TITLE OF THE INVENTION

Integrated System for Detecting Copyright Infringement, Piracy, and Plagiarism in NCERT Textbooks Using Manual Verification and AI-Based Analysis

Project Id: PCS25-16

Guide:
Dr. Ajay Kumar Shrivastava

Background

The issue of copyright infringement, piracy, and plagiarism of NCERT textbooks has been a significant concern for the Ministry of Education, particularly highlighted during the Smart India Hackathon (SIH) 2022. NCERT, the National Council of Educational Research and Training, is a premier institution responsible for the development and distribution of educational content for schools across India. However, the proliferation of counterfeit books, both in terms of physical quality and content, has led to substantial financial losses and compromised the integrity of educational materials.

Currently, there are methods to verify the authenticity of a physical book, as well as software tools available to check for content plagiarism. However, these solutions are often fragmented, expensive, and not integrated into a single, accessible platform. The lack of a unified and cost-effective solution means that many instances of infringement go unchecked, allowing fake books to circulate freely in the market, undermining the quality and credibility that NCERT has established over the years.

This project aims to address these challenges by developing a comprehensive tool that integrates both physical book verification and software-based plagiarism detection. By providing a single, user-friendly platform, this solution will enable efficient detection of copyright infringement, piracy, and plagiarism in NCERT textbooks. This tool is designed to be cost-effective, making it accessible to a wide range of users, including educators, publishers, and legal authorities, ultimately helping to protect the educational standards that NCERT upholds.

Claims

The proposed solution is an improvement over existing methods for detecting copyright infringement, piracy, and plagiarism in educational materials, specifically NCERT textbooks. The approach is comprehensive, integrating both manual verification and software-based plagiarism detection into a single, cost-effective platform. This process ensures that a book adheres to copyright laws, maintains the quality standards set by NCERT, and does not contain plagiarized content. The solution is unique in its ability to generate a detailed report that reflects both the manual inspection and software analysis, providing a robust mechanism to prevent unauthorized publication of counterfeit or plagiarized books.

1. **Integrated Approach:**

- A method that combines manual verification and software-based plagiarism detection to assess the authenticity of NCERT textbooks.
- The process includes verifying adherence to copyright laws and agreements, ensuring the physical quality of the book, and analyzing the content for plagiarism.

2. **Step-by-Step Verification Process:**

- **Step 1:** The method includes a thorough examination to ensure compliance with copyright laws and publishing agreements specific to NCERT textbooks.
- **Step 2:** The method includes a manual inspection of the physical book to verify the paper quality and the presence of NCERT watermarks on all pages.
- **Step 3:** The method uses deep learning and natural language processing-based software, leveraging BERT and transformers, to detect plagiarism in the book's content.
- **Step 4:** The method generates a comprehensive plagiarism report, with a threshold specified by NCERT, determining whether the book can be published.

3. **Cost-Effective Solution:**

- The method offers a cost-effective alternative to existing plagiarism detection software, making it accessible to a wider range of users, including publishers and educators.

4. **Comprehensive Reporting:**

- The method produces a detailed report that combines results from both manual and software-based checks, providing a holistic view of the book’s authenticity and adherence to NCERT’s standards.

5. **Unique Platform:**

- The method provides a single platform that integrates both manual book verification and software-based plagiarism detection, a feature not available in existing solutions.

6. **Threshold-Based Decision Making:**

- The method sets a plagiarism threshold as specified by NCERT, ensuring that any book exceeding this threshold is flagged and not permitted for publication.

These claims highlight the innovative aspects of the proposed solution, emphasizing its integrated approach, step-by-step verification process, cost-effectiveness, and unique reporting capabilities. The method addresses the limitations of current solutions and offers a robust, accessible tool for ensuring the integrity of NCERT textbooks.

Summary

The proposed tool is an innovative and comprehensive solution designed to address the issue of copyright infringement, piracy, and plagiarism in NCERT textbooks, a problem highlighted by the Ministry of Education during SIH'22. NCERT has been facing significant financial losses due to the proliferation of counterfeit books that either mimic NCERT's quality or contain altered content under the NCERT brand. Current solutions are either too costly or are not integrated into a single platform that addresses both manual and software-based verification.

The tool improves upon existing solutions by offering a four-step process that ensures a book's adherence to copyright laws, physical standards, and content originality:

1. **Copyright Compliance and Licensing Verification:**

- In the first step, the tool verifies that the book complies with copyright laws specific to educational materials. It also ensures that publishers have the necessary agreements and licenses to legally produce NCERT textbooks.

2. **Manual Inspection:**

- The second step involves a manual check of the physical book, focusing on the quality of the paper and the presence of NCERT watermarks on each page. This step is critical to identify counterfeit books that may replicate NCERT's quality standards without authorization.

3. **Content Plagiarism Detection:**

- The third step utilizes a deep learning and natural language processing-based software, powered by the BERT model and transformers, to analyze the book's content for plagiarism. This step is crucial in identifying unauthorized replication of NCERT content.

4. **Certified Report Generation:**

- In the final step, the tool generates a detailed plagiarism report. This report is particularly valuable to publishers, as it can be used to certify that their books are free from infringement, piracy, and plagiarism, thereby authorizing them for publication. The tool allows users to set a plagiarism threshold specified by NCERT, beyond which the book is flagged and not permitted for publication.

The proposed solution is further enhanced by the integration of a Large Language Model (LLM) powered by META's Llama 3.1, an open-source model. This feature allows users to interact with the model directly within the platform, enabling them to ask questions and clarify doubts related

to plagiarism without needing to access external resources. This makes the tool a one-stop solution for all issues related to book verification and plagiarism detection.

Additionally, the tool provides flexibility for users who may only need to check for plagiarism without requiring the full certification process. In such cases, users can simply utilize the plagiarism detection feature (step 3), without generating the certified report. This feature makes the tool adaptable to various needs, from quick plagiarism checks to full-fledged verification and certification for publishers.

By offering a cost-effective, integrated platform that addresses both manual and software-based verification, the tool not only ensures the integrity of NCERT textbooks but also provides a reliable resource for publishers to certify their books for publication. The combination of manual checks, deep learning-powered plagiarism detection, and the inclusion of an interactive LLM makes this tool a pioneering solution in the fight against copyright infringement, piracy, and plagiarism in educational materials.

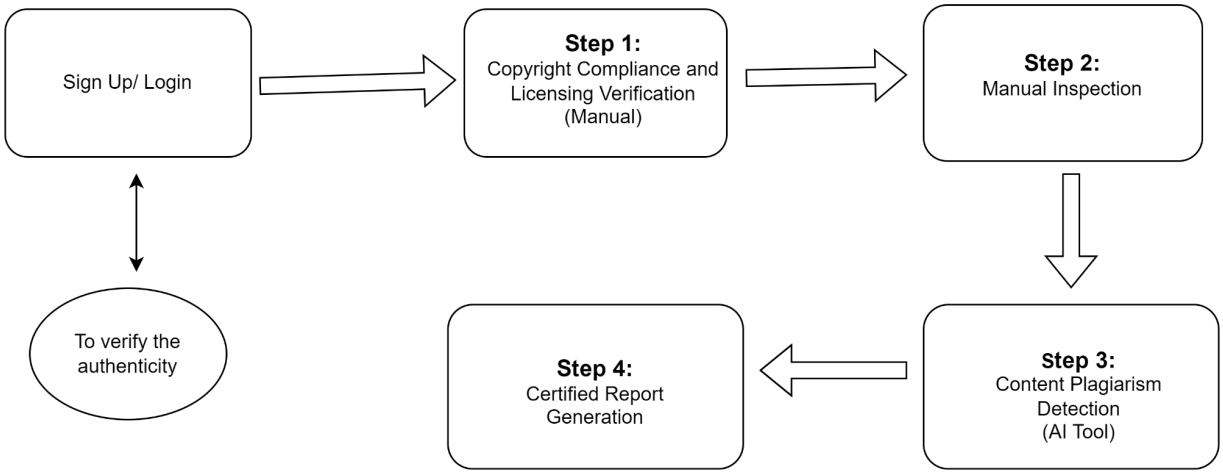
Abstract

The proposed invention provides a comprehensive solution to address the issues of copyright infringement, piracy, and plagiarism in NCERT textbooks. The tool integrates both manual verification and advanced software-based analysis into a single platform. The process begins with the verification of copyright compliance and licensing, ensuring that the book adheres to legal standards. This is followed by a manual inspection of the book's physical attributes, such as paper quality and the presence of NCERT watermarks. The third step employs a deep learning and natural language processing software, based on the BERT model, to detect content plagiarism. A certified report is generated in the final step, which serves as a certification of authenticity for publishers, confirming that the book is free from infringement, piracy, and plagiarism.

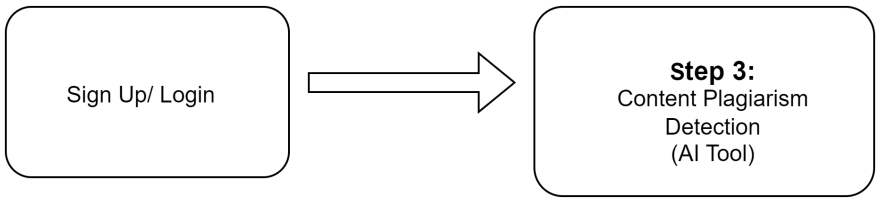
Additionally, the tool includes a Large Language Model (LLM) powered by META's Llama 3.1, allowing users to interact with the model for queries related to plagiarism. The tool is designed to be cost-effective and flexible, offering both full verification and certification services, as well as standalone plagiarism detection. This invention provides a robust, all-encompassing solution to protect the integrity of NCERT textbooks and ensure that only authorized, original content reaches the market.

Drawings

For Publishers:



For users:



Detailed Description

Introduction

The proposed tool is a comprehensive platform designed to address the pressing issues of copyright infringement, piracy, and plagiarism, specifically in the context of NCERT textbooks. This tool is an innovative solution that integrates both manual verification and advanced software-based analysis into a single, cost-effective platform, making it accessible for users and small-scale publishers alike. The platform is built with a focus on accuracy, user-friendliness, and exclusivity, ensuring that it effectively protects the integrity of NCERT textbooks while providing an affordable solution to a widespread problem.

BERT Model and Its Role in Plagiarism Detection

At the heart of the software-based analysis is the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art natural language processing (NLP) model developed by Google. BERT is designed to understand the context of words in a sentence by looking at both the preceding and following words, a process known as bidirectional training. This capability makes BERT particularly effective at tasks such as text classification, sentiment analysis, and, crucially for our tool, plagiarism detection.

The BERT model is pre-trained on a large corpus of text and fine-tuned on specific tasks, making it highly adaptable and accurate. In the context of plagiarism detection, BERT helps in identifying similarities between texts by capturing the nuances of language, including synonyms, paraphrasing, and contextual similarities. This makes it far more effective than traditional keyword-based plagiarism detection tools, which often fail to detect more subtle forms of plagiarism.

The use of BERT in our tool allows us to provide a highly accurate plagiarism detection service, which is essential for ensuring the integrity of NCERT textbooks. However, while BERT is highly effective, it is also computationally expensive, which is why we have integrated the Large Language Model (LLM) powered by META's Llama 3.1 to further enhance the tool's capabilities.

The Need for LLM and the Advantages of Llama 3.1

The integration of a Large Language Model (LLM) into the tool provides users with an interactive experience, allowing them to ask questions and seek clarifications about plagiarism directly within the platform. This feature is particularly useful for users who may not be familiar with the nuances of plagiarism detection or who may need guidance on how to use the tool effectively.

We have chosen META's Llama 3.1 as the LLM for our platform due to its open-source nature and the fact that it is currently available free of cost. Llama 3.1 is a cutting-edge model that has been extensively trained on a diverse dataset, making it highly versatile and effective for a wide range of tasks. Its open-source nature ensures that we can continuously improve and adapt the model to meet the specific needs of our platform without incurring prohibitive costs.

The LLM not only enhances the user experience by providing real-time assistance but also contributes to the overall accuracy and effectiveness of the tool by enabling more nuanced queries and providing detailed explanations of the results. This makes our platform not just a tool for detecting plagiarism but also an educational resource that helps users understand the importance of maintaining academic integrity.

Addressing the Challenges of Copyright Infringement and Plagiarism

Copyright infringement, piracy, and plagiarism are significant challenges that undermine the integrity of educational materials, particularly NCERT textbooks. The Ministry of Education raised this issue during the Smart India Hackathon (SIH) 2022, highlighting the financial losses incurred by NCERT due to the proliferation of counterfeit books in the market. These fake books often feature substandard paper quality and content that differs from the original, leading to a degradation of the standards that NCERT has set.

The proposed tool addresses these challenges by providing a four-step process that combines both manual and software-based analysis to ensure the authenticity of NCERT textbooks:

1. **Legal Compliance:** The first step involves checking that the book does not violate copyright laws and that the publisher has the necessary agreements and licenses to publish the books. This ensures that only authorized content is being published and sold under the NCERT brand.

2. **Manual Verification:** The second step involves a manual check of the book's physical attributes, including the paper quality and the presence of NCERT watermarks on all pages. This step ensures that the book meets the exacting standards set by NCERT and that any physical counterfeit can be identified.

3. **Plagiarism Detection:** The third step involves using our advanced plagiarism detection software, based on the BERT model, to analyze the content of the book. The software is trained exclusively on NCERT books, ensuring that it is highly accurate in detecting any unauthorized copying of content.

4. **Certification:** The final step is the generation of a certified report, which serves as proof that the book is free from infringement, piracy, and plagiarism. Publishers can use this certification to demonstrate that they are authorized to sell NCERT books, thereby protecting their reputation and market position.

This comprehensive approach ensures that NCERT books maintain their integrity and that the market is protected from counterfeit products.

The Cost-Effectiveness and Market Position of the Tool

One of the key advantages of our tool is its cost-effectiveness. While there are existing plagiarism detection solutions that claim accuracy rates of 98-99%, these tools are often prohibitively expensive, making them inaccessible to small-scale publishers and individual users. Our tool, by contrast, is designed to be affordable without compromising on accuracy. This makes it an ideal solution for a wide range of users, from large publishers to small sellers and individual educators.

Moreover, while there are free plagiarism detection websites available, their effectiveness is questionable as they often do not disclose the datasets on which they are trained. Our tool is exclusively trained on NCERT books, ensuring that it provides highly accurate results that are tailored to the specific needs of the NCERT ecosystem.

User and Seller Experience

The platform is designed with the user experience in mind, featuring an intuitive and easy-to-use interface. Users and sellers alike need to sign up or log in to access the platform. For users, this allows us to monitor traffic on the website and ensure that the platform is being used appropriately. For sellers, we cross-check their details with NCERT to verify that they are authorized to sell NCERT books. This dual-layered authentication ensures that only legitimate users have access to the platform, further protecting the integrity of the NCERT brand.

The platform's user interface (UI) is designed to provide the best possible experience, with clear instructions and easy navigation, making it accessible even to those with limited technical knowledge. This focus on user experience is a key differentiator that sets our platform apart from other solutions in the market.

Future Scope

The potential applications of this tool extend beyond the detection of plagiarism in NCERT textbooks. In the future, the tool could be adapted to detect plagiarism in academic assignments, research papers, and other educational materials. The integration of LLMs like Llama 3.1 could also be expanded to provide more interactive features, such as real-time feedback on writing quality, suggestions for improvement, and more.

Additionally, the tool could be developed into a comprehensive platform for ensuring academic integrity across a wide range of educational materials, making it an essential resource for educators, students, and publishers alike. The modular nature of the tool also allows for the integration of new features and updates, ensuring that it remains relevant and effective as the needs of the educational community evolve.

Conclusion

In summary, the proposed tool is a groundbreaking solution that addresses the critical issues of copyright infringement, piracy, and plagiarism in NCERT textbooks. By integrating manual verification with advanced software-based analysis, the tool provides a comprehensive, cost-effective solution that is tailored to the specific needs of NCERT. With its focus on accuracy, user experience, and future scalability, this tool is poised to become an essential resource for protecting the integrity of educational materials and ensuring that only authorized, original content is published and sold.