International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 2, Mar-Apr 2024

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Plagiarism Detection Using Document Analysis

[1]Dr.B.Koti Ratnam, [2]CH.D.SAI KUMAR REDDY, [3]A.MANIKANTA, [4]D.RAGHU BABU, [5] G.NIKHIL

[1]Associate Professor, Department of IT,
KKR And KSR Institute of Technology And Sciences, Guntur Dt., Andhra Pradesh.
[2,3,4,5]Students, Department of IT,
KKR And KSR Institute of Technology And Sciences, Guntur Dt., Andhra Pradesh.
Email: 20jr1a1249@gmail.com)

----------**********----------

Abstract:

Plagiarism Detection Systems serve a crucial role in uncovering instances of plagiarism, particularly within the educational sector involving scientific documents and papers. Plagiarism occurs when content is reproduced without proper authorization or citation from the original author.. In this project, we propose a novel approach to plagiarism detection utilizing document analysis techniques. Our objective is to develop a robust and efficient system capable of identifying instances of plagiarism by analyzing the textual content of documents. Leveraging machine learning and natural language processing methods, we preprocess the text data, extract relevant features, and train models to distinguish between original and plagiarized content. Evaluation of our system's performance is conducted using established metrics such as accuracy, precision, recall, and F1-score. Through experimentation and refinement, we aim to create a scalable and effective solution for detecting plagiarism, thereby promoting academic integrity and fostering originality in scholarly discourse. Our findings contribute to the ongoing efforts to combat plagiarism and uphold the standards of ethical scholarship in the digital age.

Keywords **—Plagiarism Detection, Document Analysis, Semantic Analysis, Machine Learning Algorithms, Formal Concept Analysis(FCA), Intellectual Property infringement, Acadamic Integrity.**

----------**********----------

I. INTRODUCTION

In the digital age, the ease of access to vast amounts of information has led to a rise in the occurrence of plagiarism. This project introduces an innovative plagiarism detection system that utilizes state-of-the-art document analysis techniques to combat this issue. The system is designed to identify and analyze patterns of text that may indicate plagiarism, including exact text matches, paraphrasing, and idea replication.The introduction of machine learning algorithms and semantic analysis into the plagiarism detection process allows for a more nuanced and comprehensive examination of documents. This project aims to create a tool that not only serves academic and professional communities by protecting intellectual property but also promotes the creation of original content and the maintenance of academic integrity.

Through the implementation of this system, we seek to provide a reliable and efficient solution for identifying plagiarism in a variety of contexts. The project will undergo rigorous testing to ensure its effectiveness and adaptability to different types of documents and plagiarism methods. Ultimately, this project aspires to set a new standard in plagiarism detection technology, offering a robust defense against the misuse of intellectual material.

II. LITERATURE REVIEW

[1]Plagiarism Detection System (EPDS) that This study introduces an External combines Semantic Role Labeling (SRL) with semantic and syntactic information to address

shortcomings in existing methods. Unlike many approaches, it accurately discerns meaning when comparing source and suspicious document sentences with similar surface text, thereby reducing erroneous matches.

By leveraging SRL, it effectively handles sentence transformations like active to passive voice. Additionally, the method employs content word expansion to identify similar ideas expressed differently, enabling detection of various plagiarism types such as verbatim copying and paraphrasing. Experimental results demonstrate superior performance compared to existing techniques, including those in PAN-PC-11.

[2] Copy detection plays a crucial role in safeguarding intellectual property and facilitating efficient information retrieval. Historically focused on program plagiarism detection, the field has shifted towards text copy detection. This study presents a thorough survey of natural language text copy detection, outlining its evolution and key developments. It reviews various existing systems and prototypes, detailing their approaches and features. Additionally, it compares key detection techniques and discusses future trends in text copy detection.

[3] This study delves into unmasking text plagiarism through syntactic-semantic natural language processing techniques. It provides a comprehensive comparison, analysis, and examination of challenges in this domain. By employing advanced syntactic and semantic analysis, the study aims to enhance plagiarism detection accuracy. The paper scrutinizes existing methodologies, identifies strengths and weaknesses, and discusses the hurdles faced in implementing these techniques effectively. Through rigorous comparison and analysis, it offers insights into improving the efficacy of plagiarism detection in textual content.

III. PROPOSED METHODOLOGY

The proposed system is a plagiarism detection tool utilizing document analysis. It is built upon Streamlit, integrating various plugins for enhanced functionality. Users can input text, and the system employs advanced language models, including Hugging Face's ChatBot API and LangChain's embeddings, to generate responses. Additionally, it incorporates a web search plugin, allowing for internet-based information retrieval to enrich responses. The system maintains a conversation history and provides accurate plagiarism detection by analyzing text and identifying sources. Furthermore, it offers features like user authentication through Hugging Face, ensuring data security and personalized experiences. The system aims to streamline the process of plagiarism detection, providing users with efficient and reliable results for their text analysis needs.

IV. PROCESS/LOGICAL DESIGN

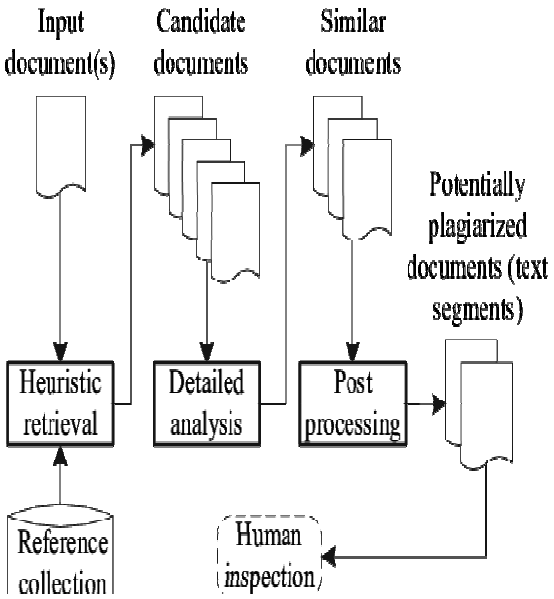

Fig1: workflow of plagiarism detection

V. MODULE DESIGN SPECIFICATION

The modules that are used to design plagiarism detection are

1.User Interface

2.Registration & Login Module

3.Text Input Module

4.Text Processing Module

5.Output Module

User interface module

The application allows users to upload documents and customize analysis settings for tailored plagiarism detection. It provides real-time feedback on analysis progress, ensuring transparency throughout the process. Once complete, analysis results are presented clearly, highlighting instances of plagiarism for easy identification. Users can download reports and export results for further review and action. Overall, the application offers a comprehensive solution for efficient and effective plagiarism detection and management.

Registration and login module

The application offers a streamlined process for new users to create accounts by inputting necessary information such as username, email, and password. It meticulously verifies user information to ensure authenticity and prevent unauthorized access, maintaining the security of user accounts. Additionally, the system provides convenient functionality for users to reset forgotten passwords and update their account credentials as needed. This ensures users have control over their account information and can easily manage their account security and accessibility. Overall, the application prioritizes user authentication and security, enhancing the user experience and safeguarding user data.

Text Input Module

The system provides users with the flexibility to upload text in various formats directly to the platform, enhancing convenience during content submission. Users can also input text by copying and pasting from diverse sources such as word processors and web browsers. Moreover, a dedicated text input area within the interface facilitates seamless manual typing or editing, accommodating different user preferences. This versatile approach ensures ease of access and input for textual content, promoting user engagement. With multiple options for text input, users can effortlessly contribute their content to the system, regardless of its original format. Overall, the system prioritizes user convenience and accessibility, enhancing the overall user experience.

Text processing module

The system begins by performing noise removal on the text data, eliminating extraneous elements such as special characters, punctuation, and HTML tags. This preprocessing step ensures that the text is clean and devoid of irrelevant information, thereby enhancing the accuracy of subsequent analyses. Subsequently, the text is split into individual words or tokens, enabling more granular analysis and processing. Additionally, inflected words are reduced to their base or root form through techniques such as lemmatization or stemming. This normalization process ensures consistency in the analysis by standardizing variations of words, facilitating more effective comparison and interpretation of the text data. Overall, these preprocessing steps lay the groundwork for robust and insightful analysis of the text content.

Output module

The system offers a comprehensive presentation of the output from the plagiarism detection analysis, ensuring clarity and transparency for users. Visualizations are incorporated to provide users with a clear representation of the extent of similarity between documents, aiding in the interpretation of results. Detailed analysis reports are generated, highlighting instances of potential plagiarism within the documents under analysis, enabling users to identify and address issues effectively. Additionally, the system allows for user interaction with the analysis results, facilitating navigation through the generated output and enhancing user engagement. Furthermore, a dedicated help resources module may be available, providing users with access to additional information and support for better understanding and utilization of the system's features. Overall, these features contribute to a user-friendly and informative experience, empowering users in their plagiarism detection and analysis endeavors.

ACKNOWLEDGMENT

"We extend our heartfelt gratitude to our dedicated project supervisor,whose invaluable guiuserdance, expertise, and unwavering support were instrumental in steering our efforts towards effectively detecting plagiarism through document analysis. We also extend special thanks to institution for generously providing access to essential resources, datasets, and facilities crucial for conducting our research. The collective commitment and collaborative spirit of our research team played a pivotal role in driving progress and achieving milestones throughout this endeavor. Moreover, we express our sincere appreciation to the wider academic community and professionals specializing in document analysis and plagiarism detection whose pioneering research and insights

have significantly informed and inspired our innovative approach. Additionally, we are profoundly grateful for the unwavering encouragement, understanding, and support extended by our friends and family, whose steadfast belief in our capabilities served as a constant source of motivation and inspiration. Furthermore, we acknowledge with gratitude the invaluable contributions and expertise shared by professionals in the fields of document analysis and plagiarism detection, whose collaboration and insights have profoundly shaped and enriched our methodology. With collective support and collaboration, we were able to realize our objectives and achieve success in our endeavors. This project's success is owed to the dedication, support, and contributions of all involved, for which we are immensely
thankful."

REFERENCES

[1] JirapondMuangprathub, Siriwan Kajornkasirat, and ApiratWanichsombat "Document Plagiarism Detection Using a New Concept Similarity in Formal Concept Analysis" Hindawi journal of applied sciences, Article ID 6662984,vol-132.

[2] A. Abdi, S. M. Shamsuddin, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, "A linguistic treatment for automatic external plagiarism detection," Knowledge-Based Systems, vol. 135, pp. 135–146, 2017.

[3] L. Ahuja, V. Gupta, and R. Kumar, "A new hybrid technique for detection of plagiarism from text documents," Arabian Journal for Science and Engineering, vol. 45, pp. 1–14, 2020.

[4] K. Vani and D. Gupta. **2016**. "Study on Extrinsic Text Plagiarism Detection Techniques And Tools," *J. Eng. Sci. Technol. Rev.*, **9**(5): 9–23, 2016, doi: 10.25103/jestr.095.02.

[5] H. A. Chowdhury and D. K. Bhattacharyya. **2018**. "Plagiarism: Taxonomy, tools and detection techniques," *arXiv*, **9**: 1–15.

[6] M. Sahi and V. Gupta, "A novel technique for detecting plagiarism in documents exploiting information sources," Cognitive Computation, vol. 9, no. 6, pp. 852–867, 2017.

[7] B. Ganter and R. Wille, "Applied lattice theory: formal concept analysis," in In General Lattice Theory, G. Grätzer, Ed., Birkhäuser, 1997.

[8] B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer Science & Business Media, 2012.

[9] R. Wille, "Formal concept analysis as mathematical theory of concepts and concept hierarchies," in Formal concept analysis, pp. 1–33, Springer, Berlin, Heidelberg, 2005.

[10] U. Priss, "Formal concept analysis in information science," Annual Review of Information Science and Technology, vol. 40, no. 1, pp. 521–543, 2006.

[11] A. E. Qadi, D. Aboutajedine, and Y. Ennouary, "Formal concept analysis for information retrieval," International Journal of Computer Science and Information Security, vol. 7, no. 2, pp. 119–125, 2010.

[12] A. Formica, "Concept similarity in formal concept analysis: an information content approach," Knowledge-Based Systems, vol. 21, no. 1, pp. 80–87, 2008.

[13] A. Formica, "Ontology-based concept similarity in formal concept analysis," Information Sciences, vol. 176, no. 18, pp. 2624–2641, 2006.

[14] C. Carpineto and G. Romano, Concept Data Analysis: Theory and Applications, John Wiley & Sons, 2004.

[15] I. Nafkha, S. Elloumi, and A. Jaoua, "Using concept formal analysis for cooperative information retrieval," The Leech, vol. 1, no. 1, 2004.

[16] M. K. M. Rahman and T. W. Chow, "Content-based hierarchical document organization using multi-layer hybrid network and tree-structured features," Expert Systems with Applications, vol. 37, no. 4, pp. 2874–2881, 2010.

[17] K. Baba, "Fast plagiarism detection based on simple document similarity," in 2017 twelfth international conference on digital information management (ICDIM), pp. 54–58, Fukuoka, Japan, 2017.

[18] J. Muangprathub, V. Boonjing, and P. Pattaraintakorn,"A new case-based classification using incremental concept lattice knowledge," Data & Knowledge Engineering, vol. 83, no. 1, pp. 39–53, 2013.

[19] F. Alqadah, "Similarity measures in formal concept analysis," in Workshops of the 11th International Symposium on Artificial Intelligence and Mathematics (ISIAM2010), Fort Lauderdale, Florida, 2010.

[20] F. Alqadah and R. Bhatnagar, "Similarity measures in formal concept analysis," Annals of Mathematics and Artificial Intelligence, vol. 61, no. 3, pp. 245–256, 2011.

[21] F. Dau, J. Ducrou, and P. Eklund, "Concept similarity and related categories in searchsleuth," in

International conference on conceptual structures, pp. 255–268, Berlin, Heidelberg, 2008.

[22] J. Saquer and J. S. Deogun, "Concept approximations based on rough sets and similarity measures," International Journal of Applied Mathematics and Computer Science, vol. 11, pp. 655– 674, 2001. Journal of Applied Mathematics 9

[23] K. X. S. de Souza and J. Davis, "Aligning ontologies and evaluating concept similarities," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems, pp. 1012–1029, Berlin, 2004.

[24] L. Wang and X. Liu, "A new model of evaluating concept similarity," Knowledge-Based Systems, vol. 21, no. 8, pp. 842–846, 2008.

[25] R. Belohlavek and V. Vychodil, "Estimations of similarity in formal concept analysis of data with graded attributes," Advances in Web Intelligence and Data Mining, vol. 23, no. 1, pp. 243–252, 2006.

[26] A. Formica and E. Pourabbas, "Content based similarity of geographic classes organized as partition hierarchies," Knowledge and Information Systems, vol. 20, no. 2, pp. 221–241, 2009.

[27] A. Tversky, "Features of similarity," Psychological Review, vol. 84, no. 4, pp. 327–352, 1977.C. Carpineto, G. Romano, and F. U. Bordoni, "Exploiting the potential of concept lattices for information retrieval with CREDO," Journal of Universal

