International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 6; June 2026
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Similarity Detection System for Assessments

Kadiyala Swathi[1], Maka Jaya Narendra Sai[1], Mandava Sai Likhitha[2], Akunuri Sri Nandini[3], Kolusu Naga Swarajya Lakshmi[4], Mallavalli Dola Rajendra[5]

[1,2,3,4,5]Department of Information Technology,
Dhanekula Institute of Engineering & Technology, Vijayawada, Andhra Pradesh, India.

Corresponding Author: Kadiyala Swathi, Maka Jaya Narendra Sai

DOI: https://doi.org/10.52403/ijrr.20260605

ABSTRACT

In an educational environment, it has become increasingly challenging to ensure the originality of students' responses in the use of online platforms for assessment submission. Manual checking of each answer to determine a comparison for similarities takes time and may even lead to inconsistency. A Similarity Detection System for Assessments, that can be considered as the proposed idea is a web-based tool, for instructors to upload an assessment and students to submit their answers typed in and in the designated time to avoid it.

It follows the step-by-step process for the similarity detection method. All submissions made by students are checked for similarity against the already submitted answers for that particular assessment. With the application of Natural Language Processing (NLP) techniques, such as text preprocessing, TF-IDF vectorization and Cosine similarity calculation for better similarity scores, submissions are either accepted, flagged or rejected based on the predefined threshold for similarities. Clear dashboards also help to display similarity percentages and submission status for each assessment.

The system provides equal, timely and automated evaluation of online assessments.

Keywords: Similarity Detection, Digital Assessments, Plagiarism Detection, Natural Language Processing (NLP), TF-IDF, Cosine Similarity, Web-Based Evaluation System, Sequential Comparison, Automated Assessment, Machine Learning

1. INTRODUCTION

Assessment is an integral part of education since it aids teachers to gauge students' understanding and application of what was learned in the class. With the rapid advancement of technology, most schools and colleges are taking their exams and assignments to online systems. Digital assessment system not only brings ease and flexibility to educators as well as students but also creates new concerns over originality and integrity in examination.

In a digital setting when majority of the students turn in their submissions to the online portal, detecting similarities or copies from their answers would be tedious if checked manually. A teacher could spend hours scrutinizing papers and yet find small common points in sentences, structure, and ideology. The standard plagiarism checkers look for similarities with content online or available documents. Therefore, they cannot compare the answers within an assessment.

To solve this problem, we propose an original and innovative system of Assessment Similarity Detection. The system is web-based whereby an educator

can create tests and students can submit text typed answers to the system within a specific timeframe. Unlike comparing papers randomly with one another, the proposed system compares one answer at a time to what has already been submitted to a specific test; this creates more structured and precise comparisons.

The system uses elementary techniques of Natural Language Processing for calculating the similarity between answers. First, it cleans and preprocesses the text. Then it converts it into a vector form using TF-IDF method and similarity is computed using cosine similarity. Finally, the similarity level can be adjusted to automatically accept, reject or flag for teacher examination depending on what the pre-set similarity limit is. Both teachers and students would have access to an intuitive interface to review the similarities found thereby maintaining trust and transparency of the entire system.

2. PROBLEM STATEMENT

With increased application of digital systems in the academic assessments and assignments, universities have encountered various new issues in managing the credibility of the education provided and academic integrity. On the positive note, online assignment/exam systems have allowed for the seamless submission of student assignments, and easy management of assignments by the teacher. On the other hand, these systems have allowed the students to cheat easily by sharing answers and plagiarism through easy coping and slight modification of other's response. Since students submit their assignment at the same time it is hard to find similarity between answers manually.

Most often the teachers are required to go through each assignment and then compare the submitted assignments from memory and/or their observation skill, which takes a large chunk of their time and effort, particularly for larger classes. This process can also lead to an error, as the judgment may vary based on person to person and context. Most traditional plagiarism detection systems are available to compare the submitted work against web contents or already published articles/contents. These are not built to perform comparison among student's submissions during an assignment evaluation. Due to these reasons, it is evident that the need for such automated similarity detection system for an assessment is high. A well-structured system of plagiarism check must provide degree of similarity among submissions for the students and instructors to refer.

3. OBJECTIVES

The primary objective of this research is to develop a feasible Similarity Detection System for digital assessments, emphasizing its ability to be Fair, Transparent, and honest, which have become highly crucial for online assignments. As submissions on the web are widely used it is essential to seek advanced alternatives to manual verification of answers. This research aims at overcoming challenges that educators encounter in the manual detection of similarities in numerous online assignments.

The objectives of the proposed system are:

- To construct a web interface for assessment creation and answer submission, with a time limit enforced.
- To apply a stepwise verification method, whereby the latest answer is verified only against the previously submitted answers of the particular assessment.
- To employ rudimentary Natural Language Processing techniques (text cleansing, TF-IDF, cosine similarity) to compare the similarity between answers.
- To formulate an automated system that categorizes the answers as accepted, flagged, or rejected based on defined similarity thresholds.
- To present dashboards allowing both students and teachers to visualize the similarity percentages and submission states.
- To establish a robust backend system that can efficiently manage parallel submissions without lag.

- By attaining these objectives, the research will lead to the development of a dependable and scalable system for efficient and consistent digital assessment.

4. LITERATURE REVIEW

The increasing use of online learning and computer-based assessment tools has created many challenges for academic integrity. Many universities and learning institutions now use web-based submission, and much research has been directed towards detecting plagiarism and assessing text similarity computationally. Typically, plagiarism detection tools rely on comparing submitted work to either internet sources or existing databases of academic work, however very little attention has been paid to identifying similarity among submitted works during the same assessment session.

A very common approach to assessing text similarity is the Vector Space Model. Salton and Buckley described the Term Frequency-Inverse Document Frequency technique, whereby it is possible to calculate the importance of a particular word within a given document with reference to a set of documents [1]. This technique is still widely used as it is effective in converting documents to vector representations which can then be compared computationally. Cosine similarity can be used to evaluate the similarity between documents. This function calculates the angle between two document vectors, which in turn is an indicator of their overlap. Manning, Raghavan and Schtze define cosine similarity as: "the cosine of the angle between two vectors ... An angle of 0 implies that the documents are identical in terms of what they discuss" [2].

The use of Natural Language Processing (NLP) techniques to pre-process text can further aid text similarity analysis. NLP operations, such as tokenisation (splitting text into words and sentences), removal of 'stop words' (e.g. And, the, that), and stemming (reducing words to their root form) help to eliminate non-essential words from the text before calculation. Jurafsky and Martin observed that these techniques improve text representation and performance of classification models [3]. Within automatic plagiarism detection literature, the use of structure-based comparison and threshold-based classification has been emphasized. Clough discusses a range of techniques that are used to detect similarities in both natural and program language and stresses the need for automated detection mechanisms [4]. Despite a firm basis of techniques within text similarity analysis, many current systems only consider checking text against external sources and lack consideration of a structured comparison within a single assessment. An assessment-based similarity detection system that fosters fair and transparent assessment practices, while remaining scalable and efficient, needs to be implemented. The system described herein builds on these natural language processing and text similarity techniques to present a structured method of comparison relevant to the requirements of a computer-based assessment tool.

5. EXISTING SYSTEM AND ITS LIMITATIONS

Most digital assessment platforms today were mainly built to help instructors distribute their assignments and gather student responses easily. Such platforms are often characterized by functionalities such as deadline settings, marking functionality and storage of submitted documents. While they aid to convenience administrative processes, such platforms often lack built-in functionalities that allow for checking the similarities among answers within the same assessment.

In order to address problems of plagiarism, many institutions will turn to external plagiarism detection systems. External plagiarism detection systems generally compare submissions against internet material, academic papers, or academic databases. Although useful for the detection of copied content from other resources, they

are not designed for comparing papers submitted by students for the same assessment or test; therefore, a student who copies or marginally reworks an answer by another student is unlikely to be detected.

A key problem with the currently used systems is the fact that marking is manually done. The assignment must be assessed in a series of submissions, which can be extremely cumbersome for large classes. Such manual checking can be time consuming and may not be consistent. Moreover, students are not often provided with a measure of similarity and may be unaware of what caused the marked similarity.

Many existing solutions are not characterized by ordered comparison (e.g. Sequential comparing) and no automation of threshold-based classification; consequently, for a large number of students, an even higher number of submissions has to be handled manually and can't offer enough support.

6. PROPOSED SYSTEM

We are presenting an online system which could resolve the common issue faced in online assessment – simple and reliable checking for similar answers. Normally, many teachers take a considerable amount of time checking student responses in written mode, and it becomes tiring as well as subject to variability. This system is based on a systematic comparison strategy together with basic machine learning techniques in order to simplify the task of assessment, making it fairer and transparent.

6.1 Web-Based Assessment and Submission Portal:

This system is a web-based portal connecting the teacher and the student at one place. The teachers would upload tests/assignments and provide details like questions, assignment name, date and time, submission limits and similarity bounds. The students can log in, select the test/assignment and submit their response within the given duration. After submissions, the submission time will be recorded and the response will be saved safely in the database.

6.2 Sequential Similarity-Based Evaluation Approach:

An important feature of our system is the approach in which it compares the submission with each other. In our system, it is based on a sequential approach. As soon as a submission is made for a test/assignment, the system compares this particular submission with all previously submitted responses for that same test/assignment.

6.3 Machine Learning-Based Similarity Analysis:

In order to evaluate the similarity of text-based submissions, we employ basic text analysis techniques. First, text cleaning is done by removing stop words and unnecessary characters. Then, text-to-vector conversion is performed using TF-IDF algorithm. After vector conversion, similarity is calculated using cosine similarity. It then represents a percentage indicating how similar one answer is to the other.

6.4 Automatic Accept, Flag or Reject Mechanism:

Based on the similarity percentage obtained for each submission, the system automatically decides its status. An answer with low similarity will be directly accepted, while that with similarity in between would be flagged for the teacher's manual check, and it would be rejected for a very high similarity. This process is automated to save the teachers time and reduce the possible inconsistency in judging subjective aspects of assignments.

6.5 Transparent Results Dashboard:

Both students and the teachers will get a clear overview in the respective dashboard. Students can check their submission's similarity percentage and status whereas teachers can view all the submissions in detail and check cases that have been flagged.

6.6 Backend Processing and Scalability:

The backend of the system will process the submitted documents sequentially ensuring

all the checks are being carried out efficiently. This allows it to handle multiple concurrent submissions in larger classes and maintain stable performance. Data is stored in a database and access is controlled in order to secure it.

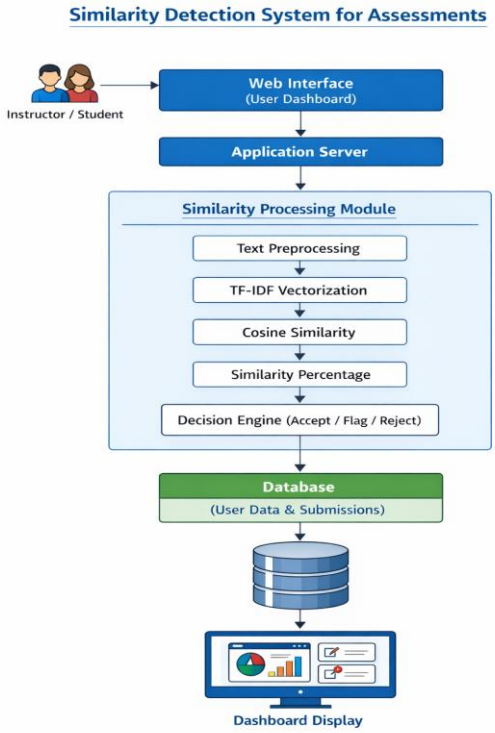
SYSTEM ARCHITECTURE

The proposed Similarity Detection System has been architected in a manner which ensures smooth operation between each component of the system, linking users, processing modules, and the database, in an organized manner. This system architecture is divided into different layers with the system's respective components residing in each layer. These components perform the required function in the specific layer, making the entire system highly efficient and allowing it to handle multiple users simultaneously.

At the top most layer is the presentation layer; where instructors and students interact with the web-based interface. Through this web-based interface, the instructor creates the assignments and its respective deadlines and can view the similarity report. Students view the status of their submission and submit their answers through the same interface. This layer mainly deals with the display of relevant information to the users.

Below the presentation layer is the application layer, which manages the entire functioning of the system. This layer includes user login, assignments creation and deadlines management, submissions management etc. It effectively acts as a mediator between the interface and the system's backend processing and is responsible for storing and processing all student submissions appropriately and sending the processed data for similarity checking.

The similarity processing layer is the core functionality of the system where a student's submitted text is initially pre-processed, eliminating unnecessary text, formatting, stop words, etc., thereafter being converted to a numeric format using the TF-IDF vector space model. The cosine similarity of the new submission with already existing submissions for the particular assignment is computed, along with a similarity percentage being generated and used along with pre-determined threshold limits, to determine if the submission needs to be rejected, flagged or accepted.



Finally, the database layer is where all the data required by the system is stored. The data included user, assignment related, students, and the computed similarity results are stored in a central database to enable quick and secured access of information.

This layered design ensures the effective management and organization of the system, the processing being done step by step as the system receives and processes the submissions and it can manage multiple users

IMPLEMENTATION DETAILS

The Similarity Detection System is implemented in a structured step-by-step manner to ensure accuracy and efficiency:

Step1: System Setup
The initial setup involves the fundamental design of the web application. This includes designing the UI using HTML, CSS and JavaScript for the front-end and developing the back-end using any of the programming languages such as Python, Node.js or Java and setting up a database. The database is used for storing all user information such as students, assessments and answers given by students.

Step 2: User Authentication Module
For security and different permissions, a different log-in system is established for students and instructors separately to distinguish between the two. This module ensures that unauthorized users do not get access to the system.

Step 3: Assessment Creation
The instructor can add all the details for an assessment, such as questions, deadlines, similarity limits, etc. Then, this information is stored in the database.

Step 4: Student Submission
The students can log in to the system and then type their answers and submit them before the submission time. The exact submission time of the answers is also recorded. This is then stored in the database.

Step 5: Text Preprocessing
Once the answer is submitted, the data undergoes processing where common words are removed from the text and punctuation, unnecessary formatting etc. Are also taken care of. Also, it is broken down into components (tokens) that are used for comparisons.

Step 6: TF-IDF Conversion
TF-IDF values are calculated and then applied on the cleaned text to convert the written text to numerical vectors.

Step 7: Cosine Similarity Calculation
Using these values, the newly submitted answer is compared with the other already submitted answers in that assessment using cosine similarity to calculate a score.

Step 8: Similarity Percentage
The obtained score is then converted into percentage so it's easier to understand the extent of similarity.

Step 9: Decision Making
Based on the percentage of similarity, the decision whether to accept or reject the answer is made:

- Low similarity – accept the answer.
- Moderate similarity – Flag the answer and ask for review.
- High similarity – Reject the answer.

Step 10: Dashboard Update and Notification
The decision made and the similarity percentage are shown on the student as well as instructor dashboard. If the answer is flagged then an alert is also sent to the instructor.

Step 11: Handling multiple submissions
This part ensures that the backend can handle several submissions at once without lagging, one after the other.

EXPERIMENTAL SET-UP

To determine how well our proposed Similarity Detection System works, an experiment was carried out, aiming to mimic a real classroom environment as much as possible. The test was performed on the web platform and several students submitted typed answers to the same assessment questions, to evaluate how the system performed under a practical academic environment.

For the test to be performed, a small set of students answers were generated. Some of the answers were entirely original, some were partially similar and some were modified forms of existing entries in the system. This was to verify if different similarity measures can be correctly detected by the system.

The setup was very basic; it consists of the web interface to submit the answer, the backend that processed the input, and the database where all information was stored. When an answer was submitted by a student, it first underwent text cleaning to remove unnecessary characters and words. After the text cleaning, it was converted into numerical form using TF-IDF. The comparison was made using cosine similarity with the previous answers

submitted to the same assessment by students.

Before the experiment was carried out, similarity limits were pre-defined, and were used to determine whether an answer should be accepted, flagged or rejected according to the percentage of similarity. The execution time was measured as well as the final classification result.

The performance of the system was evaluated in three main ways: the level of similarity detected, the consistency of the classification according to the defined limits, and the response time.

RESULTS & ANALYSIS

The Similarity Detection System was tested using various student submissions with differing similarities. Results focused on distribution of similarity, detection accuracy, threshold operation and performance. The results indicate the accuracy of the system in finding similar responses and efficient processing speed.

Similarity Score Distribution

A selection of 30 student submissions was used to test the distribution of similarity scores, percentage values for each being separated into the categories: low, moderate, high.

Similarity Range	Number of Submissions	Percentage (%)
0% – 30% (Low)	18	60%
31% – 60% (Moderate)	8	26.7%
61% – 100% (High)	4	13.3%

The majority of submissions fell in the low range. There was a smaller number of submissions with a moderate similarity and a very small number of highly similar submissions.

Detection Accuracy

The results produced by the system were compared to manual analysis to determine accuracy.

Category	System Classification	Manual Review	Correctly Classified
Accepted	18	17	17
Flagged	8	9	8
Rejected	4	4	4

Detection Accuracy = Total Correctly Classified / Total submissions) * 100

Detection Accuracy = (29/30) *100 = 96.7%

The high percentage indicates good accuracy of the system in assessing similarity when compared to human analysis.

Threshold Operation

A set of similarity thresholds were designated:

- 0%-30% Accepted
- 31%-60% Flagged
- Above 60% Rejected

Threshold Level Observed Performance

Threshold Level	Observed Performance
Low Threshold (≤30%)	Correctly identified original answers
Moderate Threshold (31–60%)	Successfully detected partially similar content
High Threshold (>60%)	Accurately detected highly similar responses

The results derived from thresholds were consistent and logical, requiring no further decisions based on individual judgments.

Processing Time Analysis

The processing time of the system was determined for multiple submissions:

Number of Submissions	Average Processing Time per Submission (seconds)
10	0.45
20	0.52
30	0.60

As seen above, there was very little change in processing time between the submission sets; therefore, the system is scalable with an efficient back end processor.

Comparison of the proposed system to manual analysis:

Evaluation Method	Time Required (30 Submissions)	Consistency	Scalability
Manual Evaluation	2–3 hours	Moderate	Limited
Proposed System	~18 seconds	High	High

The proposed system is considerably faster than manual analysis, with much greater consistency in the results and good scalability.

ADVANTAGES OF PROPOSED SYSTEM

This proposed Similarity Detection System provides a lot of practical benefits in the online assessments. One of the largest benefits is that it decreases the need for manual checking. Typically, it would take teachers many hours to compare answers among students, especially in the case where there is a large class size. However, the similarity is automatically detected, which saved the time and efforts for them, therefore it becomes much more time-saving and less stressful for instructors.

The next significant benefit is the fairness of the system. As there are predetermined similarity boundaries and a fixed procedure for comparison, students are evaluated based on the same standard regardless of their personal circumstances, thus reduces the likelihood of bias and unfairness from instructors.

Additionally, the system is transparent. Students can view their similarity percentage and also the status of their work. Instructors have access to all submitted responses, where it clearly shows which are accepted and which is flagged/ rejected. It increases the mutual trust between students and instructors because the outcome is readily accessible for both parties.

The step-by-step comparison procedure is also helpful because instead of arbitrarily comparing each response with all others, it compares each subsequent submission with prior ones, thus simplify and manage the whole system.

Also, this system can perform analysis on several submissions simultaneously, which increases the overall performance without any decrease in efficiency. Multiple students may be able to submit their responses together without causing the system to slow down. Therefore, the system can also be suitable for classroom applications.

Most of all, the system promotes academic honesty and encourages students to develop their own thoughts when creating the responses to be submitted.

In short, the system is useful and feasible for conducting similarity detection in online assessments in a fair and efficient manner.

Limitations & Future Enhancements

Limitations:

This system will likely detect textual similarity and may not accurately reflect conceptual similarity when the student paraphrases in a drastically different way. The quality of pre-processing and chosen thresholds will also influence the accuracy of the system. TF-IDF and cosine similarity will not analyze deeper meanings; rather, they use the number of times a word appears.

Enhancements:

- Incorporate more sophisticated models of semantics, such as word embeddings or transformer-based models to aid with conceptual similarity detection.

- Integrate external plagiarism checks with existing online sources and academic databases.
- Incorporate AI-based methods to detect paraphrasing.
- Real-time similarity alerts during submission.
- Provide graphical statistics for instructors to see similarity trends over time.
- Integrate other document formats such as PDF and Word documents.

CONCLUSION

The increase in web-based learning system usage has resulted in online assessment as part of the typical educational environment. Students can conveniently submit their work and teachers can manage test papers conveniently with such system; however, it also raises problems. One of the problem lies in assuring original and fairness in submissions. The manual check of answer consumes too much time especially when it involves lots of students and also might produce unreliable judgment.

As an solution for this problem, a new organised and automated approach-Similarity Detection System is proposed. The system not only provides a web interface for submission but it integrates with simple text processing method, TF-IDF, cosine similarity for comparison, it has step-wise comparison as opposed to simultaneous comparison, and automatic marking based on similarity thresholds.

The experimental outcomes show a good performance in comparing similar answers and the time consumed in the process is highly reduced. Furthermore, the clear dashboard for student and teacher to view results promotes trust and transparency.

All together this system provides an economical and efficient mechanism to promote fairness and academic honesty in online assessment and this system can also be enhanced to apply other methods for higher similarity measurement in future.

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 1988; 24(5): 513-523. DOI: https://doi.org/10.1016/0306-4573(88)90021-0
2. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008. p.112-133. Available from: https://nlp.stanford.edu/IR-book/
3. Jurafsky D, Martin JH. Speech and Language Processing. 3rd ed. Upper Saddle River: Pearson; 2023. Available from: https://web.stanford.edu/~jurafsky/slp3/
4. Clough P. Plagiarism in natural and programming language: an overview of current tools and technologies [Internet]. Sheffield: University of Sheffield; 2000 [cited 2026 May 20].

How to cite this article: Kadiyala Swathi, Maka Jaya Narendra Sai, Mandava Sai Likhitha, Akunuri Sri Nandini, Kolusu Naga Swarajya Lakshmi, Mallavalli Dola Rajendra. Similarity detection system for assessments. *International Journal of Research and Review*. 2026; 13(6): 37-45. DOI: *10.52403/ijrr.20260605*

******