International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Online Tool for Plagiarism Detection

1st Ojas Patil
Department of Computer Science Engineering
Shrama Sadhana Bombay Trust's College of Engineering & Technology.
Jalgaon, India
patilojas1@gmail.com

2nd Sanobar Shaikh
Department of Computer Science Engineering
Shrama Sadhana Bombay Trust's College of Engineering & Technology.
Jalgaon, India
sanobar.sk20@gmail.com

4th Vaishnavi More
Department of Computer Science Engineering
Shrama Sadhana Bombay Trust's College of Engineering & Technology.
Jalgaon, India
vaishmore17@gmail.com

3rd Kunika Jadhav
Department of Computer Science Engineering
Shrama Sadhana Bombay Trust's College of Engineering & Technology.
Jalgaon, India
kunikayug02@gmail.com

***Abstract*— The goal of exploration/research and goal of knowledge production is the discovery of facts and the production of knowledge aiming to improve the human situation while doing research using Scientific corruption is a result of using unethical or improper methods, which is detrimental to the creation of new scientific knowledge. [1] It is widely acknowledged that plagiarism violates the publishing ethics of most academic scholars. As a part of their academic requirements, Indian students put a great deal of effort into finishing their projects. By bringing together all project works undertaken at various levels in colleges, we will have a great source of information to be viewed and shared, as well as encourage each student community to take on unique/innovative projects.**

The idea of an integrated platform for projects is an excellent one as it would bring together the students (also of various universities and colleges) on a common platform. This would help in the sharing of ideas and knowledge and would also promote collaboration among the students. Thus, an integrated platform for projects would be a very good idea and would be of great help to the students as well as the universities and colleges with a facility for Plagiarism.

Keywords—plagiarism, automated tool, textual detection, solutions, concepts, causes.

I. INTRODUCTION

The idea of an integrated platform for projects is an excellent one as it would bring together the students (also of various universities and college/s) on a common platform. This would help in the sharing of ideas and knowledge and would also promote collaboration among the students. Such a platform would also be very helpful for the students as they would be able to get access to a wide range of projects and would also be able to benefit from other students' experiences. This would help them in planning and executing their own projects in a better way.

The platform can be used by students to post their projects and can also be used by universities and colleges to advertise their projects. This would help in getting more people involved in the projects and would also help in getting more ideas for the projects. The platform can also be used by the students to seek help from experts in the field. This would help them in getting the right guidance and would also help them in avoiding any mistakes. [1][2] Thus, an integrated platform for projects would be a very good idea and would be of great help to the students as well as the universities and colleges. The act of plagiarism is the deliberate use of someone else's work or ideas without giving them credit. This can take many forms, from copying and pasting someone else's work without attribution to paraphrasing or summarizing someone else's ideas without giving them credit. Plagiarism is a form of cheating, and it can have serious consequences for both the person who commits it and the person whose work is plagiarized.

II. THE PREVALENCE OF PLAGIARISM

A. *Plagiarism and Academia [2]*

This is a very common thing that happens in Indian colleges and universities. Sometimes, even the professors are not sure about the originality of the project work. This is a very big problem in Indian education. There are many ways to ensure that the project work is original.

- One way is to create a database of all the projects that have been submitted by the students. This way, the professors can check the originality of the work before approving it.
- Another way is to create a plagiarism checker tool. This tool can be used to check the originality of the work. If the work is found to be plagiarized, the student can be asked to re-submit the work.

B. *Scope*

The idea has been divided into two parts. The first part looks into the current scenario of the project work in Indian universities/colleges and the second part checks if the work is plagiarized or not. Project work in Indian universities/colleges is a part of the academic requirements and students have to put a lot of effort into it. The work might be imitated and plagiarized if a common knowledge platform is not created to bring all project work to the same.

There are many reasons why project work in Indian universities/colleges is vulnerable to plagiarism.

- The project work is not well defined. There is no uniformity in the format and structure of the project work.
- The project work is not well monitored. There is no central authority to check the quality of the project work.
- The project work is not well publicized. There is no platform to showcase the work of the students.

III. CENTRAL REPOSITORY

This proposed system aims to highlight the importance of students in Indian universities/colleges who put a lot of effort into the projects as a part of the academic requirements. A common knowledge platform can help avoid duplication of effort and improve quality. There is an urgent need to document and disseminate information on successful projects undertaken by students in Indian universities/colleges. A central repository of such projects can be of immense help to students, academicians, and the industry. This will help students in identifying good quality projects and also save their time and effort in searching for similar projects.

As of right now, this central repository has both benefits and drawbacks. The main drawback will be plagiarism. As a result, we began to reflect, and we decided to add a magnificent feature to this project: a facility for plagiarism.

This proposed system mainly checks plagiarism between college projects. Plagiarism is a practice that is done by many students while making Major/Minor projects in smart ways, thus they do not seem like it is copied from anywhere, and It has become a challenging task for a teacher. There are a few types of plagiarism that can be detected easily, but document plagiarism checking it is sometimes missed much content because of less copied data from the document, which result in not giving an accurate percentage of plagiarism. There is some probability in the future that natural calamities may occur or who knows about the unlikely events that may occur.

IV. LITERATURE REVIEW

A. *Plagiarism*

Plagiarism comes from the Latin word "plagiarus", which means kidnapping. Big Indonesian Dictionary defines plagiarism as "plagiarism that infringes copyright". Meanwhile, plagiarism is the act of copying or stealing another person's work, such as ideas or writings, and claiming them as your own without mentioning the original source. In addition to word-by-word plagiarism, word switch plagiarism, metaphor plagiarism, idea plagiarism, and self-plagiarism, plagiarism is divided into three categories according to the number of words taken or traced.

For example,

a) Light Plagiarism: < 30%.

b) Medium Plagiarism: 30% - 70%.

c) Heavy Plagiarism: >70%

B. *Text Pre-Processing*

Text preprocessing is the process of cleaning and preparing text data for analysis. The process involves several steps to transform raw text data into a structured format that can be analyzed using machine learning or natural language processing.

Typical steps in text preparation include the following:

1. Tokenization: This involves splitting the text data into individual words or tokens.
2. Lowercase: Converting all the text data to lowercase helps to eliminate the differences between the same words written in different cases.
3. Stop word removal: Stop words are common words that do not add any significant meaning to the text. These words are removed to reduce the dimensionality of the data and improve the processing time.
4. Stemming and Lemmatization: Stemming involves reducing words to their base form by removing suffixes and prefixes, while lemmatization involves reducing words to their base form based on their morphological analysis.
5. Removing punctuation and special characters: This involves removing all the special characters and punctuation marks from the text data
6. Indexing is a process done to build an index database of document collections. C. TF-IDF Weighting TF-IDF (Term Frequency-Inverse Document Frequency) is a weighting scheme used in information retrieval and text mining to evaluate the importance of a term (word or phrase) in a document or corpus of documents.

C. *TF-IDF Weighting*

It is a numerical measure that reflects how important a word is to a document in a collection or corpus. TF-IDF consists of two main components:

1. *Term Frequency (TF):* It measures the frequency of a term in a document. It is computed by dividing the total number of terms in the document by the frequency with which each phrase appears.

   TF = *(Number of times a term appears in a document) / (Total number of terms in the document).*

2. *Inverse Document Frequency (IDF):* It measures the rarity of a term in a collection or corpus of documents. It is calculated as the logarithm of the total number of documents in the collection divided by the number of documents that contain the term.

IDF = *log_e(Total number of documents in the collection / Number of documents that contain the term)*

$Wt,d = TFt,d \times ln\ (\ N\ /dft)$

Where Wt,d is the value of the weight of the word t in document d. The value of TFt,d is the frequency of the word t in document d. N is the total document and dft is a lot of documents containing the word t.

Once the TF and IDF values are calculated, they are multiplied to get the final TF-IDF weight for a term in a document. The higher the TF-IDF weight of a term, the more important it is to document.

D. *Cosine Similarity*

Cosine Similarity is a method for measuring the level of similarity between two vectors. Calculations in this method are done by calculating the Cosine value between two vectors in Equation.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Eaquation I

The cosine similarity between two documents (or text vectors) can be calculated using the following formula:

Cosine similarity = *(Σi Xi Yi)* / (√(Σi Xi^2) √(Σi *Yi^2))*

Where Xi and Yi are the TF-IDF weights of the ith term in the two documents, respectively. TF-IDF is a commonly used weighting scheme that measures the importance of a term in a document, based on its frequency in the document and rarity in the corpus. The formula calculates the cosine of the angle between the two vectors, which reflects the similarity between the two documents.

A cosine similarity of 1 indicates that the two documents are identical (i.e., have the same set of terms with the same weights), whereas a cosine similarity of 0 indicates that the two documents are completely dissimilar (i.e., have no common terms). This formula can be extended to calculate the cosine similarity between multiple documents or text vectors.

In this case, the formula can be written as cosine similarity = *(Σi Xi* Yj) / (√(Σi Xi^2) *√(Σj Yj^2))* where Xi and Yj are the TF-IDF weights of the ith term in the first document and the jth term in the second document, respectively. This formula computes the cosine similarity between all pairs of documents or text vectors in a collection or corpus and can be used for various clustering or classification tasks.

Overall, the cosine similarity formula is a powerful tool for measuring the similarity between documents or text vectors and can be used in a variety of research applications that involve natural language processing or information retrieval.

V. REQUIREMENT COLLECTION & IDENTIFICATION

A. *Requirement Elicitation*

Requirement Elicitation has been done by interviewing the main Actor (Institute/Teacher) and making some conclusions about factors affecting the Plagiarism Checker Tool. But it is not practically possible to comprise all the factors concluded from the conversation because of the unavailability of data for certain factors.

After Interviewing Actor(Institution/Teacher) some of the concluded factors affecting Plagiarism Checker are:

- Data Set
- Previous year's documentary of academic project.

1) Project Feature:

- Actor(Institute/Teacher) will get idea about the projects taken up by Student, it must be different from previous projects that has been uploaded by student.

2) Operating System:

- Operating System: Windows 7 or later/Linux/MacOS
- Any system with at least 2GB RAM
- Google chrome recommended web browser for useof website.

3) Assumption:

- The variables (Factors) used for prediction willgive accurate result.
- The collected data used for prediction is presumed to be consistent.
- All of the equipment will be in working condition throughout project life-cycle.

4) Functional Requirement:

Functional requirements are the functions which are expected from the software or platform. Identification of unmet requirements is aided by functional requirements and requirement analysis. They assist in precisely defining the desired system behaviour and functionality. Below Table shows functional requirement for this project.

- User should able to input the required input data.
- Checking Plagiarism Check the duplicity in data using plagiarism tool by using input of user.
- View Data and user should able to select format of output data required.

5) Non-Functional Requirement:

Non-functional Requirement is mostly quality requirement. That stipulates how well the portal does, what it has to do. Other than functional requirements in practice, this would entail detail analysis of issues such as availability, security, usability and maintainability.

- ***Response Time:*** Normal customer response shall be less than 5sec for all website.

- ***Safety:*** Information should be safely delivered to the server without being altered.
- ***Availability:*** If the internet service gets disrupted while sending information can be send again for verification.
- ***Usability:*** The system is simple to use and navigates without any delays in the manner that is intended. When that occurs, the system programme responds appropriately and promptly switches between its states.

B. Hardware & Software Requirement

Hardware interface: It is web based product. The hardware on which it resides will be any computer can have internet. The necessary hardware calls for the following setups on the system:

- ***Processor***: Intel Pentium or above
- ***RAM***: 1 GB
- ***Input device***: Standard Keyboard and Mouse.
- ***Output device:*** VGA and High Resolution Monitor.

Software interface: This project is web based model so only browser with internet connection is required from user standpoint. This product will utilize various software components for its web based functionality. The necessary hardware calls for the following setups on the system:

- Frontend: HTML, CSS, JS
- Backend: Django, Python
- Database: MySQL

VI. DESIGN

A. System Architecture

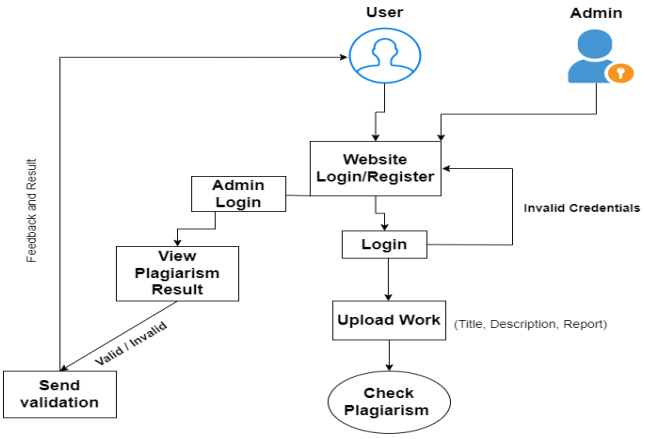
A generic discipline termed "systems architecture" is used to handle "systems"—objects that now exist or will be created—in a way that encourages analysis of their structural characteristics. The conceptual model of a system's structure, behaviour, and other aspects is called the system architecture. A formal description and representation of a system is an architecture description. It provides broad understanding of the portal. In the system architecture database provide the functionality like get information, select criteria, etc to users

Figure 6.1. System Architecture

An architectural diagram is a visual representation that maps out the physical implementation of components of a software system. In this proposed system whenever the user(students) wants to submit their project work to the college and check the validation of their work, they provide the input to the web page. Then uploaded work goes to the admin side panel then the admin sends a validation response to the user. If their workis considered valid then it checks for plagiarism with previous project work in the database. If the uploaded work is an innovative idea for the respective college then it is stored in a central repository for the college.

B. Level 2 DFD

A level 2 data flow diagram offers a more detailed look at the processes that make up an information system than a level 1 DFD does. It can be used to plan or keep track of a system's precise composition. Figure 6.2 shows Level 2 DFD of the proposed system.



Level 2 DFD

Figure 6.2 Level 2 DFD

A. Use Case

The scope and high-level functions of a system are described in use-case diagrams. The interactions between the system and its actors are also depicted in these diagrams. Use-case diagrams show what the system does and how the actors utilise it, but they do not show how the system works within. Figure 6.3 shows Use Case of the proposed system.

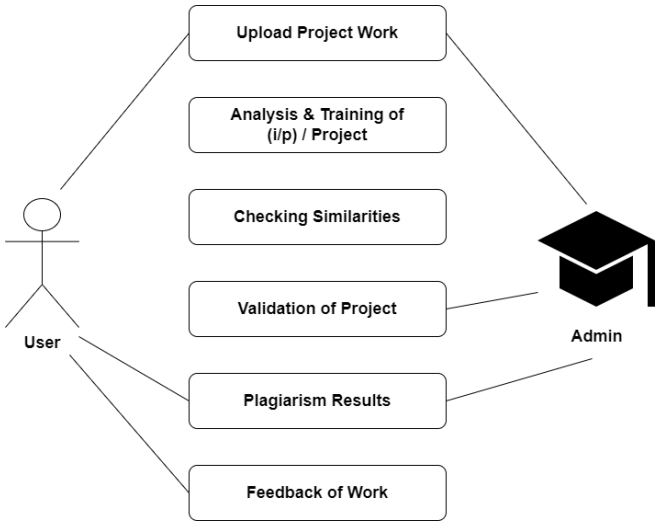


Figure 6.3 Use Case Diagram

VII. RESULT & DISCUSSION

A. *Result*

Result after uploading the input PDF file on the User Dashboard. This is the snapshot of the student/user received feedback after processing and computing of the plagiarism detection feature.

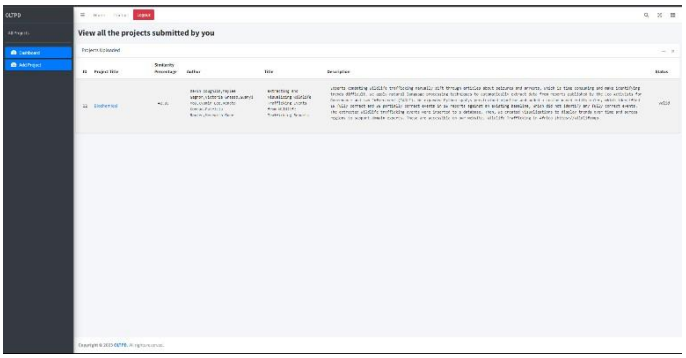


Figure 7.1 View All Projects Page

In Figure 7.2 is the User Dashboard where he/she has to upload the file and fill the required details to view the above Figure 7.1 result.

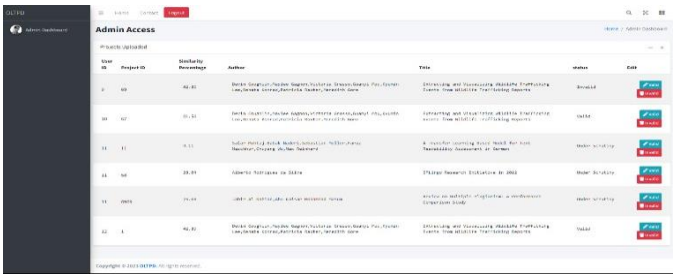


Figure 7.3 Admin Dashboard

B. *Discussion*

In order to calculate the cosine similarity, the proposal for a plagiarism detection approach takes into account vectors of every word in an entire document. The documents have been retrieved from Kaggle, where almost seven hundred thousand files related to research papers are assembled. We've compared the outcome of cosine similarity to various similar techniques, such as Jaccard comparison. The results of comparing Cosine Similarity to the second are found to be more favourable.

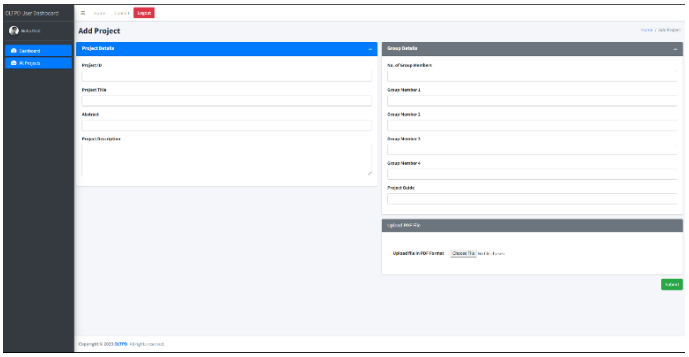


Figure 7.2 User Dashboard

In Figure 7.3 is the Admin Dashboard where the admin can see the uploaded project work of all the students and after tallying with the given threshold he/she can take action whether to make the uploaded work valid or invalid for submission. If the validation is pending, the user and admin will see the status as “Under Scrutiny”.

VIII. CONCLUSION

While working on projects as part of their academic requirements, Indian university and college students should be aware of plagiarism risks.

- Shared work should be cited and credited correctly when sharing with others.
- It is possible that all project work could be imitated and plagiarized with a common knowledge platform.
- Hence, a central repository will also help in

reducing the probability of plagiarism. A central repository of projects will also help academicians in identifying good quality projects.

REFERENCES

1. Dr. J P Patra Kavya B(2019) Plagiarism Checker And Paraphrasing Tool Using Python

2. Syed Shahabuddin(2009) International Journal of Teaching and Learningin Higher Education:PlagiarisminAcademia

3. Hoseinpourfard M.J(2012) Plagiarism: Concepts, Factors and Solutions