Smart MCQ Generation using KeyBERT and BERT based Models

Pranavesh Kumar Talupuri
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
cgame69696@gmail.com

Manjunaatt Girish Kumar
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India

Kotha Lavanya
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
kothalavanya99@gmail.com

Gudditi Chetan
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
gudditi.chetan1986@gmail.com

Nagendra Panini Challa
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
nagendra.challa@vitap.ac.in

2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON) | 979-8-3315-1859-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/DELCON64804.2024.10866615

Abstract– **This project proposes an Automated Multiple-Choice Question (MCQ) Generator that generates questions by analyzing text content using natural language processing (NLP) techniques. The system preprocesses text, extracts keywords, and associates them with appropriate sentences using Transformers, KeyBERT, and NLTK. Significant keywords are extracted by KeyBERT's BERT-based sentence embeddings, enabling direct sentence selection. Relevant sentences are chosen using KeyBERT's BERT-based sentence embeddings, which extract important keywords. To guarantee semantic accuracy, the system predicts word senses using a BERT-based MLM. MCQ generation is facilitated by blank substitution in sentences, which provides both correct answer and incorrect answers. User responses are saved for scoring and feedback. This useful application enables educators, makers, and students to quickly and efficiently create comprehension examinations. The system receives text files from users, which ensures flexibility and adaptation to a variety of educational circumstances and content domains.**

Keywords– Automated MCQ Generation, Natural Language Processing, Keyword Extraction, Word Sense Disambiguation, NLTK, KeyBERT, BERT.

I. Introduction

In today's educational environment, creating successful learning outcomes depends primarily on evaluating students' knowledge retention and comprehension of material. With the ability to assess knowledge in a variety of subjects and concepts, multiple choice questions (MCQs) have become an essential component of this assessment process for educators. But for educators and content developers, manually creating multiple-choice questions (MCQs) can be a labor- and time-intensive task. Our study proposes a unique approach to this problem by introducing an Automated MCQ Generator with Natural Language Processing (NLP) tools. By giving educators, a strong tool to speed up the generation of multiple-choice questions (MCQs) directly from textual content, this project represents a substantial advancement in educational technology. [1]

The user-provided input text is analyzed by the Automated MCQ Generator using advanced natural language processing techniques. The system automatically creates well-structured multiple- choice questions (MCQs) by means of a sequence of systematic processes that identify key concepts and extract relevant information. Our MCQ Generator stands out for its ability to provide both possible incorrect answers and right answers as alternatives, which increases the complexity and effectiveness of assessment materials. [2]

Our Automatic MCQ Generator's key components are as follows:

- Text Analysis: The system uses natural language processing (NLP) algorithms to understand and analyze the input text, dividing it into smaller and analyzable units.
- Keyword Extraction: Using natural language processing (NLP) techniques, the system identifies important words and phrases in the text that can be used as the basis for generating meaningful MCQs. [3]
- Question Generation: In order to ensure relevance and coherence with the original text, the system generates multiple-choice questions (MCQs) on its own, based on contextual understanding and the extracted keywords.
- Option Generation: The technology challenges the test-taker's knowledge and enriches the question options by providing possible wrong answers in addition to the right answer. [4]
- User Interaction: After choosing answers to the generated multiple-choice questions (MCQs), the user can interact with the assessment material by getting instant feedback on their responses.

The objective of this project is to provide educators with an effective and time-saving method for producing high-quality assessment materials. Through the use of natural language processing (NLP), we want to transform the MCQ-generating process and improve educational assessments in terms of usefulness, depth, and accessibility. By providing educators with a smooth and intelligent method for creating multiple-choice questions (MCQs) from textual content, our project, Automated MCQ Generator, contributes to the improvement of learning outcomes in contemporary educational paradigms. [5]

II. Literature Review

In the [5] proposed work the author introduces an automatic MCQ generation that focused on various techniques such as keyword extraction and sentence selection, whereas our proposed Automated MCQ Generator introduces several novel additions. Specifically, we employ advanced NLP techniques like Transformers and KeyBERT for preprocessing, keyword extraction, and sentence selection.

Research [6] focuses on generating an automated MCQ system by employing NLP for text preprocessing, keyword extraction, and MCQ generation through blank substitution. In contrast, our approach enhances this by integrating KeyBERT's BERT-based embeddings for keyword extraction and sentence selection, by combining with a BERT- based MLM for semantic accuracy. The key difference lies in utilization of advanced NLP techniques and user input.

Research [7] focuses on enhancing student engagement with computer-based tutorials by refining navigation and evaluation methods. Their methods target in addressing tutorial interaction improvement whereas our method focuses on automatic MCQ generation offering unique features such as the ability to upload multiple files and generate questions tailored to specific files.

The author of [8] surveys Nepali NLP research, organizing approaches and techniques. In contrast, our paper introduces a method for automated MCQ generation using NLP techniques like KeyBERT and NLTK, with features such as user input, score storage and answer evaluation. The key difference lies in the focus: Tej et al provide an overview of NLP research trends, while our paper offers a practical application for MCQ creation. Puneeth Thotad et al.'s Automatic Question Generator (AQG) [9] generates multiple-choice questions with answers and distractors, employing NLTK and Python. In contrast, our system prioritizes interactivity, providing an engaging learning experience with features for result analysis and feedback. Utilizing advanced natural language processing techniques such as KeyBERT and NLTK, we propose the generation of fill-in-the-blank MCQs.

III. Methodology

The project's methodology combines pre-trained language models, such as BERT, for word meaning disambiguation with KeyBERT for keyword extraction, allowing for the automated generation of multiple-choice questions (MCQs) applying pioneering natural language processing (NLP) techniques. The flow chart in Figure 1 illustrates the methodology of our project in a simple way [6].

Using the Natural Language Toolkit (NLTK), the methodology begins with text preprocessing. To make sure that the input text is coherent and clear, this phase involves several activities including lowercasing, sentence splitting, and tokenization. Text processing may be done more effectively with the help of NLTK, which offers a comprehensive toolkit. Keyword extraction is carried out using KeyBERT or comparable models after text preprocessing.

A transformer-based approach called KeyBERT is very good at finding key concepts in the text. Through keyword extraction, the system determines the key concepts that will serve as a base for creating multiple-choice questions (MCQs). This step is crucial to ensure the relevance of the questions that are generated. After identifying the key concepts, vocabulary is used to improve the accuracy of the generated MCQs. For this, pre-trained language models such as BERT are used. These models can generate the appropriate word senses from the sentences' context, ensuring that the questions that are produced are both contextually relevant and semantically accurate [7].

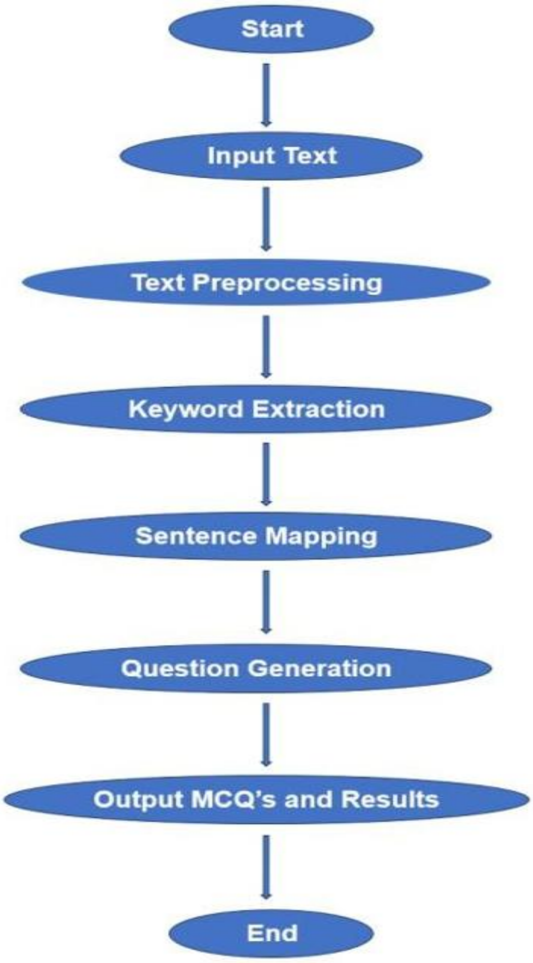


Fig. 1. Flowchart

The methodology automates the creation of multiple-choice questions (MCQs) by combining sophisticated natural language processing (NLP) tools. In order to ensure that the input is coherent, it first uses the Natural Language Toolkit (NLTK) for extensive text preprocessing, which includes lowercasing, sentence splitting, and tokenization. KeyBERT is used to extract keywords, identifying the fundamental ideas needed to formulate questions. Using a transformer-based method such as KeyBERT improves the recognition of important concepts in the text and provides a strong basis for developing multiple-choice questions. BERT and other pre-trained language models, which are skilled at inferring word meanings from context, are used to improve the quality of the queries. This two-pronged strategy guarantees the generated questions' semantic accuracy while also increasing their relevance. By combining these approaches, it is possible to produce relevant and high-quality educational resources that meet the needs of students, opening the door for more useful evaluation instruments in a variety of educational settings [8].

The system then generates multiple-choice questions (MCQs) after identifying the relevant concepts and disambiguated word senses. The multiple- choice questions are generated in a fill-in-the- blank format, with the identified key concepts represented by the blanks. This structure allows for flexibility in the generation of questions while preserving textual contextual alignment. Furthermore, incorrect answers are chosen at random from the text to provide diversity and complexity to the test participants. The generated multiple-choice questions (MCQs) are evaluated in the evaluation phase according to standards including difficulty, correctness, and relevance. To ensure the quality of the questions, user responses are evaluated and feedback on answer correctness is provided. Results are recorded for further analysis and system improvement, including MCQs, answers, and assessment results. The automatic MCQ-generating method may be improved over time with the help of this systematic recording. [9,10]

A. *Text Preprocessing*

Steps involved in text Processing: Text processing is required to make the input text data suitable for additional processing and analysis.

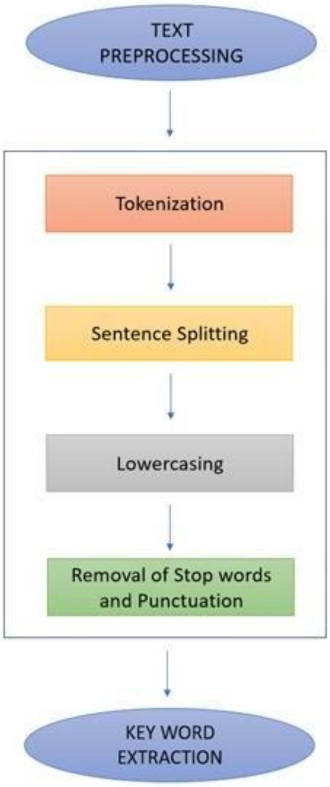


Fig. 2. Text Preprocessing

Figure 2 shows the process of Text Preprocessing.

- Tokenization, or dividing the text into discrete words or tokens, will be applied to the text data. This procedure is required to divide the text into readable portions for analysis. [11]
- Sentences will be extracted from the text after tokenization. Sentence boundaries inside the text must be identified, and this phase is crucial for several NLP activities, including sentence mapping and keyword extraction.
- To maintain text data consistency, all tokens will be changed to lowercase. This process guarantees that the text data is consistent for analysis and helps prevent duplicate tokens caused by variations in capitalization. [12]
- Common words like "the", "is", and "and" that have a minimum of meaning will be eliminated from the text data; these are known as stopwords. This allows for reducing the noise and concentrating on the most significant words and phrases for additional analysis. [13]
- The text data will be removed from punctuation, including quotation marks, commas, and periods. This is a crucial step to make sure punctuation doesn't interfere with keyword extraction or other NLP operations [14].

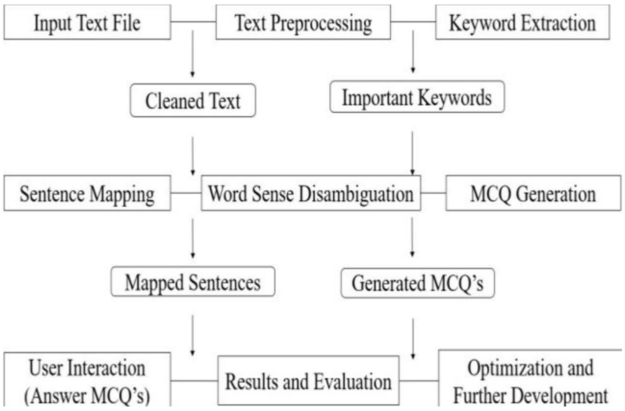


Fig. 3. Framework

Figure shows the framework of complete methodology.

B. *Keyword extraction models:*

Using natural language processing, advanced models like KeyBERT and BertForMaskedLM play crucial roles in extracting and understanding textual information. KeyBERT utilizes the Max Sum Similarity method to identify relevant keywords that encapsulate the essence of a text, facilitating the creation of meaningful multiple-choice questions (MCQs). Meanwhile, BertForMaskedLM specializes in word sense disambiguation by predicting contextually appropriate words to fill in gaps, enhancing comprehension assessment. Together, these models enable the generation of MCQs that accurately reflect key concepts and word meanings. This approach not only enriches educational content but also aids in effective knowledge evaluation [15].

1) *KeyBERT (Max Sum Similarity Method):*

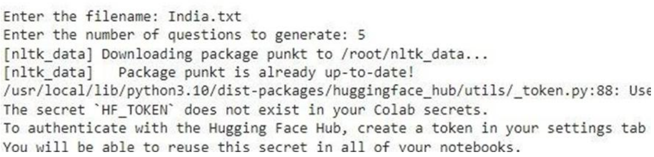
KeyBERT is a library that supports the use of complex transformer-based models, such as distilBERT, to extract relevant keywords from text. It makes use of the Max Sum Similarity method, which focuses on identifying keywords with the highest level of semantic similarity to the content of the text. To extract important keywords from the input text, KeyBERT is used. KeyBERT finds the most relevant keywords by evaluating each word or phrase's level of alignment with the overall content context. These keywords provide the basis for generating multiple-choice questions (MCQs) that properly represent the key concepts addressed in the text. [16]

2) *BertForMaskedLM (Word Sense Disambiguation***):**

BertForMaskedLM is a transformer-based model that is a variation of BERT that is specifically designed to address masked language modeling. It helps in sorting out the meaning of words in various contexts by predicting the most likely words to complete the blanks in a phrase. For word sense disambiguation, BertForMaskedLM is used in this project. It helps in recognizing the meanings of words inside phrases according to their context. BertForMaskedLM determines the correct sense of the word in the given context by substituting a mask token for the target word and predicting the best word to fill the blank. This feature is necessary to create multiple-choice questions (MCQs) that analyze understanding word meanings in the text. [17]

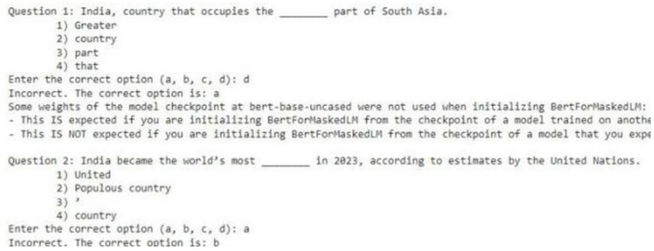
IV. RESULTS

The results of our study demonstrated the effectiveness and scalability of the suggested method in addition to comparing the quality of the generated multiple-choice questions with questions that were composed manually. We were able to evaluate and improve the system repeatedly, improving its performance over time, by saving the output of the automated MCQ generating process. Additionally, a positive way comments from users regarding the generated questions' relevancy and clarity validated the system's efficiency in automating the generation of multiple-choice questions. All of these results show how our technology may simplify training and evaluation procedures for education, providing a scalable and effective way to create MCQs that are pertinent to the context. Figures 4, 5, 6 and 7 show the successful implementation of taking user input, questions generation, correct and incorrect answers and output result file generation. [18]



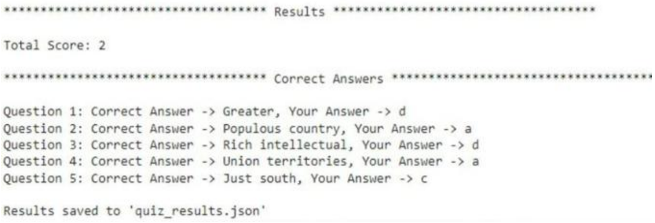
```
Enter the filename: India.txt
Enter the number of questions to generate: 5
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:88: Use
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
You will be able to reuse this secret in all of your notebooks.
```

Fig. 4. Use input of filename and number of questions



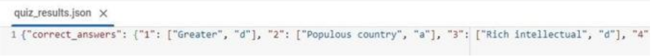
```
Question 1: India, country that occupies the _______ part of South Asia.
        1) Greater
        2) country
        3) part
        4) that
Enter the correct option (a, b, c, d): d
Incorrect. The correct option is: a
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForMaskedLM:
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on anothe
- This IS NOT expected if you are initializing BertForMaskedLM from the checkpoint of a model that you expe

Question 2: India became the world's most _______ in 2023, according to estimates by the United Nations.
        1) United
        2) Populous country
        3) '
        4) country
Enter the correct option (a, b, c, d): a
Incorrect. The correct option is: b
```

Fig. 5. Generation of MCQ's



```
*********************************** Results ***********************************

Total Score: 2

*********************************** Correct Answers ***********************************

Question 1: Correct Answer -> Greater, Your Answer -> d
Question 2: Correct Answer -> Populous country, Your Answer -> a
Question 3: Correct Answer -> Rich intellectual, Your Answer -> d
Question 4: Correct Answer -> Union territories, Your Answer -> a
Question 5: Correct Answer -> Just south, Your Answer -> c

Results saved to 'quiz_results.json'
```

Fig. 6. User's score and feedback



```
{"correct_answers": {"1": ["Greater", "d"], "2": ["Populous country", "a"], "3": ["Rich intellectual", "d"], "4"
```

Fig. 7. Quiz Results

V. CONCLUSION

In conclusion, our research offers a new approach to the automatic generation of multiple-choice questions (MCQs) from text files, providing a scalable and effective way to create MCQs for training and evaluations in education. Our method ensures the generation of contextually relevant multiple-choice questions (MCQs) by using advanced natural language processing (NLP) tools, such as BertForMaskedLM for word sense disambiguation and

KeyBERT for keyword extraction. Our method is distinct because it can extract relevant keywords, separate word senses, and generate multiple-choice questions (MCQs) based on the content of text files that users provide. Furthermore, we employ a score review system to evaluate user replies and save findings for future studies. our study generally helps to improve training processes and optimize educational assessments. Increasing adaptability by adding matching, true/false, and fill-in-the-blank question types. Question complexity can be customized by users for personalized learning. Developing a user- friendly interface for simple question editing and review. Code performance optimization to enable scalability on larger datasets. Enhancing accessibility by supporting multiple languages. Putting accessibility elements in place and encouraging inclusion. [19.20]

REFERENCES

[1] Rao, D., & Saha, S. K. (2018, December 21), "Automatic Multiple Choice Question Generation From Text: A Survey", IEEE Transactions on Learning Technologies, PP (1), 1-1. DOI: 10.1109/TLT.2018. 2889100.

[2] Kumar, S., Prasad, V., Santhanavijayan, A., Balasundaram, S. R., & Narayanan, S. (2017, January 01). "Automatic generation of multiple-choice questions for e- assessment". International Journal of Signal and Imaging Systems Engineering, 10, 54. DOI: 10.1504/IJSISE. 2017.10005435.

[3] Singh Bhatia, A., Kirti, M., Saha, S.K. (2013). "Automatic Generation of Multiple-Choice Questions Using Wikipedia". In: Maji, P., Ghosh, A., Murty, M.N., Ghosh, K., Pal, S.K. (eds) Pattern Recognition and Machine Intelligence. PReMI 2013. Lecture Notes in Computer Science, vol 8251. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642- 45062-4_104

[4] F. de Assis Zampirolli, V. Batista and J. A. Quilici-Gonzalez, "An automatic generator and corrector of multiple-choice tests with random answer keys," 2016 IEEE Frontiers in Education Conference (FIE), Erie, PA, USA, 2016, pp. 1-8, doi: 10.1109/FIE.2016.7757467.

[5] D. R. CH and S. K. Saha, "Automatic Multiple Choice Question Generation From Text: A Survey", IEEE Transactions on Learning Technologies, vol. 13, no. 1, pp. 14-25, 2020, doi: 10.1109/TLT.2018.2889100.

[6] Muthusundari, S., Vishwa, A., Kiruthik, K. S. Srinivasa, Sukesh, R., & Raj, 2021, "Automatic MCQ Generator",Annals of the Romanian Society for Cell Biology; Arad, 25(5), 2597- 2601. ISSN: 1583-6258.

[7] R. Marín, P.J. Sanz, O. Coltell, J.M. Inesta, F. Barber, D. Corella, Student-teacher communication directed to computer-based learning environments, Displays, Volume 17, Issues 3–4, 1997, Pages 167-178, ISSN 0141- 9382, https://doi.org/10.1016/S0141- 9382(96)01033-5.

[8] Shahi T. B., & Sitaula c., "Natural language processing for Nepali text: a review", Artificial Intelligence Review, 2022, 255(2), 1-29. DOI: 10.1007/s10462-021-10093-1.

[9] Puneeth Thotad, Shanta Kallur, Sukanya Amminabhavi, "Automatic Question Generator Using Natural Language Processing", Journal of Pharmaceutical Negative Results, 2022, 2759-2764.https://doi.org/10.47750/pnr.2022.13.S 10.330.

[10] Nwafor, C., & Onyenwe, I. (2021, April 30). "An Automated Multiple-Choice Question Generation using Natural Language Processing Techniques". International Journal on Natural Language Computing, 10, 1-10. DOI: 10.5121/ijnlc.2021.10201.

[11] Patil, N., Kumari, K., Ingale, D., Patil, P., & Uttarkar, A. R. (May-June 2021). A Survey on Automatic Multiple Choice Questions Generation from Text. International Journal of Scientific Research & Engineering Trends, 7(3), ISSN (Online): 2395-566X.

[12] Rakangor, S., & Ghodasara, Y. R. (January 2015). Literature Review of Automatic Question Generation Systems. International Journal of Scientific and Research Publications, 5(1), 1. ISSN: 2250-3153. [Online] Available at: www.ijsrp.org.

[13] Aldabe, I., Maritxalar, M. (2010). Automatic Distractor Generation for Domain Specific Texts. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds) Advances in Natural Language Processing. NLP 2010. Lecture Notes in Computer Science (), vol 6233. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978- 3-642-14770-8_5

[14] Musale, A. R., Gupta, S., Phadatare, M., Jadhav, I., & Vaibhav, B. O. (2020). An Automatic Generator of Multiple-Choice Question with Random Answer Key. Journal of Emerging Technologies and Innovative Research (JETIR), Volume 7, Issue 6, 830. [online] www.jetir.org. ISSN: 2349-5162.

[15] Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O., & Ghadi, Y. Y. (2022). Automatic computer science domain multiple-choice questions generation based on informative sentences. PeerJ Comput Sci, 8, e1010. https://doi.org/10.7717/peerj-cs.1010.

[16] Ahrari Khalaf, Ayesheh, Aisha Hassan Abdalla Hashim, and Akeem Olowolayemo. "Mutual character dialogue generation with semi-supervised multitask learners and awareness." International Journal of Information Technology 16.3 (2024): 1357-1363.

[17] Youness, Farida, Mohamed Ashraf Madkour, and Ayman Elshenawy. "Dialog generation for Arabic chatbot." International Journal of Information Technology 16.2 (2024): 881-890.

[18] Mehmood, Rayeesa, Rumaan Bashir, and Kaiser J. Giri. "VTM-GAN: video-text matcher based generative adversarial network for generating videos from textual description." International Journal of Information Technology 16.1 (2024): 221-236.

[19] Mulla, Nikahat, and Prachi Gharpure. "Leveraging well-formedness and cognitive level classifiers for automatic question generation on Java technical passages using T5 transformer." International Journal of Information Technology 15.4 (2023): 1961-1973.

[20] Kalra, Vandana, Indu Kashyap, and Harmeet Kaur. "Generation of domain-specific vocabulary set and classification of documents: weight-inclusion approach." International Journal of Information Technology 14.1 (2022): 275-285.