Department of Artificial Intelligence

22AIE302: FORMAL LANGUAGE AND AUTOMATA

Project Report

Plagiarism Detection System

TEAM – D07:

MALAVIKA S PRASAD	CB.SC.U4AIE23315
VIBHU SANCHANA	CB.SC.U4AIE23347
ASMI K	CB.SC.U4AIE23351
GESHNA B	CB.SC.U4AIE23360

Supervised By:

Dr. Chitra
Assistant Professor
Department of Artificial Intelligence
Amrita Vishwa Vidyapeetham

Date of submission: **Signature of the Project Supervisor:**

BONAFIDE CERTIFICATE

This is to certify that this project entitled, “Plagiarism Detection System” submitted by Malavika S Prasad, Vibhu Sanchana, Asmi K, and Geshna B is an authentic work carried out by the team under my supervision and guidance. To the best of my knowledge, the content presented in this report has not been previously submitted to any other academic institution, nor has it been utilized to fulfill any degree or diploma.

Amrita Vishwa Vidyapeetham
Coimbatore- 641112
Date:

Dr. Chitra
Dept. of Artificial Intelligence

ACKNOWLEDGEMENT

We would like to express our gratitude to our Dean, DR. K. P. SOMAN, who gave us this opportunity to do this extremely good project on the topic “Plagiarism Detection System”. This project helped us understand and learn about many new things. Secondly, we would like to thank our professor Dr. Chitra without whose guidance this project would not have been successfully accomplished. Your insights, support, and encouragement were instrumental in the completion of this project. Lastly, this project would not have been possible without the efforts and dedication of our team, therefore we thank ourselves for the amazing work.

TABLE OF CONTENTS

ABSTRACT

Increased availability of online content and massive online information databases has created a necessity for plagiarism detection as a thread of justice task for academic, corporate, and creative communities. This project outlines a complete AI-based Plagiarism Detection System compromising classical string-matching techniques augmented with sophisticated semantic knowledge based on Large Language Models (LLMs). Its main target is identifying both word-for-word copying and contextually rewritten plagiarism, which is normally overlooked by traditional detectors.

The framework utilizes a multi-layered detection pipeline that initiates with effective text preprocessing using the Natural Language Toolkit (NLTK) for normalization, noise removal, and phrasing of meaningful text. The pre-processed text then undergoes a fast Aho–Corasick automaton, a very fast multiple-pattern-matching algorithm, for fast repeated or overlapping string subsequences identification within large document sets. This traditional automata-theory-based method provides high-speed direct overlap detection at the expense of moderate computational complexity.

To go a step deeper than shallowness-based text similarity, the solution includes an LLM-driven semantic analysis module created with LangChain and driven by the Groq API. The module automatically elongates the provided topic into relevant subtopics with generative reasoning, extracts relevant contextual information from Wikipedia, and assesses the conceptual similarity of the submitted text. Through comparison of writing style, vocabulary usage, and thematic correspondence, the LLM accurately identifies paraphrased, semantically equivalent, or idea-level plagiarism that standard algorithms ignore.

Under this hybrid framework, the system attains a judicious merging of syntactic detection performance for semantic interpretability such that both literal copying and hidden rewording are aptly detected. In addition, the incorporation of LLMs adds adaptive learning functionalities such that the model accommodatingly handles varying writing styles across different subject domains. As such, the system is especially worthwhile for academic scholarly evaluation, verification of originality of contents, as well as automatic screening of publications.

Overall, this work illustrates how the interplay between automata theory and large-language models can rechure plagiarism detection through increasing efficiency, comprehension-depth, and context-sensitivity. This proposed method provides a future-proof foundation for next-generation plagiarism checkers which are scaper, intelligent, and context-sensitive, equipped for managing the changing needs of information authenticity for the modern digital era.

INTRODUCTION

Plagiarism has been a big issue in the world of easy availability and accessibility of information with the advent of the digital world. Now, with online learning portals, Research papers, blogging, people have open access to a vast amount of information in a matter of seconds. While this has proved handy, it has, in turn, made it simple for a person to lift or repurpose the work of another without due credit. That is where there is a growing necessity for decent tools that can identify plagiarism fast and accurately.

Conventional plagiarism detection tools primarily depend on String Matching or Keyword Comparison techniques. These approaches are effective for identifying direct copies but lack success when the text is Rephrased or Paraphrased for presenting a sense of originality. Forging this limitation, recent tools are gradually drifting towards Semantic Analysis, which facilitates determining the true meaning of the text instead of word comparison.

The Proposed Plagiarism Detection System intends to incorporate both such methods. It implements text preprocessing and the Aho–Corasick algorithm for fast and effective identification of direct textual overlaps. Furthermore, it includes a Large Language Model (LLM) using LangChain and the Groq API for identifying the sentiment and meaning of sentences. This makes the system identify even rewritten or conceptually equivalent stuff that is otherwise overlooked by traditional tools.

Main goals of this task are:

- To create a plagiarism detector which identifies both semantic and exact similarity.
- For increasing the speed and accuracy of detection based on automata approach.
- For the application of AI-based language models for enhanced text comprehension and topic extension.
- To develop a friendly and effective program for academic and research settings.

Overall, this work demonstrates how uniting classical algorithms with contemporary AI methods is capable of increasing the quality of plagiarism detection. It goes beyond looking at copied text recognition, but rather looking at recognizing the meaning of the text, which makes the system more intelligent as well as robust for deployment on a daily basis.

LITERATURE REVIEW

A. Traditional String-Matching Approaches

Plagiarism detection initially emerged from fundamental string-matching algorithms that compare exact textual sequences to identify overlaps between documents. Methods such as n-gram analysis, fingerprinting, and substring matching have been widely adopted due to their computational efficiency and simplicity in implementation. Research by Manning et al. [1] and others established the foundation for token-based comparison, which effectively detects verbatim plagiarism. However, these systems primarily focus on surface-level similarities, struggling to capture semantic rewording or contextual paraphrasing. Despite their speed and precision in identifying direct copying, such algorithms lack the linguistic depth to understand meaning-based similarity. This gap highlights the need for a system that maintains algorithmic efficiency while extending its scope toward semantic comprehension, a challenge addressed in this project through the integration of hybrid methods combining classical pattern recognition with contextual interpretation.

B. Semantic and NLP-Based Detection Techniques

As plagiarism became more sophisticated, researchers began applying Natural Language Processing (NLP) and semantic analysis to identify idea-level plagiarism. Approaches leveraging TF-IDF, Word2Vec, and transformer-based models such as BERT introduced the ability to assess conceptual similarity beyond exact word matches. Studies like "Semantic Similarity in Plagiarism Detection using Deep Learning" [2] demonstrated that contextual embeddings could capture nuanced meaning even in rephrased content. However, such methods often demand extensive computational resources and large training datasets, limiting their real-time applicability. Many existing NLP-based systems achieve high accuracy but compromise on speed and scalability, especially when processing large collections of documents. This project bridges that gap by implementing a balanced hybrid model that employs both semantic understanding and algorithmic pattern detection, ensuring efficiency without sacrificing depth of analysis.

C. Hybrid and Intelligent Detection Frameworks

Recent research has focused on hybrid plagiarism detection systems that combine multiple techniques like string matching, semantic similarity, and linguistic analysis to achieve comprehensive coverage. Systems like "Intelligent Plagiarism Detection using Multi-Layer Analysis" [3] have shown promising results in capturing both exact and paraphrased plagiarism. Yet, most of these frameworks lack adaptability and fail to dynamically adjust to different writing styles or topic domains. The proposed project addresses these limitations by developing a multi-layered detection pipeline that incorporates preprocessing, syntactic pattern recognition, and semantic interpretation in an integrated workflow. By utilizing natural language understanding through modern LLM-based models alongside efficient automata-based scanning, the system achieves high precision in both literal and conceptual plagiarism detection. This integration of classical algorithms with intelligent language models represents a significant step toward scalable, adaptive, and context-aware plagiarism detection.

Gaps Analysis

1. Algorithm Limitations:

   Traditional plagiarism detectors rely on exact string matching, missing paraphrased or reworded content and causing false positives with common phrases.
   Our solution: Combines refined preprocessing with efficient pattern matching and semantic analysis to detect both exact and paraphrased plagiarism accurately.
2. Semantic Understanding Challenges:

   Existing semantic methods are resource-intensive, lack interpretability, and often do not integrate well with classical approaches.
   Our solution: Integrates lightweight automata-based detection with semantic topic expansion and similarity evaluation for efficient, explainable results.
3. Integration and Automation:

   Many prior systems lack end-to-end automation, requiring manual steps and separate tools.
   Our solution: Provides a fully automated pipeline from preprocessing to semantic comparison, improving ease of use and reproducibility.
4. Adaptability Across Domains:

   Conventional tools often fail with varied writing styles or domains and lack context sensitivity.
   Our solution: Uses language models to dynamically analyze writing style and topic context, enhancing adaptability and robustness.
5. Explainability Deficit:

   Deep learning-based detectors are often black boxes offering unclear similarity explanations.
   Our solution: Combines transparent pattern matching with semantic insights for clear, interpretable plagiarism detection.

Methodology

The project began with the goal of creating a plagiarism detection system capable of identifying both exact textual copying and more challenging cases of paraphrased or semantically similar content. To achieve this, the approach was designed to combine classical pattern-matching algorithms with modern semantic analysis powered by large language models.

The initial phase focused on understanding the limitations of existing plagiarism detectors, which often relied solely on string matching and thus struggled with paraphrasing and rewording. With this insight, the project was structured as a multi-stage pipeline incorporating both syntactic and semantic evaluation to overcome those weaknesses.

The first essential step was to preprocess the input text effectively to prepare it for analysis. This involved normalizing the text like converting all letters to lowercase, removing punctuation and special characters, and cleaning inconsistent whitespace. Additionally, common stop words were removed to focus the analysis on content-rich words, reducing noise. Tokenization was used to split the text into meaningful units such as words and phrases. Unlike simple n-grams, the system extracts more meaningful phrases based on sentence boundaries and character-level n-grams for local similarity, providing a robust textual representation.

After preprocessing, the project implemented a finite-state automaton based on the Aho–Corasick algorithm. This choice was motivated by the algorithm's ability to efficiently search for multiple patterns simultaneously in linear time relative to the text length. The automaton scans large document collections to rapidly detect reused phrases or copied sentences, generating precise and explainable matches. This classical approach ensures high speed and low complexity, which is crucial for scaling to large datasets.

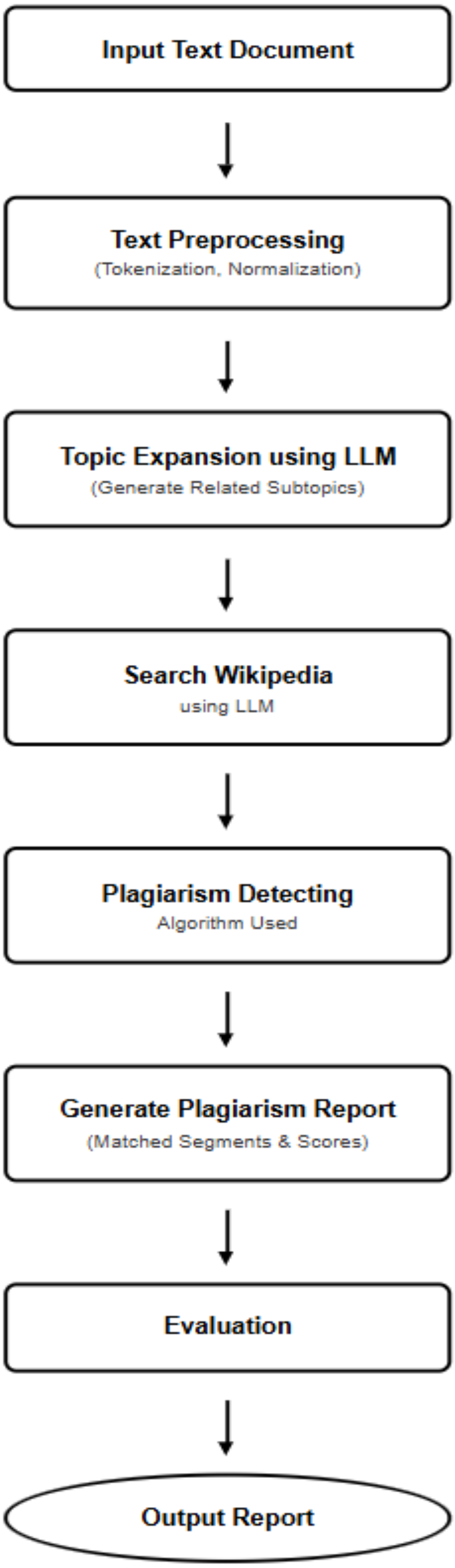
To move beyond surface-level matching, the system integrates a language model using an NLP framework that connects to a high-performance API. The model is tasked with expanding the initial topic into related subtopics by generative reasoning, enhancing the scope of content comparison. Relevant context is retrieved automatically from external sources such as Wikipedia to enrich the reference corpus. The system then uses semantic similarity metrics to compare the input text against these expanded topics, detecting paraphrasing, synonym substitution, and concept-level plagiarism that traditional methods miss. It also analyzes writing style and vocabulary use to further differentiate genuine rewording from plagiarism.

All these components i.e., preprocessing, automaton-based matching, semantic similarity analysis, and external data fetching, are integrated into a seamless, fully automated pipeline. This allows users to input raw text and receive a comprehensive report highlighting both exact matches and semantically similar content. The report presents interpretable results showing where direct copying occurs as well as where ideas are rephrased or conceptually copied, providing a clear and reliable plagiarism evaluation.

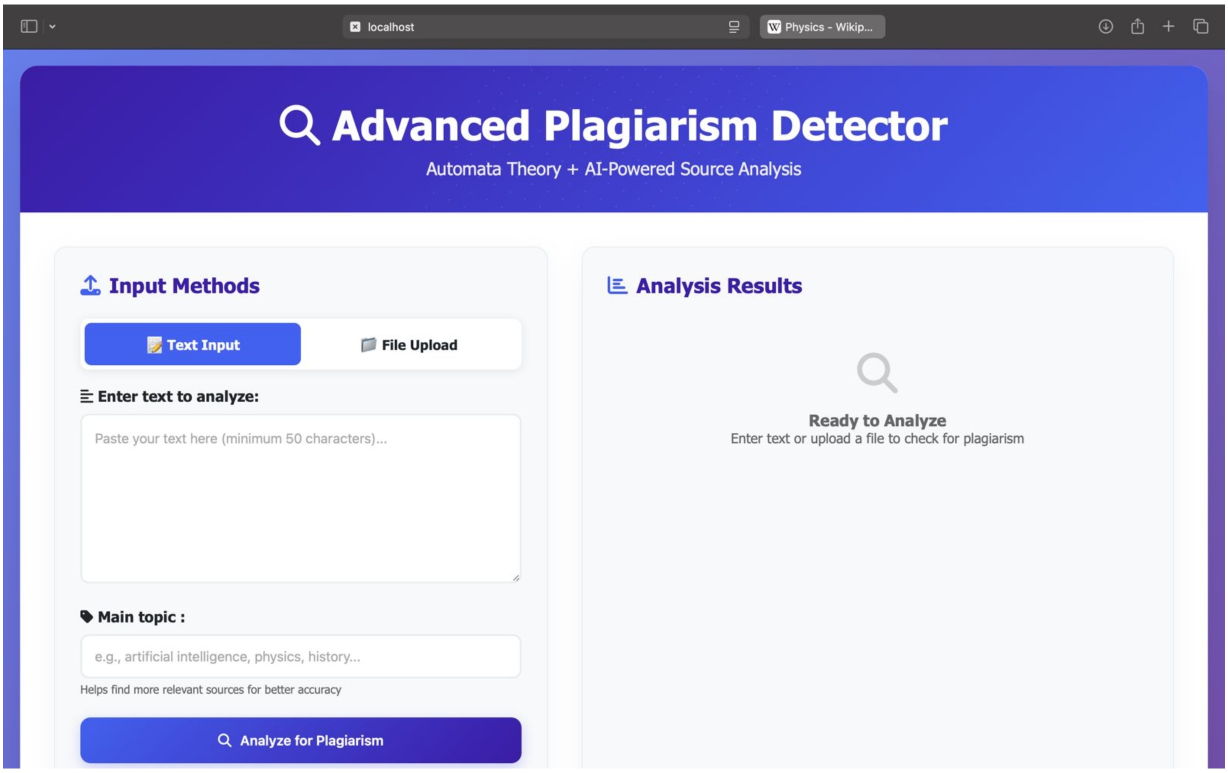
The project's methodology embodies a synergy between efficient classical algorithms and flexible AI-driven semantic understanding. Starting from the recognition of existing gaps, the system was developed to preprocess text smartly, detect exact phrase reuse swiftly, and apply contextual semantic evaluation intelligently. This layered approach results in a plagiarism detection tool that is accurate, scalable, and capable of understanding nuanced textual relationships, suitable for academic, research, and professional applications.

IMPLEMENTATION

The flowchart illustrates a linear plagiarism-detection pipeline from input to reporting, with each step clearly scoped for clarity and reproducibility. It starts by ingesting a text document and performing preprocessing such as tokenization and normalization to clean and standardize the content for analysis. Next, a user-provided topic guides an LLM-driven expansion into related subtopics, which is then used to search Wikipedia for contextual references. Using these references, the system generates a plagiarism report that highlights matched segments and assigns similarity scores, followed by an evaluation phase reporting precision, recall, and F1. Finally, the process outputs a consolidated report suitable for academic documentation and review.


Input Text Document
Text Preprocessing
(Tokenization, Normalization)
Topic Expansion using LLM
(Generate Related Subtopics)
Search Wikipedia
using LLM
Plagiarism Detecting
Algorithm Used
Generate Plagiarism Report
(Matched Segments & Scores)
Evaluation
Output Report

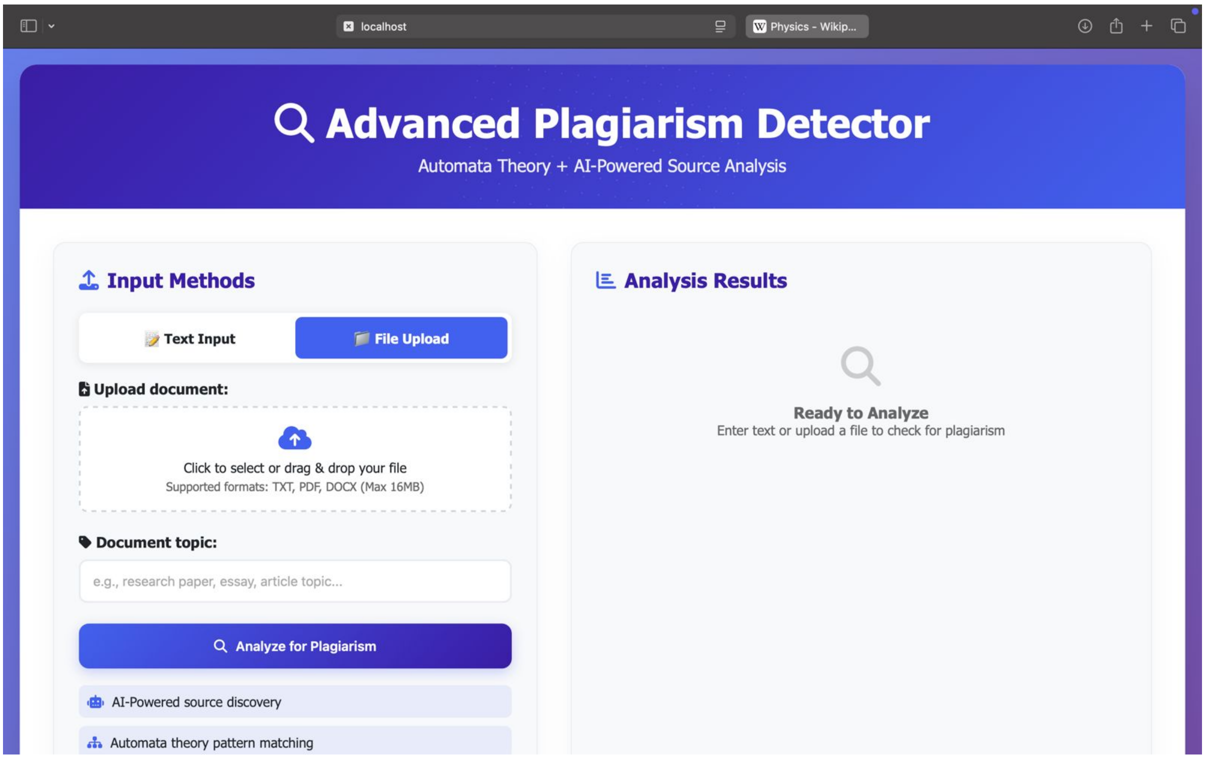
RESULTS


localhost
Physics - Wikip...
Advanced Plagiarism Detector
Automata Theory + AI-Powered Source Analysis
Input Methods
Text Input
File Upload
Enter text to analyze:
Paste your text here (minimum 50 characters)...
Main topic :
e.g., artificial intelligence, physics, history...
Analyze for Plagiarism
Analysis Results
Ready to Analyze

Helps find more relevant sources for better accuracy

Enter text or upload a file to check for plagiarism

Fig. 1: Main dashboard interface

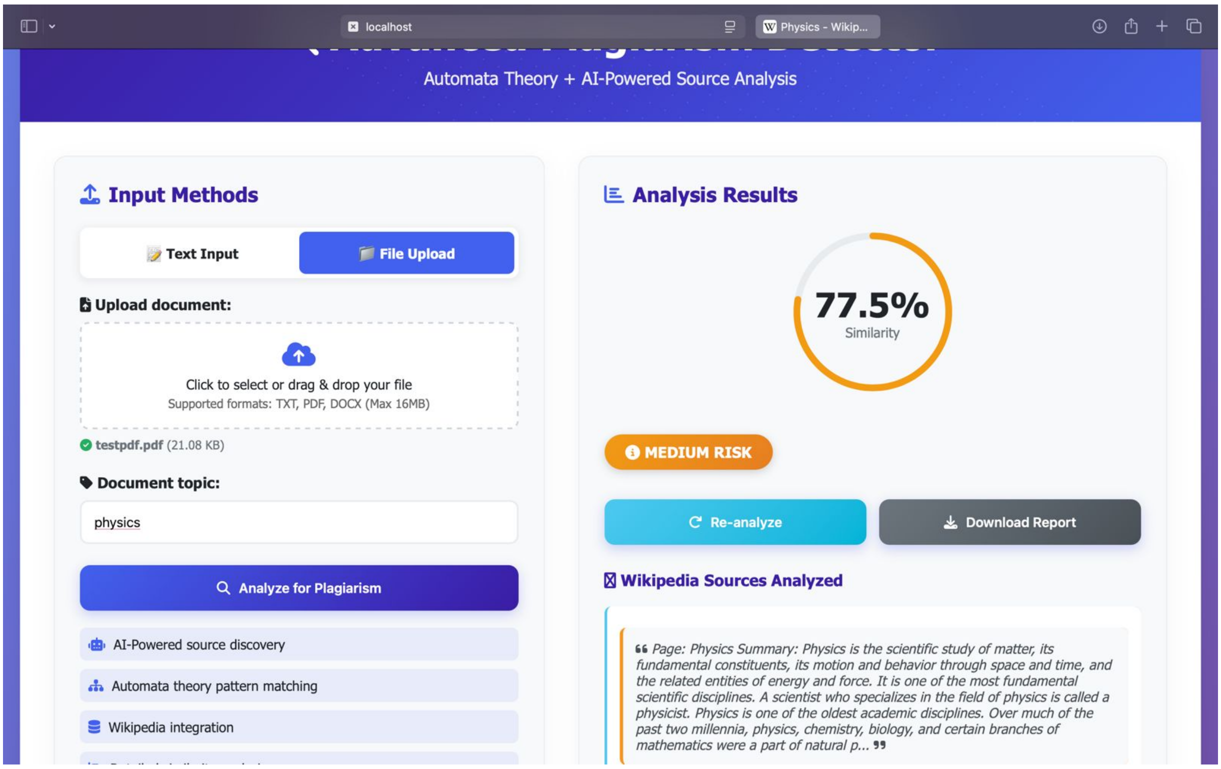

localhost
Physics - Wikip...
Advanced Plagiarism Detector
Automata Theory + AI-Powered Source Analysis
Input Methods
Text Input
File Upload
Upload document:
Document topic:
e.g., research paper, essay, article topic...
Analyze for Plagiarism
AI-Powered source discovery
Automata theory pattern matching
Analysis Results
Ready to Analyze

Click to select or drag & drop your file

Supported formats: TXT, PDF, DOCX (Max 16MB)

Enter text or upload a file to check for plagiarism

Fig. 2: File upload interface



Automata Theory + AI-Powered Source Analysis

Input Methods

Text Input | File Upload

Upload document:

Click to select or drag & drop your file
Supported formats: TXT, PDF, DOCX (Max 16MB)

✓ testpdf.pdf (21.08 KB)

Document topic:

physics

Analyze for Plagiarism

- AI-Powered source discovery
- Automata theory pattern matching
- Wikipedia integration

Analysis Results

77.5%
Similarity

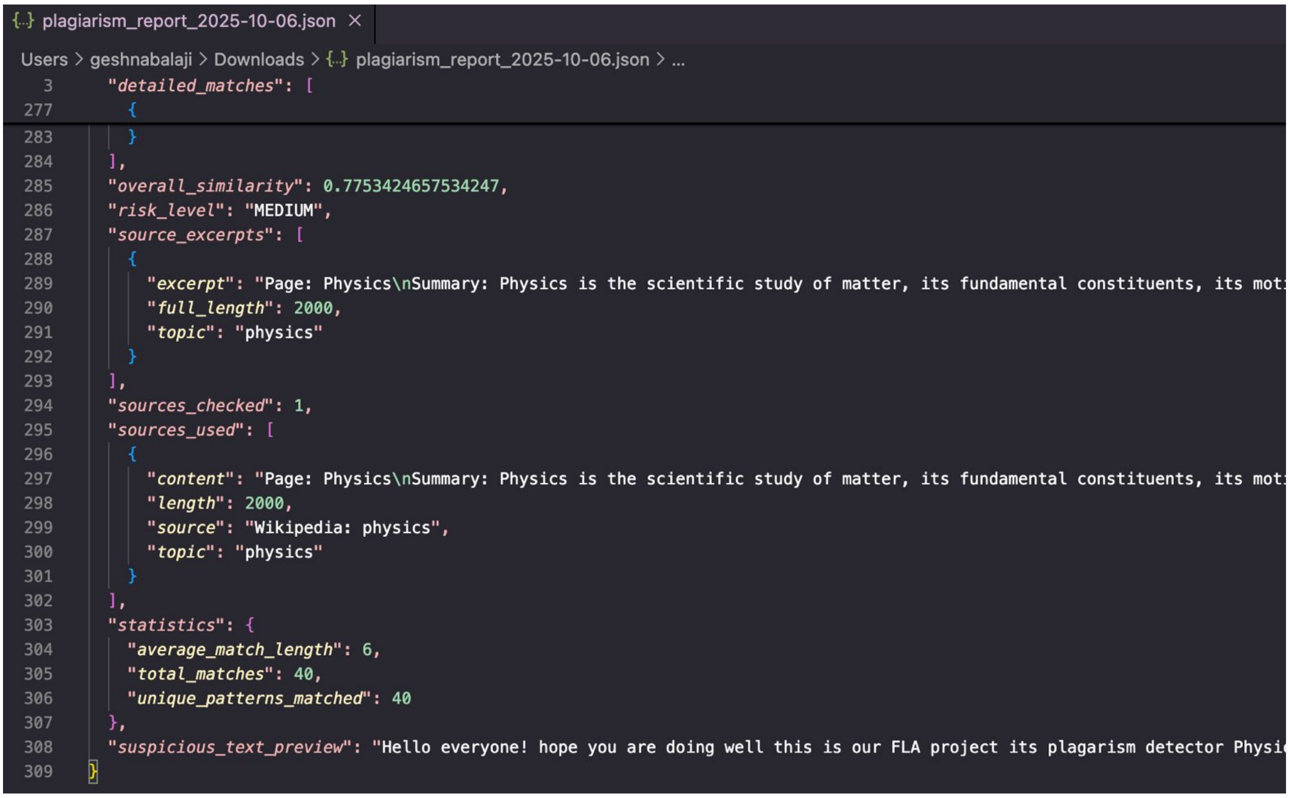
MEDIUM RISK

Re-analyze | Download Report

Wikipedia Sources Analyzed

> *“ Page: Physics Summary: Physics is the scientific study of matter, its fundamental constituents, its motion and behavior through space and time, and the related entities of energy and force. It is one of the most fundamental scientific disciplines. A scientist who specializes in the field of physics is called a physicist. Physics is one of the oldest academic disciplines. Over much of the past two millennia, physics, chemistry, biology, and certain branches of mathematics were a part of natural p... ”*

Fig. 3: Result page after plagiarism detection

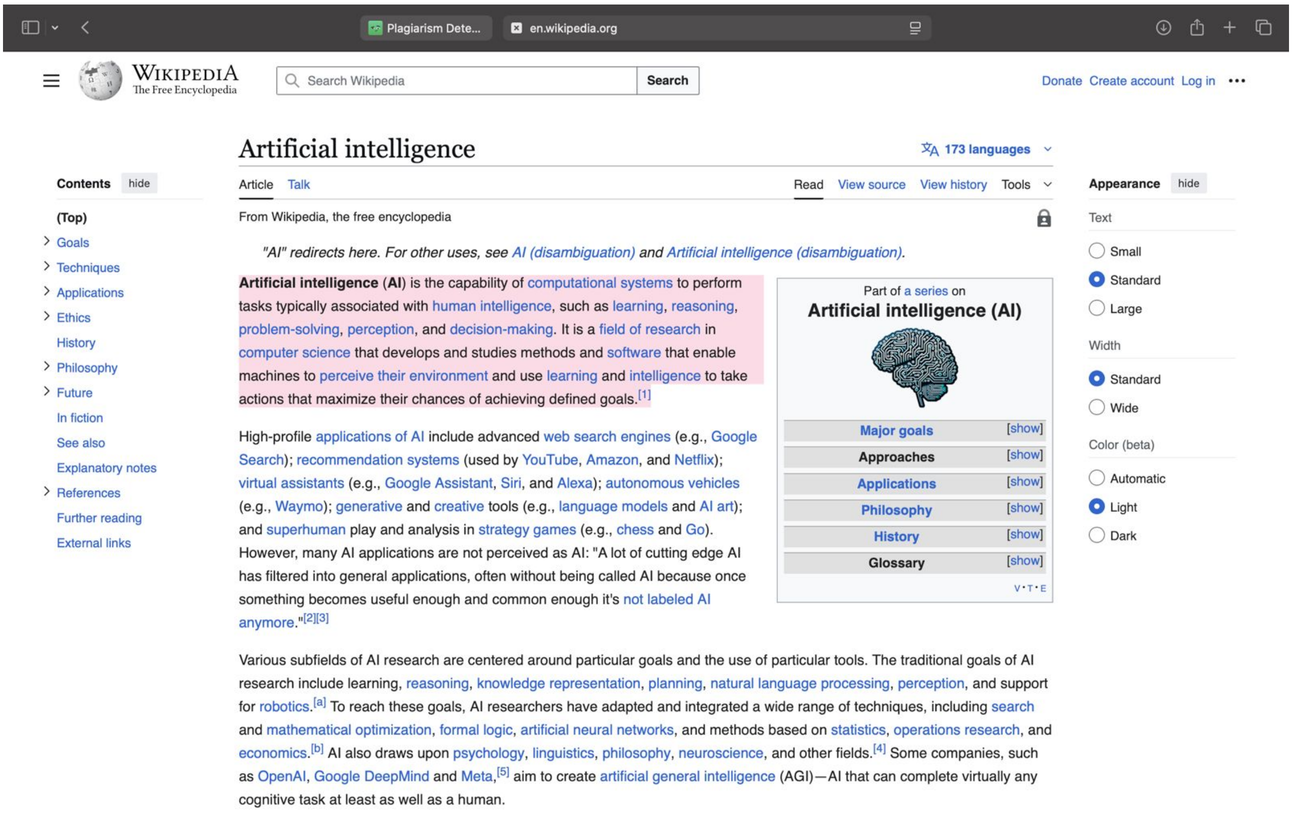


plagiarism_report_2025-10-06.json

Users > geshnabalaji > Downloads > {..} plagiarism_report_2025-10-06.json > ...

```
  "detailed_matches": [
    {
    }
  ],
  "overall_similarity": 0.7753424657534247,
  "risk_level": "MEDIUM",
  "source_excerpts": [
    {
      "excerpt": "Page: Physics\nSummary: Physics is the scientific study of matter, its fundamental constituents, its mot
      "full_length": 2000,
      "topic": "physics"
    }
  ],
  "sources_checked": 1,
  "sources_used": [
    {
      "content": "Page: Physics\nSummary: Physics is the scientific study of matter, its fundamental constituents, its mot
      "length": 2000,
      "source": "Wikipedia: physics",
      "topic": "physics"
    }
  ],
  "statistics": {
    "average_match_length": 6,
    "total_matches": 40,
    "unique_patterns_matched": 40
  },
  "suspicious_text_preview": "Hello everyone! hope you are doing well this is our FLA project its plagarism detector Physi
}
```

Fig. 4: Report Generated



Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see AI (disambiguation) and Artificial intelligence (disambiguation).

Artificial intelligence (**AI**) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It is a field of research in computer science that develops and studies methods and software that enable machines to perceive their environment and use learning and intelligence to take actions that maximize their chances of achieving defined goals.[1]

High-profile applications of AI include advanced web search engines (e.g., Google Search); recommendation systems (used by YouTube, Amazon, and Netflix); virtual assistants (e.g., Google Assistant, Siri, and Alexa); autonomous vehicles (e.g., Waymo); generative and creative tools (e.g., language models and AI art); and superhuman play and analysis in strategy games (e.g., chess and Go). However, many AI applications are not perceived as AI: "A lot of cutting edge AI has filtered into general applications, often without being called AI because once something becomes useful enough and common enough it's not labeled AI anymore."[2][3]

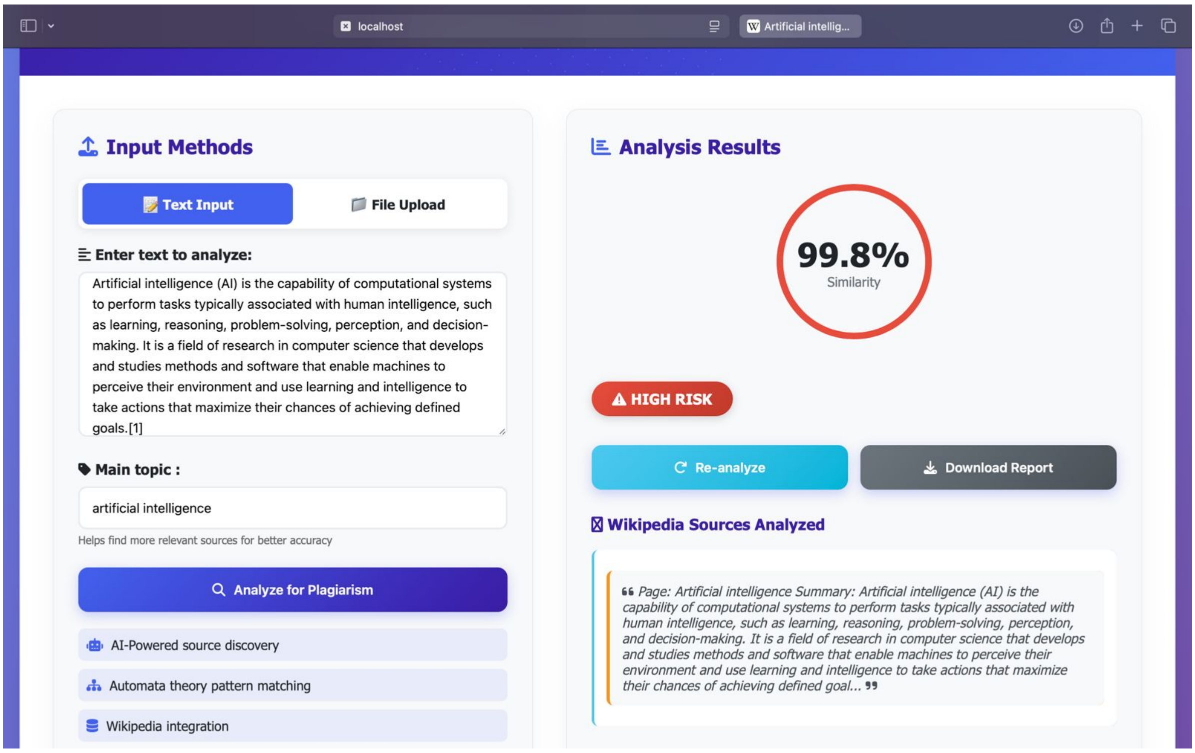
Various subfields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include learning, reasoning, knowledge representation, planning, natural language processing, perception, and support for robotics.[a] To reach these goals, AI researchers have adapted and integrated a wide range of techniques, including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, operations research, and economics.[b] AI also draws upon psychology, linguistics, philosophy, neuroscience, and other fields.[4] Some companies, such as OpenAI, Google DeepMind and Meta,[5] aim to create artificial general intelligence (AGI)—AI that can complete virtually any cognitive task at least as well as a human.

Text: ○ Small ● Standard ○ Large

Width: ● Standard ○ Wide

Color (beta): ○ Automatic ● Light ○ Dark

Fig. 5: Example text from wikeapedia



Enter text to analyze:

Artificial intelligence (AI) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It is a field of research in computer science that develops and studies methods and software that enable machines to perceive their environment and use learning and intelligence to take actions that maximize their chances of achieving defined goals.[1]

Main topic :

artificial intelligence

Helps find more relevant sources for better accuracy

99.8%
Similarity

Wikipedia Sources Analyzed

Page: Artificial intelligence Summary: Artificial intelligence (AI) is the capability of computational systems to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It is a field of research in computer science that develops and studies methods and software that enable machines to perceive their environment and use learning and intelligence to take actions that maximize their chances of achieving defined goal...

Fig. 6: Result

Conclusion and Future Work

In this project, we aimed to design and develop an AI-powered plagiarism detection system that integrates classical pattern matching algorithms with modern semantic analysis techniques. The core objective was to create a robust framework capable of detecting both exact textual overlaps and nuanced paraphrasing or idea-level plagiarism, which traditional tools often miss. Through the development process, we implemented efficient text preprocessing, applied the Aho–Corasick automaton for fast phrase matching, and leveraged advanced large language models for semantic topic expansion and contextual similarity assessment.

The completed system offers a user-friendly and automated pipeline that produces detailed plagiarism reports highlighting literal and conceptual matches. This hybrid approach ensures improved accuracy and interpretability, making the platform valuable for academic, research, and professional content originality verification. It conclusively demonstrates how combining classical algorithms with AI-driven semantic tools enhances plagiarism detection in terms of both speed and depth.

For future work, several improvements can be explored. Enhancing semantic detection accuracy through more advanced transformer-based models and multilingual support will broaden applicability. Integrating scalable vector databases for efficient retrieval over large document corpora will improve performance in real-world scenarios. Expanding the reference corpus beyond Wikipedia and incorporating adaptive context selection will enrich semantic comparison. Additionally, optimizing algorithm parallelization and incorporating explainability factors can make the system more transparent and efficient. Finally, implementing user feedback loops and deploying pilot tests will aid continuous refinement and adoption of the platform.

REFERENCES

[1] Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

[2] Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1), 45–62. https://doi.org/10.1007/s10579-010-9132-0

[3] Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2), 133–149. https://doi.org/10.1109/TSMCC.2011.2161825

[4] Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science*, 12(8), 1050–1084. https://www.jucs.org/jucs_12_8/plagiarism_a_survey

[5] Sharma, S., & Choudhary, A. (2019). Semantic plagiarism detection using Word2Vec and deep learning. *International Journal of Advanced Computer Science and Applications*, 10(3), 45–52. https://doi.org/10.14569/IJACSA.2019.0100307

[6] Alzahrani, S., & Salim, N. (2013). Intelligent plagiarism detection framework using semantic analysis. *Journal of King Saud University – Computer and Information Sciences*, 25(2), 123–132. https://doi.org/10.1016/j.jksuci.2013.02.001

[7] ScanMyEssay. (2021). Hybrid plagiarism detection methods. https://app-dev.scanmyessay.com/articles/hybrid-plagiarism-detection-methods

[8] ACM Digital Library. (2016). A hybrid approach for detection of plagiarism using NLP and text mining. https://dl.acm.org/doi/10.1145/2905055.2905061