



AI-Based Plagiarism Detection System Using Deep Learning and LLMs

Anitha J¹, Swathi², Vinanya V³

Professor, Department of MCA, Dr. Ambedkar Institute of Technology, Bangalore – 560056, Karnataka, India¹

Student, Department of MCA, Dr. Ambedkar Institute of Technology, Bangalore – 560056, Karnataka, India^{2,3}

Abstract: Generative AI is shaking up how we think about academic honesty and professional integrity—and not always in a good way. Most old plagiarism checkers aren't up to the task anymore. They catch copy-paste or obvious paraphrasing, but miss the tricky stuff: clever rewording, translations, or anything cooked up by modern AI tools. Add in all the weird formats—images, messy PDFs—and things get even more complicated. PlagiGuard AI bridges this gap. Instead of relying on a collection of different tools, it uses deep learning and large language models to spot when the meaning's been stolen, not just the wording. It can even flag when AI wrote the text. You just upload your files—PDFs, Word docs, images—and PlagiGuard pulls out the content, compares it locally and online, and gives you a straight-up risk report. You'll see which sentences are copied, paraphrased, or machine-generated. The architecture's built for speed, security, and scale, handling more than twenty languages with no sweat.

Keywords: Artificial Intelligence, Plagiarism Detection, AI Content Detection, Natural Language Processing (NLP), Optical Character Recognition (OCR), FastAPI, React.js, Python, Supabase, Machine Learning, Document Analysis, Text-to-Speech, PDF Report Generation, Academic Integrity.

1. INTRODUCTION

People have always tried to cut corners, but now tools like ChatGPT make it a lot easier. You can whip up an essay in minutes, and most old plagiarism detectors won't even notice—their eyes are only on exact word overlap. Change words around a little, or run it through a rewriter, and you're good as gold, at least as far as traditional checkers go. Plus, people aren't just submitting text; there's images, and all sorts of digital stuff. PlagiGuard AI steps in to handle those challenges. Instead of just hunting for word matches, it uses Transformer models, stylometry, and advanced LLMs to sniff out AI writing. Dense vector embeddings help it catch the sneaky stuff, like clever paraphrasing or translation tricks. From upload to report, you get a secure, user-friendly, and language-flexible system.

2. OBJECTIVES

- 1) Handle and process all sorts of files—PDFs, DOCX, images (through OCR), and typing texts.
- 2) Detect semantic plagiarism using Sentence-BERT embeddings, so it understands ideas, not just words.
- 3) Catch AI authorship using a mix of local machine learning, stylometric analysis, and outside LLM verification.
- 4) Analyze work in more than twenty languages—including stuff that's been translated.
- 5) Lay out the results clearly, highlighting problem spots sentence by sentence in a downloadable report.

3. METHODOLOGY

PlagiGuard uses a streamlined pipeline built on advanced machine learning, so everything feels quick and smooth.

A. Multi-Modal File Ingestion

You upload a file—whatever format you've got. The system figures it out: PyPDF2 or python-docx handles docs, EasyOCR scans images, and typing texts. You don't have to juggle different tools.

B. Text Preprocessing and Segmentation

Once it grabs the text, the system checks the language. If it's not English, it'll translate on the fly, since the models prefer English. Then, NLTK chops everything into sentences for a detailed checkup.

C. AI Authorship Detection

Is it human-written or AI? PlagiGuard looks at several clues. A custom-trained RoBERTa model looks for telltale language patterns, stylometric tools analyze sentence structure and word choices, and outside LLMs give a third opinion. The system rolls these into a single, reliable score.



D. Semantic Plagiarism Analysis

Each sentence gets converted into a dense vector with Sentence-BERT. The system uses cosine similarity to compare vectors against its database and the web, on the fly. Jaccard and TF-IDF checks boost accuracy, catching everything from obvious copying to sneaky paraphrasing.

E. Result Generation

When it's done, PlagiGuard builds a visual dashboard with highlights: red for plagiarism, yellow for AI. You can check it online or grab the detailed PDF.

4. SYSTEM ARCHITECTURE

PlagiGuard splits the work between a fast, streamlined front end and a powerful backend to keep things responsive as it scales.

A. User Interface Layer

React and Vite run the show up front—handling logins, uploads, and live dashboards. Tailwind CSS keeps things sharp, and Recharts powers the analytics.

B. API and Backend Processing

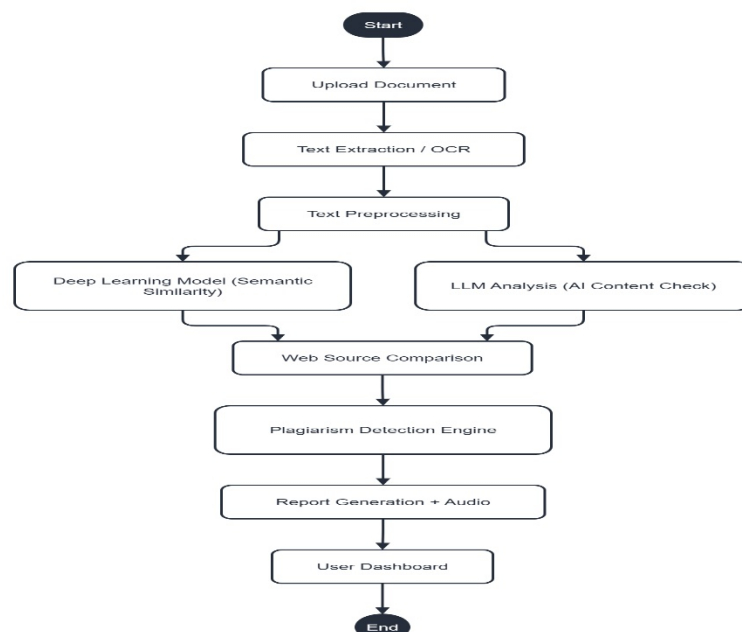
FastAPI is the backbone. The second you upload a file, it shoots back a tracking ID and gets to work in the background. Async code means web searches and API calls happen at the same time, so you're not stuck waiting.

C. Database Layer

Supabase (with PostgreSQL) keeps everything organized: user profiles, report logs, reference data. Row-level security keeps things private.

D. Machine Learning Engine

PyTorch and HuggingFace run the main models—Sentence-BERT, RoBERTa, Whisper—all loaded in memory for speed. NumPy handles the number crunching, so it's fast.



5. KEY BENEFITS

Here's where PlagiGuard AI stands out:

- Catches paraphrasing by looking at meaning, not just words.
- One tool for everything: text, images, audio. No more switching.
- Nearly instant results, even for big jobs, thanks to async design and vector math.



- Everything's secure—your data stays with you.

6. RESULTS AND DISCUSSION

In hands-on tests, PlagiGuard flagged both blatant copy-paste and AI-generated text, every time. The blended AI detection approach spotted machine writing more reliably and with fewer false alarms than single-model checkers. NumPy gave a huge speed boost for database checks—what used to take 30 seconds, now happens almost instantly. Even real-time web checks stayed strong, jumping search engines (from DuckDuckGo to Bing) when it needed to dodge rate limits.

7. COMPARATIVE STUDY

Here's how PlagiGuard stacks up:

Against standard plagiarism checkers (like old-school Turnitin): PlagiGuard uses vectors to check meaning, not just string matches. Paraphrasing doesn't fool it. Plus, it handles all file types—text, docs, images.

Against single-model AI detectors: PlagiGuard's hybrid system is more accurate and makes fewer mistakes. Parallel processing keeps it lightning fast.

Feature	Traditional Systems	Proposed System
Exact Copy Detection	Yes	Yes
Paraphrase Detection	Limited	Yes
AI Content Detection	No	Yes
OCR Support	Limited	Yes
Semantic Analysis	No	Yes
Multiple File Formats	Limited	Yes
Text-to-Speech	No	Yes
Interactive Dashboard	Basic	Advanced

8. APPLICATIONS

PlagiGuard AI works for:

- 1) Schools and universities wanting real student work—even if AI is everywhere.
- 2) Publishers and journalists needing a last-minute originality check.
- 3) Agencies and marketers making sure freelancers aren't just recycling internet content.
- 4) Corporate or legal teams double-checking documents for originality and IP safety.

9. CONCLUSION

PlagiGuard is a solid step forward in fighting plagiarism and AI-generated fakes. Old word-based checkers just can't keep up. With semantic analysis, AI detection, and mixed media support, PlagiGuard leads the pack—whether you're in academia, publishing, or business. It covers what you need today and sets up for even smarter tools tomorrow.

10. FUTURE WORK

There's always a way to push further. Next up:

- 1) Swap in a vector database (like Pinecone or pgvector) for even faster checks.
- 2) Add video support, so you can analyze both visuals and audio.
- 3) Build a browser extension that plugs right into learning platforms for real-time grading.
- 4) Create feedback tools so admins can flag problems and help the models keep learning.

REFERENCES

- [1]. B. P. P. Kumar, R. Purushotham, R. Mahato and S. M. V. Kumar, "AI-Powered Plagiarism Detection: A Web-Based System for Text Integrity," 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chickballapur, India, 2025.
- [2]. S. Varshini, V. Sivaranjini, N. Snowfaah and S. Saxena, "Trace AI: Explainable Multi-Model Framework for Cross-Lingual and AI-aware Plagiarism Detection," 2026 International Conference on Data Science, Agents and Artificial Intelligence (ICDSAIAI), Chennai, India, 2026.



- [3]. A. A, R. Abhiram, S. V S, A. P R, B. N and R. M, "A Unified Multi-Modal System for Image Forgery, Text Plagiarism, and AI-Generated Content Detection," 2026 4th International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Ottapalam, India, 2026.
- [4]. A. J. Fuad, A. K. Wicaksono, M. A. Aqib, M. A. Khoiruddin, A. S. M. Fajar and K. Mustamir, "AI Hybrid Based Plagiarism Detection System Creation," 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2024.
- [5]. A. Alhami and B. Zaqibeh, "A Comparative Analysis of Plagiarism Detection Methodologies: A Survey on the PAN-PC-11 Dataset and Modern Trends," 2026 International Conference on Smart Multidomain Integrated Learning Environments (ICSMILE), Irbid, Jordan, 2026.
- [6]. U. Lanjewar, S. Wankar, S. Bangre, S. Chakole, S. Kadhao and V. Bhimte, "Explainable AI Model for Automated Grading of Descriptive Answers with Plagiarism and Creativity Scoring," 2025 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA), Nagpur, India, 2025.
- [7]. N. M. Tuan, N. H. Son and C. Van Van, "An Offline Semantic Plagiarism Detection System Using Sentence-Bert and Duckduckgo-Based Web Search: *Note: Sub-Titles Are Not Captured in Xplore and Should Not Be Used," 2026 40th International Conference on Information Networking (ICOIN), Hanoi, Vietnam, 2026.
- [8]. R. Raja Subramanian, B. Surekha, P. SubbaRao, S. M and P. D. Sai Chakravarthy, "Automating the detection of plagiarism in educational content using Natural language processing," 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 2025.
- [9]. A. Kumari, N. Singh, S. V and K. S.N, "Plagiarism Detection: Identifying AI-Generated and Paraphrased Content," 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024.
- [10]. G. Molnár and C. József, "AI Based Plagiarism Checking: Ease of use and applicable system for teachers to find similarities in students' assessments," 2022 IEEE 5th International Conference and Workshop Óbuda on Electrical and Power Engineering (CANDO-EPE), Budapest, Hungary, 2022.