Project Synopsis
on

Solution to prevent copyright infringement/piracy/plagiarism of NCERT text books.

Submitted as a part of course curriculum for

Bachelor of Technology
in
Computer Science

Submitted by

Shantanu Mishra (2100290120152)

KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)
Department of Computer Science and Engineering
Dr. A.P.J. Abdul Kalam Technical University
2022-2023

Under the Supervision of
Dr. Ajay Kumar Shrivastava
Professor - CS

ACKNOWLEDGEMENT

We take immense pleasure in presenting the comprehensive synopsis of our B.Tech Mini Project, a significant undertaking that marked a pivotal milestone during the course of our B.Tech Third Year. Our journey through this project has been enriched by the unwavering support and guidance of Dr. Ajay Kumar Shrivastava, Professor, Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad. Dr. Shrivastava's dedication, sincerity, and steadfast perseverance have not only been a guiding force but also a constant source of inspiration, elevating our project to its current standing. We owe the successful realization of our endeavors to his conscientious efforts.

We also extend our sincere appreciation to Dr. Ajay Kumar Shrivastava, Head of the Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, for his integral role in providing full support and invaluable assistance throughout the developmental phases of our project. The collaborative efforts of the entire faculty in the department have played a pivotal role in shaping the project, and we express our gratitude for their constant assistance and cooperation.

Acknowledging the crucial role played by my colleagues, I am indebted to each one of them for their hard work, untiring efforts, and mutual cooperation, all of which contributed significantly to the success of the project. Their inspiration and encouragement have been instrumental in keeping our spirits high and ensuring the project's successful outcome.

Finally, our heartfelt gratitude goes out to our friends whose contributions, support, and camaraderie played an integral part in the completion of the project. Together, we have achieved a commendable feat, and we look forward to further collaborative ventures in the future.

Any omission in this brief acknowledgement does not mean lack of gratitude.

Team :
Shantanu Mishra (2100290120152)
Rajiv Kumar Singh (2100290120137)
Raunak Jain (2100290120140)
Md. Sahil (2100290120105)

Signature:

INTRODUCTION

1. Introduction:

In the dynamic realm of academic research and publication, the pervasive issue of plagiarism poses a formidable challenge, casting a shadow over the authenticity and integrity of scholarly contributions. Plagiarism, encapsulating the unattributed use of someone else's ideas, expressions, or textual content, has long been condemned as a grave offense within academic and publishing circles. However, the contemporary landscape witnesses a transformative era where technological advancements have ushered in sophisticated systems to detect and mitigate plagiarism, thereby safeguarding intellectual property.

This synopsis endeavors to address the complex issue of plagiarism by proposing the development of an advanced tool. This tool seeks to discern the originality of publications by meticulously analyzing distinctive paper characteristics, including watermarks, paper quality, text formatting, and the physical dimensions of books. Its overarching goal is to determine whether the contents of textbooks from various publishers have been subjected to acts of plagiarism. As the academic community grapples with nuanced forms of plagiarism, ranging from the replication of exact sentences to the misappropriation of novel ideas, there is an inherent need to explore innovative approaches for detection.

Drawing inspiration from Teddi Fishman's comprehensive definition of plagiarism and recognizing its pervasive impact on academic integrity, the proposed tool aligns with the collective effort to curb plagiarism and uphold the principles of intellectual honesty. Leveraging technological advancements and acknowledging the intricate nature of plagiarism, the tool aims to address existing limitations in detecting copied tables, figures, or formulae across diverse academic disciplines.

The chosen research methodology for this review involves a thorough exploration of existing plagiarism detection techniques. This includes an in-depth examination of their descriptions, taxonomies, strengths, and limitations. Furthermore, this paper delves into the far-reaching implications of plagiarism, emphasizing its severe consequences on the educational process. The comprehensive exploration of techniques designed to identify various forms of plagiarism serves as the cornerstone for proposing optimized and efficient plagiarism detection methodologies.

Structured to provide a panoramic view of the problem statement, the existing research landscape, and the proposed solution, this synopsis sets the stage for a detailed discussion. The subsequent sections will delve into the intricacies of the research methodology, potential threats to validity, present results and discussions, and ultimately conclude with insights that contribute meaningfully to the ongoing discourse on plagiarism detection. Through this exploration, we aim to contribute valuable perspectives that further enhance our collective understanding and strategies for tackling plagiarism in academic settings.

2. Problem Statement:

The unauthorized reproduction and plagiarism of NCERT textbooks persist as a significant challenge, threatening the integrity of educational content. Current efforts lack a robust tool to identify pirated publications based on paper characteristics, such as watermarks, paper quality, and text formatting. The absence of an effective solution allows for the proliferation of unauthorized reproductions, necessitating a comprehensive tool capable of scanning and verifying the authenticity of NCERT textbooks across various publishers, whether accessed through mobile phones, other devices, or inputted files. This gap highlights the critical need for a solution to prevent copyright infringement and plagiarism of NCERT textbooks.

3. Objective:

The major objectives of the project includes:

- Enhance the efficacy of existing plagiarism detection tools.
- Propose an optimized model leveraging machine learning and deep learning techniques.
- Develop a model that significantly reduces computational costs.
- Create a versatile product applicable to both research papers and assignments.

4. Scope:

This project delves into the critical task of mitigating the pervasive issues of copyright infringement, piracy, and plagiarism within the educational domain, with a primary emphasis on National Council of Educational Research and Training (NCERT) textbooks. In addition to addressing these challenges, the scope of the project expands its reach to encompass broader applications. The proposed model, developed through the integration of cutting-edge technologies, methodologies, and efficient algorithms, is designed not only for textbooks but also for a wider range of educational materials.

By adopting a holistic approach, the project aims to provide a comprehensive solution that transcends the boundaries of traditional learning resources. This proposed model holds the potential to extend its utility to diverse forms of academic content, including research papers, novels, and assignments. The integration of advanced technologies ensures adaptability and effectiveness across various educational materials, thereby enhancing its relevance and applicability in contemporary educational settings.

The outcomes of this project are envisioned to make significant contributions to the discourse on plagiarism prevention and intellectual property protection within the educational landscape. By extending its utility beyond textbooks to encompass a broader spectrum of academic materials, the proposed model aligns with the evolving nature of educational resources, meeting the dynamic needs of academic institutions and scholars.

LITERATURE REVIEW

Reliable plagiarism detection system based on deep learning approaches
Mohamed A. El-Rashidy, Ramy G. Mohamed, Nawal A. El-Fishawy & Marwa A. Shouman
Journal : Neural Computing and Applications (2022)

The paper investigates the emergence of scientific burglary driven by technological advancements in software and discusses potential countermeasures. It emphasizes the persisting challenges in detecting lexical, syntactic, and semantic text plagiarism, leading to the creation of a new database. This innovative database, designed for intelligent learning solutions, serves as a foundation for enhancing text plagiarism detection methods. The paper details the development of an advanced plagiarism detection system using deep learning, incorporating convolutional and recurrent neural network architectures. The long short-term memory (LSTM) approach is identified as the leading solution, outperforming contemporary systems in rigorous evaluations on benchmark datasets.
Furthermore, the paper underscores the significance of continuous adaptation and updates for the proposed plagiarism detection system to effectively combat evolving techniques employed by scientific burglars. The research contributes valuable insights into the ongoing battle against plagiarism, paving the way for future advancements in safeguarding intellectual property and ensuring the integrity of scientific research.

A Detection Method for Plagiarism Reports of Students
Daisuke Sakamoto a, Kazuhiko Tsuda b
Journal : Procedia Computer Science

The paper introduces an efficient method for detecting plagiarism in student reports by transforming texts into one-dimensional strings, using repeating shifts and mutual comparisons, and checking matching word sections. The authors suggest enhancements involving heuristics, estimating detection possibility and string compression, to improve accuracy and reduce calculation time. The method, characterized by having a single parameter for plagiarism judgment, is deemed intuitive and easy to set, providing an effective plagiarism detection solution.
Additionally, the proposed methodology is robust in detecting perfect matches of any size within the text, regardless of their position. By incorporating heuristics, the authors aim to further refine the accuracy of identifying plagiarism sections while simultaneously streamlining the computation process. The combination of these enhancements not only strengthens the plagiarism detection capabilities but also makes the method user-friendly with its simplified parameterization. Overall, this approach presents a comprehensive and efficient solution for addressing plagiarism in student reports, offering both accuracy and practicality in implementation.

Plagiarism Detection Using Semantic Knowledge Graphs
Kunal Khadilkar; Siddhivinayak Kulkarni; Poojarani Bone
Journal : IEEE Xplore

The paper discusses challenges in current plagiarism detection tools that heavily rely on string matching algorithms, which struggle to identify altered sentence structures and synonyms. To address these limitations, the paper proposes a new approach using semantic knowledge graphs. By incorporating Named Entity Recognition (NER) and assessing semantic similarity between sentences, the method aims to overcome the shortcomings of traditional tools, providing a more advanced and accurate plagiarism detection system. Additionally, the paper emphasizes the integration of linguistic rules, such as those governing active and passive voice, to enhance the robustness of the detection process and offer a more thorough assessment of textual authenticity.

Reducing Book Piracy: The Role of the HigherEducation Sector
Evelyn Chiyevo GARWE
Journal : International Research in Education

The paper examines the impact of book piracy, highlighting both positive outcomes for beneficiaries and negative consequences for copyright owners in the literary industry. It emphasizes the delicate balance between advantages to certain groups and the challenges faced by those with intellectual property rights. Additionally, the literature explores strategies to combat book piracy, specifically focusing on the role of the higher education sector. The review suggests that the sector's contribution could lead to a broader national reduction in book piracy, offering valuable insights for librarians, higher education institutions, and policymakers.

Testing of support tools for plagiarism detection
Tomáš Foltýnek, Dita Dlabolová, Alla Anohina-Naumeca, Salim Razı, Július Kravjar, Laima Kamzola, Jean Guerrero-Dib, Özgür Çelik & Debora Weber-Wulff
Journal : International Journal of Educational Technology in Higher Education

The study addresses challenges in plagiarism detection systems, highlighting their limitations in accurately identifying instances of plagiarism and often misidentifying non-plagiarized content. It proposes a realistic educational simulation approach that mimics actual educational settings, taking into account common student behaviors, instructor constraints, and the need for cost-effective solutions. The study acknowledges unexplored factors and limitations, such as untested assumptions regarding cost negotiation and pricing models. It recommends further investigation into advanced plagiarism-disguising techniques and the effectiveness of systems in handling legal and administrative aspects.
In conclusion, the study underscores the need for continuous refinement of plagiarism detection methods to address the identified challenges and unexplored factors. It emphasizes the importance of research into advanced techniques and the comprehensive evaluation of systems to ensure their efficacy in diverse educational environments. The findings aim to contribute to the ongoing improvement of plagiarism detection systems and their adaptation to the evolving landscape of academic integrity.

An effective text plagiarism detection system based on feature selection and SVM techniques
Mohamed A. El-Rashidy, Ramy G. Mohamed, Nawal A. El-Fishawy & Marwa A. Shouman
Journal : Multimedia Tools and applications

The prevalence of text plagiarism has become a widespread issue in various fields, including research manuscripts, textbooks, patents, and academic circles. Existing methods to detect plagiarism face challenges in discriminating between different types of similarity, such as lexical, syntactic, and semantic. In response, a novel plagiarism detection system is proposed, utilizing effective sentence similarity features and a hyperplane equation constructed through Support Vector Machine (SVM) and Chi-square techniques. The system undergoes three phases: preprocessing documents, traditional paragraph-level comparison, and utilizing the computed hyperplane equation. Evaluation on benchmark datasets demonstrates the system's superior performance, achieving the highest Plagdet and F-measure scores in comparison to other systems in recent years, with values of 89.12% and 92.91% on the PAN 2013 dataset and 89.34% and 92.95% on the PAN 2014 dataset.

Testing of support tools for plagiarism detection
Tomáš Foltýnek, Dita Dlabolová, Alla Anohina-Naumeca, Salim Razı, Július Kravjar, Laima Kamzola, Jean Guerrero-Dib, Özgür Çelik & Debora Weber-Wulff
Journal : International Journal of Educational Technology in Higher Research

The belief that software should easily perform tasks challenging for humans, such as detecting plagiarism, is widespread. While software can serve as a support tool for identifying text similarity indicative of plagiarism, a collaborative test involving 15 web-based text-matching systems revealed mixed results. Researchers from seven countries assessed the systems on material in eight languages, testing their effectiveness on both single-source and multi-source documents. The findings indicate that while some systems can help identify plagiarized content, they fall short of detecting all instances and may mistakenly flag non-plagiarized material as problematic. The study emphasizes the limitations and varying degrees of success among different plagiarism detection systems.

Improving plagiarism detection in text document using hybrid weighted similarity
Hamed Arabi a, Mehdi Akbari a b
Journal : Expert Systems with Applications

Plagiarism, the unauthorized use of content from other sources without proper attribution, is a growing concern for publishers and researchers. Existing methods often struggle to detect intelligent plagiarism, focusing primarily on direct copying. This study introduces two methods to identify Extrinsic plagiarism, utilizing a combination of pre-trained word embedding techniques like FastText and TF-IDF weighting at both document and sentence levels. The first method achieves 95.1% precision, while the second method reaches 93.8% precision, indicating that word embedding networks are more effective in detecting Extrinsic plagiarism compared to WordNet ontology. The use of these methods demonstrates an improvement in identifying sophisticated instances of plagiarism, addressing the limitations of traditional detection methods.

Plagiarism Detection of Images
Amirul S. Bin Ibrahin; Othman O. Khalifa; Diaa Eldein M. Ahmed
Journal : IEEE Xplore

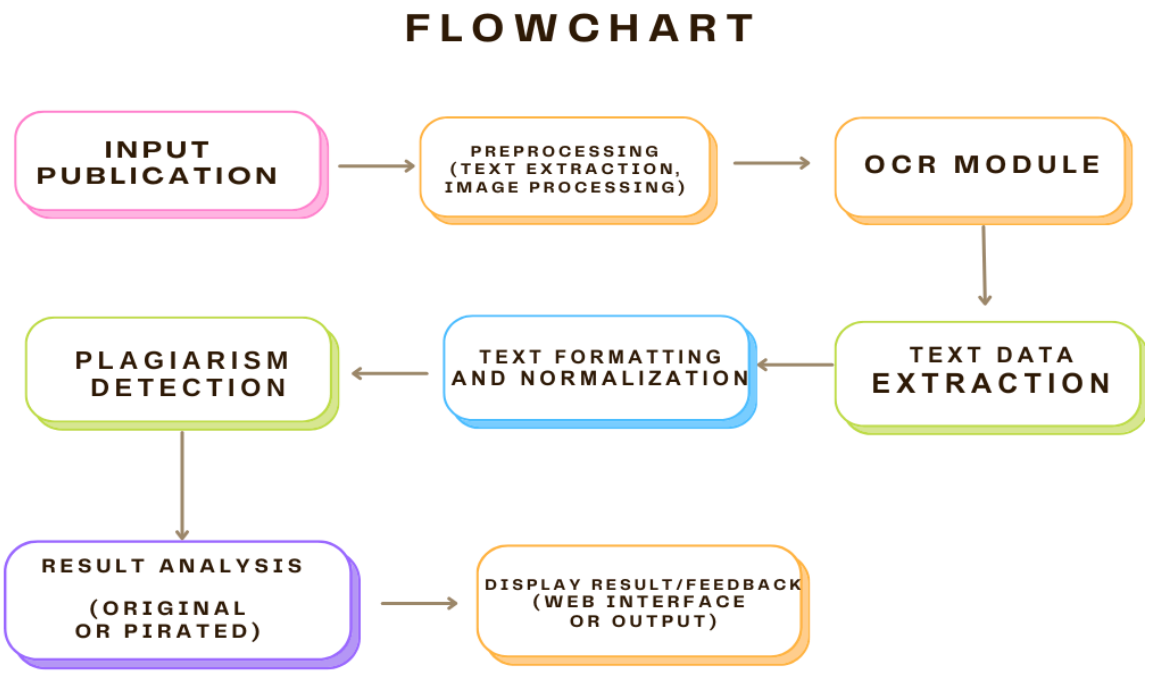
The passage discusses plagiarism and the detection process, highlighting the challenges in existing systems. It suggests a new three-stage system involving pre-processing, feature extraction, and comparison. The results show ascending similarity indexes and true/false outcomes, with a 100% accuracy in detecting unedited images but varying accuracy for edited ones, such as flipped, rotated, greyscale, and cropped images. The proposed system aims to address flaws in current plagiarism detection methods.

An Adaptive Image-based Plagiarism Detection Approach
Authors: Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel Keim, Bela Gipp
Journal : ACM

Detecting plagiarism in academic documents is crucial, and existing systems often struggle to identify disguised forms such as paraphrases and translations. To address this, a proposed image-based plagiarism detection approach analyzes images in academic documents as text-independent features. This adaptive and scalable approach integrates established image analysis methods with new similarity assessments, achieving a recall of 0.73 and precision of 1 in testing. The results suggest that this image-based method can complement other content-based approaches, providing a tool to retrieve potential source documents for suspiciously similar content from large collections. The code for this approach is open source, encouraging further research in image-based plagiarism detection.

PROPOSED METHODOLOGY

1. **Flowchart:**



2. Algorithm Proposed:

2.1. Dataset Making :

For the successful training of the deep learning model, a substantial dataset is imperative. In this project, we are in the process of compiling a dataset that encompasses all NCERT books from the 6th to the 12th standard. The objective is to create a dataset containing approximately 100,000 lines of data. Each line will consist of around 7-8 words, after the removal of irrelevant symbols and junk characters. This carefully crafted dataset forms the foundation for training the deep learning model, ensuring it is exposed to a diverse set of examples from NCERT textbooks. The removal of extraneous symbols streamlines the dataset, facilitating cleaner and more focused data for training. This dataset's creation is a crucial step in preparing the deep learning model to effectively identify instances of plagiarism and copyright infringement within the NCERT educational materials.

2.2. Model for the detection:

The model is constructed using the BERT machine learning framework, which is inherently bidirectional, allowing it access to both left and right context words. This bidirectional nature makes BERT particularly suitable for tasks like text classification. Additionally, considering the substantial size of our dataset, exploring Language Model (LLM) models is also a viable option.

TECHNOLOGY USED

Probable tech stack that is going to be used:

1. Python
2. Image Processing Libraries
3. Optical Character Recognition (OCR)
4. Machine Learning/Deep Learning Libraries
5. Natural Language Processing (NLP) Libraries
6. WebFramework
7. Database
8. Front-EndDevelopment
9. Model Deployment (Flask, Django, Node.js)

CONCLUSION

In conclusion, the development of this sophisticated tool, utilizing the BERT machine learning framework, stands poised to make substantial contributions to the realm of plagiarism detection within educational materials, specifically focusing on NCERT textbooks. The bidirectional nature of the model, endowed by BERT, enhances its efficacy in understanding contextual nuances and makes it well-suited for text classification tasks. Moreover, the consideration of Language Model (LLM) models, given the substantial dataset at our disposal, adds a layer of adaptability and potential for nuanced analysis.

As we move forward with the training phase, the fusion of advanced technologies, methodologies, and a meticulously curated dataset underscores our commitment to creating a robust and versatile tool. This tool not only addresses the pervasive issues of copyright infringement and plagiarism within NCERT textbooks but also holds promise for broader applications in the educational landscape.

The training process, leveraging the wealth of data spanning from the 6th to the 12th standard, is pivotal for ensuring the model's proficiency and adaptability. The comprehensive dataset, meticulously prepared and refined, forms the backbone of our efforts, setting the stage for a sophisticated and adept tool capable of identifying instances of plagiarism with precision.

As we embark on the next phases of this project, the outcomes aim to contribute significantly to the ongoing discourse on plagiarism prevention and intellectual property protection within educational resources. The culmination of these efforts holds the potential to not only enhance academic integrity but also redefine the landscape of plagiarism detection in the evolving realm of educational materials.

REFERENCES

[1] Altheneyan, A. S., & Menai, M. E. B. (2020). Automatic plagiarism detection in obfuscated text. Pattern Analysis and Applications, 23, 1627-1650.

[2] AlSallal, M., Iqbal, R., Amin, S., James, A., & Palade, V. (2016, August). An integrated machine learning approach for extrinsic plagiarism detection. In 2016 9th International Conference on Developments in eSystems Engineering (DeSE) (pp. 203-208). IEEE.

[3] AlSallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, 96, 700-712.

[4] El-Rashidy, M. A., Mohamed, R. G., El-Fishawy, N. A., & Shouman, M. A. (2024). An effective text plagiarism detection system based on feature selection and SVM techniques. Multimedia Tools and Applications, 83(1), 2609-2646.

[5] Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razı, S., Kravjar, J., Kamzola, L., ... & Weber-Wulff, D. (2020). Testing of support tools for plagiarism detection. International Journal of Educational Technology in Higher Education, 17, 1-31.

[6] Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR), 52(6), 1-42.

[7] Gayadhankar, K., Patel, R., Lodha, H., & Shinde, S. (2021). Image plagiarism detection using gan-(generative adversarial network). In ITM Web of Conferences (Vol. 40, p. 03013). EDP Sciences.

[8] Gupta, D. (2016). Study on Extrinsic Text Plagiarism Detection Techniques and Tools. Journal of Engineering Science & Technology Review, 9(5).

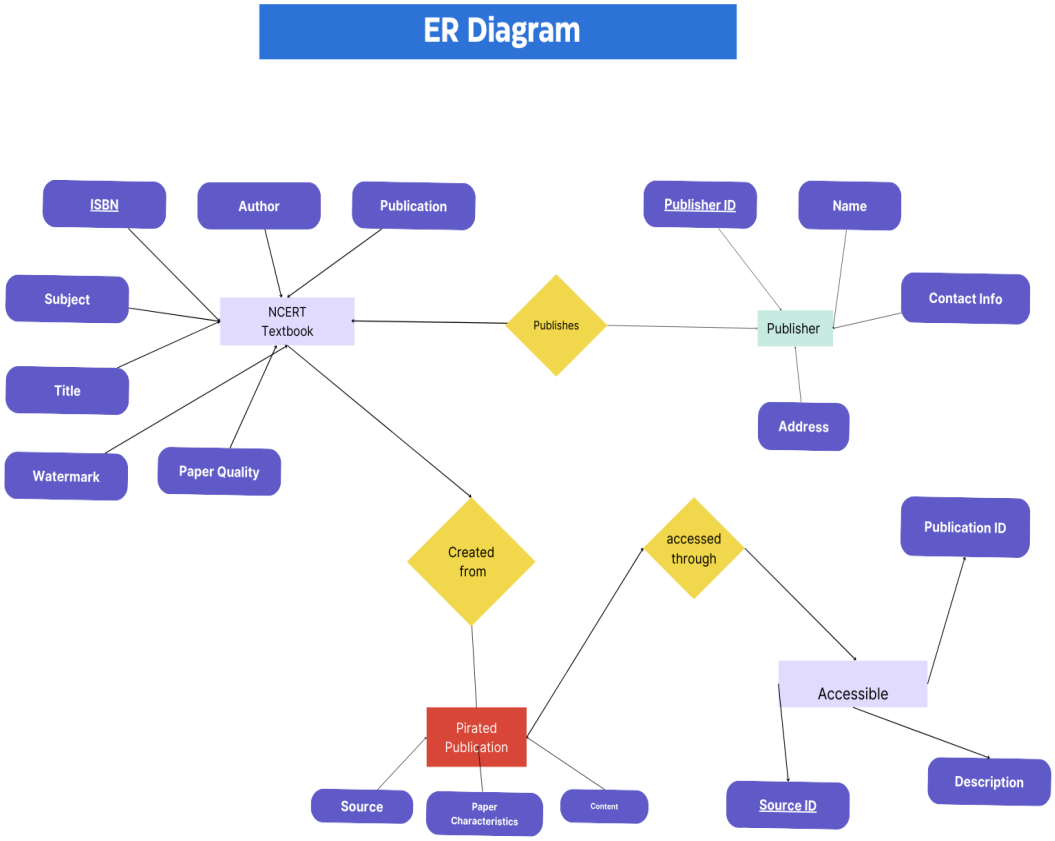
[9] Khaled, F., & Al-Tamimi, M. S. H. (2021). Plagiarism detection methods and tools: An overview. Iraqi Journal of Science, 2771-2783.

[10] Kulkarni, S., Govilkar, S., & Amin, D. (2021, May). Analysis of plagiarism detection tools and methods. In Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021).

[11] Leung, C. H., & Chan, Y. Y. (2007, October). A natural language processing approach to automatic plagiarism detection. In Proceedings of the 8th ACM SIGITE conference on Information technology education (pp. 213-218).

[12] Meuschke, N., Gondek, C., Seebacher, D., Breitinger, C., Keim, D., & Gipp, B. (2018, May). An adaptive image-based plagiarism detection approach. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (pp. 131-140).

ER Diagram :


ER Diagram
ISBN
Author
Publication
Publisher ID
Name
Subject
NCERT Textbook
Publishes
Publisher
Contact Info
Title
Address
Watermark
Paper Quality
Created from
accessed through
Publication ID
Accessible
Pirated Publication
Source
Paper Characteristics
Content
Source ID
Description