AN IMPLEMENTATION OF DATA DRIVEN ANALYSIS AND FORECAST FRAMEWORK OF COVID-19 BASED ON THE QA USER REVIEWS USING XTRACT QA MODEL

V.Nandini¹, K.C.Rajeswari², J.Jayanthi³, Anitha Elavarasi⁴
¹,² Associate Professor, Sona College of Technology, Salem, India
Email: ¹nandinivijaykumar@sonatech.ac.in, ²jayanthij@sonatech.ac.in
³,⁴ Assistant Professor, Sona College of Technology, Salem, India
Email: ³rajeswarikc@sonatech.ac.in, ⁴anithaelavarasi@sonatech.ac.in

ABSTRACT

One of the biggest hurdles that affected humankind recently is the coronavirus disease. People worldwide are being the victim in huge numbers. Health care professionals and researchers are striving in havoc to admit the change in levels of knowledge uncertainty prevailing and depth of information about the pandemic. The absence of data accessible at those early phases of the infection's spread—first about the actual spread, and later about its alleviation including the significance of wearing masks was an alarming fact. As there is typically a collection of documents based on real-world information, assessments from informal retrieval systems, forecasting data are insufficient to provide the users with genuine results to determine which documents apply to every point. Individuals overall look for precise and cutting-edge data on the quick-changing Covid pandemic. IR-related tasks, such as question-answering fact-checking and future predictions based on the QA and in general are still challenging. With this notion, an Xtract QA model is proposed that exhibits the ability of ML models to forecast the pandemic activities in a real-world semantic search engine that returns snippets containing evidence retrieval and fact-checking for the questions related to questions and claims about the COVID-19 pandemic information feeds published in health care recommendation systems, social media, and the worldwide web.

I. 1. INTRODUCTION

Forecasting is one of the most powerful statistical methods for detecting and analyzing trends and predicting future outcomes based on which timely and reducing actions can be taken all over the world in various fields of study. The standard way to deal with information retrieval system evaluation spins significance around the idea of relevant and irrelevant archives. Regarding client data reported, it needs classification either relevant or non-relevant. The crucial component in fact-checking and question-answering systems is the retrieval of appropriate passages. Most systems depend entirely on sparse retrieval to minimize recall, in particular when passages that hold few common characteristics phrases including the question sentence. These progressions in the lead and spread of science all make difficulties for Information retrieval and forecasting (IR), the logical field behind search engines. The focused objective of IR is to promptly look through a huge compilation of reports to discover relevant data to address a specific data need. The biomedical and health objectives of IR range from elevating scientific and provide clinical decision support tending to the wellbeing according to needs of customers and conflict misinformation. All of these are, obviously, profoundly important in a pandemic. More pieces of information are been posted on the internet, social media, and news feed. Amongst the huge collection of information, individual queries need to be responded to with appropriate and accurate information with answers to all questions. The facts provided with evidence are highly essential to foresee the evident information.

Many researchers are working towards the implementation of the Covid crisis and solutions. Biomedical content depends on customary AI (ML) techniques [15]. Health engineering is keenly developing for new advancements and procedures to track and control the expansion of the Covid epidemic in this global health emergency. One of the best essential worldwide innovations right presently is Artificial Intelligence (AI), which can follow the speed and distinguish the growth rate, its incarnations, recognize the danger of Corona infected patients. Since the
voluminous amount of pandemic information is being flooded at every second, it is impossible to recognize the evidence in them provided relevant information alone is to be reached based on their fears. A machine learning model is highly required at this point to anticipate the severity of patients and classify the data by analyzing the patient records. The limitations occurring in existing systems are the scarcity of trained labeled samples similar to other domains. Learning paradigms such as unsupervised learning supervised learning and semi-supervised learning are utilized for creating labels from the medical data. In this paper the main scope is to retrieve relevant QA content using NLP, cluster, and perform question-answer classification based on the level of evident information retrieved from the QA pairs, providing appropriate answers with text-based evidence focusing on ML and DL towards COVID-19 medical information.

II. RELATED WORKS

Sudhir Bhandari et al., [1] focused on a proof-based decision for assessing posterior probability distributions from covid-19 patients. Probabilities were estimated for univariate and bivariate distributions of covid-19 cases over various ages and sexual orientations in tertiary hospitals. The quantitative factors were communicated as mean and standard deviation, estimates were communicated as 95% certainty stretches, and correlation was performed utilizing a two-followed Student t-test.

Jerry Wei et al.,[2] have framed a QA classification framework for domain-specific COVID-Q model evaluation developed for classifying questions, grouping into clusters. In the data model when a new question is raised, the system retrieves the questions and the appropriate answers. The similar questions in a cluster are identified and are provided with the same answer. Annotation is performed based on the distribution of question category, clustering, and source of information for covid-specific models. The scope of the paper is limited to only less number of answers as solutions are required for the same question.

Gabriele Pergola et al., [3] have mentioned that high-quality biomedical QA datasets are a scarce resource approach to incorporate external knowledge into re-trained language models. The focus is to fine-tune the model with special characteristics for domain-specific language models. The model employs an entity-aware masking strategy to enhance the model to train the biomedical database to learn entity-centric data. The strategy substitutes a predefined set of phrases with a unique token where the model predicts appropriately the equivalent proportion of words to predict the QA tasks. The results were advantageous for biomedical QA tasks on various metrics.

Rohan Bhambhoria et al., [4] have developed a rule-based model QA framework in semi-structured medical data creating a QA pairs dataset with answers. An active learning strategy is engaged where the method selects the fewer number of samples that reduces annotation for a huge QA system. Additionally, a feedback system is provided from the subject experts. QA system is developed for text content that pairs titles with Do/Does and Is/Are pre-appended words. The system employs Medical Experts from the medical community and extends a web application to provide an efficient user interface for annotating the QA pairs generated.

Adam Poliak et al., [5] have implemented a Natural Language Processing (NLP) related chatbot application focused on thorough COVID-19 data and misinformation. Users associated with the chatbot on various stages to receive data about COVID-19, doubts, clarifications, and different subject matter of interest. The system provides authentic data as confirmed questions and answers to render answers for frequently answered questions about the pandemic. Three fundamental methods were considered

1) Producing top caliber and precise data from domain experts,

2) Repeated questions and answers that were posted frequently from online trusted sources such as official medical concerns,

3) Automatically positioning Top questions from social websites and including them in the existing QA pairs dataset.

Revanth Gangi Reddy et al., [6] have explored the QA pairs which are synthetically generated for improving the significance of information retrieval in a closed domain system. End to end QA system for covid-19 databases has been developed combining Informational retrieval (IR) and Machine Learning techniques. The systems provide the extracted passages in par with the natural language question received as query, after analyzing with a
meaning similar to related texts from a large collection of information in websites. The QA pairs are used to train the machine learning target domain to be further improved for comparison.

2.1 Deep Learning and Machine Learning models for QA pairs

Sultan et al., [7] have shown that sampling-based question creation prompts improved MRC training when compared to greedy or beam search decoding as the sample diversity is increased. A top k sampler generation is adopted where the question-answer pairs are considered as a whole rather than questions alone.

Z. Yang et al., [8] proposed a hierarchical network method for text classification. The model comprises of two attributes: (1) a progressive design that reflects the hierarchical construction of documents, and (2) two degrees of attention mechanism applied at the phrase and sentence level, empowering it to attend and enable more and less significant content while developing the document representation.

X. Zhou et al., [8] have extended the hierarchical attention model to cross-lingual sentiment classification. In each language, an LSTM network is used to model the documents. Then, classification is achieved by using a hierarchical attention mechanism, where the sentence-level attention model learns which sentences of a document are more important for determining the overall sentiment, while the word-level attention model learns which words in each sentence are decisive.

Usha et al.,[9] proposed a standardized calculation-based IDS for MAC Intrusions (NMI) to improve the IDS execution significantly. The proposed NMI incorporates two essential parts OFSNP and DCMI. The primary part is the ideal element choice utilizing NG and PSO (OFSNP) and the subsequent segment is Detecting and Categorizing MAC 802.11 Intrusions (DCMI) utilizing SVM classifier. The SSC depends on particle swarm optimization (PSO) that utilizes named and unlabeled highlights at the same time to discover a group of optimal features. Utilizing the optimal arrangement of features, the proposed DCMI uses a fast support vector machine (SVM) discovering that classes. The results show that the NMI precisely identifies and orders the explicit interruptions and it lessens the false positives and calculation by diminishing the number of features.

Weston et al., [9] developed a memory network for a handcrafted QA task in which statements are added as evidence to the question. The system retrieves one question based on the input and previously received statements.

W. Hamilton et al., [10] suggested an alternative for reducing the memory consumption of text corpus. Instead of constructing Graph Neural Network for the entire corpus, a sliding window defines separate GNN for each text during the training period. The method outperforms better results by 5% than the existing works.

A. Das et al., [11] proposed a Question Answering Model for Siamese Language, where the system, for each question it checks the semantic relation between the question and the answer pairs. A convolution Network trained to maximize the semantic similarity and the computational complexity is reduced by adapting the character level representation for each question and candidate answers.

M. Tan et al., [12] have presented a hybrid model that has utilized CNN and attention neural network for building Siamese language model on Question-answer pairs for a non-factoid type of questions and answers[13]. LSTM network is added for memory constraints. Based on the observations from the existing works Evidence-based Literature Mining systems have been a limitation widespread with question answering models that need to revolutionize search altogether. A language model has been proposed with NLP has been for detecting false claims and presenting evidence about the pandemic information's which is the need of the hour.

Sathiymoorthi et al.,[14] proposed web mining strategy that employs Web usage mining as a method of improving Web caching policies. The size, expense, frequency, maturing, cache entry time, and popularity of Web objects were factors in research. Furthermore, experimental findings show that the new policy outperforms current policies in terms of execution metrics such as hit rate and byte hit rate.

www.turkphysiotherrehabil.org
III. RESEARCH GAP

More research questions need to be addressed during havoc situations like a pandemic.

1. How can users make the distinction between the collected corpora in search of appropriate information?
2. What are the apt IR tasks like question-answering for the disaster sort of occasion?
3. What would be the powerful techniques for forecasting information based on the particular necessities of the circumstance?
4. Can the existing information be utilized with the help of machine learning to improve the search engine results?
5. Further, how probably is it that web search tools can contradict to justify a quantitative comparison during an emergency?

These questions are addressed in this initiative research which presents an effective combined performance of Question answers received from user reviews and the additional forecasting data to address in the future.

IV. PROPOSED MODEL

Information retrieval and communication innovation help in the dynamic decision-making that is solely dependent on the past information that predicts future initiatives. The size of information available for end-users is enormous. Coronavirus 2019 has been seen as an overall pandemic and a couple of analyses are being carried out using diverse mathematical models to expect the reasonable progress of this plague. The challenge lies in predicting patterns and the information with appropriate evidence is yet again a difficult chore. The predominant information about tested cases, confirmed, hospitalized, recovered and deaths across geographical for a time duration paves the way to forecast the near future.

A complete informative system depends on the accuracy of the model relates to attributes like social media, health care centers, electronic medical records, hospital, and general guidelines issued by the government. Additionally, it includes essential circumstance prevailing to prevent the transmission. AI (ML) based forecasting systems have confirmed their importance to foresee the progress and deterioration of the verdict on the future course of pandemic activities. Several prediction strategic methods are prevalently in effect to deal with forecasting issues. This study exhibits the ability of ML models through Extractive Question Answering to forecast the pandemic activities.

4.1 X-tract QA

Information Retrieval systems that have been developed are designed to retrieve results only either list of documents that possess match with the user query or extract documents related to user input. To address these issues results use a novel technique X-tractive QA similar to Semantic Text Search, which embeds both the search term and the documents in vector space, allowing them to be matched based on purpose rather than keywords. The embedding method differs in the extraction of the answer focus is on the right documents, not the keywords.

4.2 FORECASTING

Forecasting of time series data depends on current and historical information. It employs various types of data such as COVID-19 data, demographic data, mobility data techniques, and estimations to help in anticipating the number of deaths, hospitalizations, and cases. These forecasts are created independently and made available to the public. It's critical to combine these forecasts to better understand the uncertainty. Various techniques, including modeling, statistics, data mining, artificial intelligence (AI), and machine learning, are used to analyze data from the past or present to anticipate future trends.

V. SYSTEM ARCHITECTURE
The proposed model combines the QA and the forecasting module. Related to the various questions and the answers the prediction model summarizes the forecasted data. The framework Refreshes its database every half hour with recently updated QA and statistics obtained from various Covid information sources. The framework consists of

5.1 QA generator

User query/keyword is given as input and QA pairs related to the query are collected from various health-related corpus and search engines. Instead of creating new question-answer pairs, similar questions paraphrased are integrated under the roof. The reason behind integrating from reviews is that, in terms of topic and document length, review sites differ significantly from Q&A data. Additionally, in many Q&A pairs, sets include responses that are oriented toward the query aligned due to lexical and syntactic similarities.

5.2 Evident QA snippets Selector

The evident Snippet Selector generates a list of the most relevant response samples from the collected documents, with significant keywords highlighted. Given the queries, a neural QA system is constructed as an evidence selector to effectively discover the fragments of paragraphs related to a query. Given relevant paragraphs and queries, QA seeks to forecast responses or evidence. The Question answers pairs collected from health care reviews and government Covid information centers are clustered which resemble the similar answers. Those clustered QA pairs are further classified into three levels based on the severity of symptoms mentioned in the pairs as, Low, moderate and severe.

5.3 Forecast Model

The prediction and forecast model intend to include features such as symptoms level and further symptoms based on the severity level of the symptoms obtained from the question pairs. The forecasting machine learning model incorporates features for patient past data, the total number of affected patients, their behavior in the locality, health centers in and outflow of Covid patients, and the availability of resources is to be implemented. The time series model focuses on different predictions in the date-wise and location-wise views. This dynamical data is time-series data that depends on the real-world factors, policies, and decisions undertaken by the country.

VI. CONCLUSION

Since the pandemic prevailing is alarming and dependent on various information that is interrelated. With this notion, A Xtract QA model is proposed that exhibits the ability of ML models to forecast the pandemic activities
in a real-world semantic search engine. It returns snippets containing evidence retrieval and fact-checking for the questions related to questions and claims about the COVID-19 pandemic information feeds published in health care recommendation systems, social media, and World Wide Web. Additionally, the system predicts the data required for the QA pairs in terms of real-time data which is essential to keep updated about the near future. The model has been designed for the QA generator and all other modules that are theoretically designed are in progress. The results are expected to be effective significance for getting aware of the doubts and clarifications through question answer pairs and the future related data in need.

REFERENCES