SymptomID: A Framework for Rapid Symptom Identification in Pandemics Using News Reports

KANG GU, SOROUSH VOSOUGHI, and TEMILOLUWA PRIOLEAU, Dartmouth College, USA

The ability to quickly learn fundamentals about a new infectious disease, such as how it is transmitted, the incubation period, and related symptoms, is crucial in any novel pandemic. For instance, rapid identification of symptoms can enable interventions for dampening the spread of the disease. Traditionally, symptoms are learned from research publications associated with clinical studies. However, clinical studies are often slow and time intensive, and hence delays can have dire consequences in a rapidly spreading pandemic like we have seen with COVID-19. In this article, we introduce SymptomID, a modular artificial intelligence–based framework for rapid identification of symptoms associated with novel pandemics using publicly available news reports. SymptomID is built using the state-of-the-art natural language processing model (Bidirectional Encoder Representations for Transformers) to extract symptoms from publicly available news reports and cluster-related symptoms together to remove redundancy. Our proposed framework requires minimal training data, because it builds on a pre-trained language model. In this study, we present a case study of SymptomID using news articles about the current COVID-19 pandemic. Our COVID-19 symptom extraction module, trained on 225 articles, achieves an F1 score of over 0.8. SymptomID can correctly identify well-established symptoms (e.g., "fever" and "cough") and less-prevalent symptoms (e.g., "rashes," "hair loss," "brain fog") associated with the novel coronavirus. We believe this framework can be extended and easily adapted in future pandemics to quickly learn relevant insights that are fundamental for understanding and combating a new infectious disease.

CCS Concepts: • **Computing methodologies** → **Information extraction**; • **Applied computing** → **Health informatics;**

Additional Key Words and Phrases: Symptom identification, named entity extraction, BERT, novel pandemics, COVID-19, news articles

ACM Reference format:
Kang Gu, Soroush Vosoughi, and Temiloluwa Prioleau. 2021. SymptomID: A Framework for Rapid Symptom Identification in Pandemics Using News Reports. *ACM Trans. Manag. Inf. Syst.* 12, 4, Article 32 (September 2021), 17 pages.
https://doi.org/10.1145/3462441

1 INTRODUCTION

Through the past decade and more, people across the world have been affected by several novel epidemics and pandemics such as the early influenza outbreak and **severe acute respiratory**

This work is supported by the National Science Foundation (NSF award number: 2031546).
Authors' address: K. Gu, S. Vosoughi, and T. Prioleau, Dartmouth College, 9 Maynard Street, Hanover, NH; emails: {Kang.Gu.GR, Soroush.Vosoughi, Temiloluwa.O.Prioleau}@dartmouth.edu.

2158-656X/2021/09-ART32 $15.00
https://doi.org/10.1145/3462441

syndrome (SARS) [15, 34, 56]. During the onset of such events, a critical and time-sensitive task is to learn as much as possible about the novel infectious disease or epidemic to inform an appropriate public health response that will contribute to containment. For example, the emergence of today's pandemic—**Coronavirus 2019 (COVID-19)**—came with challenges associated with learning what could be known as fundamentals about the virus, including how it is transmitted, the incubation period, associated symptoms, recovery time, and so on. Early clinical papers [5, 8, 22, 24] were fundamental for gaining a basic understanding of various facets of the novel coronavirus. However, an ever-present obstacle that slows down the pace of knowledge discovery and limits the scope of such clinical research is physical recruitment of patients/subjects often confined to the geographical location of each unique research study and manual analysis of related data.

To support rapid response to pandemics, methods in **artificial intelligence (AI)** and data science can play a unique role. Latif et al. provide a comprehensive review of opportunities and efforts related to using data science in the fight against epidemics and pandemics with a particular focus on COVID-19 [23]. Examples of such effort include leveraging case data, textual data, and biomedical data for outcomes such as screening and diagnosis, simulation and modelling, logistical planning, and economic interventions. More closely related to this article are efforts that use digital text data as a source for increasing knowledge about the relevant epidemic/pandemic and associated illnesses. The most prominent source of text data used for research is extracted from social media platforms such as Twitter [1, 13**?**]. Academic publications is another common digital text format that has been leveraged for data science applications in response to pandemics. For example, in March 2020, the **COVID-19 Open Research Dataset Challenge (CORD-19)** was released [50]. This source includes over 200,000 scholarly articles on related coronaviruses and provides a growing resource for text mining and information retrieval to generate new insights about COVID-19. However, an alternative data source that is currently underutilized for data science applications in response to a pandemic is news reports. In the wake of a novel epidemic/pandemic, news reports are published at a much faster rate than academic publications. In addition, news reports often include various relevant topics and even recounts from persons who have been directly affected. This data source is primarily publicly available, not geographically bound, and does not suffer from the challenge of directly recruiting subjects for data collection.

In this article, we introduce an AI-based framework that leverages unstructured and distributed news reports (albeit from reliable sources) for rapid knowledge discovery in novel pandemics. Using today's pandemic as relevant test case, we show how the proposed framework—SymptomID—can be used to automatically learn about symptoms associated with a new epidemic/pandemic (i.e., COVID-19). Unlike prior work that uses clinical data from electronic health records [49], this research leverages publicly available and accessible news reports as a data source for insight gathering. Primary contributions of this work are as follows:

(1) We propose a modular AI-based framework for rapid identification of symptoms (or other relevant features) using publicly available news report.
(2) We present a case study that applies our framework in response to the current pandemic. Our results on symptoms identified show agreement with manually identified symptoms of COVID-19 from clinical literature; however, we also identified less-documented and currently undocumented symptoms like "brain fog" and "hair loss."
(3) We release our trained model and dataset associated with this work to further data-driven research on COVID-19.

The rest of this article is organized as follows: In Section 2, we provide an overview of the proposed framework that includes subsections on data collection, data annotation, named entity recognition, performance evaluation, entity extraction & clustering, and insights gathering.

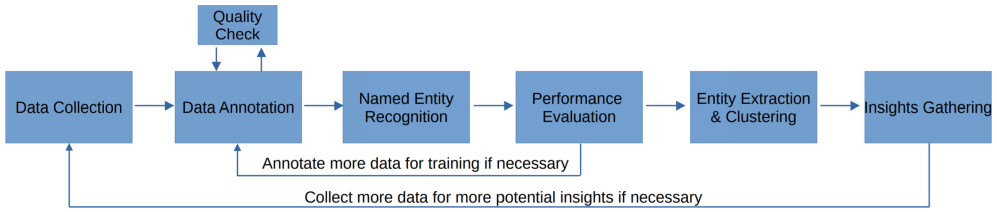


Fig. 1. Flowchart of the proposed framework.

Following this, we focus on our test case and describe a sourced COVID-19 dataset and the process for annotation in Section 3. Section 4 goes into details with our experiments and results obtained. Section 5 discusses insights gathered specifically about COVID-19. Finally, we end with a detailed discussion of the broader applicability and possible future directions of this research.

2 PROPOSED FRAMEWORK

In this work, our framework is used as a basis for insights gathering and knowledge discovery particularly about the broad range of symptoms associated with COVID-19. However, it is important to note that this framework can be easily adapted for learning about other relevant topics of interest associated with a novel pandemic.

Figure 1 shows the general pipeline for the proposed framework. It consists of six main components, namely Data Collection, Data Annotation, Named Entity Recognition, Performance Evaluation, Entity Extraction & Clustering, and Insights Gathering. More specifically, we formulate the problem of rapid symptom identification as a **Named Entity Recognition (NER)** task [33], where the target entities of interest are symptoms. However, we also show evaluations on other potential targets including location and time expressions.

2.1 Data Collection

Given that the primary goal of the proposed framework is to *rapidly* identify symptoms, we outline five important criteria that should be considered when choosing the right data sources, namely accessibility, timeliness, trustworthiness, volume, and longitudinality. It is important to note that a high-quality dataset is key to gathering reliable insights about any novel pandemic. Hence, in this work selected news reports needed to satisfy the criteria of the following:

- Accessibility: This includes news reports that are publicly available and accessible without privacy constraints. The public nature of selected news reports will facilitate reproducibility as the collated dataset can be made available to support further research efforts.[1]
- Timeliness: Given the urgency of knowledge discovery in a novel pandemic, symptoms (or other relevant targets) should be identified in a timely manner. Building on the fact that news reports are released at a faster rate than academic publications or clinical reports, news reports can serve as a unique data source for immediate analysis in the wake of a pandemic.
- Trustworthiness: In a pandemic, it is expected that a huge amount of reports will be generated by various news sources on a daily basis, and hence trustworthiness is critical. The proposed framework advocates for reputable sources on mainstream media[2] to support data reliability.
- Longitudinality: With pandemics (such as COVID-19), there can be an evolution of known symptoms associated with the relevant virus or illness. Hence, selected news reports should

[1]The collated dataset used in this work is made publicly available to further research.
[2]https://en.wikipedia.org/wiki/Mainstream_media.

be longitudinal and span across time to support identification of uncommon symptoms that may not have been reported during the initial phases.
- Volume: As with other major events, it is expected that pandemics will have large coverage from various news outlets. This is evident in today's COVID-19 pandemic and has been seen in previous epidemics like with Ebola. This large volume of data is advantageous and will also support with identification of less obvious symptoms or other topics of interest.

In this work, we leveraged the *Event Registry* [17] platform for data collection and collation of various news reports. Event Registry analyzes reports published by over 30,000 media outlets with AI methods and enables the user to filter content by keywords, entities, sources, categories, locations, and sentiment. Any other news aggregation platform can also be useful for the data collection step.

2.2 Data Annotation

The proposed framework (SymptomID) relies on supervised machine learning (i.e., NER) for analysis and knowledge discovery. Hence, a portion of the raw data needs to be annotated for building and training the relevant AI model. Standard application of NER requires for the entities of interest (e.g., symptoms, time expressions, and locations) to be determined first [33]. Following entity identification, the training dataset can then be annotated accordingly. In this work, we used the extended **Begin, Inside, Other (BIO)** [38] tagging scheme for annotating our dataset of news reports. Leveraging multiple annotators is useful to enhance the quality and reliability of the annotated dataset; however, it is also important to establish clear guidelines before beginning, because data annotation is often laborious and time-consuming. Section 3 includes more details on the data annotation process implemented in this work.

2.3 Named Entity Recognition

After the appropriate dataset has been annotated, it is time to build and train the NER model. Given that there are many variations of NER models, we propose comparing different candidates to identify the option with acceptable performance on a given dataset. In this work, we evaluated several transformer-based models for NER, such as **Bidirectional Encoder Representations for Transformers (BERT)** [19, 47], GPT [37], and XLNet [57]. Performance was the primary metric used for selecting the model of choice.

After preliminary evaluation, we chose BERT [14], the state-of-the-art **natural language processing (NLP)** model, as our approach to recognize entities of interest. Formally, we formulate NER as a multi-class token classification problem. As mentioned, time, location, and symptom entities were tagged with the BIO scheme, which generated seven classes in our dataset: **beginning of time (B-tim)**, **inside of time (I-tim)**, **beginning of location (B-geo)**, **inside of location (I-geo)**, **begining of symptom (B-sym)**, **inside of symptom (I-sym)**, and **other entities (O)**. Given an input sentence $S = \{w_1, \ldots, w_i, \ldots, w_n\}$, the goal of NER is to classify the tag t_i of word w_i, where $i \in [1, n]$. Essentially, the BERT model estimates probability distribution $P(t_1, \ldots, t_i, \ldots t_n|S)$, where $t \in T$, $T = \{$B-tim, I-tim, B-geo, I-geo, B-sym, I-sym, O$\}$.

2.4 Performance Evaluation

After model training, it is important to evaluate the performances of various models and decide whether or not to employ one as the final model of choice. In this work, we used the standard and widely-accepted F1 score, Precision, and Recall as the evaluation metrics. First, an initial batch of articles were annotated and used for training; however, the model performance was not acceptable. We then evaluated how the performance changed by using a varying number of annotated data.

This analysis revealed that we could expect improve model performance with more annotations and a larger training dataset. Following this, a second batch of news articles were annotated to increase the training dataset and this showed to improve the model performance to an acceptable level as described in Section 4.

2.5 Entity Extraction & Clustering

A reliable NER model should be able to extract all target entities in the collected dataset. However, different identified named-entities could refer to the same thing. For instance, the named-entities "difficulty breathing" and "difficult to breathe" can be extracted as different symptoms; however, they both refer to the same concept. Hence, the next step in our framework is to group similar named-entities extracted by the framework together using an acceptable clustering method. In this work, we used DBSCAN [16] for the clustering step, because it is fast and particularly advantageous for finding clusters of arbitrary density. More specifically, clustering detected entities of the same type was done by encoding them using the BERT model and then using their vector representations with DBSCAN for clustering. The symptom clusters generated from this stage are then used for analysis and insight gathering.

2.6 Insights Gathering

The primary goal of our proposed framework is to support knowledge discovery in a pandemic or about infectious diseases; this can enable the community to take appropriate actions toward containment. In this work, we focus on rapid symptom identification. Insights related to symptoms can include types of symptoms, prevalence of occurrence, timing of reporting, and so on. Note that after the initial training, the proposed framework can be used to continuously gather more insights about the relevant pandemic from news articles over an extended period of time. Through the rest of this article, we use the COVID-19 pandemic as a case study to showcase our framework in action.

3 COVID-19 NEWS CORPUS

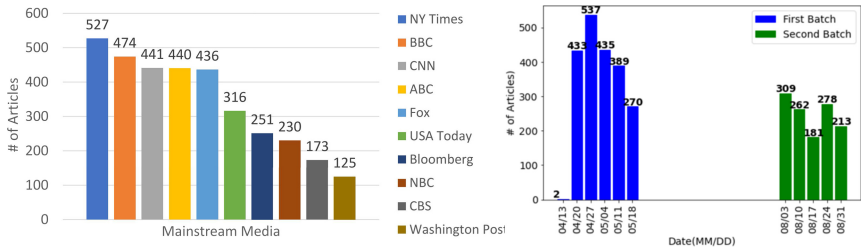
In this section, we will introduce the COVID-19 News Corpus used in this work. We describe the dataset and annotation process in detail and make both of these publicly available[3] for further research and analysis.

3.1 Data Composition

As mentioned above, we leveraged the Event Registry [17] platform for collecting and aggregating new reports for rapid symptom identification. More specifically, we searched for news articles with the keywords "COVID-19" and "symptom." To ensure data reliability and trustworthiness, only articles published in mainstream media were included in our dataset. As shown in Figure 2(a), our dataset includes a total of 3,413 news articles from 10 news sources such as *The New York Times*, BBC, CNN, ABC, Fox, and so on. The distribution of articles across weeks of the year is also shown in Figure 2(b). A first batch of 2,136 articles published between April 13 and May 18, 2020, was collected and used for model training, tuning, and testing.

Then a second batch of 1,277 articles published between August 3 and August 31, 2020, was collected to extend the full dataset for insight gathering. Note that our framework was trained and tested on the annotated subset of the first batch. Our framework was then applied to the second batch. The second batch of data is crucial for understanding how well our framework, once trained,

[3] https://github.com/KangGu96/COVID-19-News-Corpus.



(a) News article distribution of news sources (b) News article distribution of week of year

Fig. 2. A glance at the dataset composition.

Table 1. Some Sample Headlines from the Collected Dataset

Source	Date	Headline
USA Today	April 3, 2020	*I'm not dead, and each day is slightly better: The story of a senior coronavirus survivor*
NBC	April 20, 2020	*My coronavirus symptoms were 'mild'. Young people who want to end quarantine — read this first.*
BBC	May 19, 2020	*The patients who just can't shake off Covid-19*
NY Times	April 13, 2020	*We need to talk about what coronavirus recoveries look like*
CNN	May 19, 2020	*14-year-old recovering from multi system inflammatory syndrome was hospitalized with heart failure*

Table 2. Inter-annotator Reliability Score

Metrics \ Tags	Location Tags	Time Tags	Symptom Tags	Overall
Fleiss's Kappa Score	0.66	0.69	0.47	0.58

According to Reference [44], our overall agreement is moderate but close to substantial.

generalizes to new and different dataset. To provide a better understanding of our COVID-19 News Corpus, we show the headlines of 5 example news articles from our dataset in Table 1.

3.2 Annotation

Data annotation was done on 300 randomly selected articles published between April 13 and May 18. We recruited four volunteers as annotators; these were graduate (two) and undergraduate (two) students at Dartmouth College. At the time, all students were active researchers in the Computer Science Department working on projects intersecting data science and health. Each student annotator was assigned 40–80 articles to annotate according to strict guidelines provided in a training session that included a demonstration of the whole annotation process using an example article. Additional discussions were had during the annotation process to clarify confusions and ensure all annotators were following the same standard with regards to annotating the identified named-entities (i.e., symptoms, times, and locations).

Our primary goal was to annotate as many articles as possible, and hence we did not have every article annotated by multiple annotators. However, to estimate the general quality of annotations and agreement between annotators, a small subset of 20 articles were assigned to all annotators and annotations were compared both qualitatively and quantitatively. Table 2 shows the inter-annotator agreement among the four annotators calculated using Fleiss's Kappa Score [18]. Per

Table 3. Inter-annotator Agreement on the Selected News Article

News Headline: What's Going to Happen to the Junior, Now That His Mother Is Dead?[4]				
Annotator	I	II	III	IV
Location Tags	Manhattan; Queens; New York; Honduras	Manhattan; Queens; New York; Honduras	Manhattan; Queens; New York; N.Y.	Manhattan; Queens; New York; Honduras
Time Tags	March 21; March 31; early April	March 21; March 31; early April	March 21; March 31	fall; March 21; March 31
Symptom Tags	fever; headache; high fever; coughing	headache; high fever; coughing	anxiety; headache; high fever; coughing	fever; headache; high fever; coughing

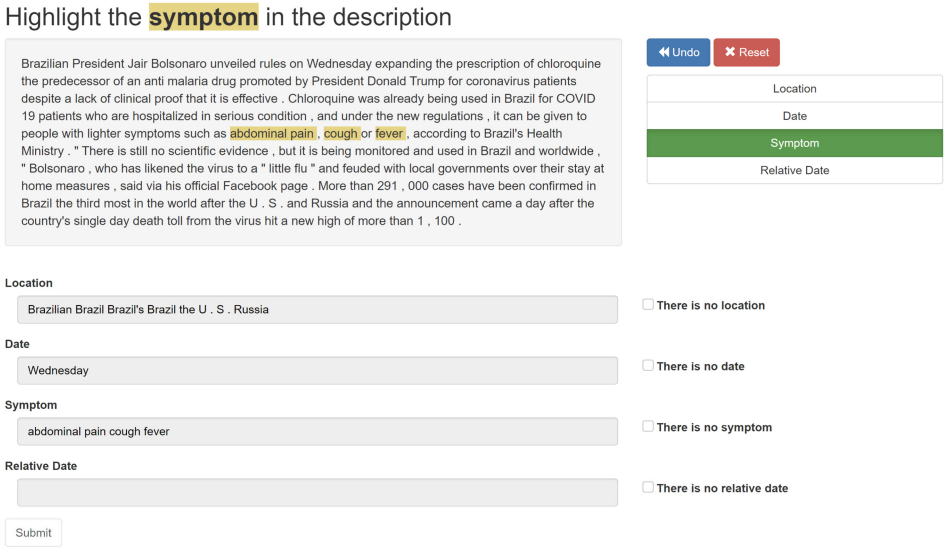


Fig. 3. Process of annotation using LocalTurk. Note that Relative Date and Date were combined into a single "time" tag after the annotation process.

Reference [44], the range of a kappa coefficient is from 0 (poor agreement) to 1 (almost perfect agreement). A score in between 0.41 and 0.60 represents moderate agreement while 0.61–0.80 represents substantial agreement. Based on this reference, Table 3 shows that there was moderate agreement between annotators for symptom tags and substantial agreement on the location tags and time tags.[4] It should be noted that such scores have limitations as they do not consider the number of categories (7 in this work) and the number of annotators/raters (4 in this work). Nonetheless, manual inspection of the annotated datasets confirmed high-quality annotations. LocalTurk [31]—a version of the well-known Amazon Mechanical Turk [3]—was the tool used for annotation in this work. Figure 3 shows an example screenshot of the annotation platform were location, date, and symptoms can be annotated accordingly. Annotation was done simply by highlighting all the words in the text under a given named-entity/category (e.g., symptom). When an

[4]https://www.nytimes.com/2020/05/07/nyregion/nyc-coronavirus-adult-disabled.html.

entity did not exist in a text/news report, annotators were instructed to click the "There is no xx" option and proceed accordingly.

4 EXPERIMENTS AND RESULTS

In this section, we describe the steps of building models for NER, performance evaluation, and entity extraction & clustering.

4.1 Implementation

4.1.1 Named Entity Recognition. As mentioned above, we opted to use BERT for NER. We utilized a pretrained BERT model provided by the Transformers API,[5] for token classification. Specifically, the "bert-base-cased" version was used, since our text had not been lowercased. The Transformers API is backed by PyTorch and Tensorflow, with seamless integration between them. As such, there are thousands of pretrained models provided to perform text analysis. Cross entropy loss was chosen as the objective for training and the model was optimized using the AdamW [32] optimizer—an improved version of the well-known Adam optimizer. The learning rate and ϵ were set to be $3e-5$ and $1e-8$, respectively. The maximum length of an input sentence (MAX_LEN) was set to 75. Furthermore, we used a batch size of 16 and train epoch of 20; these were empirically chosen through repeated training. A machine with Intel Xeon CPU of 2.30 GHz and Nvidia Telsa P100 GPU of 16 GB was used for training in all experiments.

4.1.2 Clustering. Before clustering, we used BERT to embed symptom entities into 768-dimensional vectors (this is the output size of a BERT model). Specifically, we extracted vectors from the last layer of BERT as these include more contextual information. For the clustering algorithm, we employed DBSCAN provided by Scikit-Learn API.[6] A standard euclidean metric was chosen to calculate to distance between symptom vectors. In addition, the maximum distance (ϵ) and the minimal number of samples were set to be 5 and 10, respectively. These parameters were chosen manually so that symptoms within each cluster were consistent.

4.2 Experiments

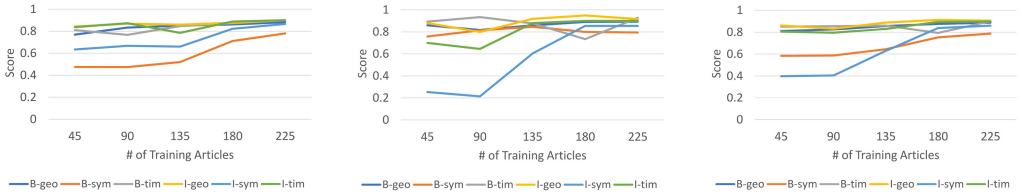
4.2.1 Performances of NER Models. We considered three types of transformer-based models, namely XLNet [57], GPT [37], and BERT for the NER task. XLNet overcame the limitations of BERT, which were relying on corrupting input with masks and requiring input of fixed length. GPT was built using transformer decoder blocks instead of the encoder blocks that BERT used. Both XLNet and GPT achieved comparable results to BERT on various NLP tasks. These were all pretrained on a large corpus and thus allowed for transfer learning with a smaller set of labelled samples. As shown in Table 4, BERT slightly outperforms GPT and XLNet on our dataset. It is important to note that there are other pretrained models in transformer family; however, this work shows the BERT can yield satisfactory results and performed better than XLNet and GPT.

4.2.2 Evaluating Sufficiency of the Training Dataset. We conducted a control experiment to evaluate the model's performance as a function of the size of the training dataset. This was done to determine whether or not more annotations would yield better performance. With the original total of 300 annotated articles in our COVID-19 News Corpus, we started by setting aside 75 articles (i.e., 25%) for testing. Then we used varying proportions of the rest of the data for training. Figure 4 shows the results of this evaluation, including the precision curve, recall curve, and F-1 curve for each class. We observed that as the number of articles used for training increased from

[5] https://github.com/huggingface/transformers.
[6] https://scikit-learn.org/stable/index.html.

Table 4. Performances of Different NER Models on Our Dataset

Model \ Performance	Precision	Recall	F1
XLNet	0.76	0.70	0.73
GPT	0.80	0.77	0. 0.78
BERT	**0.84**	**0.84**	**0.84**



(a) Precision Curve (b) Recall Curve (c) F-1 Curve

Fig. 4. Performance of varying trainset size.

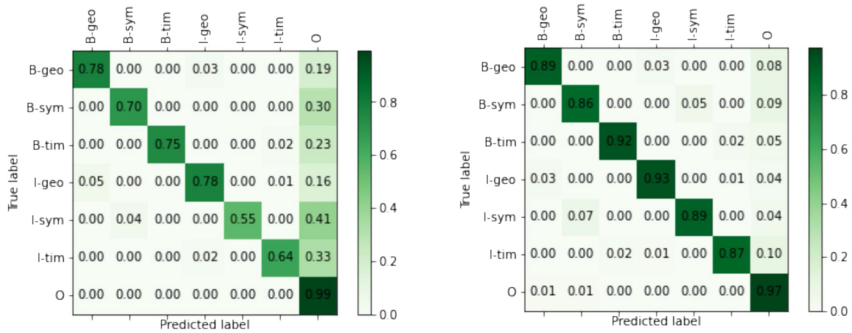
75 (i.e., 25% of the full dataset) to 225 (i.e., 75% of the full dataset), the precision scores for identification of symptoms (i.e., B-sym and I-sym) improved by more than 0.2 (≈30–40%), which is a notable improvement. Conversely, the precision for other classes was already high (up to 0.8) even with a smaller training dataset. The overall trend and precision scores suggest that the size of our annotated dataset is sufficient to fine-tune the BERT model for accurate entity extraction. Moreover, the recall scores obtained was greater than 0.8 for all classes, suggesting that BERT is able to recognize most of the targeted entities in the text.

4.2.3 The Effect of Downsampling. In NER tasks, class imbalance is common as the entities of interest sometimes occur sparsely through the text. As such, an ignorant classifier trained on such skewed data will be prone to perform poorly for minority classes. Several active learning approaches have been implemented in prior work to handle the problem of class imbalance [46]. In this work, we implemented the technique of downsampling the majority class to achieve greater balance. Specifically, sentences that with no target entities were dropped out with a probability of 0.8. Figure 5 shows a comparison of the performance before and after downsampling. Downsampling contributed to improving the accuracy score of all classes of interest. For example, the true positive scores for symptom identification classes (B-sym and I-sym) were 0.7 and 0.55 before downsampling, meanwhile these improved to 0.86 and 0.89, respectively, after downsampling. This figure also shows that the step of downsampling reduced the confusion error between entities of interest and O (i.e., the other class) significantly.

4.2.4 The Precision–Recall Curve. A standard tool for evaluation of performance for classification tasks with imbalanced classes is the precision–recall curve. This shows the tradeoff between precision and recall for varying probability thresholds. Figure 6 shows the overall curve for all classes in this work. We observe that our proposed model achieved high precision and high recall (both over 0.9) simultaneously. This indicates that our model can detect the targeted entities (i.e., symptoms, time, and location) in news articles with high performance.

5 INSIGHTS GATHERED ABOUT COVID-19

In this section, we summarize the insights learned about COVID-19 by leveraging our proposed framework (SymptomID) for automatically extracting such knowledge from publicly available news reports. We focus on symptoms as this is a particularly relevant feature to understand and



(a) Before Downsampling (b) After Downsampling

Fig. 5. Confusion matrix before downsampling and after downsampling.

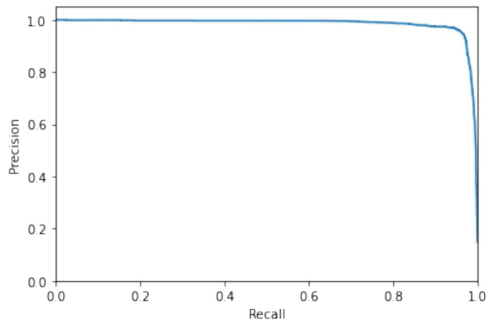


Fig. 6. Precision-recall curve. Average precision score, micro-averaged over all classes: 0.98.

quantify in the wake of any novel pandemic. In addition, we can compare the results obtained in this article with those obtained through alternative research approaches (e.g., manual data collection in clinical practice that is the most common method in literature). The objective of this section is to show that many of the insights identified through manual data collection can also be identified through AI-supported techniques that are scalable and not limited by the geographical reach of a particularly research study.

Figure 7 shows a summary of symptoms identified after entity extraction and clustering from the full set of over 3,000 new reports. For each cluster, we first extracted the major concept or phrase. Then we manually inspected every extracted phrase and abbreviated it (if necessary) before using it as the cluster label. To quantitatively evaluate the soundness of our clustering, we calculated the silhouette score [40] and obtained an average of 0.31, which is reasonable given the large number of clusters. A total of 29 unique symptom clusters were identified albeit with varying frequencies. This is in comparison to 11 symptoms currently listed on the Center of Disease Control and Prevention website as of September 30, 2020 [6]. The most prominent symptom clusters relate to "cough" and "fever," both of which are well-known symptoms of COVID-19. Additionally, other known symptoms like "headache," "breathing difficulties," and "fatigue" were also identified as have been identified in prior clinical literature [8, 22]. However, our results also identified less-documented symptoms of COVID-19 such as "cardiac complications," "rashes," "gastrointestinal issues," "blood clot," "hair loss," "high blood pressure," and "brain fog", some of which have now been identified in later work [5, 36]. However, it is important to note that some symptoms identified still remain undocumented in literature. Table 5 shows a sample of elements that were combined into a cluster for the symptoms of "loss of taste and smell" and "inflammation." Observations from this table

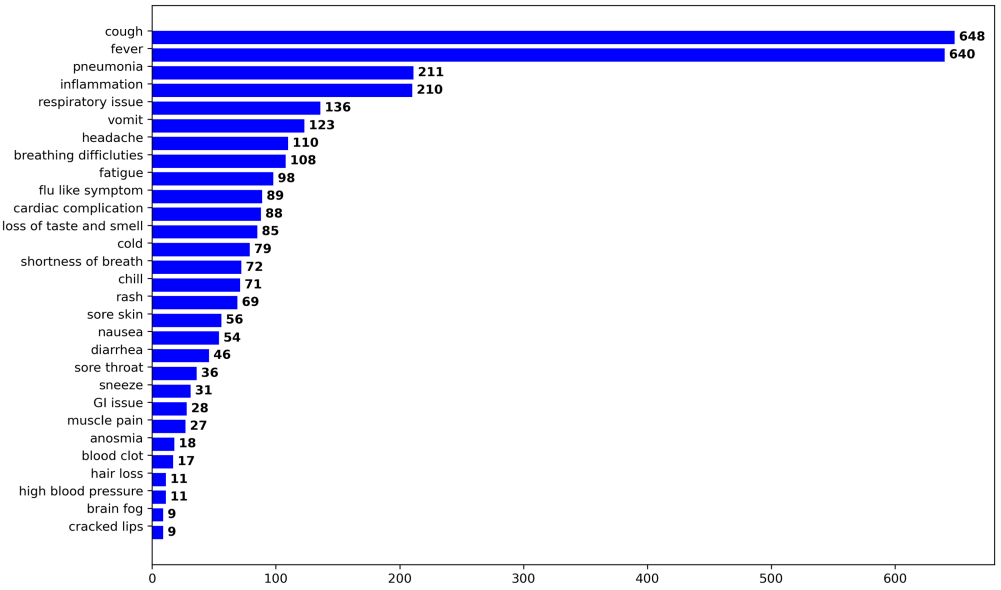


Fig. 7. Symptom clusters ranked by frequency of occurrence.

shows the clustering step is particularly important to minimize redundant outputs/findings. As mentioned above, the primary takeaway from this work is to showcase how AI-based methods such as the proposed framework (SymptomID) can be used for knowledge discovery in epidemics or pandemics. Such an approach is not intended to replace standard practice in clinical research; however, it can serve to supplement such work, contribute to finding of novel insights, and provide a rapid and scalable solution.

6 RELATED WORK

In this section, we present related research with particular focus on the topics of BERT, Named Entity Recognition, and Data Science efforts in response to COVID-19.

6.1 BERT

BERT [14] is designed to learn deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT has been applied to a wide range of NLP tasks successfully. In Sentiment Analysis, BERT outperformed previous state-of-the-art models by simply fine-tuning on a widely used sentiment analysis dataset [14]. In the Question Answering domain, Yang et al. integrated the best practices from information retrieval with a BERT-based reader to identify answers from a large corpus of Wikipedia articles [55]. BERT has also been successful in the tasks of Machine Translation [10, 12] and NER [19, 47]. For example, Conneau and Lample [12] tried to initialize the entire encoder and decoder with a multilingual pretrained BERT model and showed significant improvement could be achieved for unsupervised machine translation tasks and English-Romanian supervised machine translation tasks. Conversely, Tsai et al. [47] leveraged knowledge distillation to run a compressed BERT for NER on a single CPU, while achieving promising performance.

BERT has also been used in the health domain. For example, BioBERT [25], pretrained on large-scale biomedical corpora, was introduced and outperformed a standard BERT model on a variety of biomedical text mining tasks. Similarly, Alsentzer et al. publicly released a BERT-based model

Table 5. Unique Elements in a Few Clusters

Clusters	Unique elements (manually cleaned)
loss of taste and smell	*"loss of the sense of smell," "loss of smell and taste," "diminished ability to taste or smell," "runny sense of taste and smell," "loss of taste and smell," "loss of appetite," "loss of taste or smell and other organs," "lost their sense of taste or smell," "lack of taste and smell," "loss of taste or smell," "losing your sense of smell or taste," "loss to sense of smell or taste," "loss of taste and smell senses," "changes in smell or," "loss of smell or taste," "loss of taste," "lost my sense of taste," "lost sense of smell and taste," "impaired sense of smell," "taste and smell loss," "loss sense of smell," "losing their sense of taste and smell," "lost his sense of taste," "loss of taste or smell," "lost sense of smell," "lost a sense of taste," "lose their sense of smell or taste"*
inflammation	*"inflammatory state," "inflammatory illnesses," "inflammation of the heart," "inflammation flow syndrome," "inflammation of the arteries of the," "inflammation in their blood vessels from," "inflammatory condition," "inflammatory syndrome condition," "inflammatory illness," "inflammation of the mouth," "inflammation in their blood vessels," "inflammation toes," "inflammation in the walls of the arteries," "inflammatory syndrome," "inflammation in eyes," "inflammation of the s vessels attack," "inflammation of multiple organs including the heart," "inflammation of the heart and blood vessels," "inflammation of the blood vessels"*

trained on generic clinical notes and BioBERT model fine-tuned on discharge summaries specifically [2]. They demonstrated that using domain-specific models can yield better performance on three common clinical NLP tasks as compared to a generic model. Also, Lin et al. applied BERT with domain adaption algorithms to clinical negation detection task [30]. However, domain adaption did not improve performance over a plain BERT, which implied that BERT could already learn general representations of negation phenomena. Building on the aforementioned work, we leverage a BERT pre-trained model with a focus on a unique dataset (i.e., news reports) and a unique application (i.e., knowledge discovery for COVID-19 and/or other pandemics).

6.2 Named Entity Recognition

NER is a key component in NLP systems for Question Answering, Information Retrieval, and many other tasks. Recent advances of deep learning in NER can be found in relevant surveys on this topic [27, 54]. A noteworthy example proposed by Ji et al. is [26] a novel **meta-learning approach for domain adaptation in NER (MetaNER)**. This paper showed that MetaNER is capable of adapting to new unseen domains with a small amount of annotated data from those domains. Other relevant papers include References [28, 29].

Specific to the goal of using news articles for the NER dataset, datasets such as CoNLL 2002 [42] and CoNLL 2003 [43] are relevant. These datasets focused on four primary entities, namely, person, location, organization, and miscellaneous including all other types of entities. However, more closely related to this article are NER tasks in the health domain. One example of this is the 2010 I2B2 NER Task [48], which focused on three tasks (i.e., concept extraction, assertion classification, and relation classification) all using clinical records data. NER has also been used for extraction of drug-to-drug interactions from biomedical text data [4].

More recently and specific to COVID-19 and other closely related tasks, the U.S. White House collaborated with the National Library of Medicine, the Allen Institute for Artificial Intelligence

and other private companies to create the COVID-19 Corpus from research articles, with a set of 18 challenges for researchers to solve [50]. This dataset leverages scholarly articles as the dataset on which NER can be used. Based on the CORD-19 dataset, Wang et al. have created a CORD-NER dataset that includes 75 fine-grained entities such as genes, chemicals and diseases, as well as other entities like coronaviruses, viral proteins, evolution, materials, substrates and immune responses [53]. The CORD-NER dataset was annotated by four sources using different NER methods that achieved improved performance over alternative approaches. Likewise, Colic et al. have combined a dictionary-based system for its high recall with two models based on BioBERT for their high accuracy to perform NER [11]. At the publication time, they had processed over 25,000 abstracts from pubMed and 7,883 full-text articles from Europe PMC, which are now made publicly available. Unlike the aforementioned work, this article leverages publicly available news reports as the data source for NER and knowledge discovery, more specifically, we focus on rapid symptom identification associated with pandemics. This is a task that has not been previously explored in prior work.

6.3 Data Science for COVID-19

Since COVID-19 was declared a pandemic by the World Health Organization in March 2020, the number of cases has grown exponentially and tremendous effort is being made to curtail the spread of the virus. Data science, which covers a broad range of topics such as machine learning, statistical learning, time-series modeling, data visualisation, expert systems, and probabilistic reasoning, has attracted increasing attention from researchers to fight against COVID-19 [**?**]. Various studies have employed computer vision algorithms to speed up the process of detecting infection via medical images (CT and X-ray). Wang et al. [51] utilized deep models to extract features suitable for COVID-19 diagnosis from CT scans of confirmed cases. Chen et al. [7] leveraged UNet++ architecture [58] to detect suspicious areas on CT scans and achieved 100% accuracy on 600 scans. Similarly, Wang et al. [52] proposed a hybrid method composed of wavelet Renyi entropy, feed-forward neural network, and a three-segment biogeography-based optimization algorithm to interpret chest computed tomography images for identification of COVID-19. Conversely, Ezz et al. [20] proposed COVIDX-Net, which combined seven different CNN models, including VGG19 [45] and Google MobleNet [21], to automatically diagnose viruses in X-ray images.

Textual data has also been studied for learning more about COVID-19. Saire and Navarro [41] applied text mining methods to Twitter data to show the epidemiological impact of COVID-19 on press publications in Bogota, Colombia. Intuitively, they found that the number of tweets was positively correlated with the number of infected cases in the city. Likewise, Cinelli et al. [9] analyzed posts related to COVID-19 from Twitter, Instagram, YouTube, Reddit, and Gab. They identified different volumes of misinformation on each platform. Another related effort aims to use medical records to forecast patient admission rates [39]. Unlike the aforementioned work, in this article we have built SymptomID, a modular AI-based framework for rapid identification of symptoms using news reports. We present a case study that applies our framework in response for the current COVID-19 pandemic; however, we also envision that the proposed framework can be used in other pandemics/epidemics. To enable replication and validation of this work, we have released our trained model and dataset to support other data-driven efforts on COVID-19.

7 DISCUSSION AND CONCLUSION

Being able to identify the symptoms of a pathogen during a novel pandemic is of utmost importance. To dampen the spread of a pathogen, its symptoms need to be identified quickly so that public health officials can implement interventions to reduce the spread of the pathogen by the infected. These interventions can be as simple as asking anyone with symptoms to self-quarantine

and seek medical help. For known pathogens this is not an issue as the symptoms are well understood. However, as was the case during the COVID-19 pandemic, correctly identifying and broadcasting the symptoms of a novel pathogen can take time. This is time wasted in terms of controlling the pandemic.

Though rapid identification of symptoms of a novel pathogen is important, the sources used for this need to be trustworthy as well as fast. There is generally a tradeoff between speed and trust when it comes to reporting. Traditionally, the method through which symptoms of a novel pathogen are identified and shared has been through clinical studied and peer-reviewed scientific journals. Though extremely trustworthy, these methods can be slow as physical recruitment of subjects, data collection, and the peer-review process inevitably takes time. Even though during the early stages of the COVID-19 pandemic medical journals had a rapid track for COVID-19 research, it still took several weeks for any finding to be published. Another method that has been proposed for rapid identification of symptoms is through using self-reports on social media [9, 41]. Compared to medical journals, this method swings the pendulum to the other end in the speed vs. trust tradeoff. In other words, this method is fast but unreliable and easily manipulated by users with malicious intents (e.g., bots, trolls, etc.). SymptomID is a happy medium between these two extremes. It is based on reports of the symptoms in traditional and respected media outlets. Though not as fast as social media, news publications are faster than scientific journals as they do not go through peer-review. For the same reason, they are not as trustworthy as scientific journals; however, these publications typically do go through editorial oversight, where sources and information are checked. This makes them more trustworthy than self-reported information on social media, where there is no oversight.

SymptomID is modular and flexible, the specific methods used for each part of the framework can be replaced and the data sources can be modified or curated as needed. The system as we present it in this article relies on current state-of-the-art named entity extraction and normalization methods. As better methods are developed in the future, they can be easily plugged into our framework. Another advantage of SymptomID is that it can easily be adapted to other languages. This is important during global pandemics as a large number of reporting around the pandemic will be in different languages. Though the case study presented in this article was for English only, SymptomID can be used for most languages, as long as there are annotators available for the training phase of the framework. We propose BERT in this article for the task of named entity recognition, this model supports to 100 languages [35]. The clustering method also uses BERT to generate embeddings; once the embeddings are generated, the rest of clustering is language-independent as it works with the numerical vectors. Finally, since SymptomID relies on a pre-trained language model (i.e., BERT), it requires minimal training data to achieve high-performance (this is illustrated in Figure 4). This enables SymptomID to be trained and deployed very quickly and cheaply during pandemics.

There have been many pandemics and epidemics in the current millennia, such as SARS and MERS, [34]. COVID-19 happens to be the newest, and unfortunately, the most widespread in recent history. It is for certain that COVID-19 will not be the last pandemic that affects the world. Thus, it is only prudent to develop systems that can be used for rapid information gathering in the wake of future epidemics or pandemics. The proposed system (SymptomID) is envisioned to be one of such systems that can be used to rapidly identify the symptoms of novel pandemics using publicly available news reports. The general framework presented in Figure 1 includes steps for data collection, data annotation, named entity recognition, performance evaluation, entity extraction, and clustering, followed by insights gathering. Based on our implementation, the data collection phase (i.e., collation of relevant news reports) as described in Section 2.1 was fast and completed under 2 days, since this step leverages existing and well-established platforms for web-scraping

such as Event Registry [17]. The following the step of data annotation can be more time-consuming as this includes the laborious task of reading and curating full articles. However, our reliance on pre-trained language models (i.e., BERT) greatly reduces the need for annotated data. Thus, in this work, we directly recruited four annotators to annotate a small subset of the collected articles (300 articles). The annotations and their validation took about 2 weeks to complete. However, alternate ways to accomplish annotations faster can be through crowd-sourcing using platforms like Amazon Mechanical Turk [3]. It is important to note that the outcome of the full pipeline depends on the quality of annotations, and hence much care should be taken to validate the quality of annotations. The steps of named entity recognition (i.e., model building) and performance evaluation can be relatively quick as these leverage well-established pre-trained models such as BERT. Based on our implementation, this was completed within 1 week. Finally, the step of entity extraction, and clustering is also relatively fast, because this leverages well-established methods to generate embeddings and to perform clustering. In our implementation, this was completed within 1 day. In total, our implementation of the proposed SymptomID framework was completed in less than 30 days, but we believe future implementations can be completed within a shorter time span. However, even with limited resources (such as the ones we were operating under), our framework was operational in less than a month. This demonstrates that our framework can indeed be rapidly deployed in new and unknown pandemics in the future, even with limited resources. Note that our framework does not need to redesigned for future use, only tuned to the pandemic of interest. The tuning mainly involves collecting and annotating news articles.

As we have shown in the COVID-19 case study, SymptomID can correctly identify a large number of symptoms during a pandemic, including less-documented and atypical symptoms. In future novel pandemics, we believe that SymptomID can help speed up the identification of symptoms, allowing for rapid interventions, such as self-quarantine, to slow down the spread of the novel pathogens. Future work on improving SymptomID can be focused on integrating the three sources of information mentioned earlier: traditional media, social media, and scientific journals. As discussed, each of these data sources has its advantages in terms of volume, velocity, and veracity. By combining these data sources, SymptomID can be made faster and more reliable.

ACKNOWLEDGMENTS

The authors thank the annotators (Abigail Bartolome, Darley Sackitey, and Slyvester Coch) who contributed to preparing our training dataset.

REFERENCES

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *Proceedings of the 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS'11)*. IEEE, 702–707.

[2] E. Alsentzer, J. Murphy, W. Boag, W. Weng, D. Jindi, T. Naumann, and M. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

[3] Amazon Mechanical Turk. 2020. Retrieved from https://www.mturk.com/.

[4] I. Bedmar, P. Martínez, and M. Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'13)*.

[5] Angelo Carfì, Roberto Bernabei, Francesco Landi, et al. 2020. Persistent symptoms in patients after acute COVID-19. *J. Am. Med. Assoc.* 324, 6 (2020), 603–605.

[6] Center of Disease Control and Prevention (CDC). 2020. Symptoms of Coronavirus. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html?utm_campaign=AC_CRNA.

[7] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, and B. Zheng. 2020. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: A prospective study. *Scientific Reports* 10, 19196 (2020).

[8] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet* 395, 10223 (2020), 507–513.
[9] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. Valensise, E. Brugnoli, A. Schmidt, P. Zola, F. Zollo, and A. Scala. 2020. The COVID-19 social media infodemic. *Scientific Reports* 10, 16598 (2020).
[10] S. Clinchant, W. Jung, and V. Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*.
[11] N. Colic, L. Furrer, and F. Rinaldi. 2020. Annotating the Pandemic: Named entity recognition and Normalisation in COVID-19 Literature. Retrieved from openreview.net.
[12] A. Conneau and G. Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'19)*.
[13] Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the 1st Workshop on Social Media Analytics*. 115–122.
[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2. Retrieved from http://arxiv.org/abs/1810.04805v2.
[15] Mohamed E. El Zowalaty and Josef D. Järhult. 2020. From SARS to COVID-19: A previously unknown SARS-CoV-2 virus of pandemic potential infecting humans—Call for a One Health approach. *One Health* 9, 100124 (2020), 100124.
[16] M. Ester, H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the AAAI Annual Conference on Artificial Intelligence (AAAI'96)*.
[17] EventRegistry. 2020. Retreived from http://eventregistry.org/.
[18] L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 5 (1971), 378–382.
[19] K. Hakala and S. Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *BioNLP Open Shared Tasks@EMNLP*.
[20] E. Hemdan, M. Shouman, and M. Karar. 2020. COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images (unpublished). arXiv:2003.11055.
[21] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. Retrieved from http://arxiv.org/abs/1704.04861.
[22] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395, 10223 (2020), 497–506.
[23] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N. Kamel Boulos, Adrian Weller, et al. 2020. Leveraging data science to combat COVID-19: A comprehensive review. *IEEE Trans. Artif. Intell.* 1, 1 (2020), 85–103.
[24] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. 2020. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* 172, 9 (2020), 577–582.
[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kin, C. So, and J. Kand. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2019).
[26] J. Li, S. Shang, and L. Shao. 2020. MetaNER: Named entity recognition with meta-learning. In *Proceedings of the Annual Conference on the World Wide Web (WWW'20)*.
[27] J. Li, A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* (2020).
[28] J. Li, A. Sun, and M. Ma. 2020. Neural named entity boundary detection. *IEEE Trans. Knowl. Data Eng.* (2020).
[29] J. Li, D. Ye, and S. Shang. 2019. Adversarial transfer for named entity boundary detection with pointer networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'19)*.
[30] C. Lin, S. Bethard, D. Dligach, F. Sadeque, G. Savova, and T. Miller. 2020. Does BERT need domain adaptation for clinical negation detection? *J. Am. Med. Inf. Assoc.* 27, 4 (2020), 584–591.
[31] LocalTurk. 2020. Retrieved from https://github.com/danvk/localturk.
[32] I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*.
[33] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvist. Invest.* 30, 1 (2007), 3–26.
[34] Eskild Petersen, Marion Koopmans, Unyeong Go, Davidson H. Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili, and Lone Simonsen. 2020. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect. Dis.* 20, 9 (2020), 238–244.

[35] T. Pires, Eva Schlinger, and D. Garrette. 2019. How multilingual is multilingual BERT? arXiv:1906.01502. Retrieved from http://arxiv.org/abs/1906.01502.
[36] Temiloluwa Prioleau. 2021. Learning from the experiences of COVID-19 survivors: Web-based survey study. *JMIR Form Res* 5, 5 (2021).
[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[38] Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. arXiv:cmp-lg/9505040v1. Retrieved from https://arxiv.org/abs/cmp-lg/9505040v1.
[39] B. Roquette, H. Nagano, E. Marujo, and A. Maiorano. 2020. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Netw.* 126 (2020), 170–177.
[40] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Mat.* 20 (1987), 53–65.
[41] J. SaIRE and R. Navarro. 2020. What is the people posting about symptoms related to coronavirus in bogota, colombia?. arXiv:2003.11159. Retrieved from http://arxiv.org/abs/2003.11159.
[42] E. Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*.
[43] E. Sang and F. Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*.
[44] J. Sim and C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys. Ther.* 85 (2005).
[45] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from http://arxiv.org/abs/1409.1556.
[46] K. Tomanek and U. Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the International Conference on Knowledge Capture (K-CAP'09)*.
[47] H. Tsai, J. Riesa, M. Johnson, N. Arivazha, X. Li, and A. Archer. 2019. Small and practical BERT models for sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*.
[48] Ö. Uzuner, B. South, S. Shen, and S. DuVall. 2011. 2010 I2B2/va challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inf. Assoc.* 18, 5 (2011), 552–556.
[49] Jingqi Wang, Huy Anh, Frank Manion, Masoud Rouhizadeh, and Yaoyun Zhang. 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:2007.10286v3.
[50] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. In *ACL NLP-COVID Workshop 2020*.
[51] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, and X. Meng. 2020. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). *Eur Radiol.* 31, 8 (2020), 6096–6104.
[52] Shui-Hua Wang, Xiaosheng Wu, Zhang Yu-Dong, and Zhang Xin Tanf, Chaosheng. 2020. Diagnosis of COVID-19 by wavelet renyi entropy and three-segment biogeography-based optimization. *Int. J. Comput. Intell. Syst.* 13, 1 (2020), 1332–1344.
[53] X. Wang, X. Song, B. Li, Y. Guan, and J. Han. 2020. Comprehensive named entity recognition on CORD-19 with distant or weak supervision. arXiv:2003.12218. Retrieve from http://arxiv.org/abs/2003.12218.
[54] V. Yadav and S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. arXiv:1910.11470v1. Retrieved from http://arxiv.org/abs/1910.11470v1.
[55] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin. 2019. End-to-end open-domain question answering with BERTserini. arXiv:1902.01718v2. Retreived from http://arxiv.org/abs/1902.01718v2.
[56] Yongshi Yang, Fujun Peng, Runsheng Wang, Kai Guan, Taijiao Jiang, Guogang Xu, Jinlyu Sun, and Christopher Chang. 2020. The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. *J. Autoimmun.* (2020), 102434.
[57] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'19)*.
[58] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang. 2018. Unet++: A nested U-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

Received September 2020; revised March 2021; accepted April 2021