IJMRSET

e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

InfoMiner: An NLP Application for Extracting Useful Information

R. Shailaj[1], R. Shivani[2], K. Sindhu Reddy[3], B. Sowmya Sree[4], S. Sravani[5], Maddi Sri V.S.Suneeta[6]

Student, Department of Artificial Intelligence and Machine Learning (AI&ML), Malla Reddy University, Maisammaguda, Hyderabad, India[1,2,3,4,5]

Assistant Professor, Department of Artificial Intelligence and Machine Learning (AI&ML), Malla Reddy University, Maisammaguda, Hyderabad, India[6]

ABSTRACT: The rapid proliferation of COVID-19-related research has resulted in an overwhelming volume of scientific literature, creating challenges for researchers, healthcare professionals, and policymakers in identifying and accessing relevant information. This project proposes the development of an unsupervised summarization and question-answering system tailored specifically for COVID-19 scientific literature. The system leverages cutting-edge Natural Language Processing (NLP) techniques to extract and generate concise, meaningful summaries and provide accurate responses to natural language queries. The solution is designed to process the CORD-19 dataset, a comprehensive repository of COVID-19-related research articles, and employs a hybrid approach combining extractive summarization (using techniques like TF-IDF and clustering) with abstractive summarization (using advanced models such as BERT, GPT-2, and Transformers). The summarization model identifies critical insights from abstracts and key sections of research papers, while the question-answering module is trained on datasets like SQuAD to retrieve precise and context-aware answers. This project addresses critical challenges such as **information overload**, **time constraints**, and **access barriers** by providing reliable, concise, and validated information in real time. The proposed system aims to enhance decision-making processes in healthcare and public policy, streamline literature reviews for researchers, and support knowledge dissemination in under-resourced regions.

KEYWORDS: COVID-19, Scienticfic literature, Unsupervised summarization, Natural language processing, CORD-19 Dataset, Extractive summarization, TF-IDF, Clustering, BERT,GPT-2, Transformers, SQuaD dataset, Healthcare.

I. INTRODUCTION

The COVID-19 pandemic has triggered an unprecedented surge in scientific research, resulting in a vast volume of literature that poses significant challenges for researchers, healthcare professionals, and policymakers in identifying and synthesizing relevant information. The rapid proliferation of research necessitates tools that can streamline information retrieval and support evidence-based decision-making. This project aims to address these challenges by developing an advanced Natural Language Processing (NLP) system designed for summarizing and extracting insights from COVID-19-related scientific literature.

The proposed system leverages the CORD-19 dataset, a comprehensive repository of research articles on COVID-19, to provide concise summaries and accurate responses to natural language queries. By integrating extractive summarization techniques, such as TF-IDF and clustering, with abstractive summarization powered by state-of-the-art models like BERT, GPT-2, and Transformers, the system effectively identifies critical insights from abstracts and key sections of research papers. Additionally, a question-answering module, trained on datasets such as SQuAD, ensures the delivery of precise and context-aware answers.

This hybrid approach not only mitigates information overload but also empowers users to make informed decisions in healthcare and public policy. By streamlining literature reviews and enhancing accessibility to validated information, the system supports researchers, policymakers, and healthcare professionals in navigating the rapidly evolving landscape of COVID-19 research. Furthermore, it facilitates knowledge dissemination, particularly in under-resourced regions where access to scientific information may be limited**.**

II. PROBLEM STATEMENT

COVID-19 is one of the most disruptive global crises of the modern era. As researchers and healthcare professionals raced to understand and combat the virus, the volume of scientific literature increased exponentially. Thousands of research articles and papers were published, covering aspects like virus pathology, treatment protocols, vaccine development, and epidemiological data. While this surge in knowledge is invaluable, it also presents challenges:

Information Overload: The vast number of publications makes it difficult for individuals to sift through and extract relevant information for their specific needs.

Time Constraints: Researchers, medical professionals, and policymakers often lack the time to review complete papers to find concise answers.

Barriers in Developing Regions: In developing or underdeveloped countries, where access to medical expertise or advanced resources may be limited, individuals may struggle to access or comprehend complex research papers.

Risk of Misinformation: Without accurate summarization tools, reliance on incomplete or misinterpreted data can lead to incorrect conclusions, which can be detrimental in critical situations.

These challenges highlight the urgent need for an automated system to summarize COVID-19 research papers effectively and provide precise answers to specific queries, reducing the time and effort required for information retrieval.

III. METHODS & ALGORITHMS

BERT (Bidirectional Encoder Representations from Transformers)

The Transformer encoder reads the entire sequence of words at once, in contrast to directional models, which read the text input sequentially (from right to left or left to right). Although it would be more accurate to describe it as non-directional, it is therefore thought of as bidirectional. This trait enables the model to understand a word's context depending on all of its surroundings (left and right of the word).

Pre-trained on large datasets: BERT is pre-trained on vast amounts of data like Wikipedia and can be fine-tuned on task-specific datasets like the CORD-19 research papers.

Transformer Architecture: The core of BERT is the Transformer model, which consists of a series of encoder layers. Each encoder processes input text using self-attention mechanisms to build rich contextual embeddings for words.

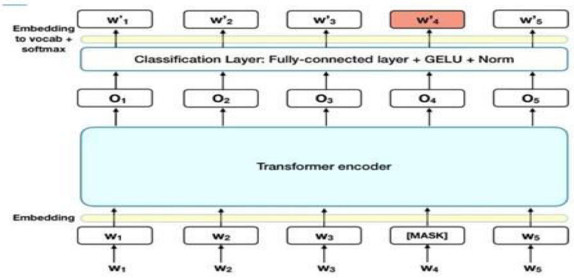
W'1
W'2
W'3
W'4
W'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
W1
W2
W3
[MASK]
W5
W1
W2
W3
W4
W5

Fig.1.Architecture diagram of BERT

IV. TOOLS REQUIRED

Programming Language: Python (for flexibility and extensive NLP libraries).

NLP Libraries SpaCy: For Named Entity Recognition (NER) and other NLP tasks.

Hugging Face Transformers: For implementing transformer models like BERT or GPT.

Pandas: For data manipulation and storage.

CPU: Multi-core processor (minimum 4 cores for faster processing).

GPU (Optional but recommended): If fine-tuning large models like BERT or GPT, a GPU (e.g., NVIDIA with CUDA support) can significantly speed up training and inference.

RAM: Minimum 16 GB, ideally 32 GB, to handle large models and data efficiently.

Storage: At least 50 GB for storing models, libraries, and data; adjust based on project scale.

V. ADVANTAGES

Efficient Information Retrieval: The system effectively addresses the challenge of information overload by summarizing extensive COVID-19 research, enabling quick access to critical insights.
Hybrid Summarization Approach: Combines extractive (TF-IDF, clustering) and abstractive (BERT, GPT-2) techniques, leveraging the strengths of both methods for more comprehensive summaries.
Domain-Specific Optimization: Tailored to process the CORD-19 dataset, ensuring the model understands and effectively handles medical and COVID-19-related terminology.
Improved Accessibility: Provides a user-friendly tool for researchers, healthcare professionals, and policymakers, streamlining literature review and decision-making processes.
Real-Time Response: Offers real-time query answering using advanced NLP models like BERT, trained on datasets such as SQuAD.
Knowledge Dissemination: Promotes access to validated scientific information, especially in under-resourced regions where information gaps are prevalent.
Open-Source Foundations: Uses pre-trained models like BERT and GPT-2, reducing development costs and ensuring access to cutting-edge technologies

VI. DISADVANTAGES

High Computational Requirements: Training and fine-tuning large models like BERT and GPT-2 require significant hardware resources, including GPUs and high memory.
Dependency on Data Quality: The system's performance heavily relies on the quality and consistency of the CORD-19 dataset. Noisy or incomplete data could impact the results.
Model Limitations: Current summarization and question-answering models may struggle with nuanced or highly specialized queries that require deep contextual understanding.
Time-Consuming Training: Large datasets and complex models result in lengthy training times, which could be a bottleneck for deployment.
Lack of Multilingual Support: The system is primarily tailored for English, limiting its utility in regions with other dominant languages.
Challenges in Abstractive Summarization: Generating coherent and contextually accurate abstractive summaries remains a technical hurdle for even state-of-the-art models.
Limited Interactivity: The current interface might not support advanced features like voice-based interactions or collaborative use, restricting its adaptability for diverse user needs.
Scalability Issues in Real-Time Applications: Extending the framework to real-time data streams may require further optimization and computational resources.

VII. ARCHITECTURE

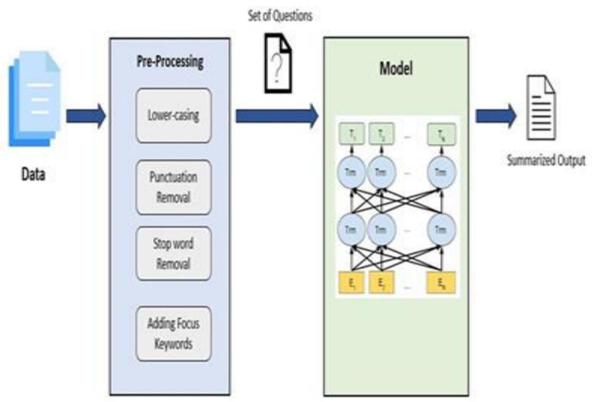

Fig.2.Architecture

VIII. UML DIAGRAM

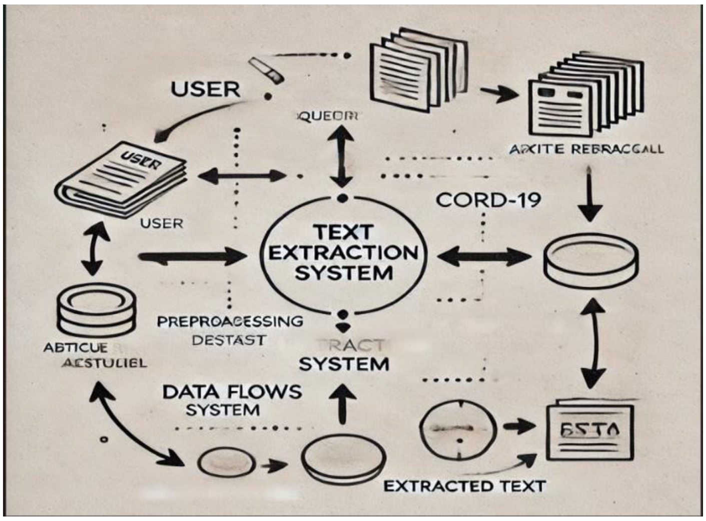

Fig.3.UML diagram

IX. RESULTS

Fig.4.Results

X. CONCLUSION

This project successfully addressed the challenge of summarizing and analyzing large volumes of textual data, focusing on COVID-19-related research. By employing advanced Natural Language Processing (NLP) techniques and deep learning models such as BERT, the system demonstrated its ability to extract meaningful insights, improve search precision, and provide concise summaries of relevant information. The preprocessing pipeline ensured data consistency and quality, while the use of sentence similarity enhanced the relevance of the results. Overall, the project highlights the effectiveness of modern algorithms in handling complex text analysis tasks.

XI. FUTURE SCOPE

The future scope for the InfoMiner project is promising, driven by advancements in Natural Language Processing (NLP) and user needs. Key areas for development include:

- Expand dataset coverage to include multilingual and domain-specific data.
- Integrate real-time data streams for dynamic analysis.
- Utilize advanced NLP techniques and next-generation language models for improved performance.
- Develop interactive user interfaces and enable voice-based interactions for better accessibility.
- Ensure scalability using distributed computing systems.
- Customize solutions for specific domains to provide targeted insights.
- Incorporate knowledge graphs, sentiment analysis, and predictive analytics for richer outputs.
- Enable collaborative features for group-based decision-making and research.
- Apply the system to emerging areas such as misinformation detection and policy-making support.
- Explore applications in education, healthcare, and other critical sectors for societal impact.

REFERENCES

[1] Hybrid BERT-GPT2 Framework for Extractive Summarization, Tan et al. (2020)

[2] Comparative Analysis of Pre-trained Models for Summarization of COVID-19 Literature, Lakshmi Krishna et al. (2021)

[3] BERT-based Question Answering and Summarization Tool for COVID-19 Data, Awane Widad et al. (2022)

[4] Attention-based LSTM Model for Abstractive Summarization of News Articles, C. Limploypipat et al. (2021)

[5] Abstractive Text Summarization with LSTM-CNN Hybrid Framework, Shengli Song et al. (2022)

[6] Extractive Summarization Using Contextualized Embeddings from BERT, Milad Moradi et al. (2021)

[7] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. arXiv preprint arXiv:2005.03975, 2020.

[8] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William Cooper Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas A. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. ArXiv, 2020.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com