



Access Online



Article DOI:

10.5281/zenodo.21409951

SMART PLAG WEB TOOL

Basavashri Hudge¹, Dr Kishan Singh¹

ABSTRACT

This paper introduces an online originality verification framework that quantifies content overlap through an adjustable cutoff parameter. The framework accepts document uploads or raw text entry, then categorizes content originality dynamically based on computed match ratios.

KEYWORDS: Plagiarism Detection, Web Development, Text Similarity, File Comparison

I. INTRODUCTION

Intellectual misappropriation occurs when an individual reproduces another person's concepts or outputs and presents them as independent contributions, omitting due credit to the original source. With the rapid proliferation of freely accessible online resources, opportunities to replicate content have expanded considerably across scholarly writing, research documentation, and digital publishing. Academic communities and professional organizations consequently need robust verification tools capable of upholding standards of authenticity.

A significant portion of currently available content-checking tools rely on rigid, inflexible cutoffs and operate as opaque mechanisms that deny users any meaningful input into comparison criteria. These fixed approaches frequently misidentify legitimately cited material as copied content, while simultaneously failing to flag subtle instances of duplication where textual overlap remains minimal.

In response to these shortcomings, the present work introduces a user-configurable originality inspection framework that grants granular control

over sensitivity parameters. Researchers and educators may fine-tune the detection boundary to accommodate varying enforcement requirements, whether for strict verbatim checking or broader overlap monitoring. Delivering this functionality through a browser-accessible interface ensures compatibility across devices and operating environments.

II. RELATED WORK

Academic investigation into content originality verification spans several decades, producing a diverse array of methodologies with differing trade-offs in precision and processing cost. Early solutions depended heavily on raw character sequence comparisons, a strategy that proved ineffective whenever source material underwent even trivial modifications. Word-token methods subsequently offered greater resilience by examining term distributions and positional patterns, yet these frameworks still struggle when writers employ synonymous vocabulary or alter clause structures.

III. SYSTEM ARCHITECTURE

A. Overview

The originality verification platform is organized as an interconnected web

¹ Department of
Electronic and
Communication,
Guru Nanak Dev
Engineering College,
Bidar (Karnataka),
India

How to Cite:

Basavashri
Hudge & Dr Kishan
Singh (2026),
Smart Plag Web
Tool, International
Educational Journal of
Science & Engineering
(IEJSE), Vol: 9, Special
Issue, May 2026
211-215

application built on discrete, independently operating components:

- User Input Module
- Text Preprocessing Module
- Similarity Calculation Engine
- Threshold Evaluation Module
- Document Repository
- Each module operates independently, improving maintainability and scalability.

B. User Input Module

Content may be provided to the platform through either of the following submission modes:

- Uploading documents (TXT, DOC, PDF)
- Entering text directly into a web form

A configurable cutoff parameter accompanies the input interface to govern the stringency of originality assessment.

C. Text Preprocessing Module

A standardization pipeline is applied to submitted content before analysis, encompassing the following steps:

- Lowercase conversion
- Removal of punctuation and symbols
- Tokenization of text into words
- Stop-word removal

The resulting normalized representation is subsequently forwarded to the comparison engine for further processing.

D. Similarity Calculation Engine

Processed input is evaluated against the stored document corpus through a shared-token counting strategy. The resulting overlap proportion is expressed as a percentage indicating how much submitted content coincides with previously indexed material.

This design balances computational efficiency against measurement reliability within practical constraints.

E. Threshold Evaluation Module

The derived overlap percentage is then assessed against the operator-specified cutoff value:

- If overlap ratio $\geq$ cutoff $\rightarrow$ Submission identified as non-original
- If overlap ratio $<$ cutoff $\rightarrow$ Submission classified as sufficiently original

This design grants operators direct authority over how

strictly the system enforces originality requirements.

IV. IMPLEMENTATION METHODOLOGY

A. Cosine Similarity

Every document is encoded as a vector of term occurrence counts. The angular alignment between document vectors is measured using:

$$\text{Similarity} = \frac{A \cdot B}{\|A\| \times \|B\|}$$

This formula evaluates the directional relationship between vectors, rendering the measure invariant to document length.

B. N-Gram Matching

Input text is partitioned into contiguous word windows of fixed width, typically spanning two or three consecutive terms. Resemblance is quantified by tallying how many of these windows appear in both documents.

C. Threshold Based Decision Model

The computed match ratio is evaluated against an operator-defined boundary value

- If Similarity $\geq T \rightarrow$ Plagiarism Detected
- Else $\rightarrow$ Original Content

This boundary is fully configurable, with representative values such as 30%, 50%, or 70% serving as common operating points.

V. RESULTS AND DISCUSSION

A. Experimental Setup

The verification platform was built as a browser-accessible application pairing a presentation layer constructed with HTML, CSS, and Bootstrap alongside a processing backend powered by Node.js and Python, supported by a persistent document repository. Users interact with the platform either by submitting files or by entering text directly. Validation was carried out across several corpora encompassing coursework submissions, published research extracts, and artificially seeded documents with predetermined overlap levels

B. Results

1) Text Input Testing

- * Input: Manually entered text
- * Output: Similarity score with duplicated phrases
- * Result: Accurate detection of partial and exact

matches

Fig. 1.1: Enter Text Selection

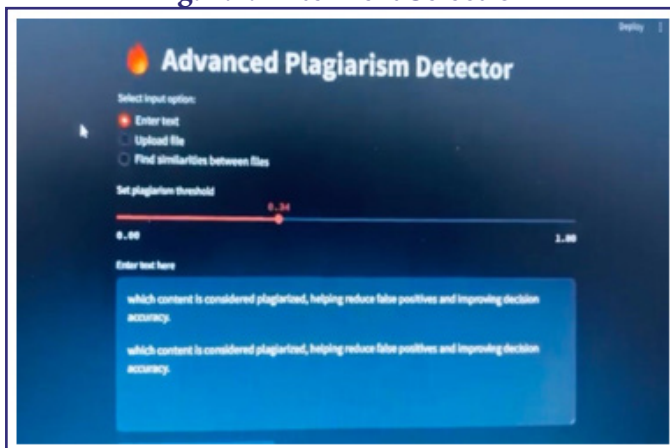


Fig. 1.2: Sentence Similarity Score

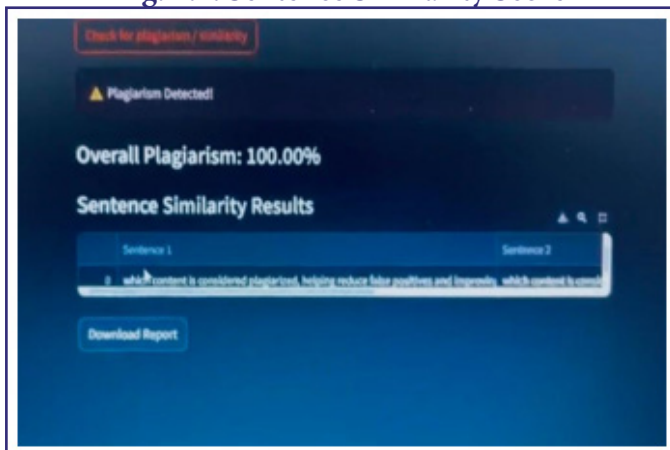


Fig. 2.1: Upload File Selection

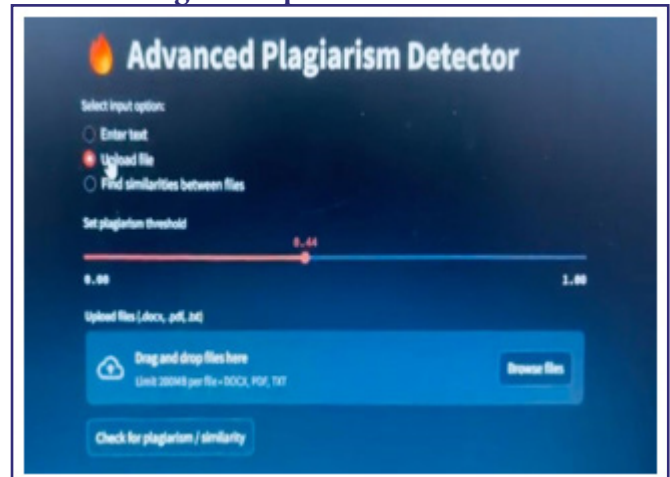


Fig. 2.2: PDF And DOCX Files Selection

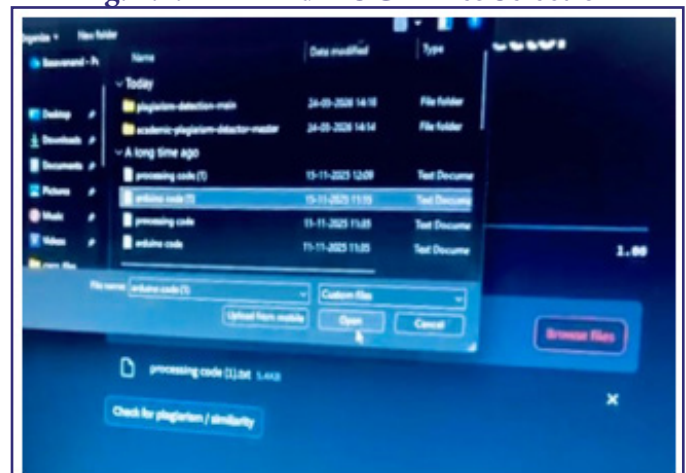
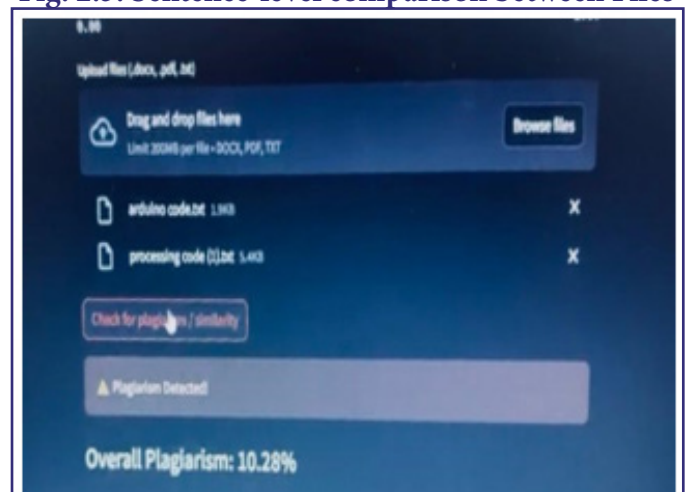


Fig. 2.3: Sentence-level comparison between Files



2) File Upload Testing

- * Input: PDF and DOCX files
- * Output: Extracted text + similarity percentage
- * Result: Efficient extraction and comparison of document content with respect to sentence when selected Upload File and comparison of document content with respect to overall file when selected Find Similarities Between Files.

Fig. 2.4: Finding Similarities Between Overall Files

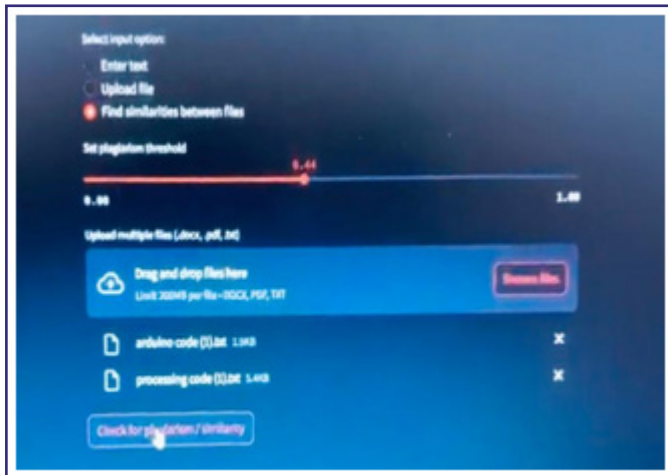


Fig. 2.4: Result Of Similarities Between Overall Files

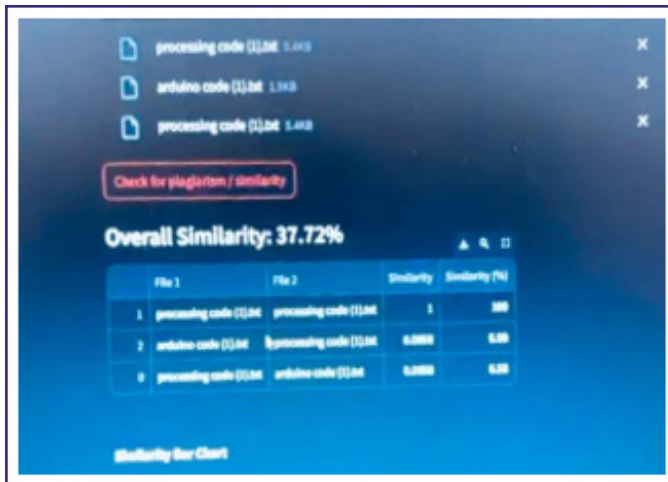


Fig. 2.5: Result Graph



C. Discussion

Empirical testing confirms that the constructed browser-based verification platform reliably identifies reproduced material. Measured overlap ratios aligned closely with expected benchmarks, validating the underlying comparison approach.

Among the platform's most notable capabilities:

- * Accept content in either raw text form or as uploaded document files
- * Return verification outcomes immediately with minimal processing delay
- * Visually annotate overlapping segments to support result interpretation

Several constraints were identified during evaluation:

- * The platform primarily flags verbatim and closely identical passages
- * Substantially reworded or meaning-preserving rewrites frequently evade identification
- * Processing efficiency decreases noticeably when evaluating exceptionally lengthy submissions

D. Comparison with Existing Systems

Benchmarked against established originality-checking platforms, the current tool offers:

- * An intuitive, uncluttered browser interface requiring no local software installation
- * Quicker turnaround for documents of short to moderate length
- * Reliable foundational overlap measurement adequate for most routine use cases

That said, AI-powered platforms incorporating deep language understanding retain a distinct advantage when identifying meaning-level reuse.

E. Summary

In summary, the platform fulfills its design objective of delivering originality verification within a browser-accessible environment at satisfactory accuracy and speed. It is well-suited to educational contexts, particularly for individual learners and small departmental deployments.

VI. ADVANTAGES

Operator-adjustable originality sensitivity

- Universal access through any standard web browser
- Auditable and interpretable overlap calculation
- Minimal processing resource requirements
- Expandable, component-based structural design

VII. LIMITATIONS

While the platform delivers meaningful value, several constraints remain:

- Meaning-level resemblance detection remains underdeveloped
- Reworded material frequently escapes accurate flagging
- A sufficiently large reference corpus is necessary for reliable analysis

VIII. FUTURE ENHANCEMENTS

Planned development directions include:

- Incorporate meaning-based comparison through language model embeddings
- Enable identification of translated and multilingual content reuse
- Add attribution tracing to pinpoint originating source documents
- Deploy learned models to dynamically calibrate sensitivity boundaries

IX. CONCLUSION

This paper describes a **configurable browser-hosted originality verification framework** built to emphasize adaptability, interpretability, and operational simplicity. Enabling operators to fine-tune similarity boundaries allows the platform to serve diverse enforcement contexts. Experimental outcomes confirm its suitability for academic communities and organizational environments alike.

The underlying structure is designed to support successive capability expansions, including deeper linguistic analysis and data-driven optimization strategies.

REFERENCES

1. M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein, "Overview of the PAN 2014 International Competition on Plagiarism Detection," in Proc. CLEF, Sheffield, U.K., 2014, pp. 1–30.
2. C. Lyon, "Textual similarity analysis for plagiarism detection," *Journal of Documentation*, vol. 67, no. 5, pp. 788–809, 2011.
3. J. Havrland and V. Kreinovich, "A simple, probabilistically sound, similarity measure based on cosine similarity," *Applied Mathematics Letters*, vol. 25, no. 7, pp. 1031–1034, 2012.
4. M. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 63–82, 2011.
5. Y. Zhang and J. Zhou, "A feature-based approach for

detecting paraphrased plagiarism," in Proc. IEEE Int. Conf. on Semantic Computing, Irvine, CA, USA, 2014, pp.345–350.