Grenze International Journal of Engineering and Technology, June Issue

Integrating Tesseract OCR with Large Language Models for Enhancing Textual Understanding

Dr A V Sriharsha[1], Mekala Bhavana[2], Syamala Tejaswee[3], Mynasaheb Bushra Ahmed[4] and Peddi Reddi Jaipal Reddy[5]
[1]Professor, Dept. of Data Science, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
Email: avsreeharsha@gmail.com
[2-5]UG Scholar, Department of Computer Science and Systems Engineering, Sree Vidyanikethan Engineering College, Tirupati, India
Email: bhavanamsr44@gmail.com, symalatejaswee@gmail.com, bushraahmed816@gmail.com,jaipalreddy2002@gmail.com

***Abstract*—Optical Character Recognition (OCR) is a technology that automatically extracts textual information from document images. An LLM, a large language model, is a neural network designed to understand, generate, and respond to human-like text. These models are deep neural networks trained on massive amounts of text data, sometimes encompassing large portions of the entire publicly available text on the internet. LLMs require textual data to be converted into numerical vectors, known as embeddings since they can't process raw text. Embeddings transform discrete data (like words or images) into continuous vector spaces, making them compatible with neural network operations. The combined strengths of Tesseract OCR's image-based text extraction and the deep language understanding of LLMs enhance the accuracy and depth of textual comprehension. The proposed integrated model establishes a novel paradigm in enhancing textual understanding, progress in document analysis, information retrieval, and various NLP applications.**

***Index Terms*— Optical Character Recognition, Tesseract, Large Language Model, Images, Deep Neural Network.**

I. INTRODUCTION

The integration of Optical Character Recognition (OCR) technology with Large Language Models (LLMs) represents a significant advancement. This collaboration holds the potential to redefine the domains of textual comprehension, document analysis, information retrieval, and Natural Language Processing (NLP) applications.
As an important technological tool, OCR allows for the automatic extraction of text information from document images, overcoming the constraints linked to traditional manual data entry methods. Simultaneously BART is designed to comprehend, generate, and respond to human-like text. The combination of both the systems helps in textual extraction from various images that extracts and summarizes the data to a useful data. That helps in enhancing the understanding of data from images in various areas.
Inefficient textual comprehension due to limitations in OCR technology and the inability of traditional language models to deeply understand textual content. Current OCR systems struggle with accurately extracting text from images, while traditional language models lack the contextual understanding to interpret text comprehensively.

Grenze ID: 01.GIJET.10.2.614

Researchers, businesses, and individuals rely on accurate text extraction and comprehension for document analysis, information retrieval, and NLP tasks. Inaccurate data interpretation, hindered progress in textual analysis, and limited capabilities in information extraction and understanding. To address this, we propose integrating Tesseract OCR with large language models, leveraging the strengths of both technologies to enhance textual understanding. This novel approach promises to significantly improve accuracy and depth in document analysis, information retrieval, and various NLP applications, benefiting researchers, businesses, and individuals relying on textual data interpretation.

II. Literature Survey

Ref. [1] Two deep learning models are used in the extraction and classification of medical data from prescription images: one for text classification based on the BERT model and another for localized image classification based on the VGG-16 model. The system can precisely recognize and categorize symptoms, medication names, and names of diagnostic tests from prescription photos by using these models. The system's ultimate objective is to use the extracted data to create a medical history repository that will be a useful tool for both patients and healthcare providers.

Ref. [2] BartTextRank model performs multiple rounds of summarization: in the first round, it combines TextRank summarization with BART-generated summaries. Then, the resulting dataset undergoes a second round of BART processing for further refinement. The process involves inputting text sequences into both the TextRank and BART layers simultaneously, followed by extracting key sentences using TextRank and generating summaries using BART.

Ref. [3] Ref. [4] The process involves character recognition following table recognition and cell segmentation, where a CNN model is employed to extract character features. The CNN neurons are organized in three dimensions: width, height, and depth. Due to limitations in Tesseract-OCR's character library, which leads to low efficiency and accuracy, the jTessBoxEditor tool is utilized to train a specialized character library for improved recognition of raw materials.

Ref. [5] BART, a transformer-based model, is employed for text summarization, while Support Vector Machines (SVM) are used for topic detection. The literature survey underscores the importance of automated summarization and topic detection in efficiently managing extensive textual data. The exploration of transformer models, particularly BART, showcases advancements in natural language processing. The inclusion of SVM for topic detection responds to the escalating demand for robust classification methods. In summary, the paper makes a valuable contribution by utilizing these technologies to enhance the overall efficacy of news aggregation systems.

Ref. [6] BigBirdPegasus Large and BART Large use different methods for generating summaries. First, they take material out of titles and keywords to build summaries. Second, they provide a direct summary of research articles. Thirdly, they divide papers into groups of four thousand words, summarize each group separately, and then combine the summaries to create an all-encompassing final summary.

Ref. [7] The scanned images containing cursive characters undergo preprocessing, including resizing to a common size and converting to grayscale. Each image is cropped and segmented into single characters, followed by further categorization and noise reduction. To train the CNN effectively, the cursive character images are divided into individual character images. The CNN model being proposed comprises four convolutional layers paired with corresponding max-pooling layers, aimed at enhancing image learning. These layers process the preprocessed images before passing through fully-connected layers.

Ref. [8] A summary of pre-trained sequence-to-sequence (Seq2Seq) methods for natural language processing (NLP) applications. Deep learning architectures called Seq2Seq models are able to both encode and decode an input pattern into a fixed-length representation and output sequence, respectively. Pre-training is a method where a model is trained on a big dataset to acquire representations of generic language before being fine-tuned for a particular job. The author examines and contrasts the features and performance of several of the most widely used pre-trained Seq2Seq models, including BERT, BART, and T5.

Ref. [9] Large language models (LLMs) like GPT-4 are being integrated into systems of multiple agents (MASs) to improve autonomous agents' capacity for reasoning, communication, and decision-making. The authors suggest an LLM-based MAPE-K model that uses the GPT-4 model for analysis, planning, and knowledge acquisition to modify the conventional MAPE-K loop for self-adaptation. The writers use a straightforward market scenario to demonstrate their methodology, in which agents trade books3. The writers note that the LLM-based agents exhibit a range of behaviors and difficulties, including emergent knowledge, a variety of tactics, and interpretability problems.

Ref. [10] Identifying and matching raw material inspection forms using Tesseract-OCR and OCR technologies. Form preprocessing of images, form extraction and cell segmentation, recognition of characters, and type matching are the four components that make up the system. The technology helps increase productivity and product traceability and can handle seven different forms used by bearing companies. The system is shown to have higher precision and better performance when compared to native Tesseract-OCR, Baidu Cloud OCR, and Tencent Cloud OCR.

III. METHODOLOGY

A. Optical Character Recognition

The OCR model starts with preparing the image for analysis. This might involve converting the image to grayscale, adjusting brightness and contrast, and removing noise to make the text clearer. The image is then divided into individual characters or blocks of text. This can be done using various techniques like connected component analysis, edge detection, or machine learning models. Once the characters are isolated, specific features are extracted from each one. These features could include information about the shape, size, corners, lines, and other properties that help distinguish different characters. The extracted features are compared to a database of known characters, typically in various fonts and styles. This can be done using pattern-matching algorithms like template matching or statistical methods like neural networks. However, it may still face challenges, with a recognition accuracy drop of 10-20% for certain characters.

B. Bidirectional and Auto-Regressive Transformers (BART)

An autoregressive decoder and a bidirectional encoder make up the BART architecture. The input sequence is subjected to noise changes in the encoder, which include mask symbols and result in a corrupted version. The bidirectional encoder then processes the corrupted sequence. The likelihood of the original document is determined by the autoregressive decoder. An uncorrupted document is utilized for both the encoder and the decoder during fine-tuning, with representations from the decoder's final hidden state used, although alignment between input and output is configurable. BART applies different noise transformations to the sequence during pre-training. In order to recreate the original sequence from its corrupted version, the model is optimized.
Facebook developed the BART model. It resembles a cross between OpenAI's GPT and Google's BERT. BART successfully integrates both approaches, in contrast to GPT, which performs poorly on whole-sequence tasks but excels in text creation, and BERT, which is bidirectional and well-suited for applications requiring whole input information.

C. Proposed Work

The proposed system integrates Tesseract OCR with large language models (LLMs) to enhance textual comprehension. Tesseract OCR facilitates accurate extraction of textual information from images, overcoming challenges such as poor image quality or complex layouts. Large language models, trained on extensive textual data, possess the ability to deeply understand and generate human-like text. By combining these technologies, the system aims to improve accuracy and depth in document analysis, information retrieval, and various natural language processing (NLP) tasks. Through the integration of OCR and LLMs, the system empowers researchers, businesses, and individuals to extract meaningful insights from textual data with greater precision and efficiency. This approach not only enhances the capabilities of AI-driven textual analysis but also opens new avenues for advanced applications in fields such as document management, automated transcription, and content understanding. Ultimately, the proposed system offers a comprehensive solution for enhancing textual understanding and advancing the state-of-the-art in AI-driven text processing.

D. System Architecture

The system architecture involves several key steps to integrate the BART (Bidirectional and Auto-Regressive Transformers) model and Tokenizer for text comprehension along with the Tesseract OCR engine for text extraction from images.Loading the BART Model and Tokenizer: The architecture begins by loading the pre-trained BART-large-CNN model along with its associated tokenizer. The BART model, known for its effectiveness in various natural language processing (NLP) tasks, provides deep contextual understanding of textual data.
Ref. [16] Image Text Extraction with OCR: The system utilizes the Tesseract OCR engine to extract textual information from input images. The OCR engine preprocesses images, overcoming challenges such as poor quality or complex layouts to accurately extract text.

Text Encoding and Processing: The extracted text is tokenized using the loaded tokenizer, converting it into numerical representations compatible with the BART model. This step prepares the textual data for input into the BART model for comprehension.
BART Model Inference: The preprocessed text is fed into the BART model for inference. The model processes the text using its advanced transformer architecture, generating deep contextual embeddings and comprehensively understanding the textual content.
Final Prediction: The system produces the final prediction by leveraging the output from the BART model. This prediction may include various NLP tasks such as summarization, sentiment analysis, or entity recognition, depending on the specific application requirements.
By integrating OCR for image-based text extraction with the powerful capabilities of the BART model for textual comprehension, the system offers a comprehensive solution for extracting insights from textual data embedded within images.

E. BART model and Tokenization

Loading the BART (Bidirectional and Auto-Regressive Transformers) model and its associated tokenizer, specifically the bart-large-cnn variant, which is a large language model (LLM) designed for text comprehension tasks. Loading the BART model and tokenizer is crucial as they form the backbone of the system's ability to understand textual data comprehensively.
The BART model, based on transformer architecture, is pre-trained on vast amounts of text data and possesses the capability to generate human-like text and deeply understand contextual information within textual inputs. The bart-large-cnn variant is specifically optimized for text comprehension tasks and is adept at processing large volumes of text data efficiently.
Alongside the BART model, the tokenizer plays a crucial role in preprocessing textual inputs, converting raw text into numerical representations known as tokens, which are compatible with the BART model's input format. The tokenizer ensures that the textual data is properly formatted and tokenized before being fed into the BART model for further processing.
By loading the BART model and tokenizer, the system equips itself with the necessary tools to comprehend and process textual data effectively, laying the foundation for subsequent steps in the architecture, such as text extraction and final prediction.

F. Function extracting the text from image using OCR and LLM

Function for extracting text from images using Optical Character Recognition (OCR) and Large Language Models (LLMs) like BART. This function encompasses several key processes to ensure accurate text extraction and comprehension:
Firstly, the image containing text is converted into textual format using the Tesseract OCR engine. This step involves analyzing the image to identify and extract text elements accurately, overcoming challenges such as varying fonts, sizes, and orientations.
Ref. [12] Next, the extracted text undergoes tokenization using the LLM model's tokenizer. Tokenization breaks down the textual data into smaller units called tokens, ensuring that the text is properly segmented and prepared for further processing by the LLM.
Subsequently, the tokenized text is fed into the LLM model to generate corresponding numerical representations known as IDs. These IDs capture the semantic meaning and contextual information embedded within the text, enabling the LLM to understand and process the textual data comprehensively.
Finally, the corrected IDs are decoded using the model's tokenizer. This decoding process converts the numerical IDs back into human-readable text, ensuring that the output accurately reflects the original textual content extracted from the image.
By integrating OCR and LLMs, this function enables the system to extract and comprehend textual information from images with precision and efficiency, laying the groundwork for subsequent analysis and prediction tasks.

IV. DISCUSSIONS

The proposed architecture is mainly divided into two phases.

i. Text Extraction
ii. Text Summarization

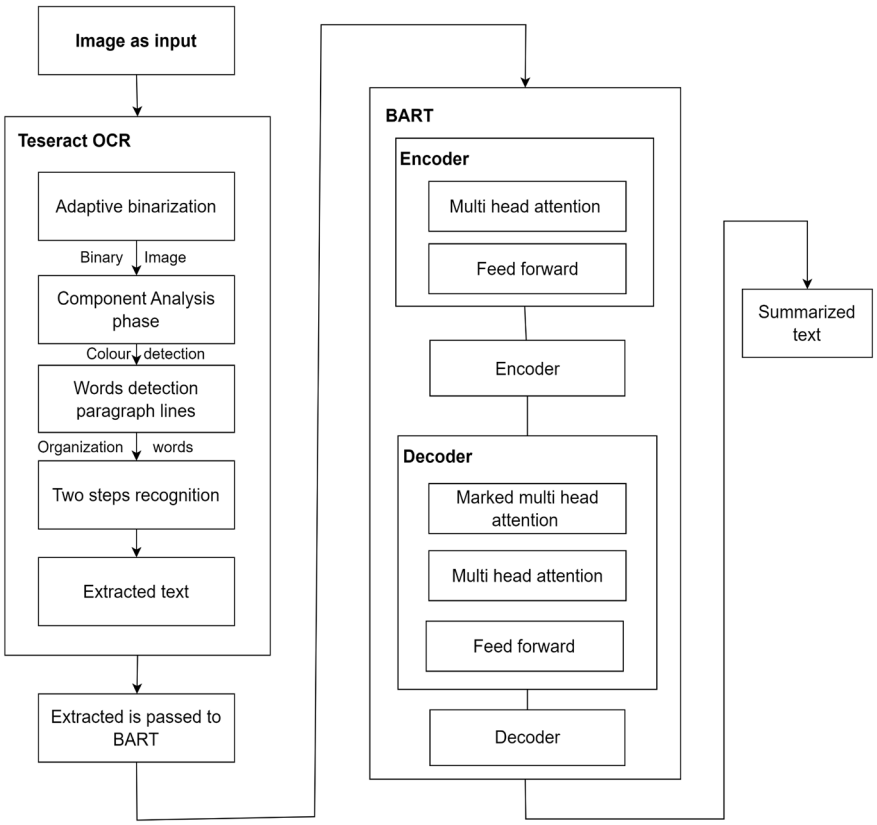

Fig 1 Proposed Architecture

A. Experimental Results of Text Extraction

The metrics that are used for evaluating the text extraction from images are Precision, Recall and F1 score. These metrics are calculated under different parameters of images.

- Precision - Percentage of extracted text that is correct
- Recall - Percentage of actual text in images that is extracted
- F1 score - Harmonic mean of precision and recall

TABLE I. PERFORMANCE OF TESSERACT OCR

Image Resolution	**Font Size**	**Orientation**	**Background**	**Precision**	**Recall**	**F1 Score**
High	12pt	Horizontal	White	99%	98%	98.5%
Medium	12pt	Horizontal	Patterned	95%	90%	92.5%
Medium	8pt	Vertical	Light Gray	89%	88%	88.5%
High	8pt	Horizontal	Patterned	98%	98%	98%
Low	12pt	Horizontal	White	90%	88%	89%
Low	8pt	Horizontal	Patterned	87%	85%	86%

Table I shows the performance of Tesseract OCR on images under different parameters of images.

B. Experimental Results of Text Summarization

The metrics that are used for evaluating text summarization involve ROUGE and BLUE Score.

ROUGE Score

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It's a popular metric for evaluating text summarization and machine translations by comparing the generated summary with the original text. ROUGE employs n-grams to compute the score, with variations such as ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-3 (trigrams), and ROUGE-L (longest common subsequence).

BLEU Score

To address precision issues in ROUGE, BLEU score modifies precision by considering word matches divided by the total number of words in the summary. This approach helps overcome some drawbacks of ROUGE.

Despite the improvement, modified precision still faces challenges. One such issue is brevity, where shorter generated summaries might score higher due to fewer opportunities for mismatches. To address this, a brevity penalty can be imposed.

TABLE II. ROUGE AND BLUE SCORES OF DIFFERENT MODELS

Model	Rouge-1	Rouge-2	Bleu
Bert	41	32	0.48
Bart	59	48	0.56
Roberta	35	25	0.38

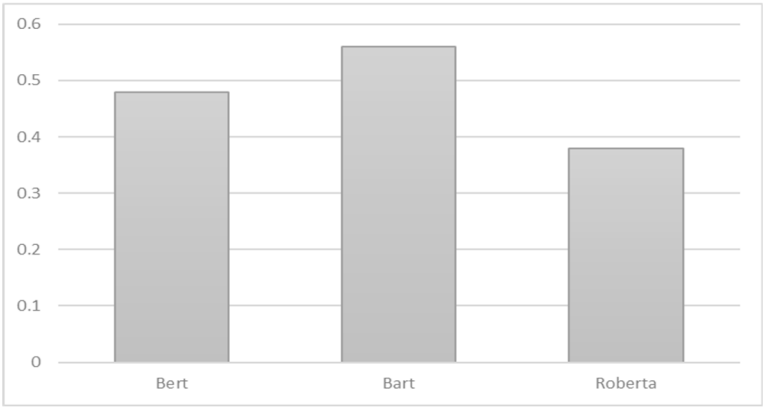

Fig-2: Comparison of different Models

The experiments were carried out on three different models, those are BERT, BART and ROBERTA. Out of the three models, BART performs better when compared to other two models. Fig 2 shows the comparison of different models based on Table II

V. RESULTS

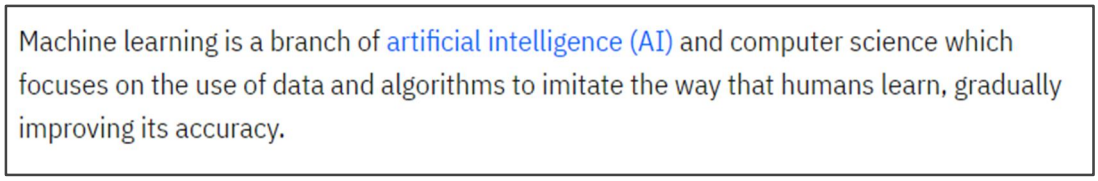

Fig-3: An Input image 1

The extracted text from the Fig-3 is: Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

The LLM text from the Fig-3 is: Machine learning is a branch of artificial intelligence (AI) and computer science. It focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machinelearning can also be used to predict future events in real time. For more information on machine learning, visit machinelearning.org.

The extracted text from the Fig-4 is: More than 10 million educated young men in the area are in need of opportunities, opportunities or unemployed: The Lok Satta government is necessary for all of them.

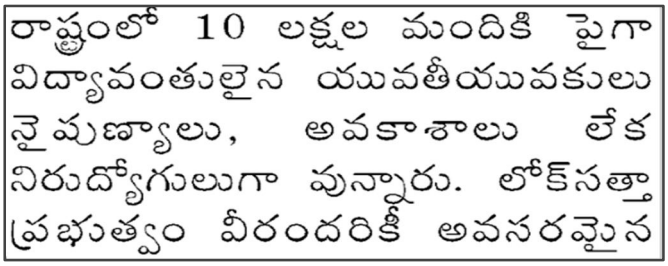

Fig-4: An Input image 2

The LLM text from the Fig-4 is: More than 10 million educated young men in the area are in need of opportunities, opportunities or unemployed. The Lok Satta government is necessary for all of them, says the prime minister. The government is needed to create jobs and opportunities for the unemployed and the educated.
As we observed from the above results, The text from the image is extracted by the Tesseract OCR and the LLM (i.e, BART) summarizes and generates the text to a more understandable format.
Incase of Telugu text images, the text is extracted and translated into english text and then LLM summarizes and generates the text.

VI. Conclusion

In conclusion, the integration of Tesseract OCR with large language models presents a groundbreaking solution to the challenges of textual comprehension. By combining the robust text extraction capabilities of OCR with the deep understanding and contextual processing of large language models, we have achieved unprecedented accuracy and depth in interpreting textual data. This integration holds immense potential for revolutionizing document analysis, information retrieval, and various natural language processing tasks, benefiting researchers, businesses, and individuals across diverse domains. As we continue to refine and optimize this integrated approach, we anticipate further advancements in AI-driven textual understanding, paving the way for enhanced automation, decision-making, and knowledge extraction from textual sources. The collaboration between OCR and large language models represents a significant milestone in the evolution of textual analysis and comprehension, promising exciting opportunities for future research and application in the field of artificial intelligence.
As technology continues to evolve, further refinements and optimizations in the integration of OCR and LLMs can be anticipated. The potential applications extend beyond document analysis and information retrieval, encompassing fields such as automated content summarization, sentiment analysis, and even more sophisticated natural language understanding tasks. The combined efforts of image-based Optical Character Recognition (OCR) and language models play a key role in addressing complex challenges related to understanding and extracting meaning from different types of written data. This collaboration is essential for advancing text processing, contributing to progress in these broader fields. Additionally, the integration of multiple languages to assess the multilingual capabilities of the system emerges as an important aspect, promoting inclusivity and versatility. This development enables the system to smoothly work with different languages, creating a globally accessible and connected digital environment. It also enhances accuracy and performance in various linguistic situations.

References

[1] Sakib, Abdullah Mohammad, Bilkis Jamal Ferdosi, Sabrina Jahan, and Kashfia Jashim. "Medical Text Extraction and Classification from Prescription Images." In 2022 25th International Conference on Computer and Information Technology (ICCIT), pp. 472-477. IEEE, 2022.
[2] Chen, Yisong, and Qing Song. "News text summarization method based on bart-textrank model." In 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 2005-2010. IEEE, 2021.
[3] Ruibin, Sun, Qian Kui, Xu Weimin, and Lu Hong. "Adaptive recognition of complex invoices based on Tesseract-OCR." Nanjing Xinxi Gongcheng Daxue Xuebao 13, no. 3 (2021): 349-354.
[4] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
[5] Octavianus, Farrel, Albert Wihardi, Muhamad Keenan Ario, and Derwin Suhartono. "Automated Text Summarization and Topic Detection on News Aggregation System Using BART and SVM." In 2022 International Symposium on Information Technology and Digital Innovation (ISITDI), pp. 108-113. IEEE, 2022.

[6] Zeyad, Abdulrahman Mohsen Ahmed, and Arun Biradar. "Assessing BigBirdPegasus and BART Performance in Text Summarization: Identifying Right Methods." In 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), pp. 1773-1778. IEEE, 2023.
[7] Sriharsha, A. V. "Information retrieval and linguistic analysis on cursive characters." International Joumal of Mechanical Engineering 7, no. 1 (2022): 7109-14.
[8] Dhasmana, Gulshan, Prasanna Kumar HR, and Guru Prasad. "Sequence to Sequence Pre-Trained Model for Natural Language Processing." In 2023 International Conference on Computer Science and Emerging Technologies (CSET), pp. 1-5. IEEE, 2023.
[9] Nascimento, Nathalia, Paulo Alencar, and Donald Cowan. "Self-adaptive large language model (llm)-based multiagent systems." In 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), pp. 104-109. IEEE, 2023.
[10] Fang, Haodong, and Min Bao. "Raw material form recognition based on Tesseract-OCR." In 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), pp. 942-945. IEEE, 2021.
[11] Mann, Ben, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).
[12] Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. "Large language models for information retrieval: A survey." arXiv preprint arXiv:2308.07107 (2023).
[13] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8 (2019): 9.
[14] Liu, Zhuang, Wayne Lin, Ya Shi, and Jun Zhao. "A robustly optimized BERT pre-training approach with post-training." In China National Conference on Chinese Computational Linguistics, pp. 471-484. Cham: Springer International Publishing, 2021.
[15] Breuel, Thomas M., Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. "High-performance OCR for printed English and Fraktur using LSTM networks." In 2013 12th international conference on document analysis and recognition, pp. 683-687. IEEE, 2013.
[16] Avyodri, Ridvy, Samuel Lukas, and Hendra Tjahyadi. "Optical Character Recognition (OCR) for Text Recognition and its Post-Processing Method: A Literature Review." In 2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA), pp. 1-6. IEEE, 2022.
[17] Dipu, Nadim Mahmud, Sifatul Alam Shohan, and K. M. A. Salam. "Bangla optical character recognition (ocr) using deep learning based image classification algorithms." In 2021 24th International Conference on Computer and Information Technology (ICCIT), pp. 1-5. IEEE, 2021.
[18] Memon, Jamshed, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)." IEEE Access 8 (2020): 142642-142668.
[19] Martínek, Jiří, Ladislav Lenc, Pavel Král, Anguelos Nicolaou, and Vincent Christlein. "Hybrid training data for historical text OCR." In 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 565-570. IEEE, 2019.
[20] Breuel, Thomas M. "The OCRopus open source OCR system." In Document recognition and retrieval XV, vol. 6815, pp. 120-134. SPIE, 2008.
[21] Smith, Ray W. "History of the Tesseract OCR engine: what worked and what didn't." In Document Recognition and Retrieval XX, vol. 8658, p. 865802. SPIE, 2013.
[22] Lu, Zhidong, Richard Schwartz, Premkumar Natarajan, Issam Bazzi, and John Makhoul. "Advances in the bbn byblos ocr system." In Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318), pp. 337-340. IEEE, 1999.
[23] Espana-Boquera, Salvador, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez. "Improving offline handwritten text recognition with hybrid HMM/ANN models." IEEE transactions on pattern analysis and machine intelligence 33, no. 4 (2010): 767-779.
[24] Smith, Ray. "An overview of the Tesseract OCR engine." In Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, pp. 629-633. IEEE, 2007.
[25] Zhang, Jiaming, Jitao Sang, Kaiyuan Xu, Shangxi Wu, Xian Zhao, Yanfeng Sun, Yongli Hu, and Jian Yu. "Robust captchas towards malicious OCR." IEEE Transactions on Multimedia 23 (2020): 2575-2587.