IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Evidence-based Recommender System for a COVID-19 Publication Analytics Service

ROLAND ORUCHE, VIDYA GUNDLAPALLI, ADITYA P. BISWAL, PRASAD CALYAM, MAURO LEMUS ALARCON, YUANXUN ZHANG, NAGA RAMYA BHAMIDIPATI, ABHIRAM MALLADI, HARIHARAN REGUNATH
University of Missouri, Columbia, MO 65211 USA
(e-mail: {rro2q2, lemusm, yzd3b, nbny6}@umsystem.edu;
{vgundlapalli2021, adityabiswal2003, amalladi017}@gmail.com;
calyamp@missouri.edu; regunanthh@health.missouri.edu)

Corresponding author: Prasad Calyam (e-mail: calyamp@missouri.edu)

ABSTRACT The rapid growth of COVID-19 publications has driven clinical researchers and healthcare professionals in pursuit to reduce the knowledge gap on reliable information for effective pandemic solutions. The manual task of retrieving high-quality publications based on the evidence pyramid levels, however, presents a major bottleneck in researchers' workflows. In this paper, we propose an "evidence-based" recommender system namely, *KnowCOVID-19* that utilizes an edge computing service to integrate recommender modules for data analytics using end-user thin-clients. The edge computing service features chatbot-based web interface that handles a given COVID-19 publication dataset using two recommender system modules: (i) *evidence-based filtering* that observes domain specific topics across the literature and classifies the filtered information according to a clinical category, and (ii) *social filtering* that allows diverse experts with similar objectives to collaborate via a "social plane" to jointly find answers to critical clinical questions to fight the pandemic. We compare the Domain-specific Topic Model (DSTM) used in our evidence-based filtering with state-of-the-art models considering the CORD-19 dataset (a COVID-19 publication archive) and show improved generalization effectiveness as well as knowledge pattern query effectiveness. In addition, we conduct a comparison study between a manual literature review process and the KnowCOVID-19 augmented process, and evaluate the benefits of our information retrieval techniques over important queries provided by COVID-19 clinical experts.

INDEX TERMS COVID-19 Publication Analytics, Literature Review Automation, Machine Learning, Recommender System, Social Networking

I. INTRODUCTION

WITH the impact of the COVID-19 causing a major societal crisis, scientific research has been crucial in terms of clinical researchers and healthcare professionals accessing publicly available medical journal databases (e.g., PubMed [1], LitCovid [2]) to perform knowledge discovery. The ability to combat pandemic-related problems through accessible literature archives can be accomplished using thin-clients such as web browsers. Such an approach allows real-time access to data resources and analysis tools that are hosted on a remote server versus having users download data and install tools to a localized hard drive to perform analysis. It also allows clinical researchers and healthcare professionals (i.e., data consumers) to pursue bold scientific research tasks. However, the mass production of related publication resources causes challenges to data consumers who seek out high-quality data and want to perform drill down analysis of facts in order to find novel insights for understanding of problems and to perform collaborative decision making on solutions. Handling continuously-growing publication databases at the researchers' disposal today is performed manually, which makes it onerous and time consuming to filter out high-quality data in literature archives.

Current online medical databases have found ways to reduce the time consumption of sorting information in the medical journals. For example, Evidence Matters [4] creates

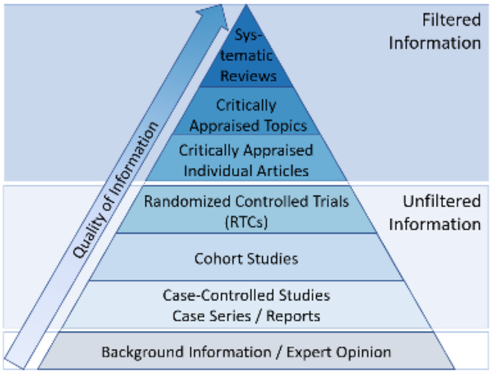

Filtered Information
Sys-tematic Reviews
Critically Appraised Topics
Critically Appraised Individual Articles
Randomized Controlled Trials (RTCs)
Unfiltered Information
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion
Quality of Information

FIGURE 1. Levels of the Evidence Pyramid [3] showing filtered and unfiltered information to help visualize both the quality of evidence and the amount of evidence available.

a clinical knowledge management system that integrates a queried user interface to retrieve latest peer-reviewed articles. The results from each article are synthesized into summaries, tables and other visuals including graphs, which provides data consumers with an "answers-first-references-second" approach. Data consumers have even sought to rely on scholarly search engines such as Google Scholar (https://scholar.google.com) and similar services like Microsoft Academic (https://academic.microsoft.com/), DeepAI (https://deepai.org/), Semantic Scholar (https://www.semanticscholar.org/) for finding relevant COVID-19 literature to propel their research objects in obtaining actionable insights. One notable platform in the COVID-19 research context, IBM Deep Search [5], leverages artificial intelligence (AI) and machine learning (ML) models to help researchers access structured and unstructured data rapidly. These models allow users to make specified queries to the collections of papers and helps them to extract critical COVID-19 related knowledge.

The above online database management systems and search engines use automated literature review and information retrieval methodologies to guide data consumers, respectively. However, these applications lack the ability to filter articles in accordance with the Levels of Evidence Pyramid [3] shown in Fig. 1. Healthcare practitioners often evaluate the quality of papers using such a hierarchical evidence structure to follow proper evidence-based practice standards. The Levels of Evidence Pyramid illustrate the inherent reduced quantity of publications with respect to the increase in high quality information (e.g., expert opinions to systematic reviews and meta-analyses). To the best of our knowledge, there are no current literature-mining web services that automates data consumers' workflows of categorizing papers according to this hierarchical evidence-pyramid structure, and allows collaborative knowledge sharing amongst the scientific data consumer community.

To overcome these limitations, we propose a novel "evidence-based" recommender system namely *KnowCOVID-19* for clinical researchers, healthcare professionals to automatically filter high-quality publications according to the evidence-based information filtering standard. KnowCOID-19 design of using a novel AI-based literature review service aims to reduce the manual burden in traditional literature review workflows of data consumers who laboriously search and filter continuously-growing publication databases for high quality information. Specifically, our approach integrates an edge computing layer to provide low-latency and direct-user consumption of publications analytics that augments the workflows of data consumers through two recommender modules. The first recommender module features *evidence-based filtering*, which observes latent knowledge patterns across the literature documents using a Domain-Specific Topic Model (DSTM) [6] that generates a Dirichlet probability distribution set of keywords in each topic. In addition, we extend the topic model through our category model to classify documents according to their clinical topic category using the Latent Dirichlet allocation (LDA) [7].

The second recommender module features *social filtering* adapts HumHub [8] (a purpose-driven social network framework), which allows for sharing filtered information insights in the form of distribution graphs, visual charts, and table formats on article information and expert user ratings. Social filtering also enables medical experts with similar objectives to find relevant collaborators in their peer communities to collectively analyze high-quality evidence to obtain crucial insights. Such an analysis for data-driven pandemic decision making and crowdsourcing via our "social-plane" allows the medical community to publish/subscribe information events to answer important clinical questions to combat the pandemic. Our edge computing service model also features a context-aware supported chatbot interface, namely *Vidura Advisor* that provides guidance to data consumers for easy navigation through our recommender module functionalities.

We validate the utility of our KnowCOVID-19 system using a collection of 10,0000 articles from the Kaggle COVID-19 Open Research Dataset (CORD-19) [9]. We conduct a two-fold case study to evaluate: (i) the effectiveness of our DSTM [6] in terms of the perplexity metric in our evidence-based filtering against state-of-the-art models to show an improved generalization performance as well as knowledge pattern query effectiveness considering the distribution of unobserved topics (e.g., drugs and genes) within the CORD-19 dataset, and (ii) the benefits of our information retrieval techniques using Term Frequency-Inverse Document Frequency (TF-IDF) [10] over clinical queries provided my expert clinicians and immunologists. In our comparison study, we detail and summarize the manual analysis in which a team of clinicians in [11] evaluate the relevance and high-quality of publications towards their fundamental research queries. We then compare the manual baseline analysis with our proposed automated analysis of journal archives according to the Levels of Evidence Pyramid. Specifically, we process their clinical queries via information retrieval techniques and filter out high quality information that significantly augments their clinical workflow in their literature review and fosters insights discovery towards evidence-based practice.

The rest of the paper is described as follows: Section II discusses related works. Section III describes the workflows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083583, IEEE Access

and requirements of data consumers when conducting literature review in accordance with the Levels of Evidence. Section IV presents an overview of our KnowCOVID-19 evidence-based recommender system along with our edge computing service model and the Vidura chatbot implementation. In Section V, we detail our evidence-based and social filtering recommender modules to automate the literature review process of data consumers. Section VI describes two case studies of generalization effectiveness and the comparison between manual and KnowCOVID-19 augmented analysis. In Section VII, we present a discussion on the latest status of KnowCOVID-19 deployment along with areas being addressed for further improvement of capabilities as well as for wider user adoption. Finally, Section VIII concludes the paper.

II. RELATED WORKS

A. AUTOMATION OF LITERATURE REVIEW

Several works have developed recommender systems to meet the demands of scientists seeking out relevant information in continuous growing archives [12] [13]. Science Concierge [12] addressed the problems of discerning scientific papers with fine-grained topics by developing a responsive, content-based recommendation system for literature search. Authors created a vector representation of the literature articles and utilized their content-based filtering approach to suggest articles based on user votes. Fast^2 [13] aimed to reduce the labor intensive task researchers face when performing a *systematic literature review* in evidence-based software engineering. The software tool implements a fast and robust start tactic for literature search, a stop rule to indicate when most papers have been found, and an error detection strategy that mitigates human error over literature archives.

In recent work, multiple groups such as [14] [15] and [16] have evaluated their models over the COVID-19 Open Research Dataset (CORD-19) [9] to help bridge the gap between researchers and the rapid growth of journal publications. In [14], authors test the efficacy of a graph-based clustering model and Bio-BERT [15] word embeddings approach for information retrieval through a Question-Answer bot related to clinical queries. CovEX [16] attempts to develop an exploratory search system that implements a content-based recommender approach using knowledge graphs and keyphrase extraction to support transparency and controllability for end-users. Other salient work that used CORD-19 dataset can be seen in [17], where a text-to-text, multi-loss training strategy was used to fine-tune pre-trained language models to perform abstractive summarization and visual inspection in domain-specific corpora. Authors in [18] use parallel one-class support vector machines which are trained on clusters of similar articles to extract and suggest the quality trends of scholarly material related to the pandemic using the CORD-19 dataset.

One notable service that attempts to help data consumers to mine literature on current pandemic-related problems is the IBM Deep Search [5]. IBM Deep Search leverages advanced machine learning models to extract content from ingested documents (e.g., Corprus Conversion Service) with high accuracy. They use their Corpus Processing Service to apply natural language processing tasks such as Named-Entity Recognition and Fact Extraction to build knowledge graphs for retrieving contextual facts and performing knowledge discovery. These graphs provide a wide range of query templates to query useful information from the graph such as *Which drugs have been used so far?* and *What are the outcomes?* Trained with supervised learning models, they focus on two types of extraction tools: (i) information that is fundamental to all literature reviews such as study design and sample size, and (ii) a range of information with shared characteristics with minimum filtering.

While these works present promising recommendation and information retrieval outcomes for scholarly articles, the foundations of these approaches lack the ability to incorporate clinical context (e.g., drugs and genes) in extracting high-quality topics pertinent to evidence-based practice standards. In an effort to show the latent relation between drug and gene topics and related literature documents, we leverage our Domain-specific Topic Model (DSTM) as a part of the KnowCOVID-19 evidence-based filtering module to filter COVID-19 publication archives and provide recommendations to foster insightful knowledge discovery.

B. AUGMENTED ANALYSIS FOR COVID-19 RESEARCH

Previous studies have addressed the critical need of healthcare professionals and end-user device clients for data analytics using edge-computing systems that provide low-latency in data access. A study in [19] aimed to combat the spread of COVID-19, which is primarily caused by human-to-human chain, by creating a smart surveillance system for effective remote monitoring of human health conditions and close interactions. Artificial intelligence-based analytics and end-user devices such as smart wearable gadgets are used to develop a framework to map the communication chain of infectious victims. The proposed framework introduces a multiple-edge layer that minimizes the latency of wearable gadgets used to process vital signs (e.g., body temperature, heart rate) and reduces the response delay of real-sensor triggering time. The authors in [20] use a similar approach of developing a mass surveillance system by integration of edge computing through a 5G wireless connectivity network. They leveraged a hierarchical edge computing system to analyze and detect raw/imagery data of potential infectious disease patients using deep learning (DL) algorithms. Their proposed framework enables edge caches to store the DL parameters from the cloud layer to provide stakeholders and beneficiaries on the user layer, fast access for health monitoring and critical decision making.

Other works address the benefits and challenges of using DL algorithms over heterogeneous edge devices that assist the mitigation of the coronavirus. A survey in [21] suggests an effective solution to overcome the limitations of insufficient COVID-19 data for deep learning models and low com-

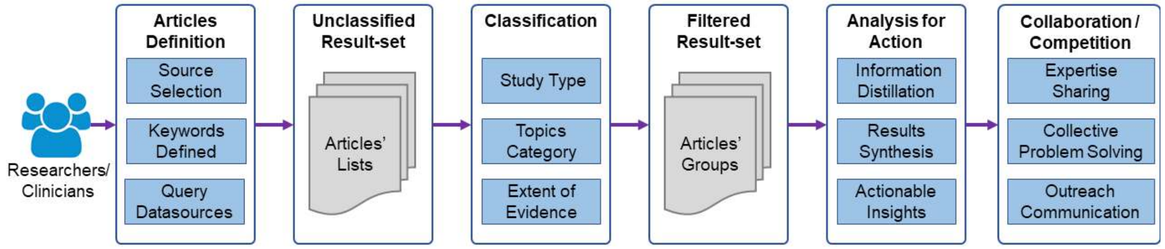


FIGURE 2. Researchers workflow during a Literature Search Process to filter information for expertise sharing

and decision making.

putational resources of edge computing through Deep Transfer Learning (DTL). This concept uses the knowledge from a previously learned model and applies it to a new model with minimal re-training or fine-tuning. While DTL over edge computing can be exploited to resolve the challenges of training over emerging datasets, very few works have proven their effectiveness for delivering instantaneous analyses for healthcare workflows. One particular study implements a DL learning model with edge computing to develop an end-to-end, mobile clinical-decision-support-system for constant patient monitoring [22]. This DL-assisted communication framework utilizes transfer learning models over X-Ray images to help novice/intermediate level physicians detect COVID-19 cases automatically. In addition, the study integrates an edge network to transfer computational resources near the edge where it is accessible for mobile device clients. Thus, the system provides an accurate detection analysis to patients and hospitals for remote tracking and real-time monitoring.

In this work of KnowCOVID-19, we integrate edge computing for developing a recommender system that adapts our trained models to provide instant access to medical journals via end-user thin-clients. The uniqueness of our approach is in our approach to utilize edge computing to provide users with pre-trained recommender modules on publication analytics for identifying latent knowledge patterns across large document collections. Furthermore, our approach has the potential to help data consumers to perform a drill-down analysis of COVID-19 literature using best practices from intelligent dashboard works such as [23] [24]. Such an approach will enable them to filter high-quality information and conduct trustworthy collaborations with relevant experts for effective clinical decision making in pandemic related response efforts.

C. EXPERTISE SHARING AND CROWDSOURCING

Social networks have played an important role for Internet users to share/express their ideas, provide expert services and connect with online communities. As more social networks evolve into purpose-driven communities, community members (e.g., scientists and researchers) advantageously crowdsource information for improving the knowledge discovery process. IntelliSearch [25] uses big data analytics integrated with crowdsourcing for developing a search engine that provides a standard ranking system for websites. The core of the system relies on active online community discussions that help refine the data and facilitate their expertise on high-quality information. The work in [26] integrates a social network that addresses the difficulty in knowledge sharing among experts to enable proper network performance expectation management. The lack of a social platform for sharing knowledge and working collaboratively makes the task of isolating and diagnosing network bottleneck events rapidly cumbersome. They defined a "social plane" that relies on recommended measurements based on content-based and collaborative filtering approaches. Based on similarity analysis, the content-based filtering facilitates network operators and other users to subscribe to useful measurements through the social plane, and the collaborative filtering promotes users to share knowledge on anomaly event symptoms to resolve bottlenecks.

The efforts from the Kaggle COVID-19 Literature Organization [9] to develop the CORD-19 dataset aim to resolve the limitations that scientists face when handling the ever-growing COVID-19 literature. Kaggle has coordinated with prestigious institutions and research companies to facilitate data analytics challenges. This data challenge platform enables medical researchers and data scientists to share/open source their findings. Consequently, this allows Kaggle subscribers in the scientific community to crowdsource information for performing knowledge discovery. Other online platforms such as CognitiveCity [27] build upon a graph-based architecture at their core, allowing stakeholders to query the data generated from a living systematic review. The authors in [27] use a social network to create a knowledge graph that connects members based on their research pursuits and interests. Their main purpose is to share knowledge for the users to explore, contribute and connect to the information they require. This is a single place to catalog the tools, datasets, analyses, and articles being generated by the global community about COVID-19, and facilitate network analytics that can drive research collaborations.

Inspired by the above works, our goal is to provide a social plane for the KnowCOVID-19 users to interact and remain connected for solving pandemic-related problems. Our novel social filtering technique has the potential to help users to share relevant information with other users to find out similar problems which facilitates discussion and collaborative mingling. We present the implementation of our KnowCOVID-19 social filtering in Section V, which uses an open-source social platform namely HumHub [8] that can be customized for launching purpose-specific social networks. The social network integration in KnowCOVID-19 provides novel features such as providing notifications and ability to include comments to any of the questions posted by COVID-19 experts.

III. WORKFLOW USING A LITERATURE REVIEW

Clinical researchers and healthcare professionals spend a significant time and effort in identifying and understanding the hierarchical organization of the evidence found in the medical literature hosted in databases such as PubMed, Scopus, MedRxiv or BioRxiv. Search engines based on keywords are a common tool for these data consumers to search, filter and analyze information that is relevant to their clinical queries and/or research objectives. Fig. 2 shows the workflow of a typical literature search process to filter publication collections for conclusions for information sharing and decision making. In the following, we detail the workflow and then summarize the requirements for performing literature review augmented with KnowCOVID-19.

A. WORKFLOW PROCESS AROUND THE EVIDENCE PYRAMID

The workflow begins by a data consumer e.g., a researcher identifying the key aspects to filter the literature they want to read in detail. First, the data consumer selects an article archive and uses a search engine functionality provided by the archive service provider to conduct the search and filtering. Specific keywords related to the topic of interest (e.g., PCR and SARS CoV 2) are used to filter down the publications to be analyzed in-depth. The filtered list of articles is then organized using preferences of spatial and temporal distributions (e.g., authors' institution, date of the published article, publication venue). Manual classification of the listed articles is then performed based on the study type, the topics category and the Levels of Evidence Pyramid in order to use as the final information source for analysis and discussion with peers.

The final output is a set of articles grouped under clinical categories to facilitate analysis for action which includes: information distillation, results synthesis and actionable insights. The analysis can include aspects of specific studies related to: methodology, sensitivity and specificity rates, advantages/disadvantages of approaches, turnaround times for the variables studied and related results. The methodology used to conduct studies help clinicians to categorize articles within the IV, V and VI Levels of Evidence Pyramid. The sensitivity and specificity rates allow clinicians to include papers which results related to these aspects fall within the valid threshold they defined for the study. The analysis step ultimately helps to create a social community for information sharing and clinical decision making amongst experts to solve focused problems. It can also help in the pursuit of strategic outreach communication to influence public policy and clinical best practices.

B. SYSTEM REQUIREMENTS TO AUGMENT WORKFLOWS

Given the complexity of the current manual process used by data consumers as shown in Fig.2, augmenting tools that help to automate and scale the workflow are necessary and valuable. Especially, when handling large collections of publications, such tools can greatly lessen the burden of finding high-quality information and also can improve the ability to correlate information across multiple articles necessary to complete a research task. Further, such tools can help provide new insights for knowledge discovery based on evidence-based information filters to improve the effectiveness of decisive actions for pandemic solutions. To create the use-inspired publication analytics in our KnowCOVID-19 system, we outline the following requirements based on the needs of clinical researchers and healthcare professionals to improve the efficiency and effectiveness of the workflow tasks.

1) Flexible and Scalable Analytics Architecture: Such a system architecture can allow for a seamless infrastructure expansion, increase the service capacity, ease the use of existing functionality, allow addition of new capabilities, and maintain service availability. Using a core cloud component in the architecture can help with the handling of heavy user loads when accessing massive publication databases and for running search and filter tasks robustly. The core cloud component can be extended with an edge computing component, which is accessible in closer locations to users (vs. cloud computing platforms). Such a component can provide low-latency drill-down analysis in the collaborative tasks and search result analyses. Moreover, it can reduce bandwidth requirements to manage high volume data management involved while following the workflow in Fig. III. Lastly, the system's front-end interface needs to be conveniently deployed at the user's end-user thin-clients to leverage the benefits provided by the core cloud and edge computing components.

2) Data Source and Timeline Management: The system should provide multiple options and allow the users to select the sources of articles (e.g., CDC, NIH, medRxiv) that are relevant for their research. Besides having the ability to choose the source of the publications, the user should be able to narrow the search by indicating the publication date window. In this way, the user will have the ability to manage a large number of short-term new publications with content relevant to the novel COVID-19, and correlate them with a long-term of historic publications with content related to previous infectious disease studies (e.g., 2012-13 MERS, 2002-03 SARS) that are related in terms of the epidemics.

3) Multiple Search and Drill-down Options: The system should allow users to apply a text mining approach that suggests publications, tools and other medical information resources in accordance with the Levels of Evidence Pyramid. Automation of the publication analytics in the literature review workflow can effectively guide users to narrow down their articles-of-interest list by selecting topics (e.g., drugs and genes). Particularly, they can filter the high-quality information to a specific clinical category (e.g., treatment, prevention, diagnosis) pertinent to their clinical queries.

When data consumers obtain actionable insights, the system should allow capture of expert feedback and help experts connect with other medical community members facing similar problems or clinical queries. Using such an approach, key insights can be shared amongst various data consumers, particularly experts to combat rapid knowledge discovery problems related to responding to the on-going pandemic crisis.

4) User-intent based Analysis/Visualization: To increase the usability, the analytic tools need to have capabilities that help them make topic associations, find similarities and visualize and share results. To meet this requirement, the system should allow users to: (a) define and manage the size of the result-set required for analysis, (b) present the result in a tabular structure including multiple metadata parameters (e.g., title of the paper, source, publication date, Levels of Evidence Pyramid, clinical category), and (c) let them to control the content's presentation by showing/hiding the parameters and filter records based on the information of parameters being selected. Users will also need convenient access to extended information about articles such as the abstract section content, the entire document retrieval from the source, or the ability to tag documents as favorites for easy future reference. Lastly, the analytic tool-set should also provide the functionality to mine knowledge patterns and correlate information of recent COVID-19 literature to previous literature on pandemics (e.g., 2012-13 MERS, 2002-03 SARS) to potentially find similarities in best practices to create innovative solutions.

5) Collaboration Tools to Crowdsource Expertise: In the midst of the dynamic situation like the current on-going COVID-19 pandemic, effective collaboration tools are needed to create a social plane for the relevant communities. The social plane can help experts to share their findings from literature reviews and can allow them to collaborate with other experts on purposeful projects to answer critical clinical questions. The system should guide experts to discern useful information from the vast and fast growing articles in journal archives by reducing the manual burden in related tasks. Moreover, the system should also help in bringing the focus of the medical community in analyzing the most trending content that is relevant to cope with various clinical research questions that need to be answered to combat the on-going pandemic response.

IV. KNOWCOVID-19 SYSTEM OVERVIEW

A. SYSTEM COMPONENTS

In this section, we detail the system components of KnowCOVID-19 shown in Fig. 3 that help in the AI-based publication analytics to help thin-client device users (i.e., clinical researchers, healthcare professionals) in sorting through filtered, peer-reviewed documents according to the Levels of Evidence Pyramid. The flexible and scalable analytics architecture presented in Fig. 3 is comprised of three layers: the cloud layer for storing and literature archives e.g., the CORD-19 Kaggle dataset, the edge computing layer that processes and trains publications and paper-tagging from the filtered knowledge bases, and the application layer that allows end-user thin clients to access the KnowCOVID-19 user interface, social plane capabilities and our Vidura Chatbot service. In response to augmenting the manual baseline workflow of clinicians of sorting through a list of articles, Fig. 4 shows an improved automated workflow that guides clinical users through a series of steps to find and filter publications, rate and tag publications based on their clinical queries, as well as to share their salient findings to an online science community in a social network setting.

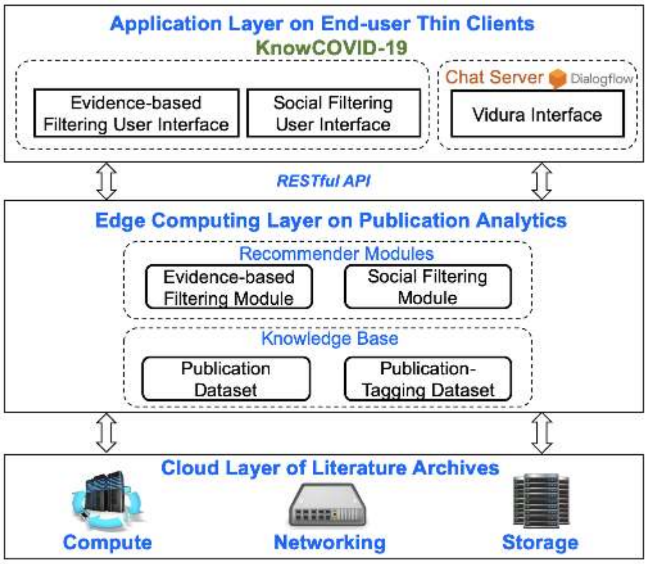

Application Layer on End-user Thin Clients
KnowCOVID-19
Evidence-based Filtering User Interface
Social Filtering User Interface
Chat Server
Dialogflow
Vidura Interface
RESTful API
Edge Computing Layer on Publication Analytics
Recommender Modules
Evidence-based Filtering Module
Social Filtering Module
Knowledge Base
Publication Dataset
Publication-Tagging Dataset
Cloud Layer of Literature Archives
Compute
Networking
Storage

FIGURE 3. KnowCOVID-19 system architecture showing the cloud layer, edge computing layer and the application layer interaction abstractions.

Cloud Layer: The cloud infrastructure consists of physical and virtual resources necessary to support networking, computing, and database storage over the 10,000 articles we have sourced from CORD-19 for services deployment. The CORD-19 dataset provides data consumers access to reliable articles (e.g., CDC, NIH, medRxiv) that are relevant to their research. We convert the peer-reviewed articles information into JSON files and store the contents including the metadata (e.g., title of the paper, source, publication date, Levels of Evidence Pyramid, clinical category) into a cloud storage instance to enable querying from the KnowCOVID-19 user interface. Users have access to valuable metadata including the publication dates of each medical publication for timeline management and for further analysis of the relation between the novelty of COVID-19 and previous infectious diseases (e.g., 2012-13 MERS, 2002-03 SARS). In addition, we have sourced other salient medical information from the *COVID-19 Vaccine Tracker* [28] and stored a set of medical terms (e.g., drugs and genes) in the Cloud Layer. This enables data consumers to perform publication analysis through analysis of the relation between the medical terms and the CORD-19

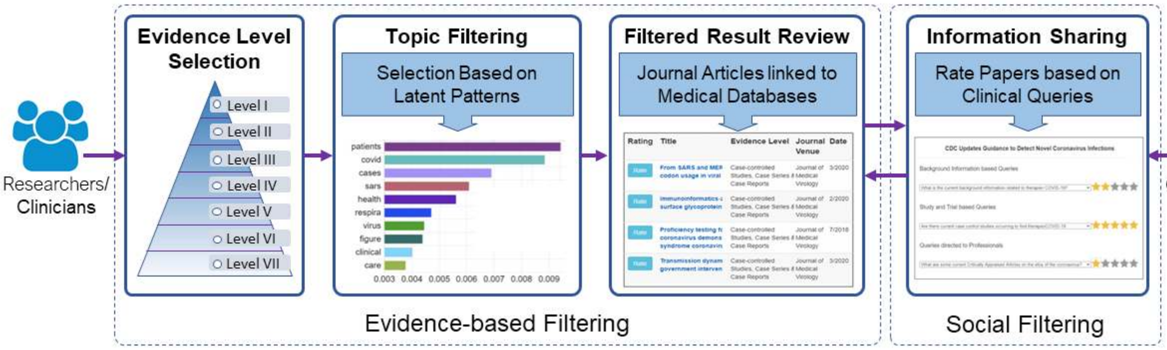


FIGURE 4. Augmented workflow steps in the KnowCOVID-19 user interface that are followed by Researchers/Clinicians to perform publication analytics to find answers to important scientific questions related to fighting the COVID-19 pandemic.

dataset using ML-based pipelines on cloud resources.

Edge Computing Layer: In facilitating workflow needs and data-intensive resources for effective publication analytics, we have implemented an edge computing configuration that filters documents and provides high-quality information for end-user consumption. Rather than relying on data-centers at remote sites for publication analytics, the edge computing layer pre-trains over stored resources from the cloud layer to provide users with reduced latency and real-time information processing. The edge layer for publication analytics stores a knowledge base on COVID-19 articles and supports user paper-tagging (e.g., publication ratings). We handle the critical need to automate the literature review for data consumers to perform multiple search and drill-down options through our *evidence-based filtering* module. Our evidence-based approach differs from traditional recommendation methods (e.g., content-based filtering, knowledge-based) as it creates a packaged analysis in accordance with the Levels of Evidence Pyramid in a manner that data consumers can further filter information for performing knowledge discovery. We also provide a set of clinical queries commonly asked in the medical community around COVID-19 for end-users to indicate how effective the articles are in answering their queries. In addition, we implement *social filtering* that utilizes collaborative filtering in publication analytics to allow users to leverage their actionable insights and share critical insights via a social plane.

Application Layer: Data consumers with end-user thin client devices (e.g., mobile phones, tablets, laptops) can access the pre-trained publication analytics modules via RESTful Application Programming Interfaces (APIs). The application layer provides a user-friendly interface with guided navigation for users to get actionable results to later filter articles, find actionable insights and/or develop trustworthy collaborations through our social filtering module. We address the need for data consumers to keep up with the latest publication trends through our KnowCOVID-19 UI that allows users to connect with each other via a social network platform built using HumHub [8]. This platform connects experts in COVID-19 within a purpose-driven community that collaborates to find solutions to help with the response to the on-going pandemic. The essential benefit of the application layer is in its ability to hide the complexities of automation of processing literature resources and enhances the workflow process user experience to enable users to more easily and more rapidly sort through high-quality publications based on Levels of Evidence Pyramid. In addition, we implement a Vidura Chatbot Assistant interface that guides users through the Google Dialogflow by pre-training on intents/queries to generate stable recommendations.

B. VIDURA CHATBOT ASSISTANT

Herein, we detail the working of the Vidura Chatbot Assistant to guide data consumers in easy navigation of the KnowCOVID-19 system functionality. Our Vidura implementation, adopted from [29], uses Google Dialogflow which is a conversational intelligence service that uses natural language processing techniques to find the intentions in the user queries. Through Google Dialogflow, we are able to develop intelligent and responsive agent interactions in our Vidura implementation that hold definitions for various functions involving 'intents', and 'fulfillments' based on user input. Vidura is trained on various intent scenarios to match the users' expressions provided by a rule-based grammar capable of restructuring, then categorizing syntactic structure of a sentence or query, while using ML matching algorithms for giving helpful responses.

Furthermore, we utilized pre-trained agents to re-train the following prescribed setup process of Vidura. The process helps to create intents for the responsive user dialogue in Vidura according to our evidence-based and social filtering recommender module entities. As shown in Fig. 5, Dialogflow is hosted as part of the KnowCOVID-19 chat server and interacts with the recommender response generator using RESTful APIs. The intended outcomes from the response generator provide information to Vidura to answer questions of clinical researchers and healthcare professionals. Such queries can include general questions about the system (e.g., *What are the Levels of Evidence Pyramid? How can I discern between the quality of evidence?*) whereas others can be more specific (e.g., *How can you help me filter information on clinical treatment with drug Remdesivir?*). The conversational

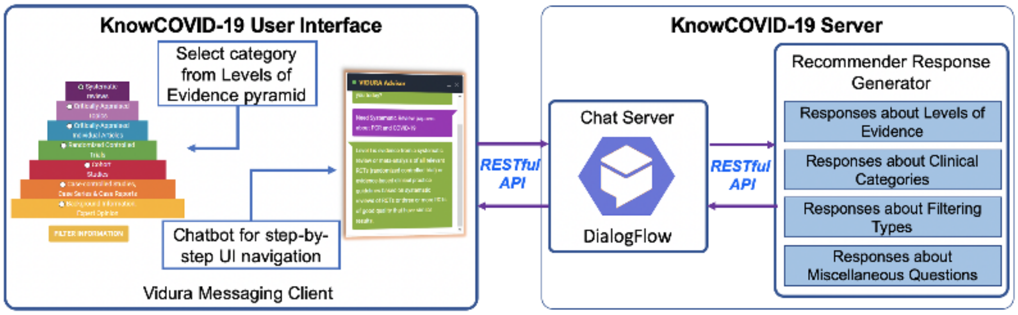

KnowCOVID-19 User Interface
Select category from Levels of Evidence pyramid
Systematic reviews
Critically Appraised topics
Critically Appraised Individual Articles
Randomized Controlled Trials
Cohort Studies
Case-controlled studies, Case Series & Case Reports
Background Information, Expert Opinion
FILTER INFORMATION
VIDURA Advisor
Need Systematic Review papers about PCR and COVID-19
Chatbot for step-by-step UI navigation
Vidura Messaging Client
RESTful API
KnowCOVID-19 Server
Chat Server
DialogFlow
RESTful API
Recommender Response Generator
Responses about Levels of Evidence
Responses about Clinical Categories
Responses about Filtering Types
Responses about Miscellaneous Questions

FIGURE 5. Vidura Chatbot assistant client-server components featuring: (i) KnowCOVID-19 client user interface, and (ii) KnowCOVID-19 chat server based on Google Dialogflow and the recommender response generator.

agent responds to user inputs through a user interface that's integrated on the KnowCOVID-19 web service through a RESTful API.

By receiving Vidura responses, the users will be able to get the guidance necessary to acquire the information they need or continue with their workflow tasks using the KnowCOVID-19 UI. Fig. 5 shows how Vidura allows the users to select a category in the Levels of Evidence Pyramid and suggests a mechanism to filter articles using our evidence-based filtering module.

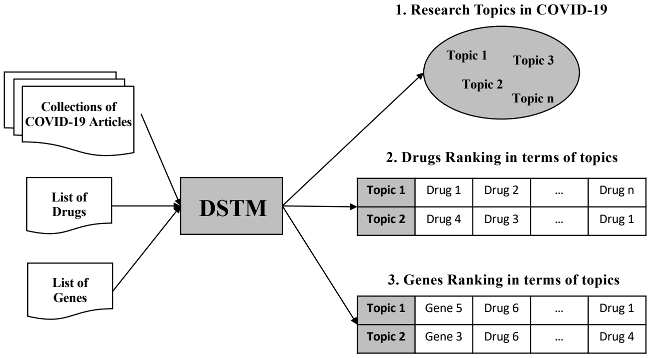
V. KNOWCOVID-19 RECOMMENDER MODULES

A. EVIDENCE-BASED FILTERING

1) Topic Model

Topic model refers to the practice of utilizing statistical algorithms to infer the topics from a collection of corpus that can aid in the discovery of hidden semantic knowledge patterns or trends. Latent Dirichlet Allocation (LDA) model [7] is the one of most popular topic models that simultaneously learns topics from corpus and topic distributions for each document. More specifically, LDA represents a topic as a multinomial distribution over the pre-defined vocabularies, and a document as a multinomial distribution over topics. LDA works by initially randomly assigning topics to documents, whilst simultaneously assigning words to topics. Following this, LDA utilizes inference algorithms such as variational inference [7] and Gibbs sampling [30] to infer those latent parameters. Finally, LDA learns the two distinct distributions: (i) topics distributions that are multinomial distributions over vocabularies, and (ii) document distributions that also are multinomial distributions over topics.

LDA learns the topics from corpus without any intents. In order to learn the domain-specific topics, we leveraged our DSTM [6] to discover the latent patterns of specific scientific topics for COVID-19. For example, it is very helpful to find relevant COVID-19 research topics based on particular drugs or genes. In the DSTM, we consider that a COVID-19 scientific paper can be decomposed into the topics in the paper, the genes explored by the authors in the paper, and the drugs used in the paper. In our DSTM, each topic is represented as a multinomial distribution over vocabularies that is same with LDA, each drug and gene are represented as multinomial distribution over topics, respectively. Finally, our DSTM learns the relationships among the research topics, drugs and genes using Gibbs sampling algorithm.


1. Research Topics in COVID-19
Topic 1
Topic 3
Topic 2
Topic n
Collections of COVID-19 Articles
List of Drugs
List of Genes
DSTM
2. Drugs Ranking in terms of topics
3. Genes Ranking in terms of topics

Topic 1	Drug 1	Drug 2	...	Drug n
Topic 2	Drug 4	Drug 3	...	Drug 1

Topic 1	Gene 5	Drug 6	...	Drug 1
Topic 2	Gene 3	Drug 6	...	Drug 4

FIGURE 6. Domain-specific Topic Model (DSTM) works as an analysis/visualization engine to discover the relationships among research topics, drugs and genes.

As shown in Fig. 6, our DSTM works as an KnowCOVID-19 engine to automatically learn the latent patterns underlying the datasets. To train a DSTM, we only need to input a collection of COVID-19 articles, a list of drugs, and a list of genes. After finishing the training phrase, the DSTM can help analyze/visualize the most popular research topics in the articles; and the DSTM will rank the most commonly investigated drugs or genes based on each topic. Our DSTM can also effectively help scientists query COVID-19 relevant drugs and genes based on their research topics, or search relevant COVID-19 topics based on specific drugs and genes.

2) Category Model

Clinical categories refer to the clinical publication type of each document. As shown in Table 1, these clinical categories referenced from [31] illustrate the best study designs and respective levels they fall under. Each clinical category topic is mapped to common research designs known from the Levels of Evidence Pyramid, which effectively narrows the list of publications pertinent to data consumers' fundamental research questions. It is important to note from a clinician perspective that not every clinical query may pertain to the highest degree of evidence (i.e., systematic reviews and meta-analyses). In that case, a clinician would then follow the next highest degree of evidence that relates to their clinical query.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083583, IEEE Access

TABLE 1. Clinical Categories and Related Levels of Evidence.

Clinical Category	Level of Evidence	Suggested Research Design(s)
All Clinical Categories	Level I	Systematic review, meta-analysis
Therapy	Levels I, II, [III, IV - secondary choices]	Randomized controlled trial (RCT), meta-analysis Also: cohort study, case-control study, case series
Etiology	Levels I, II, III, IV, [V - secondary choice]	Randomized controlled trial (RCT), meta-analysis, cohort study Also: case-control study, case series
Diagnosis	Levels II, [III, IV - secondary choices]	Randomized controlled trial (RCT) Also: cohort study
Prevention	Levels I, II, [III, IV - secondary choices]	Randomized controlled trial (RCT), meta-analysis Also: prospective study, cohort study, case-control study, case series
Prognosis	Levels IV, [V - secondary choice]	Cohort study Also: case-control study, case series

To handle the need of filtering papers according to clinical questions, our category model acts as an extension to the topic model. Similar to the topic model, the category model leverages the LDA for identifying latent topics, but trains over each document for the purposes of identifying its clinical topic. In this process, we collect a set of key terms from the LitCovid [2] medical database that were trained using LDA and mapped them to their respective KnowCOVID-19 clinical categories. We also used the LDA to list the distribution of the top k topics and its associated words within a given document. In turn, the categorization of documents can be applied by finding the maximum of the joint probability across all k topics in a document given the clinical categories terms. In other words, a document whose topics compute the highest probability with one of the clinical categories terms will be categorized under that design study.

The limitations of using a joint probability distribution between topics and clinical terms, however, can cause unstable computation across vanishing, or near zero, probabilities. Hence, this mechanism is not effective in classifying documents under a clinical category. We address this problem by calculating the logarithmic probability of each topic in a document as well as the terms under a clinical category. In turn, the logarithm achieves better results when categorizing documents under a clinical category, and supplies for better discernment of any given article (e.g., cohort study vs systematic review). This nature of category models presents a way to manage queries over a large corpus of information, making the high quality information more easily accessible in a user-friendly and practical manner.

B. SOCIAL FILTERING

1) Motivation for a Social Plane

As the COVID-19 pandemic continues to evolve, clinical researchers and medical professionals are iteratively advancing the foundations of their research areas based on their clinical queries and related research questions. To help data consumers to analyze/visualize the knowledge of the latest trends and time-relevant information about the coronavirus, we leverage our evidence-based filtering approach to engender social filtering that allows expert data consumers to share their expertise as well as data-driven insights with other experts who are working on similar pandemic challenges, and are driven by the associated clinical category inquiries.

We apply social filtering on the KnowCOVID-19 user interface that allows for end-user thin clients to redirect data consumers to a social network that allows them to publish expert comments, and also subscribe and connect with other users based on their critical information insights and tagging of relevant publications. We implement our social plane concept via HumHub [8], which is a purpose-driven social network that allows users to publish content and connect to other subscribers. Through Humhub's lightweight, powerful and user-friendly interface, the social plane thus provides an "organized" method of: (i) helping users to obtain useful resources, and further allows them to perform expertise sharing, and (ii) allows users within a scientific community to crowdsource the effort to analyze/visualize information to help answer clinical research questions.

2) Social Plane Implementation

Influenced by the work of [32] that details an effective way to use similarity measures for a 'user-based' collaborative filtering approach, we provide a "user-category" recommendation approach in which subscribers of the social plane receive notifications within their feeds based on the relevant findings of similar scholars. Paper tagging notifies target users to other scholars through a 5-star user-rating on "clinical category items." In this case, the clinical category items refer to the publications that a given user can rate on the KnowCOVID-19 user interface. The user-category recommendation approach takes an input rating of papers within their clinical categories and indicated by Levels of Evidence Pyramid that can be used to connect clinicians with other potential expert collaborators for collaborations.

Fig. 7 demonstrates our social plane implementation that allows data consumers to publish/subscribe using their user credentials. Once a user logs in, a dashboard displays a distribution chart from the evidence-based filtering that is imported from the KnowCOVID-19 web services. Data consumers are capable of inviting other expert users and they can view the latest activity of the authorized connections. The HumHub UI customization also allows data consuers to post and share their expert analysis for all their connected subscribers to view. We implemented social filtering to recommend the subscribers to look into information from

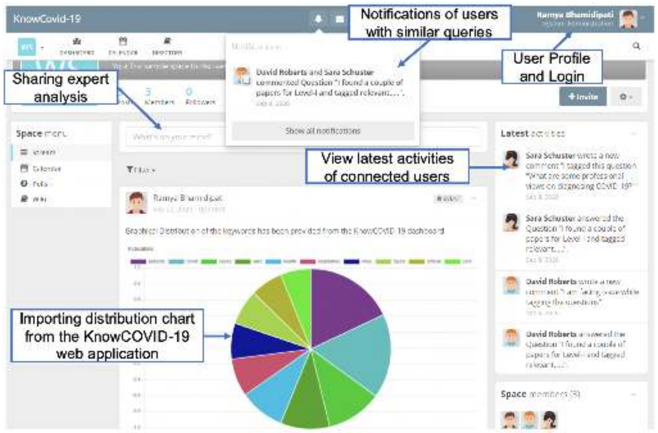

KnowCovid-19
Notifications of users with similar queries
Ramya Bhamidipati
User Profile and Login
Sharing expert analysis
David Roberts and Sara Schuster commented Question "I found a couple of papers for Level-I and tagged relevant....".
Show all notifications
+Invite
Space menu
View latest activities of connected users
Latest activities
Ramya Bhamidipati
Graphical Distribution of the keywords has been provided from the KnowCOVID-19 dashboard
Importing distribution chart from the KnowCOVID-19 web application
Sara Schuster wrote a new comment "I tagged this question "What are some professional views on diagnosing COVID-19?"
Sara Schuster answered the Question "I found a couple of papers for Level-I and tagged relevant...."
David Roberts
David Roberts answered the Question "I found a couple of papers for Level-I and tagged relevant...."
Space members (3)

FIGURE 7. Humhub-based social plane user interface in KnowCOVID-19 for collaboration and knowledge sharing.

other users that have interest in same or related research questions, clinical queries. Recommendations also prompt users to perform similar analysis using a computational workspace where they can perform and share their analysis in a manner that is helpful for other community members. Such a mechanism effectively keeps users up-to-date and connected with other data consumers/experts for the purpose of furthering knowledge on a clinical topic, and for fostering trustworthy collaborations and clinical decision making for suitable pandemic responses.

3) Crowdsourcing Clinical Queries in the Social Plane

KnowCOVID-19 has generated a set of clinical queries that are critical for finding factual evidence related to the coronavirus disease. The query tagging in the topic model and category model filtering steps are organized based on these questions for users to see aggregate ratings and relevance for the processed publications. Crowdsourcing the social filtering in expert communities via challenges allows using the tagged publications to create actionable answers that are valuable in the pandemic response. Crowdsourcing also allows creation of social networks of experts to cooperatively promote their expert insights to relevant scientific community stakeholders. The questions can be organized around trending topics within the research community or outstanding research tasks such as: (a) Background Information based Queries (e.g., are there current cohort studies occurring to find the etiology of the coronavirus?), (b) Study and Trial Based Queries (e.g., have any results of Cohort Studies proven to be successful to prevent the coronavirus from spreading?), and (c) Queries Directed to Professionals (e.g., what are some current Critically Appraised Articles on the etiology of the coronavirus?) We remark that new categories can be dynamically created based on the progress of literature analysis shared by the broader healthcare community.

Each clinical question provides relevance to experts for information sharing and crowdsourcing over analyzed data filtration. In other words, it allows data consumers to search and share evidence-based information on COVID-19 under the scope of its origins, specificity and sensitivity, as well as receiving professional advice from healthcare professionals/experts on problems concerning e.g., diagnosing COVID-19. These questions provide data consumers a direct result list of articles related to a type of writing under a specific medical topic, hence creating an efficient way for them to continue their own research thrusts and drill down facts for combating COVID-19-related challenges.

VI. EVALUATION CASE STUDY

We evaluated the effectiveness of KnowCOVID-19 by utilizing the CORD-19 dataset to compare with a manual analysis of sorting through publications without any resources provided by a team of clinical researchers. Our evaluation over the 10,000 articles measure the evidence-based filtering models that demonstrate the effectiveness of our automated KnowCOVID-19 system. We present two case studies in which: (i) we evaluate the DSTM with CORD-19 i.e., the dataset used in the COVID-19 Open Research Dataset Challenge [33], which outperforms state-of-the-art models in terms of the perplexity metric. We also demonstrate the knowledge patterns discovered by our DSTM model that can effectively guide scientists to choose suitable drugs or genes for their research problem. (ii) Finally, we compare the outcomes of a manual literature review followed by a collaborative analysis conducted by a team of clinical researchers [11] against the outcomes of the KnowCOVID-19 augmented analysis to guide clinicians' research questions in retrieving filtered information in accordance to the Levels of Evidence Pyramid through our evidence-based filtering model. In the following, we describe the manual process that clinical researchers followed for selecting papers to conduct a systematic review, and the salient findings on how our KnowCOVID-19 system augments the manual data consumers' clinical workflows based on two sets of experiments.

A. MANUAL BASELINE ANALYSIS

In a recent peer-reviewed work on the diagnostic assay used for COVID-19 [11], a team of clinical researchers and medical professionals manually conducted a cohesive systematic review on an extensive list of papers related to diagnostics tests for COVID-19. They aimed to find evidence related to the positive real-time reverse-transcriptase polymerase chain reaction (RT-PCR), serology tests and immunoassays in order to identify the limitations and clinical use for each case. To develop this study, the clinical researchers went through a workflow identical to Fig. 2 to identify the most relevant studies on the novel coronavirus. Their manual process included: search and selection, organization and labeling, detailed content review, comparison of parameters and conclusive evidence-based insights synthesized from the articles.

These clinicians have selected PubMed [1] as their source of articles and executed three main queries searching for information related to: *RT-PCR Tests*, *Serology Tests* and

TABLE 2. Four sample topics (out of 200 topics in total) extracted for the CORD-19 publications from the Kaggle dataset; Each topic is associated with 10 most likely words, 3 most likely drugs and genes that have the highest probability conditioned on that topic.

Topic 5	
Word Dist.	**Prob.**
antibodies	0.0337
antibody	0.0272
neutralizing	0.0236
vaccines	0.0169
children	0.0148
blood	0.0148
rna	0.0122
sars	0.0122
convalescent	0.0121
health	0.0121
Drug Dist.	**Prob.**
convalescent	0.9030
igm	.0845
oseltamivir	0.0025
Gene Dist.	**Prob.**
mt066156	0.2565
mt344956	0.1099
mt446312	0.1741

Topic 6	
Word Dist.	**Prob.**
led	0.0224
effects	0.0205
changes	0.0129
susceptible	0.0104
context	0.0092
discovered	0.0076
species	0.0073
directly	0.0073
sequences	0.0070
parameters	0.0070
Drug Dist.	**Prob.**
interferon	0.3418
chloroquine	0.3074
ribavirin	0.2471
Gene Dist.	**Prob.**
lc523808	0.0841
mt050493	0.0467
mn975263	0.0467

Topic 11	
Word Dist.	**Prob.**
infections	0.0328
infection	0.0262
management	0.0192
caused	0.0167
clinical	0.0159
common	0.0118
bacterial	0.0112
detection	0.0105
tested	0.0091
followed	0.0086
Drug Dist.	**Prob.**
oseltamivir	0.5880
azithromycin	0.1499
igd	0.1018
Gene Dist.	**Prob.**
mt050493	0.1701
mt344956	0.1156
mt042778	0.0884

Topic 54	
Word Dist.	**Prob.**
using	0.0357
genome	0.0338
protein	0.0283
sequences	0.0251
genomes	0.0216
sars	0.0170
strains	0.0150
phylogenetic	0.0150
proteins	0.0147
analysis	0.0134
Drug Dist.	**Prob.**
remdesivir	0.2645
convalescent	0.2280
polyclonal	0.2034
Gene Dist.	**Prob.**
mt050493	0.2590
mg772934	0.2243
ay274119	0.2087

Immune-Assays. For Nucleic Acid Amplification Tests, the terms 'PCR AND COVID-19' were typed together, and in another search SARS CoV-2 was entered. For Serology Tests, 'IgM AND 'COVID-19' were entered and 'IgG AND COVID-19' were entered on different searches. Lastly for the Immuno-Assays, Hyper-immune 'polyclonal antibodies AND COVID-19' were typed together to receive effective and relevant articles. No restriction was set for the start or end date of the publications.

Once the articles were gathered, the search outcomes were organized and labeled through the types of studies (i.e., cross-sectional studies, case series, comparative studies), methodology applied (e.g., using DNA samples and CT scans), sensitivity and specificity rates of the studies, advantages and disadvantages of the methodologies, study methods used by the authors, and also the turnaround time for the variables. By organizing important information in the form of tables for easy review and comparison of parameters, the clinicians finalized their conclusions along with evidence summaries synthesized from the selected peer-reviewed articles. An overarching conclusion was presented taking into account all the parameters evaluated along with findings on the benefits and limitations of various diagnostic tests used for COVID-19.

This workflow relies on the results of the manual queries conducted in the first step, which is driven by the limited keywords employed in the search. If in subsequent steps the researchers identify that the quality of the resulting set of papers does not fit their needs, they need to go back and refine the query by trying other combinations of keywords. Such a situation consequently delays the process or causes missing of important documents if the right search keywords are not applied. Consequently, it is important to enhance and streamline the manual workflow process providing a suitable result set that is not only based on keywords but also on the classified levels of evidence of the papers. Thus, the KnowCOVID-19 augmentation of the manual workflow process helps the researchers in expediting the time taken to obtain the desired results in their study.

B. KNOWCOVID-19 AUGMENTED ANALYSIS

1) DSTM Effectiveness Evaluation

To augment the manual clinical workflow process, we apply evidence-based filtering through the evaluation of the DSTM over the CORD-19 dataset. We collected a subset of 2100 papers from the total collection of 10,000 that are closely related to pre-defined drugs and genes terms. We split our datasets into a training dataset and a testing dataset with 80% and 20% proportions, respectively. Also, we collected a list of drugs from *COVID-19 Vaccine Tracker* [28], and a list of genes from *Virtual Incident Procurement (ViPR)* [34]. In total, our collection features 59 types of drugs and 83 types of genes. We apply DSTM to discover the latent knowledge patterns among COVID-19 research topics and drugs and genes, which can guide the scientists to choose suitable drugs or genes based on their research topics of interest.

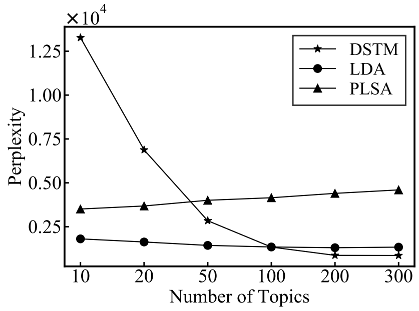


FIGURE 8. Perplexity comparison with LDA, PLSA models on the CORD-19 dataset from Kaggle considering different number of topics.

The DSTM is a probabilistic model, and the generalization effectiveness is an important factor to evaluate how well a probabilistic model predicts the unobserved datasets. To evaluate the generalization effectiveness, we use the perplexity metric to compare the performance with state-of-the-art models, such as LDA [7] and PLSA [35]. A lower perplexity score indicates better generalization effectiveness of held-out test datasets. We use 80% of the datasets to train each model, and use 20% of the datasets to test the model. All the models are trained with 50 epochs. As shown in Fig. 8, although our DSTM has the worst performance with a smaller number of topics, our DSTM achieves the best perplexity score after *topic=200*. The number of topics is the hyperparameter that needs to be tuned, and the performance of the perplexity score helps us to choose the optimal number of topics. In addition, we found that a different number of topics have only a slight impact on the LDA and PLSA's performance in comparison with LDA. This is because our DSTM is more suitable to find domain-specific topics in comparison with LDA and PLSA. For example, we input COVID-19 relevant genes and drugs to understand the COVID-19 relevant topics, and the latent patterns among topics, genes and drugs.

We also demonstrate the knowledge patterns discovered by our DSTM as shown in Table 2, which shows results on suitable drugs or genes based on their research topics. The results involve 4 samples of topics from 200 topics learned by DSTM from the CORD-19 publication dataset. Each subtable shows the top 10 words that are most likely to be generated conditioned on that topic; the top 3 most likely drugs to be used for the topic; the top 3 most likely types of genes that arise from the topic. For example, *Topic 5* refers to the research topics about antibodies, the most popular drugs investigated by this area are convalescent, IgM, oseltamivir; and the most common genes are *mt066156*, *mt344956*, *mt446312* respectively. *Topic 11* refers to the topics about infections, and drugs such *oseltamivir*, *azithromycin*, genes such as *mt050493*, *mt344956* that are commonly investigated in this area.

2) Information Retrieval Evaluation

Herein, we evaluate our evidence-based filtering approach by utilizing the same clinical queries to come up with a category model analysis for finding relevant resources pertinent to immunology, infectious diseases and epidemic/pandemic control. To draw an effective comparison between the manual literature review conducted in the aforementioned study, we analyzed papers through searching, selecting, organizing and labeling according to evidence-based practice. We remark that we only used and filtered peer-reviewed papers according to the keywords in the query inputs.

The queries were analyzed using the Term Frequency-Inverse Document Frequency (TF-IDF) used for information retrieval and search. As each query was being processed, the weight of each term was calculated according to its frequency within the text corpus of the journal venues and associated articles collected. A term with a high weight-age score indicated the rarity of the word in the query, and thus, became pertinent to the collection of documents with exact or similar terms in each clinical category. In this context, the key terms that held the highest weight-age were “serology”, “nucleic”, “immuno”, and “acid”, across the three queries. We then clinically categorized each document according to the distribution of keywords in each topic of interest.

Once the topics were generated, the user was able to then filter out information and obtain a recommended list of papers. This filtered information is organized in a table format with columns according to its clinical category, evidence levels, as well as the publication title and meta-data. This in turn allows the users to identify which clinical areas are interesting for drill-down analysis based on their query inputs. In mimicking a cohesive systematic review done by a manual process, papers related to user search queries are filtered according to their evidence level. The process of automation of the systematic review led to a speed-up of the process of finding reliable information relevant to their specified questions and research objectives (i.e., increased efficiency). This process thus eliminates the need to manually conjugate a list of papers through multiple search indexes.

3) Results of Effectiveness Analysis for KnowCOVID-19

The results of replicating the manual workflow process authors used in [11] (manifested in Section VI-A) against our evidence-based filtering approach (manifested in Section VI-B) over the CORD-19 dataset is summarized in Table 3. The ultimate goal of this experiment was show the effectiveness of the KnowCOVID-19 system by demonstrating the automation of narrowing down clinicians' search processes through relevant tools, topics and articles. In these experiments, we have exemplified a way in which the manual baseline workflow is augmented by expediting the literature research process through a sequence of automated steps.

The manual search process which included specific search terms chosen by the clinicians for different clinical queries resulted in a relatively low quantity of relevant papers i.e., no more than 12% of papers for each clinical query across the CORD-19 dataset were retrieved. Furthermore, the authors lacked any defining methods of classifying papers according to the Levels of Evidence Pyramid. Therefore, there were no capable means of identifying which levels the retrieved articles ranged in, as it is indicated as ‘N/A’ in Table 3. Consequently, we suppose that despite their relevant search terms (e.g., ‘PCR & COVD-19’, ‘IgM & COVID-19’), it is an expensive task to refine their queries *and* properly filter each article in order to fulfill their research objectives. From this observation, we conclude that the queries need to be processed to yield more effective search terms to find a higher quantity of papers and the range of evidence levels they fall under.

In contrast, we observe by utilizing TF-IDF that the KnowCOVID-19 augmented analysis extracts more significant words related to each clinical query (e.g., ‘Acid’, ‘Amplification’, ‘Serology Tests’), and thus increases the total number of documents retrieved. The results from the enhanced list of key search terms across all clinical queries show an average of over 28% of retrieved papers. Moreover,

TABLE 3. Outcomes comparison between the manual baseline analysis without the guidance of tools/resources and the KnoCOVID-19 augmented analysis.

Queries	Manual Baseline Analysis			KnowCOVID-19 Augmented Analysis		
	No. of Papers (out of 10,000)	Search Terms	Top Levels	No. of Papers (out of 10,000)	Search Terms	Top Levels
Nucleic Acid Amplification Tests: *Question 1:* What is the status of Nucleic Acid Amplification test with PCR used for COVID-19 or SARSCoV-2?	1152	PCR & COVID-19 PCR & SARS CoV2	N/A	**2668**	Acid, PCR, Amplification, SARS-CoV	[II-V]
Serology Tests: *Question 2:* What is the status of serology tests with IgM or IgG used for COVID-19?	176	IgM & COVID-19 IgG & COVID-19	N/A	**2360**	IgM, Serology Tests, IgG	[I, II]
Immuno-Assays Tests: *Question 3:* What is the current status of Immuno-assays with hyper-immune polyclonal antibodies or monoclonal antibodies used for COVID-19?	35	pAb & COVID-19 moAb & COVID-19	N/A	**3552**	Antibodies, Immuno, Monoclonal, Assays	[I, IV]

by leveraging our category model from our evidence-based filtering approach, we have indicated the 'Top Levels', or most frequent evidence levels shown for the each of the papers retrieved. In Table 3, we can see that the majority of papers retrieved that are relevant to the researchers' clinical queries range from Level I to V (e.g., Systematic Reviews to Cohort Studies). These results indicate the range of clinical categories (referenced from Table 1) that demonstrates the types of design studies being performed. Hence, KnowCOVID-19 effectively filters out the quality of information in articles in a reliable manner for data consumers to perform a drill-down analysis in order to fulfill their clinical queries.

VII. DISCUSSION

In this section, we first present a discussion on the latest status of KnowCOVID-19 deployment. In addition we present limitations, issues for generalization for other healthcare areas beyond COVID-19 and address these issues for further improvement of capabilities as well as for wider user adoption.

A. KNOWCOVID-19 DEPLOYMENT AND ENHANCING USER ADOPTION

The KnowCOVID-19 application addresses the workflow bottleneck of data consumers manually filtering high quality evidence within publication archives. While the methods are reproducible and can be applied over other datasets (e.g., publications, medical tools/terms), widespread adoption of services as KnowCOVID-19 present new challenges. From a socio-technical perspective, one important theory that explains the rate in which technology spreads among users is diffusion of innovations (DOI) [36]. The three main aspects to DOI are *adoption* at the individual level, *implementation* at the team level and *diffusion* at the community level. Till date, we have conducted a qualitative assessment with clinicians, biochemists, and veterinary doctors who have expressed their interest to adopt KnowCOVID-19 and related Vidura chatbot service for filtering high quality publications to reduce their time-consuming workflow tasks. We are continuing our implementation efforts and working towards large-scale usability study to meet community-scale user needs.

B. GENERALIZATION FOR HEALTHCARE AREAS BEYOND COVID-19

Evidence-based practice for clinical studies is commonly used across the healthcare field as a guideline for finding high quality evidence within a specific area of research [3]. While this paper proposes an automation of classifying COVID-19 publications according to the Level of Evidence Pyramid, its applicability to the medical field can be generalized to other areas of healthcare such as translational research, clinical research and biochemistry. The steps in this direction involve integration of relevant multiple data resources that are trained with our evidence-based filtering model to present new topics relevant for area-specific research tasks. In addition, to make our Vidura chatbot provide guided interfaces, dialog design also needs to be adapted to meet area-specific research task requirements.

C. ACCESSIBLE DATA RESOURCES

The KnowCOVID-19 evidence-based recommender system aligns with the principles of Open Government Data (OGD) (https://www.oecd.org/gov/digital-government/open-government-data.htm) as follows: (a) it provides the means to integrate multiple datasets, which in most cases are publicly available, and (b) once researchers have completed their custom search, analysis, and evaluation of relevant articles, they are able to make their findings available through our social plane features. Thus, the usage of KnowCOVID-19 in conjunction with open datasets and the collaboration via expertise sharing through the social plane encourages researchers to follow best practices on reproducibility and replicability.

D. DATA RELIANCE AND VALIDITY

There is rapidly growing knowledge on science advancements through publication venues such as scientific conferences, archive journals and online resources. For example, according to observations in Scopus, the number of publications have been doubling every 12 years. Given the volume of publications emerging, it is a challenge to verify data inaccuracies and also filter out vendors who organize journal issues with poor peer-review process to profit from author publication fees. Hence, there is a need for methods that help in automation to determine article data reliance and validity. While KnowCOVID-19 has demonstrated how it can filter high quality publications based on higher levels of evidence, authors need to be knowledgeable to verify the reputation of the publication source through e.g., impact scores. Existing works such as [37] have shown that they can assess the validity of publications through a subjective metric such as a 'publication pressure questionnaire'. KnowCOVID-19 could be enhanced with such methods and metrics that guide users to obtain visual cues of the critical publication portions that needs to be verified for data reliance and validity.

VIII. CONCLUSION

In this paper, we detailed a novel evidence-based recommender system namely, KnowCOVID-19 in response to meeting the needs of data consumers (e.g., clinical researchers and healthcare professionals) who are responsible to explore solutions to respond to the on-going pandemic. Specifically, we address the problem of workflow bottlenecks they face while performing literature review (e.g., long time for analysis of article collections, and discerning knowledge gaps with regards to evidence levels of articles) involving the rapidly growing COVID-19 publication databases. Our novel KnowCOVID-19 system enables the data consumers to use thin-clients for leveraging edge computing services that integrate recommender modules for publications analytics based on the Levels of Evidence Pyramid. We developed an evidence-based filtering approach to effectively sort out reliable information through a domain-specific topic model (DSTM) to observe hidden knowledge patterns and show the relation between topics (e.g., drugs and genes) and the large collections of medical literature archives. The evidence-based filtering approach also leverages our proposed category model that uses the LDA inference algorithm to filter articles based on their clinical category in the evidence-based practice standards. Furthermore, the KnowCOVID-19 supports a follow-on social filtering implemented via a social plane that allows data consumers to publish/subscribe key insights obtained from the automated literature review workflow. It also features a user interface that integrates a context-aware chatbot, Vidura Advisor that makes it feasible for clinicians to obtain guidance on how to navigate and obtain answers to their clinical questions through the evidence-based filtering process automation.

We validated our KnowCOVID-19 system by running our evidence-based filtering models across 10,000 journals from the CORD-19 Kaggle text corpus related to COVID-19, SARS-CoV-2, and related infectious diseases. We first used the DSTM probabilistic model to generalize the effectiveness in comparison to state-of-the-art models to factor how we can effectively predict unobserved topics in datasets. We then used our information retrieval technique to compare how a manual analysis in which a team of clinicians analyzed a series of COVID-19 publications without any automation tools would perform against the KnowCOVID augmented analysis. The augmented analysis substantially outperformed the pre-defined workflows of clinicians and more efficiently obtained scalable and actionable insights pertinent to their clinical queries. Our case studies in evaluation experiments demonstrated that our automated literature review approach using information classification based on the Levels of Evidence Pyramid can be helpful to filter information through a series of edge computing services. The KnowCOVID-19 system thus offers the medical research community with end-user thin-client based tools to perform rapid and detailed analysis of relevant (fast growing) literature to: (a) keep up with the latest information around the clinical categories and related information organized around evidence in published peer-reviewed data, and (b) urgently contribute solutions/guidance towards the ongoing fight against the COVID-19 infectious disease.

Our future work is to evaluate the effectiveness of our social filtering that uses a collaborative filtering approach to recommend users based on similar insights from evidence-based filtering and rated papers. In addition, future work could involve demonstrating the effectiveness of the Vidura chatbot in guiding clinical researchers to perform an effective analysis through a usability study. In addition, we plan to create mechanisms such as open datasets, model-driven tools and task-specific social collaboration features. These mechanisms will help to engage large communities of researchers/clinicians as well as data scientists to answer the bold questions using tools for sophisticated knowledge discovery and results synthesis from massive COVID-19 related publication archives.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation under the award numbers OAC-2006816 and OAC-1827177. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] National Center for Biotechnology Information, "PubMed," 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/
[2] National Library of Medicine, "LitCovid," 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/research/coronavirus/
[3] M. H. Murad, N. Asi, M. Alsawas, and F. Alahdab, "New Evidence Pyramid," *BMJ Evidence-Based Medicine*, vol. 21, no. 4, pp. 125–127, 2016.
[4] D. Timm, "Evidence Matters," *Journal of the Medical Library Association: JMLA*, vol. 94, no. 4, p. 480, 2006.

[5] "Use Deep Search to Explore the COVID-19 Corpus," 2020. [Online]. Available: https://www.research.ibm.com/covid19/deep-search

[6] Y. Zhang, P. Calyam, T. Joshi, S. Nair, and D. Xu, "Domain-specific Topic Model for Knowledge Discovery through Conversational Agents in Data Intensive Scientific Communities," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4886–4895.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[8] "The Flexible Open Source Social Network Kit," 2020. [Online]. Available: https://www.humhub.com/en

[9] M. Ekin Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle Literature Organization," *arXiv e-prints*, pp. arXiv–2008, 2020.

[10] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, pp. 1–37, 2008.

[11] D. Shyu, J. Dorroh, C. Holtmeyer, D. Ritter, A. Upendran, R. Kannan, D. Dandachi, C. Rojas-Moreno, S. P. Whitt, and H. Regunath, "Laboratory Tests for COVID-19: A Review of Peer-Reviewed Publications and Implications for Clinical UIse," *Missouri medicine*, vol. 117, no. 3, p. 184, 2020.

[12] T. Achakulvisut, D. E. Acuna, T. Ruangrong, and K. Kording, "Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications," *PloS one*, vol. 11, no. 7, p. e0158423, 2016.

[13] Z. Yu and T. Menzies, "FAST^2: An intelligent assistant for finding relevant papers," *Expert Systems with Applications*, vol. 120, pp. 57–71, 2019.

[14] D. Das, Y. Katyal, J. Verma, R. K. Ranjan, S. Dubey, A. D. Singh, S. Bhaduri, and K. Agarwal, "Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings," 2020.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[16] B. Rahdari, P. Brusilovsky, K. Thaker, and H. K. Chau, "CovEx: An Exploratory Search System for COVID-19 Scientific Literature," *University of Pittsburgh*, 2020.

[17] V. Kieuvongngam, B. Tan, and Y. Niu, "Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2," *arXiv preprint arXiv:2006.01997*, 2020.

[18] S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "Target specific mining of COVID-19 scholarly articles using one-class approach," *arXiv preprint arXiv:2004.11706*, 2020.

[19] M. U. Ashraf, A. Hannan, S. M. Cheema, Z. Ali, A. Alofi *et al.*, "Detection and Tracking Contagion using IoT-Edge Technologies: Confronting COVID-19 Pandemic," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–6.

[20] M. S. Hossain, G. Muhammad, and N. Guizani, "Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-I9 Like Pandemics," *IEEE Network*, vol. 34, no. 4, pp. 126–132, 2020.

[21] A. Sufian, A. Ghosh, A. S. Sadiq, and F. Smarandache, "A Survey on Deep Transfer Learning to Edge Computing for Mitigating the COVID-19 Pandemic," *Journal of Systems Architecture*, vol. 108, p. 101830, 2020.

[22] N. El-Rashidy, S. El-Sappagh, S. Islam, H. M. El-Bakry, and S. Abdelrazek, "End-To-End Deep Learning Framework for Coronavirus (COVID-19) Detection and Monitoring," *Electronics*, vol. 9, no. 9, p. 1439, 2020.

[23] M. Vassell, O. Apperson, P. Calyam, J. Gillis, and S. Ahmad, "Intelligent dashboard for augmented reality based incident command response co-ordination," in *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2016, pp. 976–979.

[24] D. Chemodanov, F. Esposito, A. Sukhov, P. Calyam, H. Trinh, and Z. Oraibi, "Agra: Ai-augmented geographic routing approach for iot-based incident-supporting applications," *Future Generation Computer Systems*, vol. 92, pp. 1051–1065, 2019.

[25] A. Lakhani, A. Gupta, and K. Chandrasekaran, "Intellisearch: A search engine based on big data analytics integrated with crowdsourcing and category-based search," in *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*. IEEE, 2015, pp. 1–6.

[26] Y. Zhang, P. Calyam, S. Debroy, and S. S. Nuguri, "Social Plane for Recommenders in Network Performance Expectation Management," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 97–111, 2017.

[27] B. J. Maguire, A. R. McLean, S. Rashan, E. S. Antonio, J. Bagaria, Z. Bentounsi, M. Brack, F. Caldwell, V. I. Carrara, B. W. Citarella *et al.*, "Baseline results of a living systematic review for covid-19 clinical trial registrations," *Wellcome Open Research*, vol. 5, no. 116, p. 116, 2020.

[28] Milken Institute, "COVID-19 TREATMENT AND VACCINE TRACKER," 2020. [Online]. Available: https://covid-19tracker.milkeninstitute.org

[29] A. A. Chandrashekara, R. K. M. Talluri, S. S. Sivarathri, R. Mitra, P. Calyam, K. Kee, and S. Nair, "Fuzzy-based conversational recommender for data-intensive science gateway applications," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4870–4875.

[30] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[31] Winona State University, "Levels of Evidence," 2021. [Online]. Available: https://libguides.winona.edu/c.php?g=11614&p=61584

[32] Z. Tan and L. He, "An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle," *IEEE Access*, vol. 5, pp. 27 211–27 228, 2017.

[33] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill *et al.*, "CORD-19: The Covid-19 Open Research Dataset," *ArXiv*, 2020.

[34] B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu *et al.*, "ViPR: an open bioinformatics database and analysis resource for virology research," *Nucleic acids research*, vol. 40, no. D1, pp. D593–D598, 2012.

[35] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[36] E. M. Rogers, A. Singhal, and M. M. Quinlan, *Diffusion of innovations*. Routledge, 2014.

[37] J. Tijdink, Y. Smulders, A. Vergouwen, H. de Vet, and D. Knol, "The assessment of publication pressure in medical science; validity and reliability of a publication pressure questionnaire (ppq)," *Quality of Life research*, vol. 23, no. 7, pp. 2055–2062, 2014.

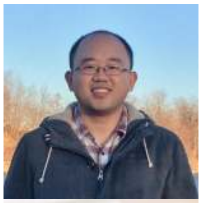
ROLAND ORUCHE received his BS in Information Technology from the University of Missouri-Columbia. He is currently pursuing his Ph.D. in Computer Science at the University of Missouri-Columbia. His research interests include machine learning, human-computer interaction, cloud computing and virtual reality.

VIDYA GUNDLAPALLI is a high school senior at Barrington High School in Barrington, Illinois. Through her time in high school, she has gained interest in researching Biochemical Pathways, specifically as they relate to insulin resistance, infectious diseases, and DHA's effect on brain function.

ADITYA P. BISWAL is a high school senior at BASIS Independent Silicon Valley in San Jose, California. Throughout his high school years, he gained an interest in machine learning and artificial intelligence, specifically in its application to computational linguistics and quantitative finance.

PRASAD CALYAM received his MS and PhD degrees from the Department of Electrical and Computer Engineering at The Ohio State University in 2002 and 2007, respectively. He is currently an Associate Professor in the Department of Computer Science at University of Missouri-Columbia. His current research interests include distributed and cloud computing, computer networking, and cyber security. He is a Senior Member of IEEE.

MAURO LEMUS ALARCON received his MS in Mathematics and Computer Science from the McNeese State University. He is currently pursuing his Ph.D. in Computer Science at the University of Missouri-Columbia. His research interests include cloud computing and healthcare data analytics.



YUANXUN ZHANG received his BE degree from Southwest Jiaotong University, China, in 2006. He is currently pursuing his PhD degree in University of Missouri-Columbia. His research interests include network performance monitoring, software-defined networking, and big data analytics.

NAGA RAMYA BHAMIDIPATI received her BE degree from Chaitanya Bharathi Institute of Technology, India, in 2017. She is currently pursuing her Masters degree in Computer Science at University of Missouri-Columbia. Her research interests include cloud computing, data analytics, and artificial intelligence.

ABHIRAM MALLADI is currently a high school student at Christian Brothers College (CBC) in Saint Louis, Missouri as an junior. He currently is pursuing his interests in Computer Science by working on many hands-on real-world projects. He also has a keen interest in robotics, machine learning, and artificial intelligence. He has previously worked on projects for his school that helped a FoodBank.

HARIHARAN REGUNATH received his MBBS and MD degrees from The Tamil Nadu Dr. MGR Medical University, India and Manipal University, India respectively. He received additional residency training in internal medicine and fellowship training in Infectious Diseases and Critical Care Medicine, both at University of Missouri-Columbia. He is currently an Assistant Professor of Clinical Medicine at the Department of Medicine – Divisions of Pulmonary, Critical Care Medicine and Infectious Diseases at University of Missouri-Columbia. His current research interests include COVID-19, ARDS, Nosocomial Infections, Infective Endocarditis, Clinical Microbiology, CMV infections and Travel medicine.