IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 3, June 2023

Plagiarism Checker using Machine Learning

Prof. Vanita Buradkar[1], Rutuja Chawale[2], Janhvi Rangankar[3],
Saptashrungi Chaudhari[4], Vijaya Zade[5]

Assistant Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4,5]
Rajiv Gandhi College of Engineering, Research & Technology, Chandrapur, Maharashtra, India.

Abstract: *“Plagiarism, the act of taking the writings of another person and passing them off as one’s own. The fraudulence is closely related to forgery and piracy-practices generally in violation of copyright laws.” Encyclopedia Britannica . 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 MerriamWebster On Line Dictionary, to “plagiarize” means 1)to steal and pass off (the ideas or words of another) as one's own 2)to use (another's production) without crediting the source 3)to commit literary theft 4)to present as new and original an idea or product derived from an existing source. "Technology has been both a miracle and a curse in terms of plagiarism. No doubt, it has become easier to find the required information and copy it. Since people often do that without attribution, it has also become easier to identify and deal with plagiarism." With free plagiarism checker tools that can search billions of documents, and find matches even if they are only a few words in length, finding plagiarism has become as easy as detecting information in Google. It is now only a matter of merely processing your query and giving you the results. "Plagiarism definition is actually straightforward. When you use someone else’s work without crediting them, it is seen as stealing their intellectual property. Just like theft, the penalties for plagiarized work are also severe all over the world. The real problem is that most people are not even aware of what they are doing." By looking the above parameter we are proposed a mechanism which will helps to workout for plagiarism checker application which will helps the peoples to find the identical content with certain papers so that the proposed mechanism will helps the people to go for there unique content*

Keywords: Plagiarism

I. INTRODUCTION

Plagiarism is defined as to take or theft some work and present it has one’s own work. This grammar and plagiarism checker system is used to analyse the plagiarism data. Plagiarism affects the education quality of the students and thereby reduce the economic status of the country. Plagiarism is done by paraphrased works and the similarities between keywords and verbatim overlaps, change of sentences from one form to other form, which could be identified using WordNet etc. This plagiarism detector measures the similar text that matches and detects plagiarism. Internet has changed the student’s life and also has changed their learning style. It allows the students to get deeper in the approach towards learning and making their task easier. Many methods are employed in detecting plagiarism. Usually plagiarism detection is done using text mining method. In this plagiarism checker software, user can register with their basic registration details and create a valid login id and password. By using login id and password, students can login into their personal accounts. After that students can upload assignment file, which will further divide into content and reference link. This web application will process the content, visit each reference link, and scan the content of that webpage to match the original content. Also, students can view the history of their previous documents. Teacher also able to check the grammar mistakes on the content and semantically plagiarism.

DOI: 10.48175/IJARSCT-11420

ISSN
2581-9429
IJARSCT

2. LITERATURE REVIEW AND RELATED

Work Internet and the World Wide Web have revolutionized information sharing and searching; it is difficult to remember what academic research was like without it, or how we could possibly live without it again. It is an awesomely powerful resource, and therein lays the rub. Never have the phrases, “With great power comes great responsibility,” and “Power corrupts and absolute power corrupts absolutely,” been so true. We could be talking about many issues with that introduction, anything from e-retail to hacking and virus dissemination, but it turns out that we’re talking about academic dishonesty. In particular we’re talking about plagiarism, or passing off another person’s work as your own. Penn State University considers plagiarism to occur when an individual

- submits a paper written by someone else,
- quotes or paraphrases another paper without proper citation, or
- presents another person’s ideas without attribution.

How serious a problem is this? A report in the Journal of Higher Education stated that 75 percent of college students admit to some form of cheating and half admit to serious cheating on written assignments (D. McCabe, L. Trevino, and K. Butterfield, “Academic Integrity in Honour Code and Non- Honour Code Environments: A Qualitative Investigation,” J.

Higher Education, vol. 70,no.2,1999). An article in the Daily Pennsylvanian quotes McCabe as saying, “Students are growing up with technology that makes Internet plagiarism simple. It is easy to use, and almost all written sources are available on the Internet Some students actually believe that they’re not doing anything wrong. They have this attitude that they’re doing research. They don’t think that they need to cite because everything on the Internet is public information” (M.C.

Peterson, “Download. Steal. Copy. Cheating at the University,” Daily Pennsylvanian, 27 November 2001). These are astonishing numbers, particularly considering that we have identified only a handful of cases at our campus, the Great Valley Graduate School. The question is, how many cases are we missing? So this is clearly a big problem for universities and schools, but it is not just an educational problem. The publishing industry, and periodical publications in particular, face the same issues. Two recent editorials addressed the increasing occurrence of plagiarism ,simultaneous submission (submitting the same work for review to multiple venues simultaneously), and republication (submitting previously published work without sufficient attribution), and how these practices are in breach of the copyright form signed by all authors, and the integrity at the heart of academic publishing

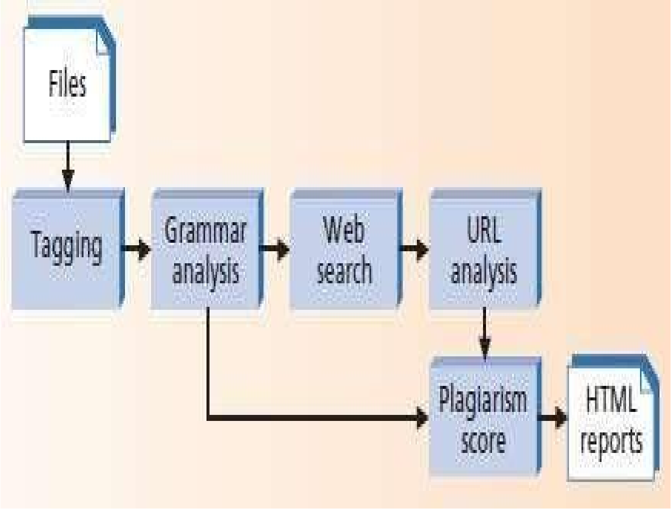
(R.L. Haupt ,“Plagiarism in Journal Articles,” IEEE Antennas and Propagation ,Aug. 2003, vol. 45, no. 4, p. 102; and W.R. Stone, “Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong!” IEEE Antennas and Propagation ,Aug. 2003, vol. 45, no. 4, pp. 47-49). In fact, combating this worry some trend has become a significant focus for all IEEE periodicals, which shows in changes at Manuscript Central, the IEEE’s online article submission system. Manuscript Central now requires authors to explicitly state that the submitted work is theirs, that every author contributed directly in its writing, and that the work has not appeared previously in another publication and is not already in submission elsewhere. Under a proposed new policy, the IEEE will ban authors caught plagiarizing from publishing for up to five years, and will identify the offending article in the archives as such. Returning to higher education, universities around the world are ramping up their efforts against plagiarism. Penn State has implemented academic integrity boards to rule on contested cases and to set standard sanctions for infractions, to enforce the inclusion of the university’s policy on academic integrity on every course syllabus that students receive, and to ensure that professors read that policy at the start of each course, including how the policy works within the context of that course. We have worked hard on the education and policy side of the problem at Penn State, but the issue of detection has not received enough attention; as we said, we have discovered only a handful of cases at our campus. The Internet is the offenders’ biggest weapon. Fortunately, the professor can also wield the plagiarizer’s weapon of choice. If students use search engines to find the material to copy, professors can use them to find those original sources. Unfortunately this can be very time consuming. The student is only working on one paper, but the professor must grade and verify all of the students’ papers. To do this by hand can take more than an hour per paper, and this is in addition to the standard grading activities. Thankfully, services and tools are available to aid in the fight. We have also developed our own tool.

DOI: 10.48175/IJARSCT-11420
ISSN 2581-9429 IJARSCT

III. AVAILABLE TOOLS

Turnitin (http://www.turnitin.com) is the most well known plagiarism detection service. Originally plagiarism. com, it is a commercial service that I Paradigms developed for registered individual educators or institutions. Professors and teachers submit papers to the site and receive results a day or so later . The site compares papers against an index of Internet content as well as large databases of “paper-mill” essays (essays available for purchase on the Web for use by students as term papers at school and university) and previously submitted papers. Recently a student at McGill University challenged using this service partly on the grounds that Turnitin subsequently adds the paper to its in-house database of material, constituting an economic benefit to Turnitin without compensation to the student. Despite this potential setback, the Joint Information Systems Committee (JISC), an organization representing all higher education institutions in the UK, recently selected Turnitin as its plagiarism detection service in the form of http://www.submit.ac.uk. Word CHECK (http://www.wordchecksystems.com) is a stand-alone application that detects collusion between students in a course rather than plagiarism of external source material To use the application, the professor or teacher loads all the documents into the system’s internal archive. The system compares all papers to detect copying within the class The document comparison is based on keyword profiles (a type of linguistic fingerprint) and phrase matching. Although this system is not strictly detecting plagiarism it could if the internal archive includes paper mill essays and similar content. Unfortunately, a 2001 review of the tool commissioned by the JISC (http://online.northumbria.ac.uk/faculties/art/information_stud ies/Imri/Jiscpas/docs/jisc/Detection_ Technology.pdf) found its detection performance unsatisfactory EVE2 (http://www.canexus.com/eve) is a commercial application that downloads to a user’s desktop and determines if students have copied material from the Internet. For each paper, the application generates a report, including the percentage that contains plagiarism, the list of URLs, and an annotated copy of the paper with the copied sections highlighted in red. The tool accepts several file formats, including plain text and Microsoft Word documents, but it will only generate annotated copies of papers from the plain text files. Essentially the tool is an interface for Web searches, but this simplicity does not limit its effectiveness. The only drawback, which the 2001 JISC report identifies, is that the searches are only against HTML Web content, and much of the material on the World Wide Web is in alternate formats. W Copy Find (http://plagiarism.phys.virginia.edu/Wsoftware.html) is a freely available collusion detection tool that Prof. Lou Bloomfield developed at the University of Virginia . At least two versions exist, but the most useful version features a simple graphical interface that lets users load the set of documents into the internal archive of the tool, just as with Word CHECK. The tool compares these documents against each other and optionally against a separate archive of files (that the professor might have collected over the years) for matching phrases. The tool presents the results as HTML files and hyperlinks common phrases between documents to indicate which students in a class were colluding. Although it cannot search against Internet content, the tool is fast, very easy to use, and the results are clear

IV. PLATFORM ARCHITECTURE

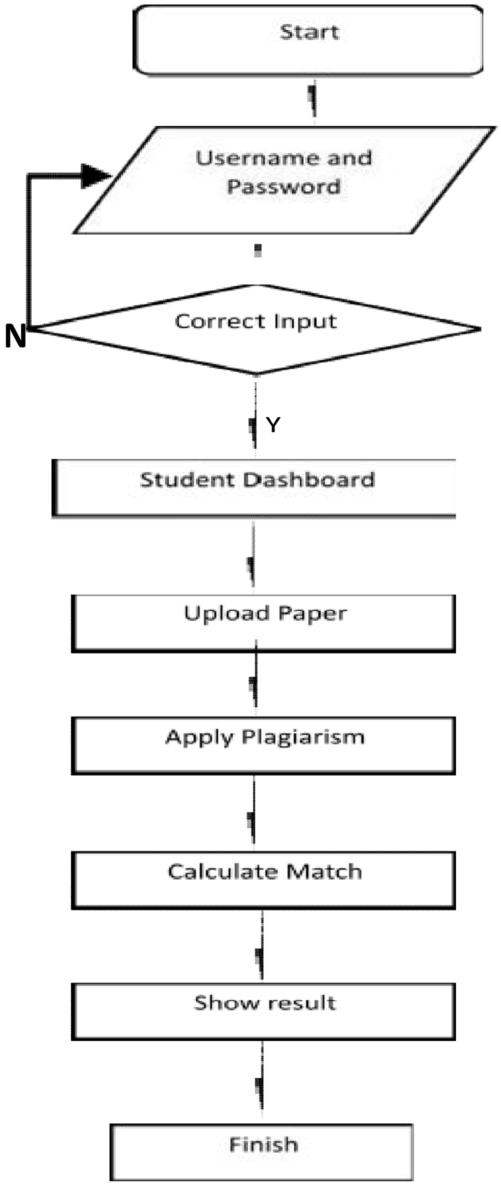


DOI: 10.48175/IJARSCT-11420
ISSN 2581-9429 IJARSCT

V. OVERVIEW

Plagiarism can be defined as "the use or close imitation of the language and thoughts of another author and the representation of them as one's own original work" . Many students make (intentionally or unintentionally) some type of cheating and plagiarism in their assignments. Usually it is difficult for teachers to detect plagiarism in student assignments by hand. The detection process becomes easier, faster and more efficient if it's performed automatically (i.e. via a computerized system). The plagiarism detection "is the processes of locating instance of plagiarism within a document wither its text or code" . Detection can be either manual or computer-assisted. Manual detection requires effort, and it is impractical in cases where many documents to be compared. Modular Object-Oriented Dynamic Learning Environment (Moodle) is an open source Course Management System (CMS) , it's used by educators as a dynamic web site for their students; it's best tool to manage learning. Moodle can support large deployments and hundreds of thousands of students, so it can be used for schools. It can be used to conduct fully online courses while others used it to augment face-to-face courses. Some teachers use Moodle to deliver material to students, assignments and quizzes, that makes learning more richly and collaborative.

VI. FLOWCHART



VII. WORK PROPOSED

The main goal of the project is to create an online plagiarism detection system that can be used by teachers at SGBAU to detect cheating and plagiarism cases in student submitted assignments. The system should be integrated as web based solution which will be user friendly and helpful for integration with any platform.

VIII. OBJECTIVE

The project objective mentioned in Section 1.2 can be achieved by specifying the following goals:
1. Studying, analysing, and comparing existing open-source plagiarism detection software and algorithms.
2. Picking one of the analysed algorithms to be used in this system, or designing a new one from the scratch.
3. 1. Studying, analysing, and comparing the existing open-source plagiarism detection software and their algorithms.
4. a. Searching for related algorithms and open- source software.
5. b. Studying and analysing these algorithms.
6. c. Comparing the algorithms and identifying their advantages and disadvantages.
4 . Picking one of the analysed algorithms to be used in this system, or designing a new one from the scratch. a. Based on the comparison among the analysed algorithms, the best (most suitable / adaptable one to the system requirements) will be picked to be implemented in the system.

IX. IMPLEMENTATION

Each and every system designed has its own peculiar principle upon which it works, but there is one principle which is common to all systems. For instance, every system (Computer application program) designed must work on the principle of human operator, input, output, and process. The peculiarities or differences exist only in the role of human operator and in what, how and when data is supplied as input and processed as output. When a system is designed on paper by an engineer and fabricated or put into an actual physical form, it has to be tested and implemented into the actual work it is designed for. The process of producing or manufacturing a new system does not just start and end there, still even after having the system doing what is it produced for properly, the producer (engineer) needs to accompany his product with a written detail of how it works, how it is operated and how it is maintained. That written detail, in a form of a book, is what is called manual. In Software engineering too, the same thing is applicable. When a program designed, written, compiled, debugged and deployed, it is tested and implemented and finally, a detail written explanation of how it is works and how to use it, is coupled with it as help instead of manual.

Advantage:

1. Proposed system will help us to detect duplicate content before releasing paper
2. Provides the similarity percentage.
3. Useful for writing original pieces within a short time.
4. Helps in checking your paraphrasing prowess
5. Helps in staying within the regulatory and ethical limits.

X. CONCLUSION

Development of any country is mostly originated from academic research institutions or centres ; these are places where most innovations, inventions are originated. Plagiarism discourages this development. This study provides means of detecting plagiarism both on almighty internet and saved documents on user’s local folders to discourages this unethical academic behaviour , self-cheating and gear up in uplifting the academic integrity of academic institutions and facilitates the development of our under-developed countries.

REFERENCES

[1] Sanjay Goel, Deepak Rao et al.: Plagiarism and its Detection in Programming Languages, December 15, 2005.
[2] http://www.php.net , (accessed: 23/10/2009)
[3] http://docs.moodle.org/en/Talk:About_Moodle, (accessed: 23/10/2009)

[4] Manuel Freire, Manuel Cebrian and Emilio del Rosal: An Integrated Source Code Plagiarism Detection Environment, Escuela Politecnica Superior, Universidad Autonoma de Madrid, 28049 Madrid, Spain.

[5] Shared Information and Program Plagiarism Detection, Xin Chen, Brent Francia, Ming Li, Brian Mckinnon, Amit Seker, University of California, Santa Barbara, December 13, 2003.

[6] Paul Clough: Plagiarism in natural and programming languages: an overview of current tools and technologies, Department of Computer Science, University of Sheffield, July 2000.

[7] JPlag: Finding plagiarisms among a set of programs, University at Karlsruhe D76128 Karlsruhe, Germany

[8] Manuel Freire, Manuel Cebrifn and Emilio del Rosal Escuela: AC: An Integrated Source Code Plagiarism Detection Environment, Polit_ecnica Superior, Universidad Aut_onoma de Madrid.

[9] http://software.bioinformatics.uwaterloo.ca/SID/servlets/Index Page , (accessed: 18/11/2009)

DOI: 10.48175/IJARSCT-11420

ISSN
2581-9429
IJARSCT