ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing

Qingyu Chen,* Robert Leaman,* Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA; email: zhiyong.lu@nih.gov

Annu. Rev. Biomed. Data Sci. 2021. 4:313–39

First published as a Review in Advance on May 14, 2021

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-021821-061045

*These authors contributed equally to this article

Keywords

artificial intelligence, natural language processing, text mining, COVID-19, pandemic control

Abstract

The COVID-19 (coronavirus disease 2019) pandemic has had a significant impact on society, both because of the serious health effects of COVID-19 and because of public health measures implemented to slow its spread. Many of these difficulties are fundamentally information needs; attempts to address these needs have caused an information overload for both researchers and the public. Natural language processing (NLP)—the branch of artificial intelligence that interprets human language—can be applied to address many of the information needs made urgent by the COVID-19 pandemic. This review surveys approximately 150 NLP studies and more than 50 systems and datasets addressing the COVID-19 pandemic. We detail work on four core NLP tasks: information retrieval, named entity recognition, literature-based discovery, and question answering. We also describe work that directly addresses aspects of the pandemic through four additional tasks: topic modeling, sentiment and emotion analysis, caseload forecasting, and misinformation detection. We conclude by discussing observable trends and remaining challenges.

INTRODUCTION

Background

Since the initial reports of an outbreak of "pneumonia of unknown cause" in Wuhan, Hubei province, China (1), the acute lack of knowledge surrounding COVID-19 (coronavirus disease 2019) (2) has driven intense investigations into its potential impact on society and interventions likely to reduce it. Once the cause was identified as a novel coronavirus, provisionally named 2019-nCoV (2019 novel coronavirus) (2), the focus of these investigations shifted to the essential questions surrounding its transmissibility and the prognosis of those infected. The virus was officially renamed SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) (3) and spread worldwide in the following weeks as public health officials sought to identify measures to slow its advance and frontline healthcare workers urgently called for the available treatments to be tested. As disruptive public health measures were enacted and the death toll rose, researchers responded with a flood of articles addressing many aspects of COVID-19 and SARS-CoV-2 (4, 5). Yet the difficulty of identifying reliable and actionable knowledge specific to a particular context caused a sort of second epidemic: information overload (6, 7), exacerbated by the evolving understanding of the disease and a wave of article retractions from even well-respected journals (8, 9). Meanwhile, members of the public faced the severe psychological stress of changing public health measures, heavy economic impacts, and health uncertainties (10) while experiencing their own information overload through news and social media, aggravated by inconsistent messaging and deliberate misinformation campaigns (11, 12). At every step, success in the fight against COVID-19 has been driven by access to the right amount of reliable information and the willingness to act upon it.

Figure 1 summarizes the statistics of paper growth between January and November 2020 in LitCovid, a literature database keeping track of COVID-19 related papers in PubMed (13, 14). Since May, the number of papers in LitCovid has been growing at about 10,000 articles per month, accounting for over 7% of articles in PubMed. Each day, there are about 320 new articles that are related to COVID-19. Such dramatic growth significantly increases the burden of manual curation, analysis, and interpretation. It is pressing for natural language processing (NLP) techniques to address the burden.

Natural Language Processing: Introduction, Text Genres, and Pipelines

NLP, also known as text mining, allows automated processing and analysis of unstructured texts, such as extracting information of interest and representing it in a structured format appropriate for computational analysis or applying transformations like summarization or translation that make the text more digestible for human readers (15). NLP thus enables analyses to be faster and use larger scales than manual analysis allows and has long been recognized as a way to alleviate information overload in biomedical research (15–17). Tools for identifying a relevant subset of documents in the scientific literature [information retrieval (IR)] have been particularly successful (18), although the trend has been toward more specific information within individual articles and more comprehensive results across the literature (19). NLP methods have matured significantly in recent years (20), with successful applications in a variety of applications, including literature-based discovery (LBD) (21), facilitating analysis of high-throughput (gene expression/genome-wide association) data (22), and pharmacovigilance (23), among many others.

NLP methods have been applied to a wide variety of textual sources, which we categorize broadly into four domains. First, biomedical literature summarizes the findings of studies in publications, such as articles in PubMed and PMC (PubMed Central). Second, clinical notes, found in electronic health records (EHRs), describe a variety of clinically important aspects of an encounter with a patient at a particular visit, including their history, symptoms, diseases, test

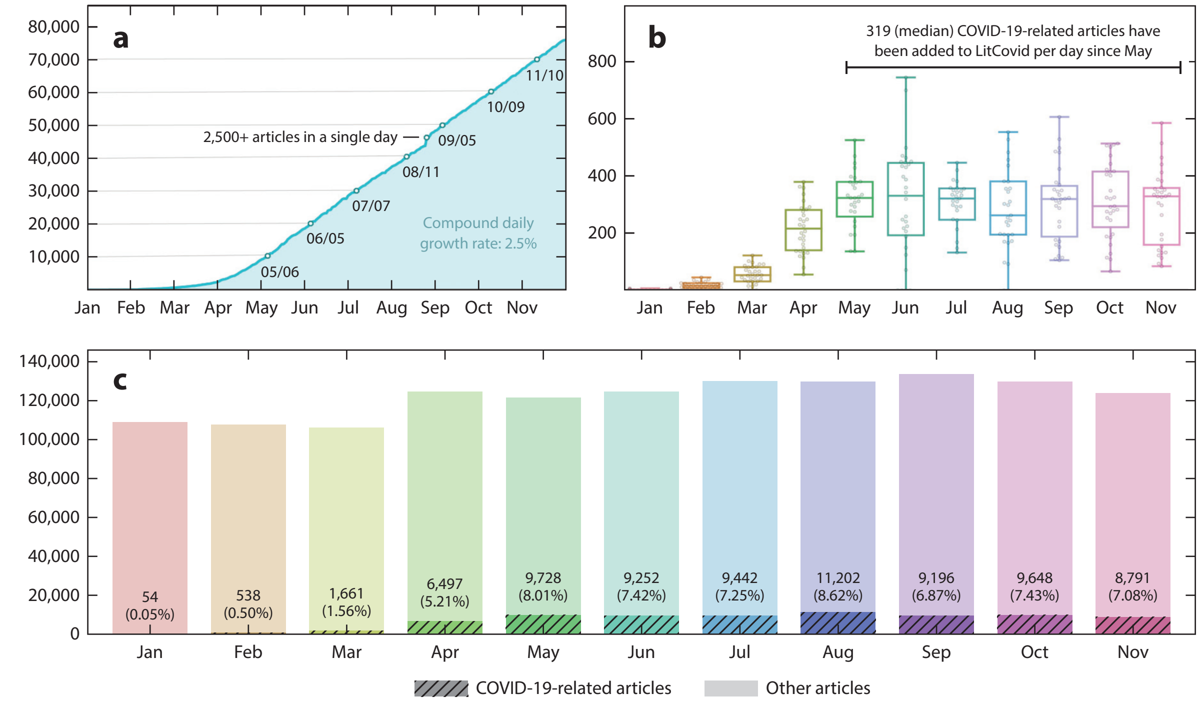

Figure 1

Growth of LitCovid, a database of biomedical publications related to COVID-19, from January to November 2020. (*a*) The accumulated growth of LitCovid. (*b*) The number of articles added to LitCovid daily, summarized by month. (*c*) The monthly ratio of articles related to COVID-19 compared with all PubMed articles.

results, and interventions. Third, social media, such as posts on Twitter, reflect personal opinions, statuses, or emotions. Fourth, news articles contain descriptions of current events intended for the general population. While the results of research studies are typically published as scientific articles, applying NLP to text from other domains can provide answers to questions the literature does not yet address; for example, social media and news can be used to address public health issues (24). Due to the contrasting purposes of these source types, however, the text found in these sources exhibit dramatically different characteristics. Both clinical notes and social media, for example, exhibit much higher rates of spelling and grammar errors than the published literature or news articles. The content of clinical notes can vary widely between health care systems or even between individual hospitals. There are also significant differences between abstracts and full-text articles within the published literature, motivating differences in the NLP methods applied (25). NLP for clinical notes (EHRs), social media, and news is primarily discussed below in the section titled Pandemic-Oriented Applications.

NLP tasks are typically addressed in a pipeline, with more fundamental tasks (such as IR) performed before intermediate tasks [such as named entity recognition (NER), that is, identifying mentions of critical concepts such as genes and diseases], which in turn are performed before high-level tasks (such as knowledge discovery) (26). **Figure 2** shows an example of an NLP pipeline for question answering (QA), illustrating the use of both IR and NER as subcomponents. Many of these tasks are well studied, allowing existing NLP systems to be adapted to address their task in the context of the COVID-19 pandemic with a relatively small effort. However, the pandemic also presented several specific challenges. Many NLP methods use existing resources, such as a

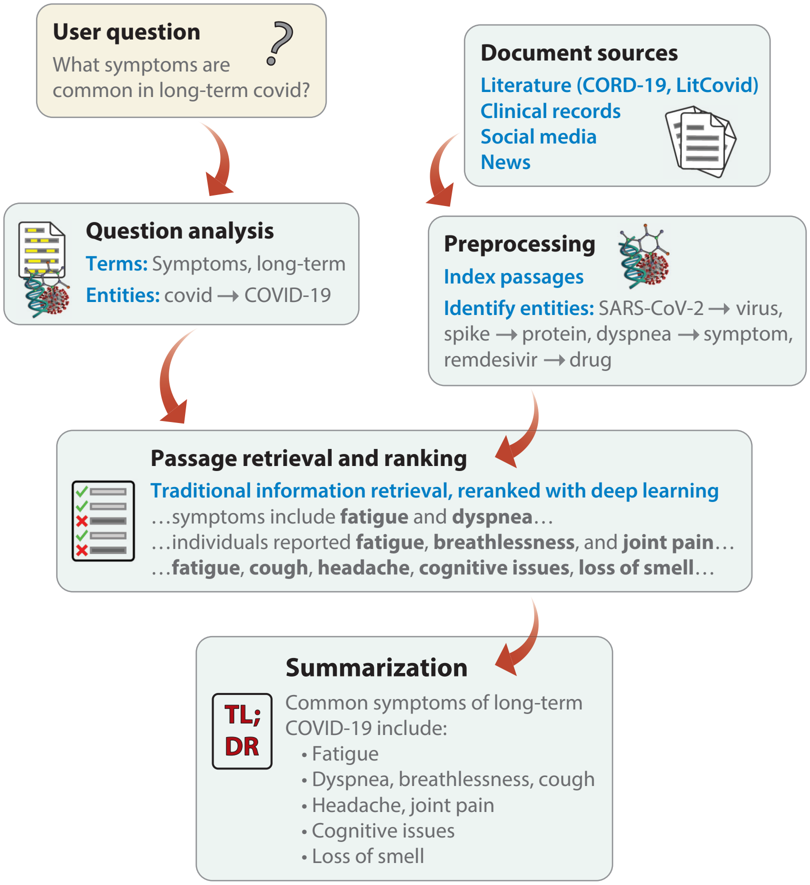

Figure 2

Overview of a natural language processing (NLP) pipeline for COVID-19 question answering, illustrating the use of other NLP tasks as submodules. In this pipeline, information retrieval is used to retrieve passages that may contain answers and named entity recognition is used to identify entities in both the question and the source documents. This task is described in detail in the section titled Question Answering.

large dataset of related text (corpus) or a list of concepts, names, and relationships, to help provide meaning and context. Because both the virus and disease were new, there was initially very little relevant text available, and the relationships between the related concepts were largely unknown. The pandemic introduced completely new terminology (e.g., "COVID-19") that were provided by updates to ontologies such as ICD-10 (International Classification of Diseases, 10th Revision) and MeSH (Medical Subject Headings) relatively early in the pandemic (27, 28) and existing NLP systems then had to be updated to include them. Less conspicuously, the pandemic also caused significant changes in the frequency of existing terms (e.g., hydroxychloroquine, remdesivir), called a domain shift, which also necessitated system updates to avoid performance degradation. Even identifying text that mentions COVID-19 or SARS-CoV-2 presents a challenge: Authors initially used descriptions because names were not available, the name of the virus was controversially changed from "2019-nCoV" to "SARS-CoV-2" (29), and a surprisingly high number of new name variations appear in the literature each week (30). At a higher level, NLP pipelines typically first address identifying the evidence and assertions from single articles, with later processes or even the human reader tasked with integrating the information collected, considering its accuracy, or identifying which information is most current. However, many critical aspects of the disease—its etiology, transmissibility, symptoms, mechanisms of action, prognostic indicators, disease course, fatality rate, and long-term effects, as well as questions about which interventions for prevention or treatment would be most effective—were addressed in the literature multiple times, with competing hypotheses, potentially contradictory evidence, and conclusions that were later revised or refuted. Useful knowledge about COVID-19 or SARS-CoV-2 is therefore often diffused across many articles. Finally, the pandemic is a worldwide concern, with different expert audiences (e.g., researchers, pharma, clinicians, public health officials) all requiring information that addresses their specific needs, but much of the same information needing to be disseminated and understood by nonexperts in government, the media, and the public in every language. Thus, a major need is the ability to transfer knowledge encoded using precise medical terms into information understandable by nonexperts, possibly in another language.

The Organization of the Review

In this review we consider NLP work found on PubMed and Google Scholar at the end of August 2020 that has used published biomedical research as input or that has been used to gather data to inform biomedical research. We describe NLP tasks, use cases, and datasets for addressing information needs in the context of the COVID-19 pandemic. We first describe several traditional NLP tasks in the context of the COVID-19 pandemic, and then consider several tasks specific to the pandemic. Specifically, we detail work on four core NLP tasks: IR, NER, LBD, and QA. We also describe work that directly addresses aspects of the pandemic through four additional tasks: topic modeling, sentiment and emotion analysis, caseload forecasting, and misinformation detection. For each task, we describe its scope, related datasets, methods, and future work, which is also summarized in **Figure 3**.

LITERATURE SEARCH AND INFORMATION RETRIEVAL

Introduction

IR is a set of techniques allowing users to easily and rapidly access relevant information contained in large text collections. With the rapid increase of unstructured textual information on the web in recent decades, this field has become increasingly significant, powering search engines such as Google and biomedical portals such as PubMed. However, recent events have made IR more relevant than ever. Indeed, the outbreak of the COVID-19 pandemic resulted in an explosion of scientific literature, reaching a peak in June 2020 with more than 8% of all scientific output dedicated to COVID-19 (31). More than 50,000 publications dedicated to the new coronavirus appeared in PubMed alone, as of October 2020. This so-called infodemic (32) led to the emergence of numerous resources dedicated to helping doctors, biomedical researchers, and the general public stay informed without being overwhelmed by the ever-increasing number of new publications. Among the first was LitCovid (13, 14), introducing an easy-to-use topics-based navigation, quickly followed by many others. These resources satisfy a variety of information needs. First, they keep clinicians up to date with their peers' experiences with various treatments (33), management strategies, and diagnostic methods (34). They also allow researchers to share important information on viral mechanisms (35), leading to the development of new vaccines. Finally, they allow public officials to plan and assess the efficacy of various preventive measures (36) while offering the general public a trusted source for current information on the pandemic. Here we present a representative collection of these systems (**Table 1**) and describe the different tasks

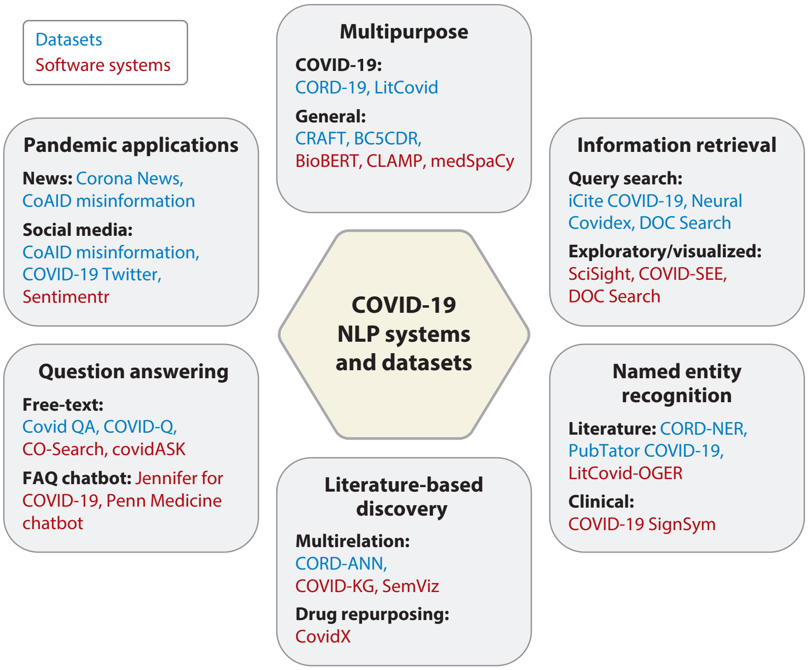

Figure 3

Selected natural language processing (NLP) datasets and systems addressing COVID-19. Expanded listings are provided in **Table 1**, **Table 2**, and **Supplemental Table 1**.

Supplemental Material >

involved in their creation. These tasks include (*a*) selecting all relevant and only relevant information, (*b*) efficiently preprocessing and storing the information, and (*c*) allowing users to quickly and easily access relevant documents.

Datasets

The first step of any literature resource is to decide on the scope of the collection of the relevant documents. This decision will impact the balance between exhaustiveness (allowing access to as many publications as possible) and focus (decreasing the number of publications users need to go through daily). For example, resources such as CORD-19 (COVID-19 Open Research Dataset) (37) agglomerate publications related to several coronaviruses [SARS-CoV, MERS (Middle East respiratory syndrome), SARS-CoV-2], while others such as LitCovid (13, 14) are strictly dedicated to SARS-CoV-2. In addition to the biological scope, literature portals differ in the type of publications integrated. Some such as LitCovid (13, 14) are strictly dedicated to published literature; others such as CORD-19 (37) supplement published literature with preprints from bioRxiv and medRxiv, while others still, such as DOC Search (5), include news sources and clinical trials. As a way to combine both exhaustivity and quality, tools such as COVID Scholar (38) allow the user to filter results to show only published research or only publications specific to COVID-19. The creation of CORD-19 by the Allen Institute was an important milestone, as it allowed existing search engines to be applied to the COVID-19 literature without each system needing to collect the relevant literature themselves and keep it updated. Following the release of CORD-19, several teams applied their existing search engines to CORD-19 data, including the Neural Covidex (39)

Table 1 A summary of representative datasets and tools commonly used in the general or biomedical domain and applied to COVID-19 applications

Dataset/system	Description	COVID-19 application examples
BC5CDR (28)	A dataset with 1,500 articles for disease and chemical NER in the literature	NER: training models to identify entities (29, 30)
BioBERT (31)	A transformer model pretrained on the articles from the general and biomedical domains	QA: identifying answer snippets from retrieved documents or passages (32)
BioSentVec (33)	A sentence embedding pretrained on PubMed abstracts and MIMIC-III clinical notes	IR: searching for related documents/passages (34, 35)
BioWordVec (36)	A word embedding pretrained on PubMed abstracts and MeSH terms	LBD: generating word representations (37)
CLAMP (38)	A toolkit for creating NLP pipelines for clinical text	Pandemic applications: identifying symptoms of COVID-19 with modifiers (body location, severity, etc.) (39)
CRAFT (40)	A manually annotated dataset of 67 full-text biomedical journal articles with 7 biomedical concepts	NER: training models to identify entities (41)
medSpaCy	A clinical NLP tool for named entity recognition and modifiers such as negation built on the spaCy framework	Pandemic applications: identifying COVID-19 status in electronic health records (42)
PubTator Central (43)	A web-based tool with preannotated concepts (e.g., gene, disease, chemical, cell line, species, variant) among all of PubMed/PMC	NER: recognizing biomedical concepts in text (44, 45) LBD: identifying biomedical concepts for knowledge discovery (46, 47)
RoBERTa (48)	A transformer model pretrained on articles from the general domain	Pandemic applications: feature extraction for forecasting COVID-19 case counts from news reports
SciBERT (49)	A transformer model pretrained on the articles from computer science and biomedical domains	LBD: generating word representations (50) Pandemic applications: identifying symptoms appearing before COVID-19 diagnosis (51)
Sentence-BERT (52)	A transformer model adapted to sentence pair–related tasks	IR: searching for related documents/passages (53)
Sentimentr (54)	A tool for calculating text polarity sentiment at the sentence level	Pandemic applications: classifying sentiments and emotions (55)
SQuAD (56)	A dataset of over 100,000 question–answer pairs from over 500 articles in the general domain	QA: training question-answering models (57)
UniLM (58)	A transformer model for natural language understanding and generation tasks	QA: generating answer snippets (32)

Abbreviations: BERT, bidirectional encoder representations from transformers; IR, information retrieval; LBD, literature-based discovery; MeSH, Medical Subject Headings; MIMIC-III, Medical Information Mart for Intensive Care; NER, named entity recognition; NLP, natural language processing; PMC, PubMed Central; QA, question answering.

and LIA (Ludwig Initiative Against COVID-19) (40). In contrast, resources that do not rely on CORD-19, such as LitCovid, require a significant curation effort to identify relevant publications.

Methods

Once all relevant publications have been collected, it is necessary to process them to optimize search and browsing. For users interested in a specific topic, such as a drug (e.g., remdesivir) or a group of patients (e.g., pediatrics), free-text search represents a natural access point to the growing literature. Some systems such as iSearch (41), 2019nCoVR (2019 Novel Coronavirus Resource) (42), or LitCovid (13, 14) use existing longstanding Lucene-based search solutions such as Solr or ElasicSearch. Search performance is improved by simple techniques such as stemming or lemmatization, which transform each term into its root form (allowing, for example, "treatment" to match "treatments") and the use of domain-specific dictionaries of synonyms. LitCovid (13, 14), for example, uses synonyms extracted from MeSH, allowing a query term such as "cancer" to match documents containing terms such as "tumor." Other resources such as COVID Scholar (38), the Neural Covidex (39), or LIA (40) use custom search engines. The Neural Covidex (39), for example, employs a traditional Lucene-based search engine Anserini (43) for the document retrieval step and BM25 ranking, and then reranks candidate documents using a modified T5-base deep learning model (44). In some tools such as DOC Search (5), the formulation of the search query is facilitated by smart autocomplete recognizing biomedical concepts such as drugs, species, anatomical parts, and genes.

In addition to search, facets offer an intuitive and easy way to explore available data. While most websites such as COVID-SEE (COVID Scientific Evidence Explorer) (45) provide basic facets easily extractable from publications' metadata such as journal, authors, and year of publication, more advanced faceting requires applying an NER system during the preprocessing step. LitCovid (13, 14), for example, offers faceting on the countries and chemicals mentioned in the article abstract, while DOC Search (5) includes types of clinical studies, age groups and genders, treatments, and patient outcomes. COVID-19 Intelligent Insight (14) goes even further by adding filters for coronavirus strains, proteins, genes, or species. The SciSight system (46) is designed exclusively around facet-powered exploratory search, where instead of starting with a free-text query, users immediately select and manipulate facets. Publications can also be assigned high-level categories for better organization, such as the "treatment," "diagnosis," and "prevention" categories used in LitCovid or the "prevention and mitigation measures" category in the COVID-19 World Information Aggregation system (47). The Global Literature on Coronavirus Disease system (48), provided by the World Health Organization, also allows the user to filter results by the main subject of the publication and the type of study. The annotations necessary to provide these features can be achieved using automated systems, manual curation, or a combination of both.

The lack of test collections specific to COVID-19 in the early months made evaluation problematic, leading some researchers to express frustration with their inability to quantify the quality of the results provided by their system (39). Shared tasks such as the TREC-COVID (Text Retrieval Conference-COVID) (49) addressed this by allowing participants to apply their tools to the same set of documents and queries and then have their results evaluated manually. However, while results ranking performance are important, it is not an absolute criterion for literature portals comparison since the presence or absence of specific features, coverage of the literature, and the ease of use of the user interface might have more impact on the user experience than variations in document ranking (39).

Applications

In addition to search and browsing capabilities, many portals offer distinct visualizations to facilitate the exploration of the COVID-19 literature. LitCovid (13, 14) displays a world map of countries mentioned in abstracts, and DOC Search (5) shows a matrix of co-occurrence among various drugs and patient symptoms and outcomes, helping clinicians to quickly identify the best treatment for their patients. iSearch (41) clusters publications and represents them as a FoamTree, while Neural Covidex (39) highlights the most pertinent passages in retrieved documents with the help of a BioBERT deep learning model (50) (**Table 2** and **Supplemental Table 1**). Resources also differ in the way they present search results. While most of the systems return the results as a list of publications, others such as LIA (40) return a list of matched sentences. Finally, in addition to interactive website access, most resources offer batch downloads and APIs (application programming interfaces) for automated data retrieval. For example, iSearch (41) allows users to export search results into CSV (comma-separated value) or Excel format with custom columns, and LitCovid (13, 14) allows users to subscribe to personalized RSS (really simple syndication) feeds based on a search query.

Supplemental Material >

Summary

The dramatic increase in COVID-19-related literature has led to the creation of numerous literature portals, leveraging NLP techniques to improve the performance of search and browsing and allowing users fast and easy access to the vast number of COVID-19 publications at this critical time. However, many challenges remain.

First, the faster review process for coronavirus-related publications is subject to abuse, with several manuscripts quickly retracted by journals. One of the most famous examples is the controversial publication linking 5G networks to the new coronavirus (51), which was highly criticized and quickly retracted by the journal. Coronavirus-dedicated literature portals, therefore, need to not only add new publications regularly but also rapidly update or remove publications if needed. Some resources delayed marking the controversial 5G publication as retracted, resulting in public backlash online. Second, scientists are overwhelmed with the sheer quantity of new articles (52). The expedited review process led to an increasing amount of short, low-value publications, mixed with high-quality reviews. Indeed, most publications have the lowest levels of evidence (53), and most publications are commentaries, with many authors simply sharing their views and opinions (53). To help users navigate the information overload, portals need to develop methods of prioritizing higher-quality publications. However, quality is inherently subjective, making it difficult to automate and any specific method is likely to be controversial. Finally, some journals have not made the full text of coronavirus-related articles freely available (52), limiting the amount of literature available.

NAMED ENTITY RECOGNITION

Introduction

NER is a fundamental step in NLP to downstream text-processing tasks. As one of the components in a traditional text-mining pipeline, NER is usually the first step to provide semantic interpretations of the unstructured text by locating and classifying the concept mentions. Early NER systems used rules and dictionaries (54). Later, machine learning (ML) approaches were widely applied to most bioconcepts [e.g., disease/chemical (55), gene (56)]. The NER task consists of two steps. The first step is to recognize the boundaries of the concept spans in the text. The second step is to link the concept spans with the specific concept in the dictionary [e.g., NCBI (National Center for Biotechnology Information) Gene (57)]. In general, every span would be assigned with an accession number [e.g., NCBI gene identifier] to represent the corresponding concept. The NER task incorporates multiple subtasks, including tokenization, sentence splitting, entity ambiguity, entity variation, etc. Different concept types encounter different levels of difficulty on each subtask. For example, the main challenge in recognizing chemicals is the high degree of variation in the concept names and chemical formulas, and the main challenge for gene concepts is their high degree of ambiguity due to the species variety.

Table 2 A summary of representative COVID-19-specific datasets and tools

Dataset/system	Type/source	Primary features	URL
IR			
LitCovid (59)	Published papers on SARS-CoV-2	Includes browsing by topics, chemical annotations, location extractions, and world map and weekly histogram	**https://www.ncbi.nlm.nih.gov/research/coronavirus/**
CORD-19 (60)	Published papers and preprints on SARS-CoV-2 and other coronaviruses	Includes over 130,000 articles	**https://cord-19.apps.allenai.org/**
iSearch (61)	Published papers and preprints on SARS-CoV-2	Clusters publications and provides for visualization as FoamTree	**https://icite.od.nih.gov/covid19/search/**
Neural Covidex (62)	Published papers, preprints, and clinical trials on SARS-CoV-2 and other coronaviruses	Highlights most pertinent passages in matching documents	**https://covidex.ai**
NER			
PubTator COVID-19	Published papers on SARS-CoV-2	Provides annotations for genes/proteins, drugs/chemicals, diseases, cell types, species, and genomic variants on CORD-19 and LitCovid	**https://github.com/ncbi-nlp/PubTator-Covid19**
CORD-NER (63)	Published papers and preprints on SARS-CoV-2 and other coronaviruses	Includes 75 fine-grained entity types on CORD-19	**https://xuanwang91.github.io/2020-03-20-cord19-ner/**
LitCovid-OGER-BB/LitCOVID-PMC-OGER-BB (64)	Published papers on SARS-CoV-2	Provides annotations on LitCovid using OGER and BioBERT, using the ontologies annotated in CRAFT and an additional dictionary of manually curated COVID-19 terms	**https://covid19.nlp.idsia.ch/**
COVID-19 SignSym (39)	Published papers on SARS-CoV-2	Provides a tool to identify COVID-19 symptoms in clinical records	**https://clamp.uth.edu/covid/nlp.php**
LBD			
CORD-ANN (65)	Published papers and preprints on SARS-CoV-2 and other coronaviruses	Consists of 500 manually annotated and over 10,000 automatically annotated sentences for entities and relations to build knowledge graphs	**https://github.com/knowledge-learning/cord19-ann**
COVID-19-Chemicals-Diseases (47)	Published papers on SARS-CoV-2	Lists the chemicals and diseases mentioned in LitCovid, including term and frequency	**https://github.com/amir-karami/COVID-19-Chemicals-Diseases**
SemViz (66)	Published papers and preprints on SARS-CoV-2 and other coronaviruses	Consists of a platform for the semantic visualization of multiple types of entities and relations extracted from the CORD-19 dataset	**https://www.semviz.org/**
COVID-KG (67)	Published papers and preprints on SARS-CoV-2 and other coronaviruses	Provides a knowledge discovery framework to extract fine-grained multimedia knowledge elements (entities, relations, and events) from scientific literature	**http://blender.cs.illinois.edu/covid19/**

(*Continued*)

Table 2 *(Continued)*

Dataset/system	Type/source	Primary features	URL
QA			
CovidQA (68)	Question–answer	Includes 124 COVID-19 question–article–answer triplets from 85 articles	**https://github.com/castorini/pygaggle/**
COVID-Q (69)	Question–category	Consists of 1,690 COVID-19-related questions annotated into 15 broad categories and 207 detailed subcategories	**https://github.com/JerryWei03/COVID-Q**
MQP (70)	Question–question	Provides 3,048 MQPs annotated as similar or different	**https://github.com/curai/medical-question-pair-dataset**
covidASK (57)	QA search engine	Includes efficient indexing, which enables deep learning QA models employed in real-time, and biomedical NER	**https://covidask.korea.ac.kr/**
Jennifer for COVID-19 (71)	FAQ chatbot	Consists of question–answer pair gold standard annotations via crowdsourcing; both English and Spanish supported	**http://bit.ly/jenniferai**
Pandemic applications			
Corona News	News	Includes 141,208 COVID-19-related news headlines	**https://systems.jhu.edu/research/public-health/ncov/**
CoAID	Social media and news	Provides misinformation ground truth labels	**https://github.com/cuilimeng/CoAID**

A detailed version of this table is provided in **Supplemental Table 1**. Abbreviations: FAQ, frequently asked questions; IR, information retrieval; LBD, literature-based discovery; MQP, medical question pairs; NER, named entity recognition; NLP, natural language processing; OGER, OntoGene's Entity Recognizer; PMC, PubMed Central; QA, question answering; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Datasets

Supplemental Material >

Since May 2020, PubTator (58), a text-mining system with high-quality automatic bioconcept annotations, has provided automatic named entity annotations for most relevant concepts (e.g., gene, disease) for the two major well-known COVID-19-related datasets [i.e., CORD-19 (37) and LitCovid (13, 14)]. PubTator updates the annotations for these corpora daily for both abstracts and full text when available. In addition, the CORD-NER corpus (59) provides automated annotations on the CORD-19 dataset, which covers 75 fine-grained entity types (including gene, disease, and chemical). Details and links for both sets are listed in **Table 2** and **Supplemental Table 1**.

Methods

In most cases, the methods of the two steps of the NER are highly independent, since the tasks of recognizing the mention boundary and linking the mention to the concept can be treated individually. However, each of the steps can utilize information from the other. The methods for recognizing mentions in text can be summarized as follows.

1. Dictionary-based methods use name lists or dictionaries containing synonyms to recognize matches in the text.
2. Rule-based and regular expression methods are usually applied for named entities that follow a constant nomenclature [e.g., tmVar (60) for genomic variants] or named entities with highly frequent co-occurring strings, prefixes, or syntactic tags.

3. ML methods treat the challenge as a sequence labeling problem to recognize the boundaries of the mentions. The method needs a set of documents with manual annotations to train a model. The most successful ML approach for biomedical concept recognition was conditional random fields (CRF) (61), which has been heavily applied to many biomedical concepts, although a wide variety of ML models have been used, such as the semi-Markov model used by TaggerOne (62). In the last few years, deep learning approaches have been used to optimize NER performance. The two most common methods are biLSTM (bidirectional long short-term memory)+CRF (63) and BERT (bidirectional encoder representations from transformers) (64). Unlike CRF, biLSTM+CRF and BERT take advantage of language modeling to address the probability distribution and provide context to distinguish among words and phrases.

Unlike NER, the development of a normalization method is highly dependent on the specific characteristics of the bioconcepts in question, a problem that is more difficult to address than NER. For example, species assignment is the main challenge of gene normalization but is not critical for other bioconcepts. Thus, normalization steps in COVID-19 research have mostly applied existing tools [e.g., TaggerOne (62)] or dictionary lookup [e.g., COVID-19 SignSym (65)]. The primary challenges for normalization are term variation and ambiguity. Focusing on articles on COVID-19 can significantly narrow the scope of the candidate concepts to the recognized span in the text. Therefore, a few studies (59, 66) have normalized spans to self-defined categories instead of using specific concept identifiers. As an example, in CORD-NER, cough and vomiting are both categorized as symptoms of COVID-19.

System

Supplemental Material >

To support downstream NLP research on COVID-19, several studies have addressed NER specifically on the CORD-19 or LitCovid datasets. As shown in **Table 2** and **Supplemental Table 1**, Wang et al. (59) used SciSpacy with distant supervision to extract multiple concepts from the CORD-19 dataset, including gene, disease, and chemical concepts. Colic et al. (67) used OGER (OntoGene's Entity Recognizer; a dictionary-lookup approach with fuzzy matching) and BioBERT for the concept recognition in LitCovid. The COVID-19 SignSym system (65) extracts mentions of signs and symptoms from clinical text (EHRs), along with eight associated attributes, and normalizes them to LOINC (Logical Observation Identifiers Names and Codes) codes using a dictionary-lookup approach. The performance of the NER is highly dependent on the evaluation method. Existing tools (e.g., TaggerOne) achieve 80–85% in F-measure, but if normalizing the spans to the self-defined categories, the performance is above 90% (65).

Summary

Due to the labor required to create training corpora, developing a training corpus for a customized NER tagger for COVID-19 in a short period of time is challenging. Most studies have preferred using existing tools (e.g., PubTator) and methods (e.g., BioBERT) or a dictionary lookup approach for NER tasks in COVID-19-relevant articles rather than developing new methods. A detailed description of these tools and methods is shown in **Table 1**. The performance of these tools and methods for COVID-19 still has room for improvement. A practical idea for quickly developing a customized system without manual curation is to apply distant supervision (59) to quickly expand the entity dictionary or training corpus. Another solution for creating a training corpus is to manually refine automated annotations made by existing tools. Our previous work (68) demonstrated that preannotated entities can significantly improve the efficiency and accuracy of manual annotation.

LITERATURE-BASED DISCOVERY

Introduction

LBD, also known as hypothesis generation, is the discovery of new knowledge based on known facts derived from the literature. Specifically, LBD is usually described as the process of connecting two pieces of already known knowledge previously regarded as unrelated (69). For example, in the Swanson ABC co-occurrence model (70), if text is found that explicitly states the knowledge that "A is associated with B" and "B is associated with C," then the implicit knowledge of "A may be associated with C" is discovered. In the biomedical domain, LBD enables the discovery of implicit knowledge that may advance biomedical research, making it an important high-level task for biomedical NLP. The computational techniques used for LBD in the biomedical domain have been recently surveyed (71). Here, we focus on LBD works addressing COVID-19, including related datasets, methods, and applications.

Datasets

The CORD-19 and LitCovid datasets, as shown in **Table 2**, are widely used for LBD studies on COVID-19, but these datasets do not provide annotations suitable for training LBD systems. However, CORD-ANN (artificial neural network) is a manually annotated corpus (72) of COVID-19 research created for training LBD systems. This corpus contains 500 sentences selected from CORD-19, with a total of 10,201 entity annotations and 9,444 relation annotations.

Methods

Most work on LBD adopts the following outline: concept extraction, hypothesis generation, and evaluation of results. First, biomedical entity recognition tools [e.g., PubTator (58) and CORD-NER (59)] are used to identify the entity concepts from text. Then, the discovery models are used to find new associations between the target concepts. Finally, the novel hypothesis is evaluated and analyzed. These LBD methods can be divided into two types: Co-occurrence methods directly use co-occurrences in text as relationships between concepts, while distributional methods first represent concepts using context vectors and then find implicit relationships in vector space via vector operations and nearest-neighbor search.

Most work on LBD for COVID-19 is based on co-occurrence. The first concept co-occurrence methods used concept co-occurrences to generate linking and target terms. For example, Pinto et al. (73) first used PubTator to annotate the disease, gene, and species concepts from 100,000 COVID-19 related papers, and then the human gene–disease co-occurrences supported by at least four papers were retained. They analyzed these associations and found that angiotensin-converting enzyme 2 was highly expressed in the lungs of patients with comorbidities associated with severe COVID-19. Tarasova et al. (74) discovered 46 proteins related to both HIV-1 and COVID-19 from biomedical texts using the concept co-occurrence method. Karami (75) used frequency analysis to understand the significance of chemical and disease concepts; co-occurrence analysis could then assist in identifying relationships between entities in COVID-19 literature.

In addition to the co-occurrence methods, deep learning–based distributional embedding methods have been proposed for LBD. In these methods, distributional embedding methods [e.g., BioWordVec (76) and SciBERT (77)] are first used to construct vector representations of terms, which can learn co-occurrence information from a large amount of text. Then the semantic similarity measures can be leveraged to derive new scientific knowledge from already existing knowledge. For example, contextual embeddings from the SciBERT model that was pretrained on a large multidomain corpus of scientific publications and fine-tuned on the CORD-19 dataset have been leveraged to discover latent COVID-19 therapy targets in the scientific literature (78). BioWordVec has also been used to mine the extensive biomedical literature for treatments to SARS that may also then be appropriate for COVID-19 (79).

Applications

To facilitate knowledge discovery from literature, researchers have developed some applications to offer distinct visualization analyses. For example, SemViz (80) is a platform for semantic visualization of multiple types of entities and relations extracted from the CORD-19 dataset. These entities and relations can be visualized in many ways, e.g., as word clouds, heat maps, or graphs. This system enables the discovery of novel inferences over relations in COVID-19-related data. Yeganova et al. (81) analyzed the LitCovid collection by applying state-of-the-art NER, classification, and clustering techniques. These analyses have produced a comprehensive, synthesized view of COVID-19 research to facilitate knowledge discovery from literature. Furthermore, LBD has important applications in drug repurposing, the process of finding new applications for existing drugs. The core objective is to discover drug–gene–disease interaction evidence from biomedical literature for drug repurposing. For example, Gates & Hamed (82) used NLP techniques to construct a drug–entity network from biomedical literature, and then proposed a novel ranking algorithm, CovidX, to recommend existing drugs for potential repurposing. Patel et al. (83) built a disease–gene–drug tripartite network from biomedical text and then, using the network to identify potential new purposes for drugs already approved, revealed six drugs likely to treat comorbid symptoms of COVID-19 patients. A novel and comprehensive knowledge discovery framework, COVID-KG, was developed to extract fine-grained multimedia knowledge elements (entities, relations, and events) from scientific literature, and then was used to generate a comprehensive report for drug repurposing (84). Successful LBD applications can provide essential help for drug repurposing, which may enable the development of potential drugs for the cure of COVID-19.

Summary

Most COVID-19-related work on LBD uses the co-occurrence model, but it is difficult to capture the complexity of biomedical processes with this simple model. Alternatively, deep learning–based distributional embedding methods have also been used to capture complex semantic associations, achieving better performance. However, deep learning models are limited by high interpretability requirements, as it is essential that novel hypotheses are explainable. Enriching LBD technologies with explainable biomedical context is important and challenging. Many existing LBD methods and systems have been applied to the drug repurposing task, but obtaining reliable and convincing drug hypotheses in real application settings remains challenging.

QUESTION ANSWERING

Introduction

QA is the NLP application that accepts a question as input, often in natural language, and outputs a ranked list of related answers or a summarized answer snippet (85). It is a joint task, merging techniques from IR (retrieving related documents or passages for a given question), text summarization (summarizing answers among related passages), and LBD (finding related entities mentioned in a question over knowledge bases).

There are two broad categories of QA systems: IR based and knowledge based (86). The main difference is that the former retrieves answers over free text whereas the latter is on structured databases such as ontologies and knowledge bases. To date, most QA systems on COVID-19 are IR based. IR-based QA systems generally consist of three modules (85). The first module is question processing: The question is preprocessed or reformatted. The question type might be classified at this step as well (e.g., is the question asking about COVID-19 transmissions or treatments?). The second module is document/passage retrieval: finding related documents or passages regarding to the question. The third module is answer generation: constructing the final answers from the retrieved documents or passages.

QA is a critical tool for addressing information needs during the COVID-19 pandemic, reducing the load on healthcare professionals by automatically answering questions 24 hours a day and providing expert-curated FAQs (frequently asked questions) to combat misinformation. QA is also a mature NLP application, allowing QA systems to be quickly applied to the pandemic. For example, the Perelman School of Medicine at the University of Pennsylvania created a COVID-19 chatbot and made it publicly available within two weeks (87). In this section, we survey QA-related datasets, methods, and systems specifically addressing COVID-19.

Datasets

Supplemental Material >

Table 2 and **Supplemental Table 1** describe the existing COVID-19-related QA datasets. The datasets can be categorized into three groups based on the data type: (*a*) question–answer, where each instance is a question–answer pair, the most common QA data type; (*b*) question–category, where each instance is a question and its annotated question type; and (*c*) question–question, where each instance is a pair of questions and whether they are semantically similar. Here we describe representative datasets. For instance, CovidQA (42) consists of 124 question–article–answer triplets derived from 85 unique articles in CORD-19, covering 27 categories. Five curators provided annotations by synthesizing questions from the categories provided by the CORD-19 Kaggle task organizers, and then manually identifying the relevant documents and locating answers.

In contrast, COVID-Q (88) consists of 1,690 COVID-19-related questions that are annotated into 15 broad categories and 207 detailed question classes. Multiple curators annotated the dataset in three phases. First, two curators discussed and assigned questions into categories. Second, an external curator verified and suggested category changes if required. Third, questions assigned to more than four question classes were further sampled and assigned to three Mechanical Turk workers. The majority vote was used for validation.

Another dataset of a different type is MQP (89). It consists of 3,048 question pairs collected from the medical domain (i.e., not specific to COVID-19) that are manually labeled as similar or different by medical doctors. Two doctors participated in the annotation with an agreement of over 85% on the 836 question pairs in a test set. While these question pairs are from the general medical domain, we included this dataset because its creators have constructed it for the purpose of matching COVID-19 questions and already deployed a system using the trained models. Given a COVID-19 question, the system finds the most similar question in the FAQ pages from popular COVID-19-related websites and returns the answers accordingly.

Methods and Systems

We surveyed the existing COVID-19 QA systems and outlined them in **Table 2** and **Supplemental Table 1**. There are four QA search engines and two FAQ chatbots publicly available. The methods used by FAQ chatbots are relatively more straightforward. These methods use a corpus of curated question–answer pairs. Given a user question, the system finds the most similar question in its corpus (often by string matching) and returns the answer to that question if it is similar enough (90). In contrast, QA search engines work beyond question–answer pairs. As mentioned, the general pipeline to develop a QA search engine consists of question-processing, document/passage retrieval, and answer-generation modules. The detailed methods for each module are described below.

Question Processing

The question-processing methods used in the COVID QA search engines can be categorized into three groups. First, traditional text-processing methods, arguably the most common methods, are applied to user questions. Most systems use case folding (e.g., convert all the words to lower case), lemmatization (transforming a word into a valid base word; e.g., "testing" becomes "test"), and removing stop words (common words such as "a," "an," and "the"). The CAiRE-COVID system also applies sentence simplification, where complex sentences are transformed into several shorter and simpler sentences (91).

Second, model-driven processing methods are also used where a question needs to be processed into a compatible format for the later ranking models. This is particularly for the COVID QA search engines using BERT models. BERT models need the inputs that are processed in a specific format such that each word is mapped to an ID and special symbols are used to denote the start and end of a question.

Third, domain-specific processing methods are applied where the method is tailored to the biomedical domain. For example, covidASK used BEST (Biomedical Entity Search Tool) (92), an NER tool, to annotate the biomedical entities mentioned in a user question.

Document/Passage Retrieval

The document/passage retrieval module retrieves the documents/passages that are relevant to the processed questions. The methods used are essentially the same as those described in the section titled Literature Search and Information Retrieval. The existing COVID QA search engines used (*a*) traditional IR methods, such as BM25 in CAiRE-COVID; (*b*) deep learning–based methods, such as Sentence-BERT (93) in RECORD; and (*c*) both, such as BM25 and Sentence-BERT in CO-Search.

Answer Generation

The answer-generation module produces final answers as output by identifying answer snippets—the text probably containing the answers—from the documents or passages retrieved. The existing COVID QA search engines use BERT-related models to identify the answer snippets. The shared approach is training a BERT model on a QA dataset from other domains but has many more instances than the existing COVID QA datasets. Most of them use SQuAD (Stanford Question Answering Dataset) (94), consisting of over 100,000 question–answer pairs from over 500 articles in the general domain. Other datasets include HotpotQA (consisting of over 110,000 question–answer pairs from Wikipedia articles) and PubMedQA (95) (consisting of 1,000 manually annotated and over 250,000 automatically annotated question–answer pairs from PubMed articles). The models are trained in these datasets and then applied directly.

In addition to identifying the answer snippets, CAiRE-COVID and CO-Search also automatically generate a summary from the selected answer snippets. In general, there are two types of text summarization methods: extractive summarization, where the key sentences from the original text are selected as a summary, and abstractive summarization, where the original text is significantly rewritten. CAiRE-COVID uses both methods. For extractive summarization, it selects the top three answer snippets based on the cosine similarity between a question and the identified answer snippets using the BERT model. For abstractive summarization, similar to identifying answer snippets, it trained two deep learning models, UniLM (96) and BART (97), on other larger datasets and then applies them to the COVID-19 context. In contrast, Co-Search uses BERT and GPT-2 (98) for abstractive summarization.

Summary

We have provided a detailed summary of NLP methods to tackle COVID-19 QA. It is impressive that QA systems were in production within two weeks to respond to the pandemic. The range of publicly available QA datasets and systems specific to COVID-19 represents a significant community effort.

We identify two primary challenges for further improvement. First, to date, the scale of COVID-19 QA datasets is still too limited to directly train deep learning models. The existing QA systems have directly applied the models trained from datasets of other domains. This is arguably the main bottleneck for identifying the most relevant answer snippets. A potential solution is to use a combination of manual and automatic annotations to generate weakly supervised instances for training, similar to the PubMedQA dataset (95). Second, the COVID-19 QA systems lack evaluations of the user experience. For instance, most of the COVID-19 QA systems use deep learning–based models for searching for relevant documents/snippets and identifying answer snippets. Critically, these models take significantly more time for inference than traditional models, which is a concern for production systems. More thorough evaluations that incorporate other aspects of the user experience are needed.

PANDEMIC-ORIENTED APPLICATIONS

Introduction

The wide variety of information needed during the COVID-19 pandemic has motivated a diverse assortment of task-specific NLP applications. The specific purpose of these applications and the methods they use vary widely, but they are broadly differentiated by whether they seek to inform biomedical or public health research and whether the information sought is specific, such as a forecast of the number of COVID-19 cases, or open ended, such as the reasons for noncompliance with social distancing orders. These applications supply many case studies in applied NLP, ranging from crucial information provided by well-known classical approaches to novel tasks addressed by cutting-edge NLP methods. As the COVID-19 pandemic is a worldwide concern, this section highlights research in languages besides English where possible.

Methods

Since so many of the difficulties caused by the COVID-19 pandemic are new or poorly understood, many of the information needs are open ended. Topic modeling is an NLP method that can provide a qualitative summary of a text dataset by identifying subjects that appear frequently; these can then be further analyzed if needed by stratifying the data, for example, by time or geographic location. While topic modeling is commonly used in social media or news, one recent article applied it to the scientific literature to identify topics that have received less attention in

SARS-CoV-2 research compared to the research on other coronaviruses (99). This study found that the SARS-CoV-2 literature is comparatively heavy on topics related to public health and clinical care rather than basic microbiology such as pathogenesis and transmission, suggesting that these may represent research opportunities. Topic modeling has also been used to identify useful unpublished clinical knowledge from social media posts by physicians (100). This study identified eight topics; the most common topic was actions and recommendations, followed by warnings about misleading information.

Applications of topic modeling for public health include a study that identified reasons for noncompliance with social distancing orders by using a straightforward form of topic modeling based primarily on word counts (101); the study found that the most common reasons include nonessential work and concerns over the mental and physical health implications of social distancing. A well-known but more sophisticated method, latent Dirichlet allocation, has been used to categorize public policy trends in India to understand the effectiveness of different policy changes (102). This study found that the aspects of public health messaging most strongly associated with behavior change are its consistency and breadth.

Many of the challenges caused by the COVID-19 pandemic are not caused directly by the SARS-CoV-2 virus and thus may be difficult to study. For example, social distancing mandates to prevent the spread of COVID-19 cause disruptions in day-to-day living, financial hardship, and isolation, which can all have serious negative effects on mental health (103). A study using Twitter posts quantified stress levels in the United States over time using text classification but also used topic modeling to identify the primary causes (104). This study demonstrated a strong correlation in major US cities between increased stress levels and the number of COVID-19 cases; stress levels were initially driven by fear of infection and widespread panic but later shifted primarily to financial concerns. Another study used latent Dirichlet allocation to extensively analyze topic trends in posts on Reddit, a news aggregation and discussion website (105). This study found that both anxiety and suicidality increased substantially across the site during the pandemic; however, the increase within mental health communities was particularly strong.

In addition to topic modeling, sentiment analysis has also been used to identify emotions and understand public opinion surveillance (106). This study identifies the emotions elicited from the COVID-19 comments on Reddit. Similar studies were also performed in different social media platforms, such as Twitter and Weibo (a Chinese social media site) (107–111). Due to the lockdown in many areas, the Internet has become a primary place to express feelings and opinions, and the number of COVID-19-related posts has increased dramatically. **Table 2** and **Supplemental Table 1** summarize related datasets and methods on COVID-19 sentiment analysis. For instance, there were over 5 million tweets in May related to workplace and school reopening (112). Such a massive amount of data burdens manual interpretation and leads to the development of related NLP tools. Automatically tracking the emotions and sentiments over the period and across regions would facilitate how COVID-19 impacts people's well-being and lead to more effective decision-making (111, 113).

Supplemental Material >

Many important clinical questions can be addressed by analyses that build on NER, often over a relatively limited set of concepts. For example, one study attempted to identify early symptoms indicative of COVID-19 by applying the SciBERT system to identify symptoms recorded in EHRs during the week prior to a COVID-19 diagnosis (114). This study noted strong associations with anosmia/dysgeusia and fever/chills. Another study determined whether symptoms could be used to decide whether an individual should be tested for COVID-19 by applying a rule-based system to EHRs (115). This study, in contrast, concluded that despite the strong association with anosmia and dysgeusia, which are both uncommon, initial symptoms of

COVID-19 are typically nonspecific, diminishing the effectiveness of symptom-based screening. The COVID-19 SignSym system identifies COVID-19 signs and symptoms in hospitalized patients using an adaptation of the clinical NLP tool CLAMP (Clinical Language Annotation, Modeling, and Processing) (116) but also extracts the attributes body location, severity, temporal expression, subject, condition, uncertainty, negation, and course (117). Another study improved follow-up by identifying individuals whose COVID-19-positive status was recorded in the EHR free-text narrative rather than the structured attributes by adapting the medSpaCy system (118). Fries et al. (119), in contrast, introduced a novel weak supervision approach for NER to identify COVID-19 symptoms in EHR records. These authors also considered the novel task of identifying exposure to COVID-19, reporting a final F1 score of 80.1%. Finally, an analysis of the symptom progression of COVID-19-positive individuals over many weeks using Reddit and Facebook groups found that symptoms persist for 90 days or more in many individuals and often include symptoms not officially associated with COVID-19 (120).

Applications

Effectively managing clinical resources during the pandemic requires an accurate forecast of the number of cases within a geographical area. One group used a list of keywords to identify posts related to COVID-19 on Weibo, and then created a classifier to differentiate between posts that report a SARS-CoV-2 infection and posts that merely mention COVID-19 (121). These methods are well known, but they were used to accurately predict daily case counts earlier than the official statistics by up to 14 days. Another system demonstrated that caseload forecasts can be improved by incorporating multiple data sources (122), using the RoBERTa (Robustly Optimized BERT Pretraining Approach) transformer language model to extract features from news reports, and then combining these in a long short-term memory model with other information sources to model the caseload forecast.

Combating misinformation, whether created deliberately or by mistake, has become an important task during the pandemic. Inaccurate information can spread quickly through social media networks, and most users are ill equipped to identify science-based misinformation, especially during a rapidly evolving crisis. However, automated misinformation detection is a complex task, and we identified a wide variety of approaches. One study stopped short of detecting misinformation directly and instead attempted to predict how many times a post on the social media site Weibo will be reposted, in order to allow these to be prioritized for review. This method employed supervised classification, using features such as words conveying emotion (123). Another approach detected misinformation by training a language model on a large amount of text about COVID-19 from a reliable source, and then used the model to calculate the perplexity of the potentially misinformative text, which is a measure of surprise (124). Misinformation will typically have a high perplexity because it uses vocabulary and phrases that differ significantly from the reliable text used to train the model. Other work trained a deep learning model to directly differentiate between reliable and unreliable assertions about COVID-19 (125). Another system identified YouTube videos containing conspiracy theories about the origin of SARS-CoV-2 (such as being caused by 5G cellular networks) by analyzing the transcript of the video using a supervised ML approach (126). One sophisticated approach for identifying fake news converts facts extracted from a news item into logical propositions, combines this with the logical form of an ontology of reliable COVID-19 information, and then evaluates the joint logical system for inconsistencies (127). Finally, CoAID is a COVID-19 misinformation dataset containing annotations of accurate and inaccurate information, extracted from both news and social media posts (128).

Summary

The far-reaching disruptions caused by the COVID-19 pandemic have created a variety of information needs that can be addressed using NLP. In this section, we described applications of topic modeling, sentiment and emotion analysis, and analyses created with NER over a relatively limited set of concepts. We also described NLP studies addressing caseload forecasting and misinformation detection. The methods used in these studies vary widely in sophistication, with some critical information provided by well-understood classical methods, while others demonstrated the adaptability of cutting-edge methods on novel tasks.

CONCLUSION

The COVID-19 pandemic has had a widespread impact on society through increased mortality and morbidity, disruptions to everyday life, and general uncertainty. Many of these difficulties are novel in type, scale, or cause, and one of the primary ways to address them is with access to better information, that is, the right amount of accurate information at the point where it can be put into action. In this review we have surveyed NLP research either using biomedical research as input or being used to inform biomedical research. This review also covers NLP on social media and EHR data; however, these are not the primary focus and more comprehensive reviews on these domains are needed.

Considering NLP work to address the COVID-19 pandemic as a whole, we see several overall themes. First, many of the existing tasks in NLP can directly address information needs during the COVID-19 pandemic. IR and QA approaches directly address information needs identified by the user; IR primarily by providing documents and QA by enabling both queries and results in natural language. NER provides a foundation for semantic interpretation of text. LBD identifies potentially new knowledge by combining information extracted from scientific articles, often augmented with additional knowledge bases. Topic modeling identifies the most common subjects in a text while sentiment analysis quantifies positive or negative affect, and both enable further analysis through stratification and summarization. Misinformation detection is a complex, high-level NLP task that has found important direct applications during the pandemic.

Second, we see a trend toward addressing aspects of the COVID-19 pandemic through the rapid adaptation of existing systems. This adaptation has been achieved with a variety of methods, ranging from classical methods—which often require no training—to complex ML systems retrained with a combination of manual and automatic annotations. We observe that both extremes of this dichotomy have advantages for rapid adaptation: Simple methods such as dictionaries are easily extended, while state-of-the-art ML methods employing either weak supervision or transfer learning methods require relatively few manually annotated examples. A notable class of these ML systems are deep learning transformer models, which primarily train on unlabeled data and require relatively little manually annotated data for fine-tuning.

Third, we see a strong trend toward creating datasets that address aspects of the COVID-19 pandemic. The largest among these datasets consist of text that is related to the pandemic, such as collections of scientific articles or social media posts. These are typically gathered with support from automated NLP tools, and derived datasets often provide additional automated annotations. Unsurprisingly, datasets requiring significant manual effort are less common, such as those for complex tasks like QA and LBD or large datasets for tasks like NER or sentiment analysis.

Finally, we see many efforts to address tasks that the COVID-19 pandemic brought to the forefront or that have novel aspects. These tasks include identifying topics understudied in the literature, analyzing long COVID cases from social media, identifying mentions of exposure to

SARS-CoV-2, forecasting caseloads, providing feedback on public health initiatives, quantifying mental health effects of the pandemic, and addressing misinformation. Despite some novel aspects, these efforts build upon a significant amount of existing NLP work, often by mixing classical methods and cutting-edge techniques.

However, significant challenges remain. While many projects have addressed different aspects of the COVID-19 pandemic, these efforts have largely remained fragmented, although notable exceptions include high-profile resources such as CORD-19. NLP systems frequently use a pipeline architecture so that systems that address high-level tasks typically incorporate other systems. While newer deep learning models provide end-to-end training for some tasks, these architectures remain under development for many important NLP tasks. Discovering projects that provide useful software or data that can be incorporated and reused takes time, as do characterizing and integrating the systems. While this review has provided a useful overview of much of the work to date, improved interoperability and easier methods to discover relevant software and data artifacts would be beneficial.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research is supported by the NIH Intramural Research Program, National Library of Medicine.

LITERATURE CITED

1. WHO (World Health Organ.). 2020. *Pneumonia of unknown cause—China*. Dis. Outbreak News, WHO, Geneva. **https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/**
2. WHO (World Health Organ.). 2020. *Novel coronavirus (2019-nCoV): situation report, 22*. Tech. Rep., WHO, Geneva. **https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf**
3. WHO (World Health Organ.). 2020. *Novel coronavirus (2019-nCoV): situation report, 10*. Tech. Rep., WHO, Geneva. **https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf**
4. Stegmann J. 2020. MeSH descriptors indicate the knowledge growth in the SARS-CoV-2/COVID-19 pandemic. arXiv:2005.06259 [cs.DL]
5. Nowakowska J, Sobocińska J, Lewicki M, Lemańska Ż, Rzymski P. 2020. When science goes viral: the research response during three months of the COVID-19 outbreak. *Biomed. Pharmacother.* 129:110451
6. Brainard J. 2020. New tools aim to tame pandemic paper tsunami. *Science* 368:924–25
7. Valika TS, Maurrasse SE, Reichert L. 2020. A second pandemic? Perspective on information overload in the COVID-19 era. *Otolaryngol. Head Neck Surg.* 163(5):931–33
8. Soltani P, Patini R. 2020. Retracted COVID-19 articles: a side-effect of the hot race to publication. *Scientometrics* 125:819–22
9. Yeo-Teh NSL, Tang BL. 2020. An alarming retraction rate for scientific publications on coronavirus disease 2019 (COVID-19). *Account. Res.* 28(1):47–53
10. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, et al. 2020. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *Lancet Psychiatry* 7:547–60

11. Tasnim S, Hossain MM, Mazumder H. 2020. Impact of rumors and misinformation on COVID-19 in social media. *J. Prev. Med. Public Health* 53:171–74
12. Rathore FA, Farooq F. 2020. Information overload and infodemic in the COVID-19 pandemic. *J. Pak. Med. Assoc.* 70:S162–65
13. Chen Q, Allot A, Lu Z. 2020. Keep up with the latest coronavirus research. *Nature* 579:193
14. Xu J, Kim S, Song M, Jeong M, Kim D, et al. 2020. Building a PubMed knowledge graph. *Sci Data* 7(1):205
15. Jensen LJ, Saric J, Bork P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* 7:119–29
16. Rzhetsky A, Seringhaus M, Gerstein M. 2008. Seeking a new biology through text mining. *Cell* 134:9–13
17. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, et al. 2008. Text mining for biology-the way forward: opinions from leading scientists. *Genome Biol.* 9:S7
18. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, et al. 2018. Best match: new relevance search for PubMed. *PLOS Biol.* 16:e2005343
19. Markowetz F. 2017. All biology is computational biology. *PLOS Biol.* 15:e2002050
20. Zhao S, Su C, Lu Z, Wang F. 2020. Recent advances in biomedical literature mining. *Brief. Bioinform.* 2020:bbaa057
21. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. 2019. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* 16:505–7
22. Huang L-C, Ross KE, Baffi TR, Drabkin H, Kochut KJ, et al. 2018. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Sci. Rep.* 8:6518
23. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, et al. 2015. Utilizing social media data for pharmacovigilance: a review. *J. Biomed. Inform.* 54:202–12
24. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J. 2020. Challenges and opportunities for public health made possible by advances in natural language processing. *Can. Commun. Dis. Rep.* 46:161–69
25. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, SJPcb Brunak. 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Comput. Biol.* 14:e1005962
26. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. 2012. Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* 13:829–39
27. WHO (World Health Organ.). 2020. COVID-19 coding in ICD-10. Slideshow, WHO, Geneva. **https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf**
28. NLM (U.S. Natl. Lib. Med.). 2020. New MeSH supplementary concept record for coronavirus disease 2019 (COVID-19). *NLM Technical Bulletin*, Feb. 14. **https://www.nlm.nih.gov/pubs/techbull/jf20/brief/jf20_mesh_novel_coronavirus_disease.html**
29. Jiang S, Shi Z, Shu Y, Song J, Gao GF, et al. 2020. A distinct name is needed for the new coronavirus. *Lancet* 395:949
30. Leaman R, Lu Z. 2020. A comprehensive dictionary and variability analysis of terms for COVID-19 and SARS-CoV-2. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, ed. K Verspoor, KB Cohen, M Conway, B de Bruijn, M Dredze, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
31. Odone A, Galea S, Stuckler D, Signorelli C, Univ. Vita-Salute San Raffaele COVID-19 Lit. Monit. Work. Group. 2020. The first 10,000 COVID-19 papers in perspective: Are we publishing what we should be publishing? *Eur. J. Public Health* 30(5):849–50
32. Gazendam A, Ekhtiari S, Wong E, Madden K, Naji L, et al. 2020. The "Infodemic" of journal publication associated with the novel coronavirus disease. *J. Bone Joint Surg. Am.* 102:e64
33. Rabby MII. 2020. Current drugs with potential for treatment of COVID-19: a literature review. *J. Pharm. Pharm. Sci.* 23:58–64
34. Siordia JA Jr. 2020. Epidemiology and clinical features of COVID-19: a review of current literature. *J. Clin. Virol.* 127:104357

35. Srivastava S, Verma S, Kamthania M, Kaur R, Badyal RK, et al. 2020. Structural basis for designing multiepitope vaccines against COVID-19 infection: in silico vaccine design and validation. *JMIR Bioinform. Biotech.* 1:e19371
36. Keeling MJ, Hollingsworth TD, Read JM. 2020. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J. Epidemiol. Community Health* 74:861–66
37. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, et al. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, ed. K Verspoor, KB Cohen, M Dredze, E Ferrera, J May, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
38. Trewartha A, Dagdelen J, Huo H, Cruse K, Wang Z, et al. 2020. COVIDScholar: an automated COVID-19 research aggregation and analysis platform. arXiv:2012.03891 [cs.DL]
39. Zhang E, Gupta N, Nogueira R, Cho K, Lin J. 2020. Rapidly deploying a neural search engine for the COVID-19 open research dataset: preliminary thoughts and lessons learned. arXiv:2004.05125 [cs.CL]
40. Ludwig. 2020. Welcome to LIA: Ludwig initiative against COVID-19. **https://covid19.ludwig.guru/**
41. NIH OPA (Natl. Inst. Health Off. Portf. Anal.). 2020. *iSearch COVID-19 portfolio*. Web Resour., NIH, Bethesda, MD. **https://icite.od.nih.gov/covid19/search/**
42. Zhao WM, Song SH, Chen ML, Zou D, Ma LN, et al. 2020. The 2019 novel coronavirus resource. *Yi Chuan* 42:212–21
43. Yang P, Fang H, Lin J. 2017. Anserini: enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1253–56. New York: Assoc. Comput. Mach.
44. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, et al. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910.10683 [cs.LG]
45. Verspoor K, Šuster S, Otmakhova Y, Mendis S, Zhai Z, et al. 2020. COVID-SEE: scientific evidence explorer for COVID-19 related research. arXiv:2008.07880 [cs.CL]
46. Hope T, Portenoy J, Vasan K, Borchardt J, Horvitz E, et al. 2020. SciSight: combining faceted navigation and research group detection for COVID-19 exploratory scientific search. arXiv:2005.12668 [cs.IR]
47. Aizawa A, Bergeron F, Chen J, Cheng F, Hayashi K, et al. 2020. A system for worldwide COVID-19 information aggregation. arXiv:2008.01523 [cs.CL]
48. WHO (World Health Organ.). 2020. *Global research on coronavirus disease (COVID-19)*. Tech. Rep, WHO, Geneva. **https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov**
49. Roberts K, Alam T, Bedrick S, Demner-Fushman D, Lo K, et al. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *J. Am. Med. Inform. Assoc.* 27(9):1431–36
50. Lee J, Yoon W, Kim S, Kim D, Kim S, et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36:1234–40
51. Fioranelli M, Sepehri A, Roccia MG, Jafferany M, Olisova OY, et al. 2020. RETRACTED: 5G technology and induction of coronavirus in skin cells. *J. Biol. Regul. Homeost. Agents* 34(4). **https://doi.org/10.23812/20-269-E-4R**
52. Brainard J. 2020. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*, May 13. **https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat**
53. Chan J, Oo S, Chor CYT, Yim D, Chan JSK, Harky A. 2020. COVID-19 and literature evidence: Should we publish anything and everything? *Acta Biomedica* 91:e2020020
54. Goulart RRV, de Lima VLS, Xavier CC. 2011. A systematic review of named entity recognition in biomedical texts. *J. Braz. Comput. Soc.* 17:103–16
55. Li J, Sun Y, Johnson RJ, Sciaky D, Wei C-H, et al. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016:baw068
56. Lu Z, Kao H-Y, Wei C-H, Huang M, Liu J, et al. 2011. The gene normalization task in BioCreative III. *BMC Bioinform.* 12:S2
57. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, et al. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43:D36–42
58. Wei C-H, Allot A, Leaman R, Lu Z. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* 47:W587–93

59. Wang X, Song X, Li B, Guan Y, Han J. 2020. Comprehensive named entity recognition on CORD-19 with distant or weak supervision. arXiv:2003.12218 [cs.CL]
60. Wei C-H, Harris BR, Kao H-Y, Lu Z. 2013. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* 29:1433–39
61. Lafferty J, McCallum A, Pereira FC. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282–89. New York: Assoc. Comput. Mach.
62. Leaman R, Lu Z. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* 32:2839–46
63. Huang Z, Xu W, Yu K. 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991 [cs.CL]
64. Devlin J, Chang M-W, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]
65. Wang J, Pham HA, Manion F, Rouhizadeh M, Zhang Y. 2020. COVID-19 SignSym: a fast adaptation of general clinical NLP tools to identify and normalize COVID-19 signs and symptoms to OMOP common data model. arXiv:2007.10286 [cs.CL]
66. Dong X, Li J, Soysal E, Bian J, DuVall SL, et al. 2020. COVID-19 TestNorm: a tool to normalize COVID-19 testing names to LOINC codes. *J. Am. Med. Inform. Assoc.* 27(9):1437–42
67. Colic N, Furrer L, Rinaldi F. 2020. Annotating the pandemic: named entity recognition and normalisation in COVID-19 literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, ed. K Verspoor, KB Cohen, M Conway, B de Bruijn, M Dredze, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
68. Wei C-H, Harris BR, Li D, Berardini TZ, Huala E, et al. 2012. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database* 2012:bas041
69. Bekhuis T. 2006. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *J. Biomed. Digit. Libr.* 3:2
70. Swanson DR, Smalheiser NR. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intel.* 91:183–203
71. Gopalakrishnan V, Jha K, Jin W, Zhang A. 2019. A survey on literature based discovery approaches in biomedical domain. *J. Biomed. Inform.* 93:103141
72. Piad-Morffis A, Estevez-Velarde S, Estevanell-Valladares EL, Gutiérrez Y, Montoyo A, et al. 2020. Knowledge discovery in COVID-19 research literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, ed. K Verspoor, KB Cohen, M Conway, B de Bruijn, M Dredze, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
73. Pinto BG, Oliveira AE, Singh Y, Jimenez L, Gonçalves ANA, et al. 2020. ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. *J. Infect. Dis.* 222(4):556–63
74. Tarasova O, Ivanov S, Filimonov DA, Poroikov V. 2020. Data and text mining help identify key proteins involved in the molecular mechanisms shared by SARS-CoV-2 and HIV-1. *Molecules* 25:2944
75. Karami A. 2020. Investigating diseases and chemicals in COVID-19 literature with text mining. *JMIR Preprints* 18/06/2020:21503. **https://preprints.jmir.org/preprint/21503**
76. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* 6:52
77. Beltagy I, Cohan A, Lo K. 2019. SciBERT: pretrained contextualized embeddings for scientific text. arXiv:1903.10676 [cs.CL]
78. Martinc M, Škrlj B, Pirkmajer S, Lavrač N, Cestnik B, et al. 2020. COVID-19 therapy target discovery with context-aware literature mining. arXiv:2007.15681 [cs.CL]
79. Kuusisto F, Page D, Stewart R. 2020. Word embedding mining for SARS-CoV-2 and COVID-19 drug repurposing. *F1000Research* 9:585
80. Tu J, Verhagen M, Cochran B, Pustejovsky J. 2020. Exploration and discovery of the COVID-19 literature through semantic visualization. arXiv:2007.01800 [cs.CL]
81. Yeganova L, Islamaj R, Chen Q, Leaman R, Allot A, et al. 2020. Navigating the landscape of COVID-19 research through literature analysis: a bird's eye view. arXiv:2008.03397 [cs.DL]

82. Gates LE, Hamed AA. 2020. The anatomy of the SARS-CoV-2 biomedical literature: introducing the CovidX network algorithm for drug repurposing recommendation. *J. Med. Internet Res.* 22:e21169
83. Patel JC, Tulswani R, Khurana P, Sharma YK, Ganju L, et al. 2020. Identification of pulmonary comorbid diseases network based repurposing effective drugs for COVID-19. *Res. Square*. **https://doi.org/10.21203/rs.3.rs-28148/v1**
84. Wang Q, Li M, Wang X, Parulian N, Han G, et al. 2020. COVID-19 literature knowledge graph construction and drug repurposing report generation. arXiv:2007.00576 [cs.CL]
85. Jurafsky D, Martin JH. 2009. *Speech and Language Processing*. London: Pearson
86. Athenikos SJ, Han H. 2010. Biomedical question answering: a survey. *Comput. Methods Programs Biomed.* 99:1–24
87. Herriman M, Meer E, Rosin R, Lee V, Washington V, et al. 2020. Asked and answered: building a chatbot to address Covid-19-related concerns. *NEJM Catalyst Innovations in Care Delivery*, June 18. **https://catalyst.nejm.org/doi/full/10.1056/cat.20.0230**
88. Wei J, Huang C, Vosoughi S, Wei J. 2020. What are people asking about COVID-19? *A question classification dataset*. arXiv:2005.12522 [cs.CL]
89. McCreery CH, Katariya N, Kannan A, Chablani M, Amatriain X. 2020. Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3458–65. New York: Assoc. Comput. Mach.
90. Li Y, Grandison T, Silveyra P, Douraghy A, Guan X, et al. 2020. Jennifer for COVID-19: an NLP-powered chatbot built for the people and by the people to combat misinformation. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, ed. K Verspoor, KB Cohen, M Dredze, E Ferrera, J May, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
91. Narayan S, Gardent C, Cohen SB, Shimorina A. 2017. Split and rephrase. arXiv:1707.06971 [cs.CL]
92. Lee S, Kim D, Lee K, Choi J, Kim S, et al. 2016. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLOS ONE* 11:e0164680
93. Reimers N, Gurevych I. 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv:1908.10084 [cs.CL]
94. Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. arXiv:1606.05250 [cs.CL]
95. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. 2019. PubMedQA: a dataset for biomedical research question answering. arXiv:1909.06146 [cs.CL]
96. Dong L, Yang N, Wang W, Wei F, Liu X, et al. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pp. 13063–75. **https://papers.nips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html**
97. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, et al. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461 [cs.CL]
98. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, et al. 2020. CO-Search: COVID-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv:2006.09595 [cs.IR]
99. Doanvo A, Qian X, Ramjee D, Piontkivska H, Desai A, Majumder M. 2020. Machine learning maps research needs in COVID-19 literature. *Patterns* 1(9):100123
100. Wahbeh A, Nasralah T, Al-Ramahi M, El-Gayar O. 2020. Mining physicians' opinions on social media to obtain insights into COVID-19: mixed methods analysis. *JMIR Public Health Surveill.* 6:e19276
101. Moore RC, Lee A, Hancock JT, Halley M, Linos E. 2020. Experience with social distancing early in the COVID-19 pandemic in the United States: implications for public health messaging. medRxiv 2020.04.08.20057067. **https://doi.org/10.1101/2020.04.08.20057067**
102. Debnath R, Bardhan R. 2020. India nudges to contain COVID-19 pandemic: a reactive public policy analysis using machine-learning based topic modelling. *PLOS ONE* 15:e0238972
103. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, et al. 2020. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *Lancet Psychiatry* 7:547–60

104. Li D, Chaudhary H, Zhang Z. 2020. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. *Int. J. Environ. Res. Public Health* 17(14):4988
105. Low DM, Rumker L, Talker T, Torous J, Cecchi G, Ghosh SS. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: an observational study. *J. Med. Internet Res.* 22(10):e22635
106. Jelodar H, Wang Y, Orji R, Huang H. 2020. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. arXiv:2004.11695 [cs.IR]
107. Aslam F, Awan TM, Syed JH, Kashif A, Parveen M. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. *Humanit. Soc. Sci. Commun.* 7:23
108. de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez JI. 2020. Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. *Int. J. Environ. Res. Public Health* 17:5542
109. Drias HH, Drias Y. 2020. Mining Twitter data on COVID-19 for sentiment analysis and frequent patterns discovery. medRxiv 2020.05.08.20090464. **https://doi.org/10.1101/2020.05.08.20090464**
110. Samuel J, Ali GG, Rahman M, Esawi E, Samuel Y. 2020. COVID-19 public sentiment insights and machine learning for tweets classification. *Information* 11:314
111. Zhou J, Yang S, Xiao C, Chen F. 2020. Examination of community sentiment dynamics due to COVID-19 pandemic: a case study from Australia. arXiv:2006.12185 [cs.SI]
112. Ahmed ME, Rabin MRI, Chowdhury FN. 2020. COVID-19: social media sentiment analysis on reopening. arXiv:2006.00804 [cs.SI]
113. Han X, Wang J, Zhang M, Wang X. 2020. Using social media to mine and analyze public opinion related to COVID-19 in China. *Int. J. Environ. Res. Public Health* 17:2788
114. Wagner T, Shweta F, Murugadoss K, Awasthi S, Venkatakrishnan AJ, et al. 2020. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. *eLife* 9:e58227
115. Callahan A, Steinberg E, Fries JA, Gombar S, Patel B, et al. 2020. Estimating the efficacy of symptom-based screening for COVID-19. *NPJ Digit. Med.* 3:95
116. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, et al. 2018. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* 25:331–36
117. Wang J, Pham HA, Manion F, Rouhizadeh M, Zhang Y. 2020. COVID-19 SignSym: a fast adaptation of general clinical NLP tools to identify and normalize COVID-19 signs and symptoms to OMOP common data model. arXiv:2007.10286 [cs.CL]
118. Chapman AB, Peterson KS, Turano A, Box TL, Wallace KS, Jones M. 2020. A natural language processing system for national COVID-19 surveillance in the US Department of Veterans Affairs. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, ed. K Verspoor, KB Cohen, M Dredze, E Ferrera, J May, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
119. Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, et al. 2020. Trove: ontology-driven weak supervision for medical entity classification. arXiv:2008.01972 [cs.CL]
120. Picone M, Inoue S, DeFelice C, Naujokas MF, Sinrod J, et al. 2020. Social listening as a rapid approach to collecting and analyzing COVID-19 symptoms and disease natural histories reported by large numbers of individuals. *Popul. Health Manag.* 23(5):350–60
121. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. 2020. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: observational infoveillance study. *J. Med. Internet Res.* 22:e19421
122. Zheng N, Du S, Wang J, Zhang H, Cui W, et al. 2020. Predicting COVID-19 in China using hybrid AI model. *IEEE Trans. Cybern.* 50:2891–904
123. Li L, Zhang Q, Wang X, Zhang J, Wang T, et al. 2020. Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo. *IEEE Trans. Comput. Soc. Syst.* 7:556–62
124. Lee N, Bang Y, Madotto A, Fung P. 2020. Misinformation has high perplexity. arXiv:2006.04666 [cs.CL]

125. Elhadad MK, Li KF, Gebali F. 2021. An ensemble deep learning technique to detect COVID-19 misleading information. In *Advances in Network-Based Information Systems*, ed. L Barolli, KF Li, T Enokido, M Takizawa, pp. 163–75. Cham, Switz.: Springer Int.
126. Serrano JCM, Papakyriakopoulos O, Hegelich S. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, ed. K Verspoor, KB Cohen, M Dredze, E Ferrera, J May, et al. Stroudsburg, PA: Assoc. Comput. Linguist.
127. Groza A. 2020. Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology. arXiv:2004.12330 [cs.AI]
128. Cui L, Lee D. 2020. CoAID: COVID-19 healthcare misinformation dataset. arXiv:2006.00885 [cs.SI]