Deep Learning Intelligent IR Related with COVID Pandemic

Priti Khodke[1], M.S.Ali[2], Kiran A. Dongre[3]

[1]*Associate Professor, Department of Computer Science & Engineering, PRMCEAM – Badnera, Amravati, Maharashtra, India*
[2]*Principal, PRMCEAM – Badnera, Amravati, Maharashtra, India*
[3]*Associate Professor, Department of Electrical & Engineering, PRMCEAM – Badnera, Amravati, Maharashtra, India*

---***---

Abstract - *A number of research publications have suddenly cropped up due to the sudden coronavirus outbreak across the globe. This pandemic forced researchers to study, understand, keep track of the advancements in the domain and apply technology to tackle the scenario. Though Technology cannot prevent the pandemics; however, it can definitely be used to create awareness, help prevent and track the spread, create awareness, warn and empower those on the ground to be aware of the situation, and noticeably lessen the impact. This paper summarizes how technology have proven to be a boon to control, if not end the pandemic. The technological advancements in the domain of AI, NLP, Robotics, digital payments has surely helped a lot during the pandemic for increasing the visibility, suggesting drugs, tracing and sharing information, contact less daily need deliveries to name a few. Ample of literature is available, the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 195,000 scholarly articles, including over 87,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This has been made openly available to researchers to apply AI techniques with data mining to retrieve information and develop tools to help the medical community. This paper presents how Artificial Intelligence - Information retrieval and mining has proven to be a helpful tool during COVID-19*

Key Words: COVID-19, Information retrieval, data mining, CORD 19, Deep learning, Artificial Intelligence

1. INTRODUCTION

The COVID 19 pandemic has become a threat for life. Since 2019 humans are learning and adapting to new ways of dealing and handling the pandemic. Technology has played a very important role during this scenario. This paper mainly deals with the role of Artificial Intelligence in the COVID 19 Scenario. [1] discussed the role of digital and 4.0 technologies in relation to their impact on industry, labour market and society during Covid 19 pandemic. It also discussed a quantitative analysis of rapidly changing behavior of the technologies in this pandemic scenario using a data driven relational structure.

Suddenly too much literature on COVID poured on to the web. The need is to intelligently use the available literature to handle the scenario. Though Technology cannot prevent the pandemics; however, it can definitely be used to create awareness, help prevent and track the spread, create awareness, warn and empower those on the ground to be aware of the situation, and noticeably lessen the impact.

2. ROLE OF TECHNOLOGY DURING COVID PANDEMIC

In development, growth and evolution, in particular for companies and activities involved in the process of producing goods for sale, digital technology is playing a main role. The present work [1] discussed the role of digital and 4.0 technologies in relation to their impact on industry, labour market and society during Covid 19 pandemic. It also discussed a quantitative analysis of rapidly changing behavior of the technologies in this pandemic scenario using a data driven relational structure. It uses the methodology where firstly the suitable data sources are defined for gathering information, it then extracts COVID related literature, digital technologies and finally establishes a connection on how technology is connected to the topics identified in the documents.

2.1 Role of Artificial Intelligence and Big Data

[2] QUOC-VIET PHAM et.al. Presented a survey on the state-of the-art solutions against the COVID-19 pandemic. The paper aimed at emphasizing importance of artificial intelligence (AI) and big data in responding to the COVID-19 outbreak and preventing the severe effects of the COVID-19 pandemic. It 's main purpose to show the effectiveness of AI and big data for finding fast and effective approaches that can effectively combat the COVID-19 disease and review state-of-the-art solutions using these technologies. They reviewed the applications of AI for detection and diagnosis, tracking and predicting the outbreak, infodemiology and infoveillance, biomedicine and pharmacotherapy. The applications of big data for the COVID-19 disease have been also presented, including outbreak prediction, virus spread tracking, diagnosis and treatment, and vaccine and drug discovery. The work additionally talked about the challenges needed to overcome for the success of AI and big data in fighting the COVID-19 pandemic.

2.2 Data Sets for COVID - 19

2.2.1 The CORD Dataset

[3] The paper describes the research, construction and the contents of the CORD -19 dataset, which is widely used for COVID -19 related literature. It also talks about the challenges faced during construction of the dataset and also mentions the Roadmap for CORD19 further. Literature form arXiv, bioRxiv, medRxiv, PubMed and the World Health Organization's Covid-19 Database is sourced in CORD-19.

"The CORD-19: The COVID-19 Open Research Dataset", authored by Lucy Lu Wang, Kyle Lo, et. al. has presented an overview of different aspects of the CORD-19 dataset. This dataset include several 100 scientific papers related to COVID -19, and related information about historical corona viruses such as SARS and MERS.

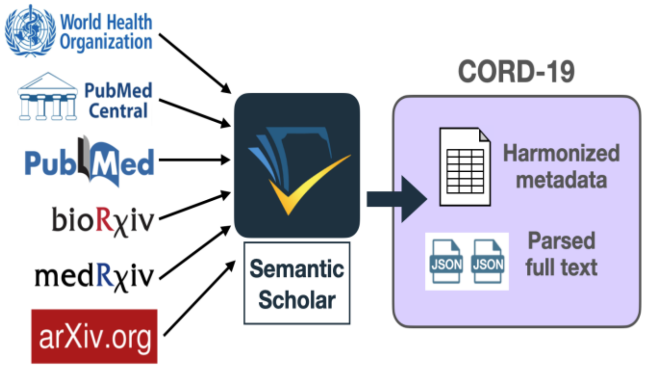

Figure 1: Sources for CORD-19 Data

It is used for text mining and information retrieval or discovery systems. The dataset is updated continuously on a daily basis with papers from new sources and the latest publications. Figure 1. Papers and preprints are collected from different sources, and the deduplicated harmonized metadata and full text are released as part of CORD-19.

The research is aimed to bring together the biomedical experts, the computing community, and policy makers for COVID-19. AI-based techniques are used to retrieve the information and natural language processing is used to extract useful information.

CORD-19 can be used directly by clinicians and clinical researchers, or as a tool to assist the clinicians, or is used to support further text mining and Natural Language Processing (NLP) research, and also for shared tasks and competitions.

One drawback of this dataset is that documents apart from research papers and preprints, such as technical and statistical reports, white papers, foreign language papers, and government published information are not curated within. Also, the papers could be made available in other than PDF format such as JSON, XML or HTML.

2.2.2 Multilingual COVID 19 Dataset

The Authors [4] in the paper, "NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset" published a multilingual dataset consisting of more than 20 million microblogs related to COVID-19.

The languages covered are Japanese and Chinese apart from English. They have collected English and Japanese microblogs from Twitter, and Chinese microblogs from Weibo within the duration 20 January 2020 to 24 March 2020. Key-word based search method is used for data collection. The retweets in twitter are filtered out using "filter: retweets" operator, and only "original microblogs" are included from Weibo to ensure uniqueness of data. The authors have demonstrated one of the possible utilizations of their dataset through qualitative and quantitative analysis. The quantitative analysis through daily microblog count, and the qualitative analysis by applying the word cloud base analysis, are used to summarize the results about the trends in social media towards the concern for this global corona virus pandemic. A deeper analysis based on different aspects can be done using this data set which may contribute in extracting useful clinical information, or studying the emotional/ social impact of the social media communication, or render hints about efficient broadcasting of the clinical information, and more.

2.3 Artificial Intelligence for Information Retrieval in COVID Literature

[5] In "CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research", the authors Dan Su1, Yan Xu1, Tiezheng Yu, Farhad Bin Siddique, Elham J. Barezi, Pascale Fung, have proposed a deep learning-based system. The natural language processing (NLP) question answering (QA) techniques combined with summarization for mining the available scientific literature related to COVID- 19 is used. The proposed CAiRe-COVID system is made of three major modules. First, the long or complicated queries are paraphrased into simple and system comprehendible ones. These updated queries are then fed to the IR module, and then into question answer modules, which extract the relevant snippets and question answer models from the full text articles. The third module is used to summarize the top 3 most relevant paragraphs and provide a summary or an extractive and abstractive summary to the given query.

Bhrugesh Joshi et. Al [6] have designed deepMINE ,which is the AI-NLP based system to assist the researchers for screening the thousands of published research and making the comprehension in a short amount of time. The system can mine the relevant articles and give a short article summary, which can make ease in fast and efficient comprehension for researchers. . The goal of proposed system deepMINE, is to provide quick and efficient access of the openly available research articles.

Meaningful research articles can be extracted from huge set and brief technical summary can be produced for research articles as per user interest with the application of the literature mining system. The deep natural language processing-based text summarization used by the proposed system for generating detailed technical summary of the input research article.

[7] presents a COVID Search Engine CO-Search for providing rigorous search and results over the huge literature available on Coronovirus. The author used the dataset CORD-19 along with TREC-COVID competition's evaluation dataset. The CORD 19 data set was used for training the system and the evaluation was done with the help of the TRECCOVID dataset. The CO-Search is a retriever-ranker semantic web crawler which that takes search inquiries in common language – usually in question form and retrieves logical articles over the coronavirus writing. It uses Semantic model for retrieval. The retriever which is built from a Siamese-BERT encoder is composed of a TF-IDF vectorizer and reciprocal-rank fused with a BM25 vectorizer. The question-answering module makes up the ranker which uses summarizer working on multi-paragraph to adjust the retriever scores. A bipartite graph of document paragraphs and citations is generated.

In All the Search Engine Works with the following steps

- Indexing – In this phase the documents are fragmented by paragraphs and image captions and then embedded with a Siamese BERT network which is pre-trained and finally saved as an index with the TF-IDF & BM25 vectors of the entire documents.

- Retrieval – Here the TF-IDF and SBERT retrieval scores computed in the indexing phase are then combined, further combined via reciprocal ranked fusion with the retrieval scores of BM25.

- Ranking – Here the query and the documents retrieved in the Retrieval phase are then parsed through a question answering model and abstractive summarizer before they are ranked on the bases of answer match, summarization match and the retrieval scores.

The below Figure 2 summarizes the process used by the CO-Search.

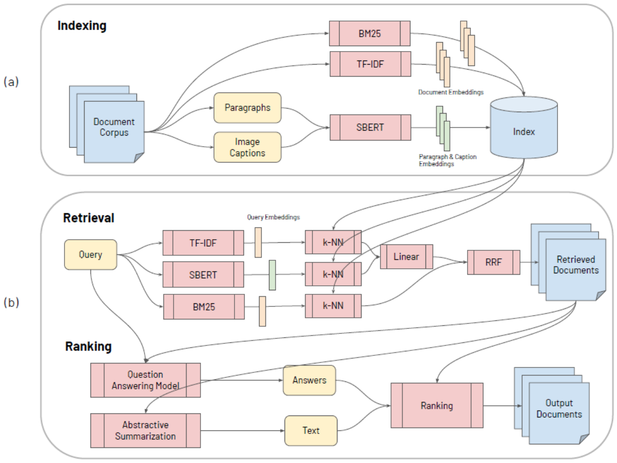

Figure 2: System Architecture – Co-Search

[8] uses Machine learning approach for Information retrieval related to Corona virus. For mining the COVID 19 articles author uses classification, clustering and one class support vector machine (OCSVMs) thus extracting various happenings and trend related to the coronavirus literature searches. The dataset used for experimentation is the COVID-19 open research dataset (CORD-19). Firstly, the features of the research documents and defined tasks are generated using doc2vec, which are later used for classification and clustering.

Based on the domain knowledge, the literature is classified; then clustered using clustering techniques (here k-means, DBSCAN and HAC) and then parallel OCSVMs is used for assignment of the tasks. The author claims better results using parallel OCSVMs, preceded by k-means clustering. The overall approach used is bottom-up. In order to answer the queries related to corona virus literature, initially tasks defined using the domain knowledge, then the tasks are mapped to similar clusters. The overview of the approach is as shown in Figure 3. consisting of steps –

1. Document embedding (DE),

2. Articles clustering (AC),

3. Dimensionality reduction (DR) and visualization

4. one-class classifier (OCC) from your manuscript in different sections.

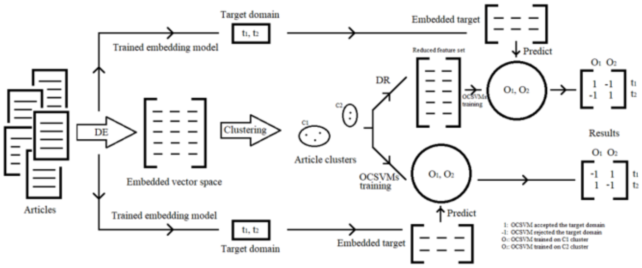

Figure 3 : Schematic diagram for target specific mining using one class approach

No proven medicine and/or vaccine is available for Novel coronavirus (COVID-19) however, patients are recovering with a number of antibiotics as well as Vitamin supplements. In order to tackle the spread of the disease a number of AI techniques were used. The author [9] developed data mining models to predict recovery of COVID-19 infected patients' recovery. The models were developed by applying a number of algorithms - K-nearest, the decision tree, support vector machine, random forest, logistic regression and naive Bayes were applied on the dataset. The model predicted a number of things including: patients at high risk – their age group, the time duration required for COVID patients for recovery and so on. The author concluded that the model with decision tree data mining algorithm was more efficient in predicting the patients recovery from the COVID -19 as compared to models built using other algorithms like k-nearest , support vector Machine, Random forest , Naïve bayes and logistic regression with the best accuracy of 99.85%.

The author[10] created a web based system EVIDENCEMINER which allows users to fire their query using natural language and the system retrieves textual evidence from CORD.

EVIDENCEMINER uses the Named entity recognition (NER) and Open information extraction Figure 4. It follows three steps

1. Pre-processing

2. Corpus Indexing

3. Textual Evidence retrieval

During pre-processing, the EVIDENCEMINER does the named entity recognition and extracts the patterns in the corpus indexing step the three offline indexes namely Word indexing, entity indexing and pattern indexing. For fast retrieval of results these entities and meta-patterns are computed prior and stored or indexed. Now with the users query and the indexes the evident sentences are retrieved and ranked based on the best match using the confidence score it calculates word score, entity score and pattern score.

The systems performance is compared with BM25 [11](Robertson et al., 2009) and search engine, sentence-level named LitSense [12] (Allot et al., 2019).

The author claims that the system can help build applications for the COVID related activities. The System block diagram is shown below

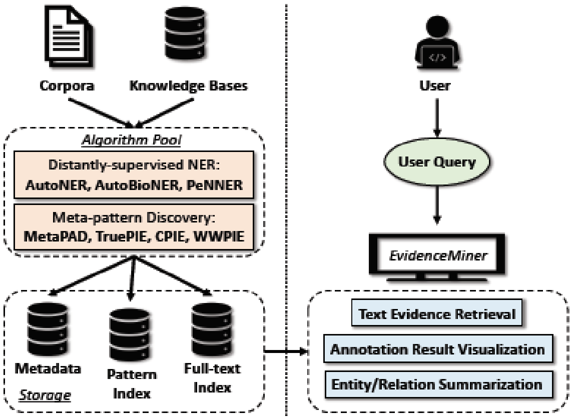

Figure 4 : System architecture of EVIDENCEMINER

The authors [13] after analyzing the existing literature and articles on COVID 19 (the CORD-19), establishes relationship between them, further applying cartography in order to mine structured knowledge. The author shows how throughout history; during similar epidemic or pandemic, researchers have responded in similar manner

The data mining method named Association Rule Text Mining (ARTM) is used for connecting interesting terms and their relationship for forming the association rules. Further Information cartography ; from the association rules mines out the structured knowledge. Information cartography basically creates metro maps from the structured knowledge making it easy to visualize and provide directions.

Li Bai et.al. in [14] aimed to observe the COVID-19 Intelligent Diagnosis and Treatment Assistant Program (nCapp) primarily based on the Internet of Things (IoT) medical science to conduct medical work for the duration of the COVID-19 epidemic, specifically for outpatients, and quality control (QC) will help the diagnosis and treatment, and acquire early identification, isolation, and treatment of sufferers with COVID-19. This consensus is suitable for unique experts at all stages of hospitals and even managers at all stages of hospitals, nearby community improvement corporations, and public fitness centers. It will allow them with smart help to work in the well timed discovery, isolation, and management of sufferers who are confirmed, suspected, and suspicious to have the disease via the nCapp.

Medical IoT (MIoT) goals to establish a decision-oriented big data analysis model supported by way of data science such as communication, electronics, biology, and medicine. MIoT can additionally be used for the prevention and control of COVID19. It can establish a three-level linkage nCapp system primarily based on the medical concept and technology of the IoT to diagnose and treat COVID-19. The IoT nCapp cloud medical system platform includes the simple features of the IoT and has a core graphics processing unit (GPU). Cloud computing systems linked to present digital medical records, image archiving, and image archiving and verbal exchange can higher aid in deep mining and intelligent diagnosis.

3. CONCLUSIONS

The greatest risk today; the pandemic COVID 19 an infectious virus killing people. Artificial Intelligence has played a great role in providing transparency in this current COVID-19 situation. It is evident that after the COVID-19 outbreak; technological innovations from AI to robotics will help to manage the epidemic and better equip to fight future public health emergency in a timely, systematic, and calm manner.

References

[1] N. Melluso, S. Fareri, G. Fantoni, A. Bonaccorsi, F. Chiarello, E. Coli,V. Giordano, P. Manfredi, and S. Manafi, “Lights and shadows of covid-19, technology and industry 4.0,”arXiv preprint arXiv:2004.13457,2020.

[2] Q.-V. Pham, D. C. Nguyen, W.-J. Hwang, P. N. Pathiranaet al.,“Artificial intelligence (ai) and big data for coronavirus (covid-19)pandemic: A survey on the state-of-the-arts,” 2020

[3] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide,K. Funk, R. Kinney, Z. Liu, W. Merrillet al., “Cord-19: The covid-19 open research dataset,”ArXiv, 2020.

[4] Z. Gao, S. Yada, S. Wakamiya, and E. Aramaki, “Naist covid: Multilingual covid-19 twitter and weibo dataset,”arXiv preprintarXiv:2004.08145, 2020.

[5] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. J. Barezi, and P. Fung, “Caire-covid: A question answering and multi-document summarization system for covid-19 research,”arXiv preprint arXiv:2005.03975, 2020.

[6] B. P. Joshi, V. D. Bakrola, P. Shah, and R. Krishnamurthy, “deepmine-natural language processing based automatic literature mining and research summarization for early stage comprehension in pandemic situations specifically for covid-19,”bioRxiv, 2020.

[7] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev,and R. Socher, “Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization,” arXiv preprint arXiv:2006.09595, 2020.

[8] S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, “Target specific mining of covid-19 scholarly articles using one-class approach,”arXivpreprint arXiv:2004.11706, 2020.

[9] L. Muhammad, M. M. Islam, U. S. Sharif, and S. I. Ayon, “Predictive data mining models for novel coronavirus (covid-19) infected patients recovery,” 2020.

[10] X. Wang, W. Liu, A. Chauhan, Y. Guan, and J. Han, “Automatic textual evidence mining in covid-19 literature,”arXiv preprintarXiv:2004.12563, 2020.

[11] S. Robertson and H. Zaragoza, The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc, 2009.

[12] A. Allot, Q. Chen, S. Kim, R. Vera Alvarez, D. C. Comeau, W. J. Wilbur,and Z. Lu, “Litsense: making sense of biomedical literature at sentence level,”Nucleic acids research, vol. 47, no. W1, pp. W594–W599, 2019

[13] I. Fister Jr, K. Fister, and I. Fister, “Discovering associations in covid-19related research papers,”arXiv preprint arXiv:2004.03397, 2020.

[14] L. Bai, D. Yang, X. Wang, L. Tong, X. Zhu, N. Zhong, C. Bai, C. A.Powell, R. Chen, J. Zhouet al., “Chinese experts’ consensus on the internet of things-aided diagnosis and treatment of coronavirus disease2019 (covid-19),”Clinical eHealth, vol. 3, pp. 7–15, 2020.