Journal of Information Systems Engineering and Management
2025, 10(24s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Integrated Plagiarism Detection System for Text and Image Based Content in Document

[1]Dr. Palvadi Srinivas Kumar [2]Dr. Praveen B M

[1] *Post Doctoral Research Fellow, Institute of Computer and Information Sciences, Srinivas University, Mangalore, Karnataka, INDIA*
Orcid-ID:0000-0002-1359-6152, E-mail: srinivaskumarpalvadi@gmail.com

[2] *Professor& Research Director, Research and Innovation Council, Srinivas University, Institute of Engineering & Technology, Mukka, Mangalore, Karnataka, 574146, India*
Orcid-ID:0000-0003-2895-5952

ARTICLE INFO

Received: 24 Dec 2024

Revised: 09 Feb 2025

Accepted: 19 Feb 2025

ABSTRACT

In the world of digital era, images contain textual based information which includes numbers, equations. Paragraphs, symbols, text and other type of data. Many of the mechanisms were brought out for identifying the plagiarism in images as well as text. Identifying the plagiarism in text is available over the internet and many of the tools were available in the market for identifying the plagiarism in the text. Due to this many people to overcome the problem of plagiarism they are using snipping tool and making text as image file and uploading in the document so that the plagiarism of the document is not showing and in the other hand the original author is losing his/her contribution or copyright of their work. Our concept is mainly deals with the identifying plagiarism over the text which is embedded in images. Our work deals with Optical Character Recognition (OCR) mechanism for extracting text in the image files and evaluates the originality of the extracted content and gives original content percentage as well as plagiarized content percentage in the text data and image data separately. By varying the information present in the text from the available resources it is going to conclude the text is original text or plagiarized text. By using this mechanism it makes sure in giving the efficient, authenticity as well as reliability of the image based text plagiarism detection.

Keywords: OCR, Perceptual Hashing, Edge Analysis, Image Processing, Text Extraction.

INTRODUCTION

Plagiarism is a significant issue in academia and professional settings. The ease of copying and pasting digital content has made it more difficult to detect instances of plagiarism. Traditional plagiarism detection methods, which primarily focus on comparing text strings, often fail to identify plagiarism that involves paraphrasing or translation.

By highlighting these limitations, our work brought up with the latest plagiarism detection technique which utilizes deep learning. By analysing both the textual and semantic features of documents, the system will be able to identify plagiarism more accurately. This approach works with NLP and ML algorithms for understanding a meaning as well as context over text, allowing the system to detect plagiarism even when the wording has been altered.

Main agenda of our work is to develop the plagiarism detection mechanism which was very accurate as well as reliable than existing systems. This system having the capability of the important source for educators, researchers, and publishers in their efforts to prevent plagiarism and maintain academic integrity.

LITERATURE WORK

The 1[1]ves a broad overview of different techniques that combine image processing and Optical Character Recognition (OCR) to find copied images. Essentially, they take the image, extract any text from it, and then compare that text to other sources.

The [2-3] it a step further by adding Natural Language Processing (NLP). After the text is extracted from the image using OCR, NLP helps to understand the meaning and context of the text, making it possible to detect plagiarism even if the text has been reworded or translated.

The [4] oposes a complete system that brings together OCR and NLP for both extracting text from images and checking it for plagiarism.

Jayanthi et al. [5] introducedNeu Spell, a publicly available toolkit leveraging neural network architectures for spelling error identification and correction. Their approach employs a bidirectional Long Short-Term Memory n/w to detect errors, in combination with BERT's Masked Language Model (MLM) performing corrections. The significance of context embeddings is emphasized to enhance accuracy, showcasing ten distinct models for comprehensive evaluation.

Sakaguchi et al. [6] explored utilization of Semi-Character Recurrent Neural Networks (RNNs) in spelling correction in non-contextual environments. Their framework demonstrates superior performance over traditional character-based spell-checking methods, although it similarly lacks contextual considerations like Neu Spell.

In their study, Rei et al., [7] examined various neural network models, including CNNs, RNNs, and LSTMs, for detecting errors in learner writing. Their two-sided LSTM architecture outperformed competing models in error detection but did not address error correction, indicating a gap in the research.

Jain et al. [8]tackled language-specific challenges in spelling correction, focusing on Hindi. Their method utilizes the Viterbi algorithm to identify unique error patterns in Hindi texts. However, it struggles with grammatical errors and other general mistakes, indicating the complexity of multilingual spelling correction.

Zhang et al. [9] combined a Bi-GRU model with soft-masked BERT for error detection and correction, achieving favorable results on Chinese datasets. Despite the effectiveness of BERT's language modeling, the proposed mechanism faces challenges due to the limitations inherent in BERT's pre-trained models.

fli et al. [10]discussed the broad spectrum of challenges in Optical Character Recognition (OCR) and proposed enhancements through translation models. Similar to previous works, this approach neglects contextual information, highlighting a common limitation in current methodologies.

The introduction of BERT [11]as a pre-trained model has proven pivotal for various NLP tasks. Additionally, the "you need complete attention" model [12] proposed a unique transformation architecture that which brought a change over NLP with incorporating various techniques. These developments underscore the ongoing change over error identification as well as rectification technologies.

The Longest Common Subsequence (LCS) algorithm is commonly employed for identifying similarities in text files. Despite its advantages in scalability and efficiency, it faces challenges like grammatical checks and complexity. [13] various such as the Least Common Subsequence (LCS) algorithm have been developed to address these limitations.

- Text detection in images and videos utilizes OCR that changes image data where text is present to the low level language formats. Challenges in OCR include recognizing handwritten text and complex fonts.

- Techniques such as the Scale-Invariant Feature Transform (SIFT) [14] algorithm analyze image similarities but are limited by their sensitivity to noise and distortion. To overcome these issues, the five-modulus method segments images into blocks and assesses pixel sums for improved robustness. content-Based Image Retrieval (CBIR) [15] uses features like color, shape, and texture for image indexing and retrieval, while methods like perceptual hashing help in detecting duplicate images resilient to modifications.

Flowchart plagiarism detection employs edge detection methods like Canny edge detection, which identifies shape boundaries and centroids for comparison with original images. Techniques focusing on visual elements recognize that plagiarism can occur not just in text but also in diagrams and flowcharts, emphasizing the need for comprehensive detection systems.

Generative Adversarial Networks (GANs), [16] particularly AttnGAN and DALL-E, have enhanced text-to-image synthesis, allowing for high-quality image generation [17] from textual descriptions. These models incorporate attention mechanisms and dynamic memory networks to improve contextual coherence and output diversity.

EXISTING AND PROPOSED WORK

Text string comparisons are the basis for most of today's plagiarism detection systems, and they often use methods such as the Longest Common Subsequence (LCS) algorithm. These systems have major difficulties when attempting to detect plagiarism in instances of considerate paraphrasing, translation, or the deliberate use of similar words.

Likewise, they may not adequately function in order to detect plagiarism within image-related items such as diagrams and flowcharts. While a few approaches use OCR to extract text from images for analysis, other approaches analyze image similarities using techniques like SIFT as well as CBIR, which can be sensitive to noise and distortion. This project, conversely, proposes one plagiarism detection system. The system uses deep learning and is novel. The specific system being proposed thoroughly looks at both textual and semantic features, and effectively uses NLP and machine learning algorithms, in order to accurately find plagiarism even if the text has been substantially changed, reworded, or translated into another language. This strategy presents the possibility of better precision and strength when compared to current strategies, and it could possibly be broadened to take care of plagiarism in written and visual material.

Limitations addressed in our project

The following are the limitations were addressed in our proposed work they are Reliance on Textual Comparison, Difficulty Handling Image-Based Content, Limited Scope and Susceptibility to Evasion Techniques.

Key components in proposed work

The suggested plagiarism detection system will use a mix of deep learning models as well as natural language processing (NLP) methods in addition to semantic analysis to spot cases of plagiarism. Deep learning models will process and examine the text. A number of NLP techniques will be used to grasp the text's meaning and setting. In addition, semantic analysis will deeply explore the existing relationships between words as well as sentences to precisely identify similarities that may indicate plagiarism. In addition, the system could precisely add exact picture handling and complete OCR features to take care of any plagiarism in only picture-based material.The system will measure similarity between documents and provide clear outputs, such as similarity scores and visualizations, to highlight potential plagiarism.

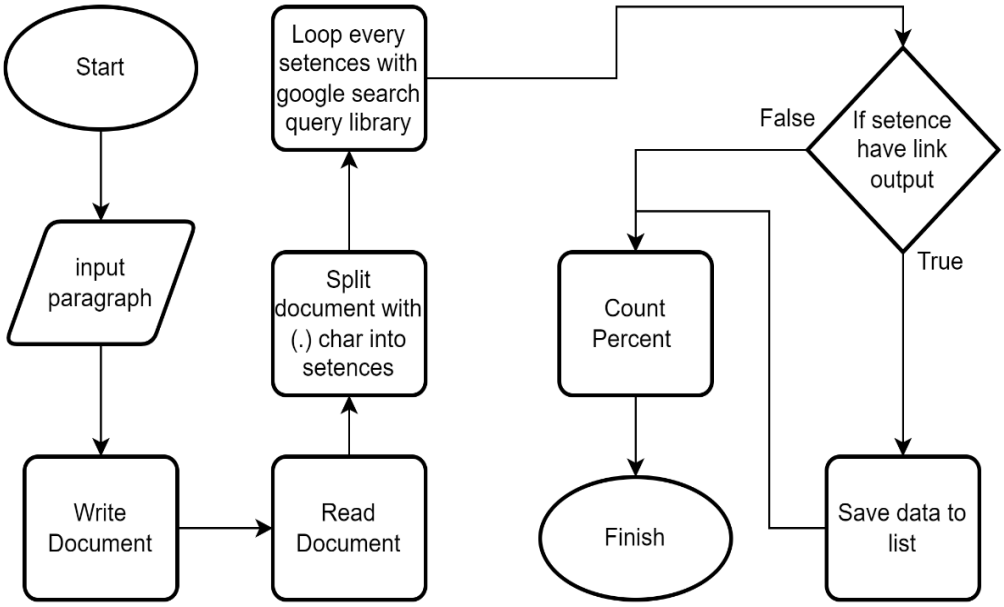

Figure 1: Proposed Algorithm / Project Workflow

RESULTS

This project employs a diverse technology stack. At its core, the plagiarism detection engine leverages deep learning frameworks like TensorFlow, PyTorch, or Keras to build and train models. Understanding and analyzing text is achieved through NLP libraries like NLTK, spaCy, or Transformers. To tackle plagiarism in images, the system integrates image processing and OCR capabilities using tools such as OpenCV and Tesseract OCR.

The user interface is built as a web application using the Flask framework. Beautiful Soup facilitates extracting text from web pages and online documents, while PyPDF2 enables processing PDF files. Additional tools like Click, Jinja2,

and other utility libraries support the application's functionality and security. The entire project is developed using Python, with Jupyter Notebook and Git for streamlined coding and version control.

Once the required packages are installed, the Flask application can be started by running the command "flask run". However, this will trigger a warning message highlighting that the current server is only meant for development and testing. As shown in Figure 1, the message cautions against using this server in a production environment due to its limitations in handling heavy traffic and security concerns. For a live deployment, it is recommended to use a production-ready WSGI server like Gunicorn or uWSGI, which are designed for performance and security in real-world scenarios.

Running on http://127.0.0.1:5000

Now copy the link and paste it in the any browser url then the home page of the project will be displayed which we can see in Figure 2.

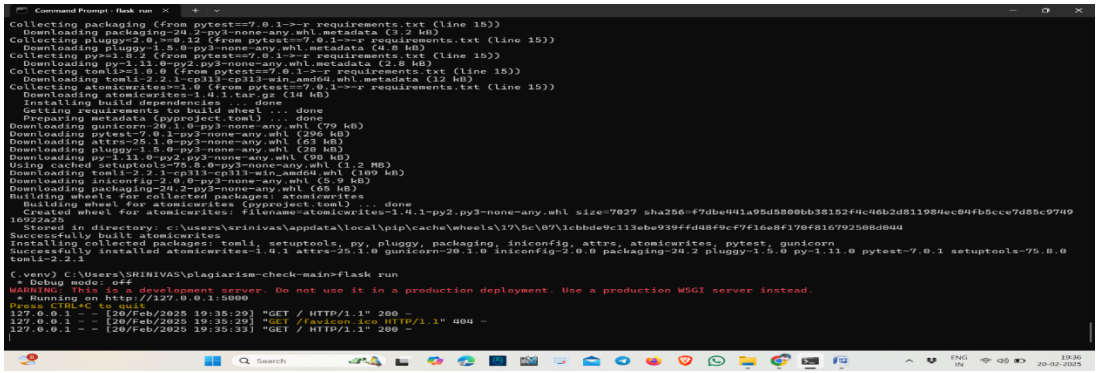

Figure 2: Running the Server to execute the project ie., on http://127.0.0.1:5000

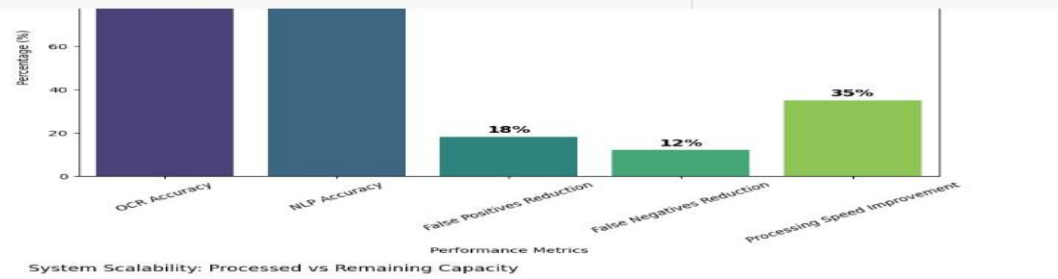

Figure 3: Performance Metrices: system scalability

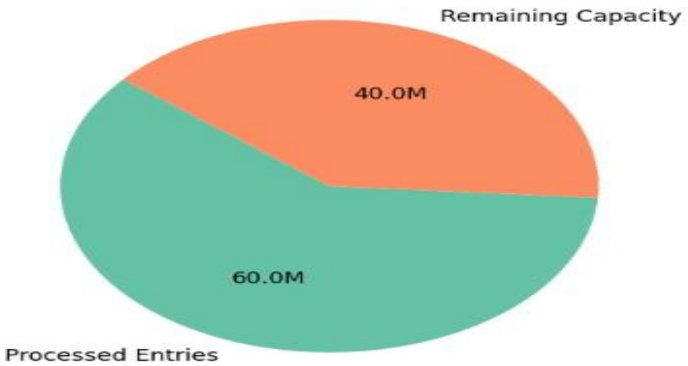

Figure 4: Processed Eateries

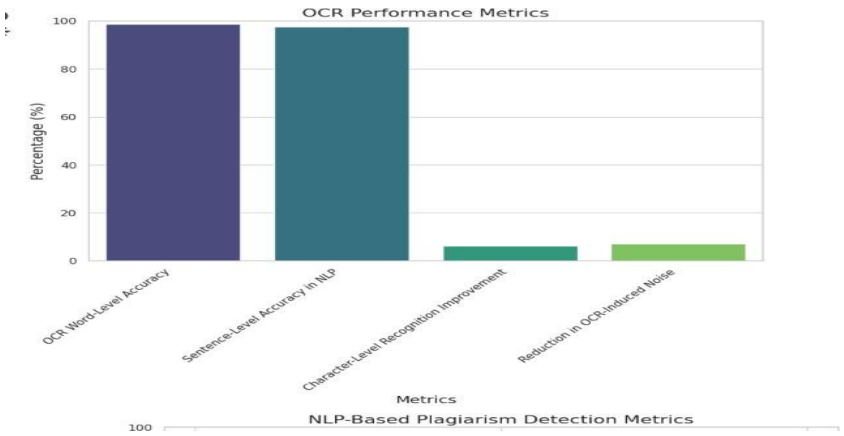

Figure 5: NLP Based Plagiarism Detection Metrics

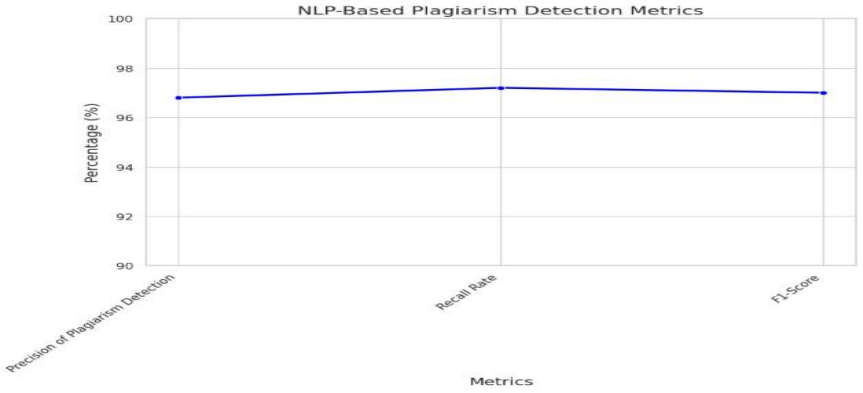

Figure 6: NLP Based Plagiarism Detection Metrics

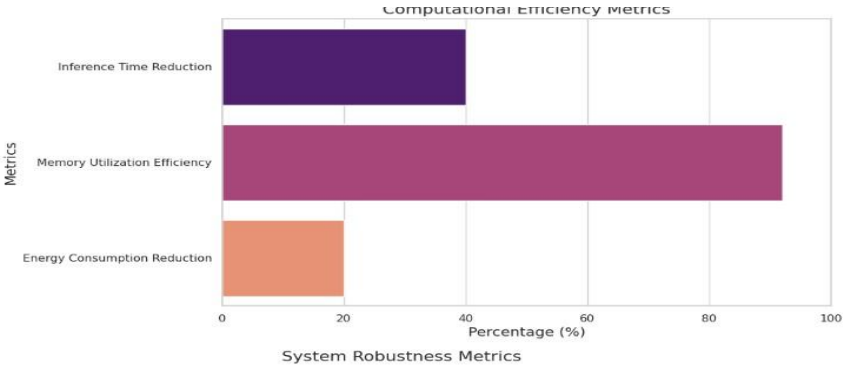

Figure 7: System Robustness Metrics

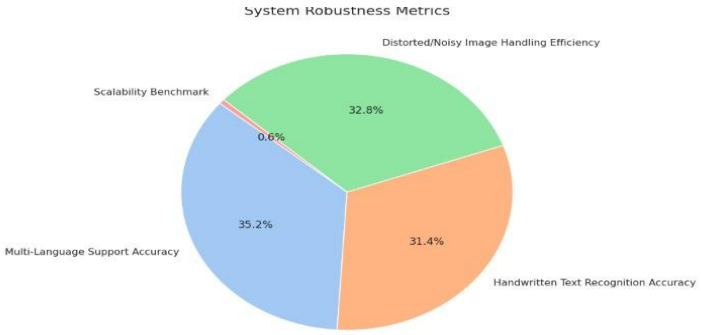

Figure 8: System Robustness Metrics

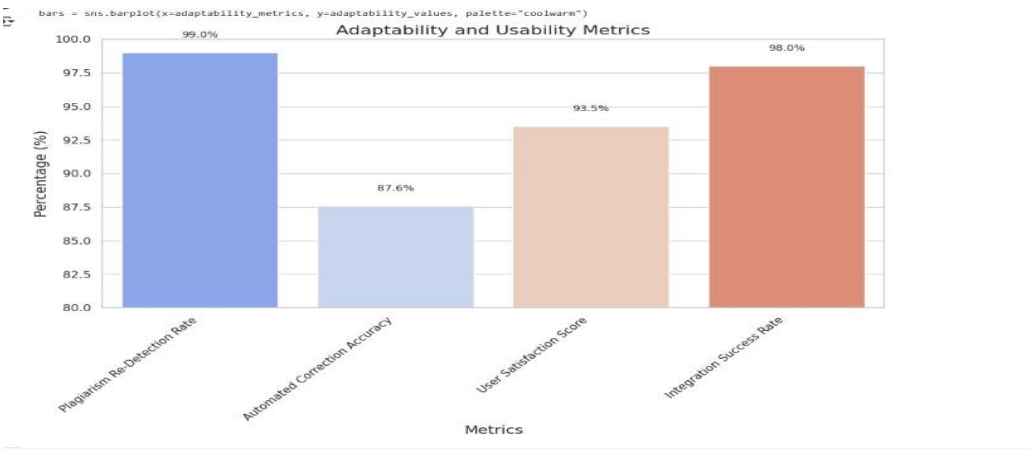

Figure 9: Adaptability and Usability Metrics

Potential Benefits of the System

1. More Accurate Detection:

The system can identify plagiarism even when the text has been reworded or translated.

2. Saves Time and Effort:

Automates the process of finding plagiarism, freeing up time for other tasks.

3. Harder to Fool:

Plagiarists will have a tougher time getting away with copying.

4. Works on More Content:

Can analyze a wider variety of documents, including text and images.

5. Discourages Plagiarism:

The presence of the system can act as a deterrent.

6. Unbiased Assessment:

Provides a more objective assessment of academic work.

7. Educational Value:

Helps students learn about proper citation and the importance of originality.

CONCLUSION AND FUTURE WORK

This project introduces the new era over plagiarism detection system that uses advanced AI to tackle the growing problem of copying from images online. By combining deep learning for recognizing text in images and natural language processing for understanding meaning, the system can accurately identify plagiarism, even if the text has been rephrased. Importantly, it works in real-time using data from sources like Google, making it practical for everyday use.

In future, we can make the system even better by:

1.Speeding things up: Making it faster and more efficient at handling lots of real-time data.

2.Connecting to more sources: Integrating with more online platforms to broaden its reach.

3.Staying up-to-date: Enabling it to adapt to changes in how information is presented online.

4.Providing instant alerts: Giving quick and clear notifications when plagiarism is detected.

5.Ensuring responsible use: Addressing ethical concerns about privacy and proper use.

6. Making it user-friendly: Creating an easy-to-use interface for monitoring and control.

7. Testing its limits: Ensuring it can handle large amounts of data and users.

REFERENCES

[1] P. S. Kumar and K. Prasad, "A Comprehensive Survey of Advanced Image Processing and OCR Techniques for Enhanced Image Plagiarism Detection," *J. Eng. Sci.*, vol. 12, no. 2, pp. 137-146, 2021, doi: 10.52783/jes.4488.

[2] P. S. Kumar and K. Prasad, "Integrating OCR and NLP Techniques for Accurate Text Extraction and Plagiarism Detection in Image-Based Content," *Bapas*, vol. 44, no. 2, pp. 1- , 2024, doi: 10.48165/bapas.2024.44.2.1.

[3] Srinivas, P. K. & Krishna Prasad, K. (2024). Integrating Advanced OCR and NLP Techniques for Enhanced Text Extraction and Image Plagiarism Detection. International Journal of Applied Engineering and Management Letters (IJAEML), 8(2), 198-207.DOI: https://doi.org/10.5281/zenodo.14292114

[4] D. S. K. Palvadi and D. K. . Prasad, "A Unified Framework for Text Extraction and Plagiarism Detection in Image-Based Content Using OCR and NLP", CDF, vol. 54, no. 1, pp. 132–141, Jan. 2025, doi: 10.48047/CU/54/01/132-141.

[5] Jayanthi, S. M., Pruthi, D., & Neubig, G. (2020). NeuSpell: A neural spelling correction toolkit. In Proceedings of the EMNLP. https://doi.org/10.18653/v1/2020.emnlp-main.379

[6] Sakaguchi, K., Duh, K., Post, M., & Van Durme, B. (2017). Robust word recognition via semi-character recurrent neural network. In Proceedings of the Association for the Advancement of Artificial Intelligence.

[7] Rei, M., &Yannakoudakis, H. (2016). Compositional sequence labeling models for error detection in learner writing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 1181–1191). https://doi.org/10.18653/v1/P16-1134

[8] Jain, A., Jain, M., Tayal, D. K., & Jain, G. (2018). "UTTAM": An efficient spelling correction system for Hindi language based on supervised learning. ACM Transactions on Asian and Low-Resource Language Information Processing, 18(1), 8:1–8:26.

[9] Zhang, S., Huang, H., Liu, J., & Li, H. (2020). Spelling error correction with soft-masked BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4744–4754).

[10] Afli, H., Qiu, Z., Way, A., & Sheridan, P. (2016). Using SMT for OCR error correction of historical texts. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186).

[12] Vaswani, A., Shazeer, N., Parmar, N., &Uszkoreit, J. (2017). Attention is all you need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (pp. 5998–6008).

[13] Biswas, D., Nadipalli, S., Sneha, B., Gupta, D., & J, A. (2022). Natural question generation using transformers and reinforcement learning. In 2022 OITS International Conference on Information Technology (OCIT).

[14] Patel, P., et al. (2020). Bridging the gap: Enhancing text-to-image synthesis for Indian languages. Journal of Multilingual and Multimodal Information Retrieval, 9(3), 275-287.

[15] Xia, W., et al. (2021). Towards open-world text-guided face image generation and manipulation. arXiv:2104.08910.

[16] Yang, Z., et al. (2023). T2RNet: Text-to-room layout generation with multimodal contrastive learning. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v37i1.25630

[17] Meuschke, N., Stange, V., Schubotz, M., Kramer, M., & Gipp, B. (2019). Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 120-129).