंकते रहे। इसका प्रयोग कभी निजी कार्य में तो कभी युद्ध के समय करते थे। आज गोला फेंक प्रसिद्ध ओलंपिक खेलों में से एक है। जोकि विश्व के विभिन्न देशों के एथलीटों द्वारा खेला जाता है।

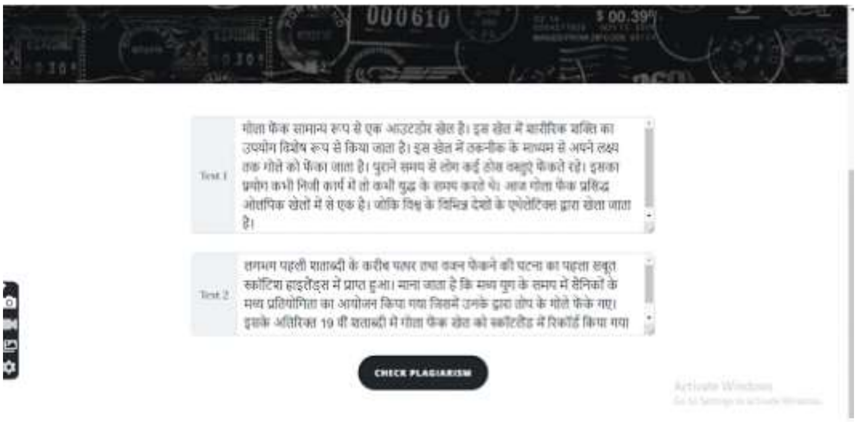
Figure 2: Hindi text used for checking plagiarism



Plagiarism Result

The paragraphs contains 100.0 % similarity.

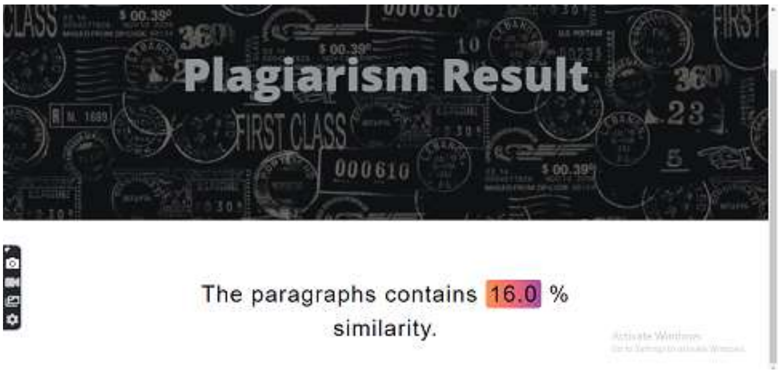
Figure 3: Plagiarism report for first Hindi text



Text 1 गोला फेंक सामान्य रूप से एक आउटडोर खेल है। इस खेल में शारीरिक शक्ति का उपयोग विशेष रूप से किया जाता है। इस खेल में तकनीक के माध्यम से अपने लक्ष्य तक गोले को फेंका जाता है। पुराने समय से लोग कई ठोस वस्तुएं फेंकते रहे। इसका प्रयोग कभी निजी कार्य में तो कभी युद्ध के समय करते थे। आज गोला फेंक प्रसिद्ध ओलंपिक खेलों में से एक है। जोकि विश्व के विभिन्न देशों के एथलेटिक्स द्वारा खेला जाता है।

Text 2 लगभग पहली शताब्दी के करीब पत्थर तथा वजन फेंकने की घटना का पहला सबूत स्कॉटिश हाइलैंड्स में प्राप्त हुआ। माना जाता है कि मध्य युग के समय में सैनिकों के मध्य प्रतियोगिता का आयोजन किया गया जिसमें उनके द्वारा तोप के गोले फेंके गए। इसके अतिरिक्त 19 वीं शताब्दी में गोला फेंक खेल को स्कॉटलैंड में रिकॉर्ड किया गया

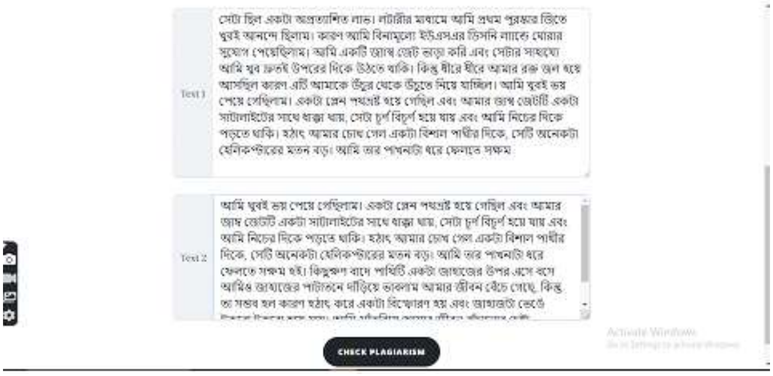
Figure 4: Another Hindi text used for checking plagiarism



Plagiarism Result

The paragraphs contains 16.0 % similarity.

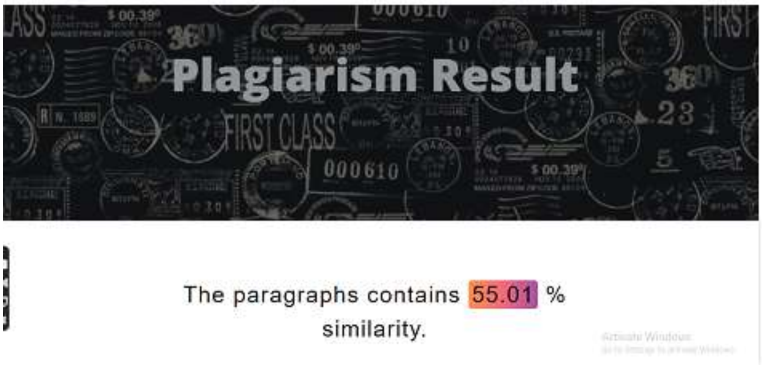
Figure 5: Plagiarism result of 2nd Hindi text



Text 1 সেটা ছিল একটা অপ্রত্যাশিত লাভ। লটারির মাধ্যমে আমি প্রথম পুরস্কার জিতে খুবই আনন্দে ছিলাম। কারণ আমি বিনামূল্যে ইউএসএর ডিসনি ল্যান্ডে ঘোরার সুযোগ পেয়েছিলাম। আমি একটি জাম্প জেট ভাড়া করি এবং সেটার সাহায্যে আমি খুব দ্রুতই উপরের দিকে উঠতে থাকি। কিন্তু ধীরে ধীরে আমার রক্ত জল হয়ে আসছিল কারণ এটি আমাকে উঁচুর থেকে উঁচুতে নিয়ে যাচ্ছিল। আমি খুবই ভয় পেয়ে গেছিলাম। একটা প্লেন পথভ্রষ্ট হয়ে গেছিল এবং আমার জাম্প জেটটি একটা স্যাটেলাইটের সাথে ধাক্কা খায়, সেটি চূর্ণ বিচূর্ণ হয়ে যায় এবং আমি নিচের দিকে পড়তে থাকি। হঠাৎ আমার চোখ গেল একটা বিশাল পাখির দিকে, সেটি অনেকটা হেলিকপ্টারের মতন বড়। আমি তার পাখনাটি ধরে ফেলতে সক্ষম

Text 2 আমি খুবই ভয় পেয়ে গেছিলাম। একটা প্লেন পথভ্রষ্ট হয়ে গেছিল এবং আমার জাম্প জেটটি একটা স্যাটেলাইটের সাথে ধাক্কা খায়, সেটি চূর্ণ বিচূর্ণ হয়ে যায় এবং আমি নিচের দিকে পড়তে থাকি। হঠাৎ আমার চোখ গেল একটা বিশাল পাখির দিকে, সেটি অনেকটা হেলিকপ্টারের মতন বড়। আমি তার পাখনাটি ধরে ফেলতে সক্ষম হই। কিছুক্ষণ বাদে পাখিটি একটা জাহাজের উপর এসে বসে আমিও জাহাজের পাটাতনে দাঁড়িয়ে ভাবলাম আমার জীবন বেঁচে গেছে, কিন্তু তা সম্ভব হল কারণ হঠাৎ করে একটা বিস্ফোরণ হয় এবং জাহাজটি ভেঙে

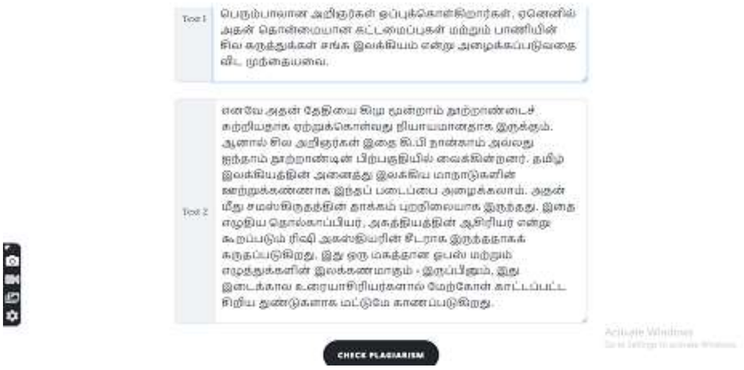
Figure 6: Bangla text used for checking plagiarism



Plagiarism Result

The paragraphs contains 55.01 % similarity.

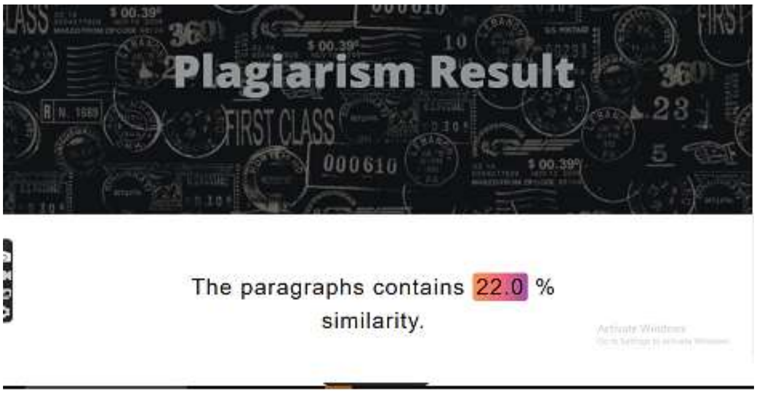
Figure 7: Plagiarism report for Bangla text



பெரும்பாலான அறிஞர்கள் ஒப்புக்கொள்கிறார்கள், ஏனெனில்
அதன் தொன்மையான கட்டமைப்புகள் மற்றும் பாணியின்
சில கருத்துக்கள் சங்க இலக்கியம் என்று அழைக்கப்படுவதை
விட முந்தையவை.

எனவே அதன் தேதியை கிமு மூன்றாம் நூற்றாண்டைச்
சுற்றியதாக ஏற்றுக்கொள்வது நியாயமானதாக இருக்கும்.
ஆனால் சில அறிஞர்கள் இதை கி.பி நான்காம் அல்லது
ஐந்தாம் நூற்றாண்டின் பிற்பகுதியில் வைக்கின்றனர். தமிழ்
இலக்கியத்தின் அனைத்து இலக்கிய மாநாடுகளின்
ஊற்றுக்கண்ணாக இந்தப் படைப்பை அழைக்கலாம். அதன்
மீது சமஸ்கிருதத்தின் தாக்கம் புறநிலையாக இருந்தது. இதை
எழுதிய தொல்காப்பியர், அகத்தியத்தின் ஆசிரியர் என்று
கூறப்படும் ரிஷி அகஸ்தியரின் சீடராக இருந்ததாகக்
கருதப்படுகிறது. இது ஒரு மகத்தான ஓபஸ் மற்றும்
எழுத்துக்களின் இலக்கணமாகும் - இருப்பினும், இது
இடைக்கால உரையாசிரியர்களால் மேற்கோள் காட்டப்பட்ட
சிறிய துண்டுகளாக மட்டுமே காணப்படுகிறது.

Figure 8: Tamil text used for checking plagiarism



Plagiarism Result

The paragraphs contains 22.0 % similarity.

Figure 9: Plagiarism report for Tamil text

LIMITATIONS AND FUTURE SCOPE

The "Indian Plagiarism System using Python" project lays the foundation for a robust plagiarism detection system tailored to the unique challenges of India. The future scope of the project encompasses several potential avenues for enhancement and expansion:

1. Integration of Additional Indian Language: Expand language support to include more Indian languages, ensuring that the system becomes even more inclusive and capable of detecting plagiarism across a broader linguistic spectrum.
2. Enhanced Machine Learning Algorithm: Further refine and expand machine learning algorithms to improve the system's ability to identify complex patterns, including advanced paraphrasing and translation-based plagiarism.
3. Real Time Database Update: Implement a mechanism for real-time or frequent updates to the plagiarism database by seamlessly integrating with new academic and professional content, keeping the system's knowledge current and comprehensive.
4. Collaboration with educational Institute: Establish collaborations with educational institutions, publishers, and research organizations to promote the adoption of the system as a standard tool for ensuring academic integrity.
5. Development of an API for integration: create an Application Programming Interface (API) that allows seamless integration of the plagiarism detection system with existing learning management systems, content management systems, and other educational platforms.
6. Research collaboration for continuous advancements: Foster collaboration with research institutions and experts in the fields of natural language processing, machine learning, and plagiarism detection to stay abreast of the latest advancements and incorporate cutting-edge technologies.

By pursuing these future enhancements, the "Indian Plagiarism System using Python" project can evolve into a comprehensive and indispensable tool for maintaining academic integrity and promoting originality in a rapidly evolving educational and professional landscape.

CONCLUSION

In conclusion, the "Indian Plagiarism System using Python" project has successfully tackled the pervasive issue of plagiarism in India's academic and professional domains. By incorporating advanced natural language processing algorithms, adapting to the linguistic diversity of Indian languages, and emphasizing cultural and educational nuances, the system offers a robust and culturally sensitive solution. Its adaptability to diverse educational standards, integration with local content sources, and incorporation of advanced plagiarism detection techniques contribute to a comprehensive tool. With a user-friendly interface catering to various skill levels and a commitment to legal and ethical considerations, the project not only addresses the immediate need for plagiarism detection but also promotes a culture of academic integrity and originality in the Indian context. This project marks a significant stride toward enhancing the quality and authenticity of academic and professional content creation in India.

REFERENCES

1. Ganesh Kumar Soni, “Plagiarism Detection and Prevention: A Study,”. International Journal of Library & Information Science, 7(1)
2. Dr. JP Patra, Kavya B, Shrishti swarnakar, Nandita Sahu, “Plagiarism checker and Paraphrasing tool. In Wesleyan Journal of Research,” (May 2021).
3. Anil Kumar Jharotia, “Plagiarism Detection Through Software in Digital World”.
4. R. R. Naik, M. B. Landge, C. N. Mahende, “A review on plagiarism detectiontools,” International Journal of Computer Applications 125 (11)
5. Sudhir S Patil, Dr.Hemant Yeole ,“Overview of Plagiarism Checkers and Plagiarism Detection Tools: A Study,”
6. Harshall Lamba, Sharvari Govilkar, “A Survey on Plagiarism Detection Techniques for Indian Regional Languages,”
7. https://github.com/orvil1026/Plagiarism-in-Hindi