Plagiarism Detection

Prof. Vrushali Awale[1], Sumit Bambole[2], Sujal Fethfulwar[3], Drushya Bhaisare[4], Guatam Kumar[5]

Department of Computer Science Engineering[1-5]

Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India

Abstract: *We are in Information age and there is lots of information available over internet. Most of the information over internet is produced as duplicate by the different authors. In order to maintain credentials of the original work of the author and improve the quality of research, plagiarism checking software may be useful. With help of plagiarism checking software, we can avoid the duplicacy in research and academic writings. This paper gives an overview of various effective plagiarism detection methods that have been used for plagiarism detection to identify that how much work is original and how much work is illegal and copied from others' original works*

Keywords: plagiarism detection; machine learning; artificial intelligence; deep learning; neural network.

I. INTRODUCTION

In this age of information technology, the Internet has brought tremendous changes to the academic and research world. The Internet has had a dramatic effect on the way information is created, organized, accessed, and disseminated. Now, information can be accessed easily through databases, the Internet, electronic lists, and various other formats. In the academic world, though academic dishonesty is not a new phenomenon, it is widely acknowledged that access to online information has made it easier and more tempting to copy or quote another person's original work without attribution.

Plagiarism can be considered as one of the electronic crimes, like (computer hacking, computer viruses, spamming, phishing, copyrights violation and others crimes). Plagiarism defined as the act of taking or attempting to take or to use (whole or parts) of another person's works, without referencing or citation him as the owner of this work. It may include direct copy and paste, modification or changing some words of the original information from the internet books, magazine, newspaper, research, journal, personal information or ideas.

According to the Merriam-Webster Online Dictionary, to "plagiarize" means:

- To steal and pass off (the ideas or words of another) as one's own.
- To use (another's production) without crediting the source.
- To commit literary theft.
- To present as new and original an idea or product derived from an existing source.

Also according to Turnitin.com, plagiarism.org and Research Resources, considered plagiarism as:

Turning in someone else's work as your own.

Copying words or ideas from someone else without giving credit.

Failing to put a quotation in quotation marks.

Giving incorrect information about the source of a quotation.

Changing words but copying the sentence structure of a source without giving credit.

Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on "fair use" rules).

II. MOTIVATION

The practice of plagiarism is not a strange thing anymore. Students, programmers and even lecturers plagiarize or copy the source code from different sources before submitting it to the evaluator. This isn't just limited to schools and colleges but also in the industries. Detecting plagiarism practices is a solution that should be done so that the fraudulent actions can be minimized. Detection of plagiarism clusters is very important to find out how many students or groups accomplishing program homework independently. It gives instructors more opportunity to enhance or modify education. Using the software can be a deterrent for students to plagiarism. However, using this software does not

provide the final answer, which is why the authors have come up with an idea of using two approaches and comparing them to find out which performs better, the first technique is to use machine learning techniques like XGBoost algorithm, SVM classifier and the second technique is to make use of Artificial Neural Network and backpropagation on the dataset to identify if there is plagiarism in code

III. LITERATURE REVIEW

The literature surrounding plagiarism detections underscores their pivotal role in the academic landscape, with an increasing emphasis on technological solutions to counteract the rise of plagiarism. Scholars have extensively explored the methodologies underpinning these tools, investigating the efficacy of various algorithms in detecting similarities and ensuring the accuracy of results. Considerations of precision and recall metrics have been central, aiming to strike a balance between identifying instances of plagiarism and minimizing false positives.

Furthermore, researchers delve into the challenges posed by evolving forms of plagiarism, such as paraphrasing and mosaic writing, urging the continual refinement of detection mechanisms to address these nuances.

The impact of plagiarism detections on educational practices is a recurring theme in the literature, as educators grapple with maintaining academic integrity in a digital age. Studies delve into the integration of these tools into educational workflows, discussing their influence on teaching strategies and student learning outcomes. Scholars also scrutinize the ethical implications of plagiarism detection, considering issues of privacy, fairness, and the potential for unintended consequences. The literature reveals an ongoing discourse on the role of plagiarism detections in fostering a culture of originality, discouraging academic dishonesty, and shaping pedagogical approaches.

While the literature acknowledges the progress made in plagiarism detection technology, it consistently advocates for continued research and development. Comprehensive evaluations of existing tools, comparative analyses of different systems, and investigations into their impact across diverse academic disciplines remain crucial focal points. The evolving nature of plagiarism and the dynamic academic landscape necessitate ongoing efforts to enhance the precision, adaptability, and user-friendliness of plagiarism detections, ensuring their effectiveness in upholding academic integrity.

IV. METHODOLOGY

A plagiarism detection typically employs a multifaceted methodology to identify and flag potential instances of plagiarism within a given text. The process involves several key steps.

Firstly, the system utilizes advanced algorithms to break down the text into smaller units, such as sentences or phrases. These units are then compared against an extensive database containing a wide range of academic papers, articles, books, and other sources. The comparison involves analyzing the linguistic and structural features of the text units, looking for similarities with existing content. Sophisticated algorithms assess not only exact matches but also variations in wording and sentence structure. Additionally, the system considers the contextual relevance of the identified similarities to distinguish between common phrases and potential instances of plagiarism. Furthermore, many plagiarism detections integrate machine learning techniques to enhance their accuracy over time. They continuously learn from new data, adapting to evolving language patterns and ensuring the system remains effective against various forms of plagiarism.
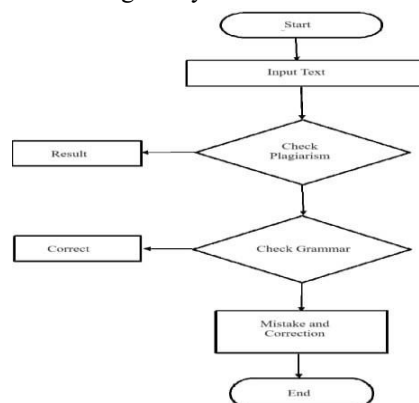
Overall, the methodology combines linguistic analysis, pattern recognition, and database comparisons to provide a comprehensive assessment of the text's originality. It's important to note that while plagiarism detections can efficiently identify potential matches, human judgment is often necessary to interpret the results and determine whether the similarities constitute plagiarism or permissible overlap in content

V. WORKING

A plagiarism detection is a tool designed to identify and flag instances of plagiarism in written content. Its functioning involves several key steps. First, the user submits the text or document to be checked. The plagiarism detection then breaks down the content into smaller units, such as sentences or phrases, and creates a unique fingerprint for each unit using algorithms Next, the tool compares these fingerprints against a vast database that contains a wide range of sources, including academic papers, articles, books, and websites. The database may be compiled from publicly available information or licensed from various publishers and online repositories. The comparison helps identify similarities between the submitted content and existing sources.

Plagiarism detections use advanced algorithms, including text-matching and natural language processing techniques, to assess the level of similarity. Some tools provide a percentage of similarity, indicating how much of the content matches existing sources. Additionally, they often highlight specific passages or sentences where similarities are found, allowing users to review and address potential plagiarism to enhance accuracy, plagiarism detections may employ machine learning algorithms that continuously learn and adapt to evolving patterns of plagiarism. Regular updates to the database also ensure that the tool remains effective in identifying content from new publications and online sources.

Institutions such as educational organizations, publishers, and businesses use plagiarism detections to uphold academic integrity, protect intellectual property, and maintain content authenticity. It's essential to note that while these tools are valuable in identifying potential plagiarism, human judgment is crucial for interpreting results and considering factors like proper citation and fair use. Overall, plagiarism detections play a vital role in promoting ethical writing practices and ensuring the originality of content in various domains.
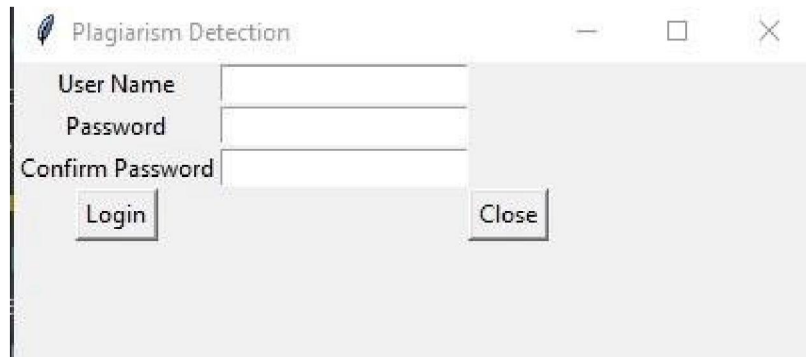
VI. CONCLUSION

Plagiarism is a crime. However, unintentional plagiarism is to be taken on another level. With the development of web based resources, plagiarism detection is now as easy as plagiarizing a document. Most of the cases of plagiarism are the result of lack of awareness about the copyright rules and the ethics related to the use of information. The students consider copying and pasting information from the Internet for writing assignment easy and convenient and they are not aware that it is against copyright law. Proper information literacy training is essential to change the situation. Plagiarism, however, is to be detected and discouraged at all levels. Making the information users aware about the judicious use of information and pattern of acknowledging even at the stage of learning is essential.

Librarians have a responsibility in building up a better scientific community by involving themselves in Information Literacy endeavors. They must educate and instruct students on ethical issues involved in the research process. As mentors, librarians must teach students the research process and citation styles. They should inculcate ethical values among the readers, and teach them how to give credit to the ideas of other scholars. Librarians must generate a publish-or-perish mentality in researchers by uprooting plagiarize-and-publish mentality. The awareness about the plagiarism detection software will help to avoid plagiarism to a great extent
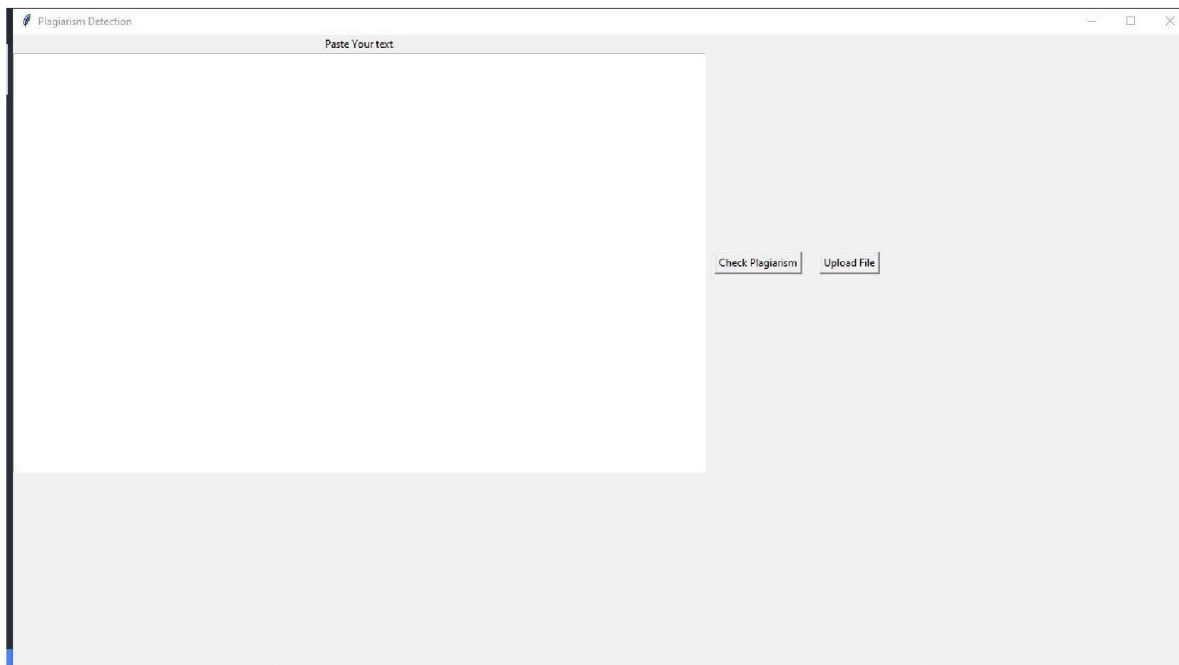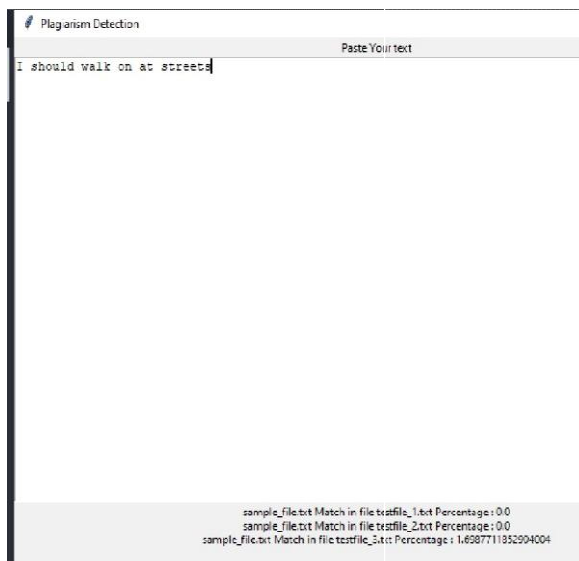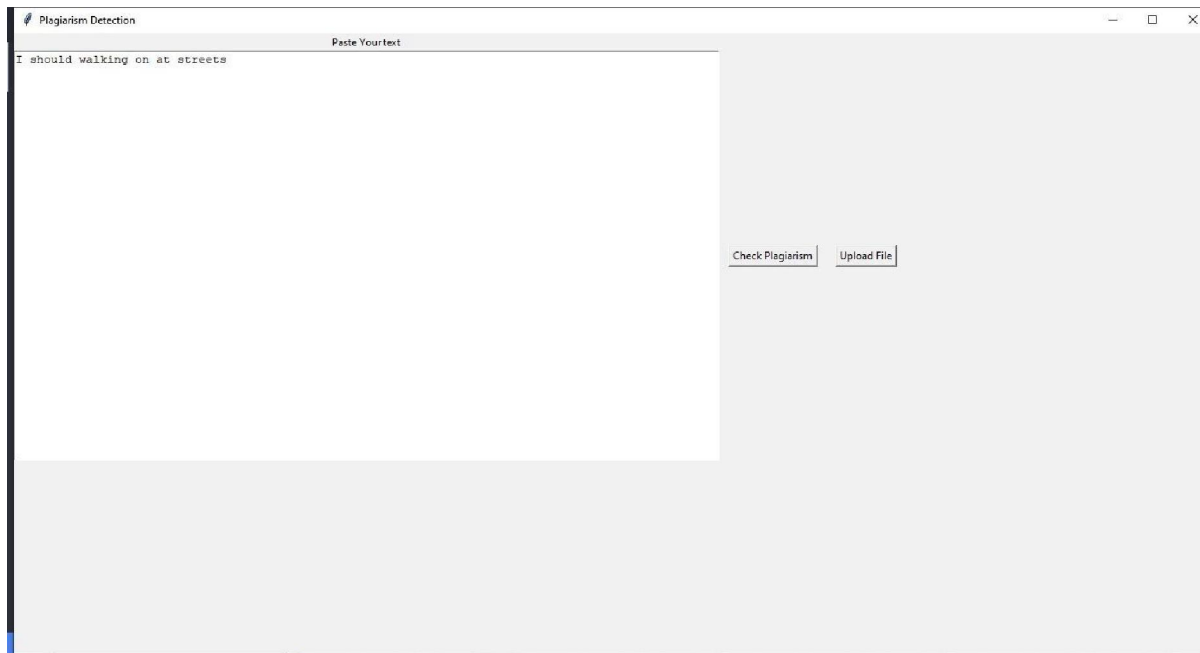
VII. OUTPUT

REFERENCES

[1]. http://www.ala.org/ala/mgrps/divs/acrl/standards/informationliteracycompetency.cfm#stan (Accessed on 22nd November, 2010).

[2]. Longman dictionary of contemporary English. 5th ed. (2009). England. Pearson Education, 1320.

[3]. www.wikipedia.org (Accessed on 24th November,2010)

[4]. http://www.scanmyessay.com (Accessed on 10th April, 2015).

[5]. http://www.plagiarismdetect.com (Accessed on 10th April, 2015).

[6]. http://www.dupliDetection.com (Accessed on 10th April, 2015).

[7]. http://www.plagiarismDetection.com (Accessed on 10th April, 2015).

[8]. http://copytracker.ec-lille.fr ((Accessed on 10th April, 2015).

[9]. http://www.plagium.com (Accessed on 10th April, 2015).