Extracting a Knowledge Base of COVID-19 Events from Social Media

Shi Zong[1] **Ashutosh Baheti**[2] **Wei Xu**[2] **Alan Ritter**[2]
[1]University of Waterloo [2]Georgia Institute of Technology
s4zong@uwaterloo.ca, abaheti95@gatech.edu
{wei.xu, alan.ritter}@cc.gatech.edu

Abstract

We present a manually annotated corpus of 10,000 tweets containing public reports of five COVID-19 events, including positive and negative tests, deaths, denied access to testing, claimed cures and preventions. We designed slot-filling questions for each event type and annotated a total of 28 fine-grained slots, such as the location of events, recent travel, and close contacts. We show that our corpus can support fine-tuning BERT-based classifiers to automatically extract publicly reported events, which can be further collected for building a knowledge base. Our knowledge base is constructed over Twitter data covering two years and currently covers over 4.2M events. It can answer complex queries with high precision, such as *"Which organizations have employees that tested positive in Philadelphia?"* We believe our proposed methodology could be quickly applied to develop knowledge bases for new domains in response to an emerging crisis, including natural disasters or future disease outbreaks.[1]

1 Introduction

Since December 2019, the novel coronavirus rapidly spread across the world, and consequently, a flood of COVID-19 related information has appeared on social media. This includes reports on public figures who have tested positive/negative for the virus, which often break first on Twitter, such as Bill Gates's announcement as shown in Figure 1. Besides public figures, individual users and organizations on Twitter also report COVID-19 events around the world. For example in January 2021, many sources in different countries reported an increasing number of new cases exported from the UK (Figure 2). Being able to gather this information can potentially help experts and the general public to quickly identify issues and assess the situation near real-time, complementing officially reported data which may take longer to obtain, and does not include information at the same level of granularity as that reported in natural language on news and social media.

Figure 1: Example tweet that contains a self-reported TESTED POSITIVE event.

In this paper, we present an empirical study on the extraction of large quantities of structured knowledge related to an ongoing pandemic from Twitter. To achieve this, we construct a corpus of 10,000 tweets with rich linguistic annotations, covering five event types: positive tests, negative tests, denied access to testing, deaths, claimed methods of cure and prevention. More specifically, we annotate fine-grained semantic information for each event type by designing slot-filling questions and asking annotators to highlight text spans as answers. We show that our corpus can support training BERT-based classifiers to extract structured information automatically from Twitter. While slot F1 scores vary from 0.3 to 0.9 in individual tweets (most F1 scores are greater than 0.5), we show it is possible to achieve very high accuracy by aggregating extractions over a large corpus, exploiting redundancy of information that arises when events are widely discussed on Twitter. Although many Twitter datasets have emerged after the COVID-19 outbreak, to the best of our knowledge, our work is the first to provide complex linguistic annotations to support structured information extraction.

To demonstrate the utility of our dataset, we built COVIDKB, a knowledge base that supports

[1]Our corpus (with user-information removed), automatic extraction models, and the corresponding knowledge base are publicly available at https://github.com/viczong/extract_COVID19_events_from_Twitter.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3810–3823
October 12–17, 2022.

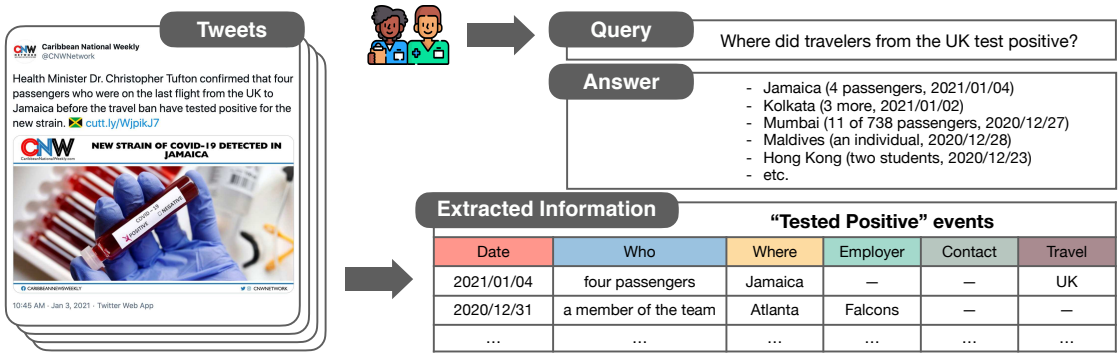


Figure 2: Overview of our COVID-19 event extraction system, which continuously extracts and indexes structured information about publicly reported events from Twitter. Users can enter structured queries to retrieve relevant tweets, such as `{location:?, travel:UK}` to find test positive cases that are exported from the UK.

structured queries over COVID-19 events, by indexing events extracted by our model over millions of tweets. Our system allows users to execute structured search queries over the extracted events, answering questions such as *"Which organizations in Houston have reports of employees who tested positive?"* or *"Who tested positive that had close contact with Boris Johnson?"* (see Figure 2). We envision COVIDKB could help address the issue of information overload for professionals (Zhang et al., 2020) who need to stay on top of recent developments related to COVID-19, including journalists (Karmakharm et al., 2019), epidemiologists and public policymakers. Our extractor can also detect claims about methods of cures and prevention of the disease, which could be useful in helping to track online misinformation (Thorne et al., 2018; Stefanov et al., 2020; Hossain et al., 2020).

2 Related Work

Event Extraction from Twitter. There has been much interest in extracting events from Twitter. For example, Ritter et al. (2012) built a system for open domain event extraction. Recent work also explored extraction of cybersecurity events (Ritter et al., 2015; Chang et al., 2016), including denial of service attacks (Chambers et al., 2018) and software vulnerabilities (Zong et al., 2019). Zhou et al. (2017) use a nonparametric Bayesian mixture model for event extraction. In this work, we design event types and attributes that are specific for COVID-19 and develop automatic NLP tools for extracting structured information from tweets.

Existing COVID-19 Datasets. There have been many datasets that collect tweets related to COVID-19 (Chen et al., 2020; Banda et al., 2020). However, most are either unlabeled or provided with general-purpose NLP model predictions, rather than structured linguistic annotations of COVID-specific information, as in this work. For example, Twitter officially releases a stream with predicted entities (such as *person* and *place*) and topic labels (such as *sports* and *movies*). Qazi et al. (2020) released a COVID-19 collection of geo-located tweets that contain COVID relevant keywords and hashtags. Dimitrov et al. (2020) put together 8 million tweets with automatically generated entity linking and sentiment scores. Hu et al. (2020) presented a large-scale dataset of 40 million raw posts from Weibo with no annotations. There also exist a few datasets that contain human annotations at the time of writing. For example, Hossain et al. (2020) annotated 5,000 tweets for studying COVID-19 misconceptions. Nguyen et al. (2020) classified 10,000 tweets as informative and uninformative. Amini et al. (2021) annotated a dataset of mechanism relations from COVID-19 related scientific papers. Compared to prior work, we provide more fine-grained human annotations on text spans with predefined slots for COVID-19 events. Our annotations can support training supervised learning models that are capable of extracting structured information (Adrian Bejan and Harabagiu, 2014; Venugopal et al., 2014), similar to other influential datasets in information extraction and question answering, such as KBP (Ji et al., 2011) and SQuAD (Rajpurkar et al., 2016).

Social Media Monitoring for Public Health. Analyzing social media and other user-generated web data for monitoring public health has been an active research area. For example, Google Flu Trends (GFT) uses search engine query data to detect influenza epidemics (Ginsberg et al., 2009). Paul

et al. (2014) use the Twitter message content to forecast influenza rates. GFT has been found to over-estimate influenza-like illness (Lazer et al., 2014). In contrast to GFT, our main focus is to develop methods that process large quantities of raw tweets into a *structured* format to help people find specific information, rather than forecasting or nowcasting official statistics.

3 An Annotated Corpus for COVID-19 Event Extraction

To extract structured knowledge from tweets, we formulate the problem as a supervised slot filling task (Jurafsky and Martin, 2000; Benson et al., 2011; Ji et al., 2011). Specifically, given a tweet, annotators are asked to first identify whether it contains a relevant event, then highlight the text spans of answers that correspond to a list of pre-defined questions for each event type (detailed questions are in Table A2).

3.1 Data Collection

We consider five event types related to COVID: TESTED POSITIVE, TESTED NEGATIVE, CAN NOT TEST, DEATH, and CURE & PREVENTION. The design of these event types is inspired by the statistics reported in Johns Hopkins COVID-19 dashboard, which are of interest to the public and epidemiologists.[2] The first four types aim to extract structured information about events related to COVID-19, many of which are news stories about public figures. We have been continuously collecting Twitter data related to COVID-19 since 2020/01/15 by tracking relevant keywords using the Twitter API, such as *tested positive* for TESTED POSITIVE events (see Table A1 for a full list of our carefully selected keywords). As we will shown in Section 6.1, our fixed set of keywords are able to track the evolution of pandemic even over a period of two years, although a dynamic selection of keywords is promising to explore in future work.

Preprocessing. In this work, we mainly focus on English tweets, identified by using `langid.py` (Lui and Baldwin, 2012). We remove retweets and other duplicates, keeping the tweet that was posted earliest. Before de-duplication process, all URLs and user mentions are removed. We also use Jaccard similarity with a threshold of 0.7 to remove near-identical tweets that are posted same-day.

[2] https://coronavirus.jhu.edu/map.html

Event Type	# Anno. Total	# Event Specific	# Slots
TESTED POSITIVE	3,000	2,146	9
TESTED NEGATIVE	1,700	893	8
CAN NOT TEST	1,700	680	5
DEATH	1,800	626	6
CURE & PREV.	1,800	832	3
Total	10,000	5,177	31

Table 1: Statistics of COVID-19 Twitter Event Corpus.

3.2 Annotation Process

We randomly sample 10,000 tweets from five event types to annotate. The train and dev sets consist of 7,500 annotated tweets, that were published between 2020/01/15 and 2020/04/26. To construct the test set, we annotated 2,500 tweets, 500 for each event type, that were published from a later time period between 2020/04/27 and 2020/06/27. This simulates a real-world scenario that a model is trained on historical records and then applied to future data. Table 1 shows the overall statistics of our labeled corpus.

3.2.1 Two-phase Annotation

Given a tweet, annotators are asked to first identify whether it contains a relevant event, then highlight the text spans of answers that correspond to a list of pre-defined questions for each event type in Table A2. We hire crowd workers on Amazon's Mechanical Turk to annotate our full dataset. Each of the 10,000 tweets is annotated by 7 crowd workers in two steps. We paid crowd workers $0.4-0.5 per HIT and gave extra bonuses to annotators with high annotation quality. The hourly pay was approximately $8.55. The main portion of our annotation interface is shown in Figure A1.

Part 1: Event Specificity. Although tweets have been filtered by keywords for each event type, many of them are generic news reports, such as, *"37% of those tested under 17 for Coronavirus in California tested positive"*. Since we are interested in capturing tweets with detailed information, we first ask the annotators to judge whether a tweet refers to a specific event. For example, for tweets about positive tests, we ask the annotators whether a tweet is about an individual or a small group of people testing positive. Annotators proceed to the next step only if they answer yes to this question.

Part 2: Slot Filling. In the second step, we ask a set of pre-defined questions specifically designed for each event type, as listed in Table A2. The annotators are provided with candidate answers, which include all noun phrases and named entities

extracted by a Twitter-specific NLP tool (Ritter et al., 2011),[3] in a drop-down list. We also combine noun phrases if they are adjacent or separated by a preposition.[4] We include *author of the tweet* as an additional option for the WHO questions.[5] For each tweet, annotators have an average of 10 to 11 possible answers to choose from, and are allowed to choose more than one answer for WH-questions.

3.2.2 Inter-annotator Agreement

During annotation, we track crowd workers' performance by comparing their annotations with the majority vote of other workers and remove workers' qualifications if their F1 scores fall below 0.65.[6] For the first step of annotation on specificity, the inter-annotator agreement between crowdsourcing workers is 0.68, measured by Fleiss κ (Artstein and Poesio, 2008). We observe a 0.62 F1 score for selected text spans between annotators in our slot filling task, by using each Turker's annotation in turn as the prediction, and then compare it against answers from all other workers. Same method to calculate inter-annotator agreement for text spans has been used in Yang et al. (2018) and Lee and Sun (2019).

To further validate the quality of slot-filling annotations from the crowdsourcing workers, we hired an experienced in-house annotator to carefully re-annotate the test set (2,500 tweets total, with 500 from each event; see Section 3.1 for details). The in-house annotator is paid $15 per hour. By comparing crowdsourcing workers with our in-house annotator, we find individual annotators do miss some examples, which is similar to previous reports on linguistic annotations on relations and events, such as ACE 2005 (Min and Grishman, 2012). However, by aggregating annotations from multiple crowdsourcing workers,[7] we observe high agreement (an average of 0.72 F1 score) with our in-house annotator. We also ask the in-house annotator to examine a sample of tweets to find answer spans that are not identified as candidates by the automatic NLP tool. We find this scenario occurs in less than 2% of tweets in our dataset.

3.3 Corpus Analysis

Basic Statistics. Our annotated tweets have an average length of 34.6 tokens with a standard deviation of 15.6 tokens. We note 41.42% of the tweets have external links and 29.64% include hashtags. Examples of our annotated tweets are in Table A3.

Bots and Organizational Accounts. Among all the 9,656 unique users, 2.4% are potentially bots, as identified by the Botometer API (Varol et al., 2017). We also note that 4.1% of tweets about CURE & PREVENTION are potentially posted by bots. Estimated by the Humanizr (McCorriston et al., 2015), 18.5% of user accounts in our data belong to organizations, rather than individuals.

4 Automatic Event Extraction

We now use our annotated corpus to train and evaluate supervised learning methods for automatic COVID-19 event extraction. Each slot filling question is treated as a binary classification task: given a tweet t and the candidate span c, the classification model $f_{e,s}(t, c) \rightarrow \{0, 1\}$ predicts whether c correctly answers the question for the slot s of event type e.

4.1 Experimental Settings

Baselines. We conduct experiments with two methods for automatic COVID-19 event extraction:

(1) *Logistic Regression.* We implemented a basic logistic regression classifier using bag-of-ngram features (n = 1, 2, 3). The target chunk c is replaced with a special token before computing n-grams.

(2) *Fine-tuning BERT.* We also fine-tune a BERT based classifier (Devlin et al., 2019) that takes a tweet t as input and encloses the candidate phrase c in the tweet with a pair of special entity start `<E>` and end `</E>` markers. The BERT hidden representation of token `<E>` is then fed as input to a linear layer to produce the binary prediction. Since our dataset consists of COVID-19 related tweets, we use COVID-Twitter-BERT (CT-BERT; Müller et al., 2020), an uncased $\text{BERT}_{\text{large}}$ model pre-trained on 22.5M in-domain tweets, related to COVID-19 (0.6B tokens).

Implementation Details. By design, many slots within an event are semantically related. For example, the `age` slot is directly related to the `who` slot.

[3] github.com/aritter/twitter_nlp

[4] We notice in some cases these noun phrases are not perfect and may include extra words. Annotators are instructed that a candidate answer should only be chosen when it contains no more than three extra words.

[5] These annotations are used to develop classifiers that can detect and remove instances where users publicly report information about themselves.

[6] For more discussions on managing workers on Amazon Mechanical Turk, we recommend reading: https://homes.cs.washington.edu/~msap/notes/turking-tips.html.

[7] We consider to include a span annotation for slot-filling task if 3 out of 7 MTurk annotators agree.

During development, we found it beneficial to train the final linear layers of all slots for a given event using the shared CT-BERT parameters. All shared CT-BERT models are fine-tuned with a 2e-5 learning rate using Adam (Kingma and Ba, 2015) for 4 epochs. This model has about 345M parameters.

4.2 Results

We evaluate our model performance for event type identification and slot filling on the test data, which consists of 2,500 tweets. Event types can be directly derived from the slot-filling predictions: an event is identified if text spans are extracted for any of the pre-defined slots associated with the event types by our models. Table 2 presents F1 scores on classifying event specific tweets on the test set. Table 3 presents slot filling results of the Logistic Regression, $\text{BERT}_{\text{large}}$ and CT-BERT models, as measured by precision, recall and F1 metrics.[8]

We observe that CT-BERT gives the best overall performance, which outperforms the bag-of-ngrams baseline. CT-BERT has F1 scores ranging from 0.3 to 0.9, depending on the slot for extracting events from individual tweets. The F1 score for most slots is greater than 0.5 and the final micro average F1 achieved by CT-BERT is 0.67. While we do notice some slots have low F1 scores, these slots are normally associated with few annotations in the train set. Besides, we will show in Section 5 that the performance of our CT-BERT model is sufficient to support the development of a knowledge base, which achieves much higher accuracy for COVID-19 event extraction from Twitter by aggregating extractions over a large volume of tweets.

Event Type	BERT	CT-BERT
TESTED POSITIVE	0.90	0.89
TESTED NEGATIVE	0.72	0.77
CAN NOT TEST	0.72	0.73
DEATH	0.73	0.79
CURE & PREVENTION	0.64	0.70

Table 2: F1 scores for classifying event specific tweets.

5 COVIDKB Knowledge Base

We have built models that can extract structured information related to COVID-19 from individual tweets. To demonstrate the utility of our annotated dataset and models, we create a knowledge base (Figure 2) that enables structured search over COVID-19 events that are automatically extracted from Twitter.

[8]We omit reporting results for a few slots with less than 20 annotations in test set, such as the `duration` slot for TESTED NEGATIVE and the `when` slot for CAN NOT TEST.

TESTED POSITIVE		Logistic	BERT	CT-BERT		
Slot	**#**	**F1**	**F1**	**P**	**R**	**F1**
who	375	.48	.82	.86	.82	**.84**
close contact	61	.02	.44	.65	.61	**.63**
relation	21	0.0	.51	.83	.48	**.61**
employer	121	.15	.44	.65	.54	**.59**
recent travel	27	0.0	**.36**	.44	.26	.33
when	22	.05	.38	.47	.36	**.41**
where	176	.27	.60	.91	.49	**.64**
TESTED NEGATIVE		Logistic	BERT	CT-BERT		
Slot	**#**	**F1**	**F1**	**P**	**R**	**F1**
who	274	.23	.67	.78	.68	**.73**
close contact	27	0.0	0.0	.24	.48	**.32**
relation	56	0.0	**.55**	.77	.41	.53
where	49	0.0	**.44**	.36	.55	**.44**
when	27	0.0	0.0	.35	.41	**.38**
CAN NOT TEST		Logistic	BERT	CT-BERT		
Slot	**#**	**F1**	**F1**	**P**	**R**	**F1**
who	153	.16	.57	.77	.58	**.66**
relation	70	.08	.37	.69	.34	**.46**
symptoms	52	.06	.43	.55	.62	**.58**
where	30	.20	.44	.55	.40	**.46**
DEATH		Logistic	BERT	CT-BERT		
Slot	**#**	**F1**	**F1**	**P**	**R**	**F1**
who	139	.29	.68	.83	.76	**.79**
relation	37	0.0	.59	.96	.65	**.77**
when	33	.26	**.75**	.66	.82	.73
where	65	.22	.54	.70	.60	**.64**
age	33	.18	.78	.89	.94	**.91**
CURE & PREVENTION		Logistic	BERT	CT-BERT		
Slot	**#**	**F1**	**F1**	**P**	**R**	**F1**
opinion	152	.08	.66	.85	.59	**.69**
what	261	.22	.66	.83	.64	**.72**
who	235	.08	**.51**	.87	.37	**.51**
Micro Average F1		.25	.62		**.67**	

Table 3: Slot-filling results on the test set for logistic regression, $\text{BERT}_{\text{large}}$ and CT-BERT based classifiers. # is the count of gold annotations in the test data for each slot type. F1 in bold are highest in their row.

5.1 COVIDKB Overview

COVIDKB Statistics. Until 2022/04/01 (start dates are in Table 1), our COVIDKB knowledge base has contained around 4.2M extracted events from over 20M raw tweets and is continuously growing by processing tweets daily. Events are extracted from deduplicated tweets, which follow the same pre-processing steps in Section 3.1. Breakdowns of our extracted events are listed in Table A4.

Interacting with COVIDKB. COVIDKB supports a simple structured query interface where a user specifies one or more text-filters as a query (see Figure A2). This includes two SQL operators,

Simple Queries	P@10	P@20	P@50	P@100
(S-1) Who tested positive on 2021/06/15?	100	100	100	99
(S-2) Who is promoting cures or preventions?	90	90	96	91
(S-3) Where were people not able to access testing?	100	100	100	100
(S-4) How long did people wait for negative test results?	100	85	82	82
(S-5) Which organizations have employees who tested positive?	90	90	90	94
Advanced Queries	**P@5**	**P@10**	**P@20**	**P@50**
(A-1) Who tested positive that had close contact with Boris Johnson?	80	70	60	58
(A-2) Who tested positive that has a recent travel to Japan?	100	100	100	96
(A-3) What methods of cure and prevention do people think are effective?	80	90	85	88
(A-4) Where did people test positive who traveled from the UK?	100	100	100	100
(A-5) Which organizations have employees that tested positive in San Francisco?	100	100	90	92

Table 4: Queries used to evaluate results returned by our knowledge base, reported using Precision@K. The queries are presented here in natural language for improved readability. Simple queries can be realized as a single `GroupBy` operation; advanced queries contain both `GroupBy` and `Select`. For example, the structured query for A-1 is `{who:?, contact:'Boris Johnson'}`. All queries use the default time range (from 2020/01/15 to 2022/03/01) unless explicitly specified.

`Select` and `GroupBy`. For the event slot queried by the user, using a special token "`?`", our system returns a list of all unique answers, which were extracted from tweets that match the search criteria and sorted by mention frequency. For example, a user might enter the query `{employer:?, location:'San Francisco'}`, and the system will return a list of organizations located in San Francisco where one or more employees tested positive. This simple interface enables a rich set of informative queries over events that were automatically extracted by our classification models.

Table 4 shows a list of example queries supported by COVIDKB. Queries are randomly generated by the authors of this paper. Note that throughout this paper we present queries to our system using natural language questions for the sake of readability. In each case, translation to a structured query is straightforward. The user specifies zero or more fields to filter on (`Select`) and a single field to group the results by (`GroupBy`). As our knowledge base is continuously updating, users can further combine above structured queries with different time ranges (e.g., query S-1 in Table 4 sets the start and end dates as 2021/06/15). We do not address the problem of automatically mapping natural language questions to structured queries (Suhr et al., 2018) in this work, though there is significant prior work on this topic (Artzi and Zettlemoyer, 2011; Berant et al., 2013).

5.2 COVIDKB Evaluation

Precision of Top Extractions. We evaluate the accuracy of answers returned by our knowledge base using 10 sample queries and manually inspect the correctness of the top K extractions, sorted by frequency (tweets have been deduplicated as mentioned in Section 5.1). As reported in Table 4, our knowledge base has high precision for nearly all queries, including queries involving slots with few annotations. For example, the `duration` slot is excluded in Table 3, because there are fewer than 20 instances in the test set, whereas COVIDKB still achieves good performance on queries involving this slot, thanks to the redundancy of information in Twitter. Table 5 present outputs returned by our knowledge base.

Extracted Answer Types. In Table 6, we also show a manual analysis of the types of answers, which are correctly extracted by our system for queries that target the `who` slot. We define two answer types: (1) **Specific** entities, which are clear referents to people (mostly public figures), such as *Boris Johnson* and *Dominic Cummings*; (2) **Generic** entities, which are typically nominal references, such as *a woman*. We observe that the percentage of generic answers varies heavily depending on the query. For example, query A-1 about people who had close contact with Boris Johnson consists almost entirely of references to specific public figures, whereas A-2, about people who tested positive after traveling from Japan yields only generic references.

5.3 Error Analysis

We perform an error analysis to understand the types of errors our knowledge base contains. Two authors of this paper carefully conducted manual in-

(S-1) Who tested positive on 2021/06/15?
Teofimo Lopez tests positive for COVID-19, entire Triller PPV card pushed back to August (by @mookiealexander) https://t.co/DoaHNb9Z4T
Vaccinated Hawaiian resident tests positive for Delta coronavirus variant https://t.co/0IJ8QfpYS9
Royal Caribbean cruise ship launch, sailings postponed after crew members test positive for COVID-19… https://t.co/VVrOdS6uEX
(A-1) Who tested positive that had close contact with Boris Johnson?
#news PM Boris Johnson in self-isolation after coming into contact with a lawmaker who tested positive for COVID-19 https://t.co/Kcy2X3M6vJ
Jair Bolsanaro has tested positive for Covid-19. Noval Djokovic and Boris Johnson had it. Life sometimes comes a full circle very fast.
WH says Trump spoke with Boris Johnson and "wished him a speedy recovery" after the British PM tested positive for coronavirus.
Boris Johnson's senior adviser, Dominic Cummings, is self-isolating at home after developing #coronavirus symptoms. http://bbc.in/2WQhbsZ Last week, the PM and Health Secretary Matt Hancock both tested positive for #Covid19. WATCH: https://bbc.in/2Jv55xj #Newsnight
(A-3) What methods of cure and prevention do people think are effective?
Very good indeed but you need also to remind them keeping social distancing, another basic protective measure to prevent the spread of #covid19.
Just like washing your hands is necessary to prevent from Coronavirus, inspecting your personal protective equipment https://t.co/xjY7FRgsV1
Two men in Georgia drank disinfectants in efforts to prevent COVID-19, officials say http://a.msn.com/01/en-us/BB13kJMw?ocid=st…

Table 5: Examples of correct extractions and errors returned by our knowledge base for sample queries. We use different colors for marking the types of extracted text spans (see Section 5.3 for more details for the error types): **correct extraction**, **classification errors**, **segmentation errors**, and **ambiguous cases**.

Query ID	# Corr / # All	Specific	Generic
S-1	99 / 100	63.6%	36.4%
S-2	91 / 100	75.8%	24.2%
A-1	29 / 50	100.0%	0.0%
A-2	48 / 50	6.2%	93.8%

Table 6: Analysis of answer types in response to the queries (where applicable) in Table 4. The percentage of generic answers varies significantly.

spections for all the returned results of our sample queries in Table 4. 67 incorrect extractions were identified in 750 extractions, which can be grouped into four major categories: classification errors (58.2%), segmentation errors (37.3%), ambiguous cases (13.9%) and others (4.5%). We present some examples of these errors in Table 5.

Classification Errors. We notice our BERT based model struggles with slots that may involve subtle inferences, such as relation or close contact, although the limited number of annotations for these slots might also be a factor in this type of error. For example, in the second tweet of query A-1 in Table 5, the tweet does not imply that *Jair Bolsanaro* was in close contact with *Boris Johnson*; in the third tweet of query A-1, the model fails to identify that *Boris Johnson* and *the British PM* refer to the same person.

Segmentation Errors. In some cases the extracted items contain extra tokens because of chunker errors, for example *georgia drank disinfectants* was extracted as a cure method. We also notice our choice of only extracting noun phrase chunks does not capture verb phrases for the CURE & PREVENTION category. For example, instead of extracting *washing your hands* and *don't touch your face* as prevention methods, our system only extracts *your hands* and *your face* (see query A-3 in Table 5).

Ambiguous Cases. In some cases, it is debatable whether an extraction is correct without additional context. For instance in the last tweet of query A-1 in Table 5, we do not know if *Dominic Cummings* tested positive, although the tweet seems to indicate that he might have been infected. We consider the extraction to be an error in this case, since the tweet did not specifically mention that he tested positive.

6 Case Studies

6.1 Correlation with Official Data Sources

To investigate whether statistics of events in COVIDKB correlate with official data sources, we plot the reported global positive cases and the number of extracted tested positive events from our knowledge base over time in Figure 3. Global reported positive numbers are from Center for Systems Science and Engineering at Johns Hopkins University.[9] We use 7-days moving average when drawing two time series curves. We observe that for both two waves in 2021 and current Omicron wave (highlighted in grey in Figure 3), our extracted events follow similar trend as actual reported cases globally and also show peaks. This analysis provides evidence to support quality of the

[9] https://github.com/CSSEGISandData/COVID-19

extracted information in COVIDKB, and suggests our knowledge base may contain information that could be used to analyze emerging dynamics of the pandemic. However as mentioned previously, the main use-case for COVIDKB is to enable semantic search to help journalists, epidemiologists or other professionals quickly analyze information posted on social media.

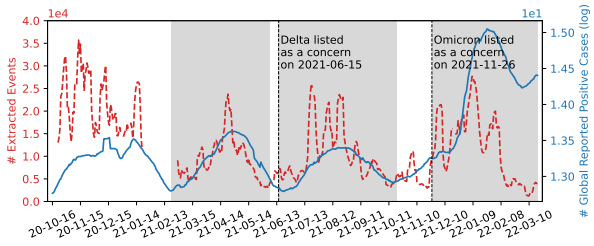


Figure 3: Number of extracted positive events and the actual global reported positive cases (log) show the similar trends in three waves (in grey). Data from 2021/01/21 to 2021/02/26 is missing due to technical issues.

6.2 Analyzing Claimed Cures and Preventions

Public's Attention Shifts over Time. Our knowledge base could also be helpful in monitoring public attention shifts regarding potential treatments and preventative measures over time. To demonstrate this, we analyze the top frequently mentioned potential cure and prevention methods that people believe are effective within different time ranges (a visualization of top 15 results are in Table A5). Time ranges are roughly divided to follow the global trends of the pandemic shown in Figure 3.

We observe people's opinions regarding certain cure and prevention methods remain unchanged throughout the whole pandemic, including *social distancing*, *hydroxychloroquine*, *(wash) your hands* and *masks*. As time proceeds, there is more focus on medical treatments. For example, *vaccine* and *vaccination* are more frequently discussed. Drugs also draw attention, especially in the last time range (from 2021/10/16 until now): we notice a variety of drugs appear in our knowledge base, including *fluvoxamine*, *monoclonal antibodies*, *AstraZeneca antibody drug* and *Israeli drug*.

We note not all above methods are actually effective for coronavirus. Researchers hold a mixed view for treatments such as *hydroxychloroquine* and *ivermectin*.[10] This type of automatically extracted information in COVIDKB could be helpful to track the spread of misinformation online.

Who is promoting cures? We also analyze the returned results from query S-2 to understand who is promoting cures. A variety of people and organizations are observed, most frequent 10 of which are *Donald Trump*, *China*, *scientists*, *CDC*, *White House*, *Jim Bakker*, *Pfizer*, *Madagascar*, *Dr. Fauci*, and *Bill Gates*.

7 Conclusion

In this paper, we presented a corpus of 10,000 tweets annotated with 5 types of events and 28 slots. We showed that our corpus supports automatic extraction of COVID-19 events using supervised learning. By aggregating extractions over millions of tweets, our approach can accurately answer a range of structured queries about events that are publicly reported in real-time on Twitter. Our knowledge base could be a useful tool for epidemiologists, journalists and policymakers to more efficiently track the spread of this new disease. This work also presents a case-study on how an information extraction system can be rapidly developed for a new domain in response to an emerging crisis. For example, our methodology could be applied to develop knowledge bases for natural disasters (Spiliopoulou et al., 2020) or future disease outbreaks.

Ethical Considerations

This study was conducted under the approval of the Institutional Review Board (IRB) of our university and complies with Twitter's terms of service. Following Twitter's policy for content redistribution, we will only release our annotated corpus that contains Tweet IDs (not Tweet Objects) and a list of character offsets corresponding to the annotated mentions. We will not release any user information or demographic data. Our event extractors produce structured representations of information that was explicitly and publicly stated. We do not derive or infer any potentially sensitive characteristics or health information that may violate users' privacy. Almost all events that are currently indexed by our knowledge base come from public news reports.

[10] For example, Ivermectin has been used in clinical trials: https://www.covid19treatmentguidelines.nih.gov/therapies/antiviral-therapy/ivermectin/. However, it is not approved or authorized by FDA: https://www.fda.gov/consumers/consumer-updates/why-you-should-not-use-ivermectin-treat-or-prevent-covid-19.

To further protect users' privacy, we specifically designed two slot-filling questions during annotation in order to detect and remove cases where users publicly report information about themselves, or a person with whom they have a close relationship.

Our knowledge base should be used with caution, as we note the Twitter users are not representative samples of the total population; posts from Twitter users are also not necessarily representative samples of public opinions (Wojcik and Hughes, 2019). As Twitter Stream API provides only 1% of all public tweets, our knowledge base naturally is not able to index all reported cases online. Our extractors may contain other unknown biases due to data collection process, for example they might perform worse on African American English. All these limitations should be taken into consideration in any application that makes use of our data.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions. This material is based in part on research sponsored by the NSF (IIS-1845670) and IARPA via the BETTER program (2019-19051600004), DARPA via the ARO (W911NF-17-C-0095) in addition to an Amazon Research Award. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation therein.

References

Cosmin Adrian Bejan and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics*.

Aida Amini, Tom Hope, David Wadden, Madeleine van Zuylen, Eric Horvitz, Roy Schwartz, and Hannaneh Hajishirzi. 2021. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*.

Yoav Artzi and Luke Zettlemoyer. 2011. Bootstrapping semantic parsers from conversations. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 421–432.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration. *arXiv preprint arXiv:2004.03688*.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Nathanael Chambers, Ben Fry, and James McMasters. 2018. Detecting denial-of-service attacks from social media text: Applying NLP to computer security. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ching Yun Chang, Zhiyang Teng, and Yue Zhang. 2016. Expectation-regulated neural model for event mention extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. TweetsCOV19 – a knowledge base of semantically annotated tweets about the COVID-19 pandemic. *arXiv preprint arXiv:2006.14492*.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 at EMNLP 2020*.

Yong Hu, Heyan Huang, Anfan Chen, and Xian-Ling Mao. 2020. Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Heng Ji, Ralph Grishman, and Hoa Dang. 2011. Overview of the TAC2011 knowledge base population track. In *TAC 2011 Proceedings Papers*.

Daniel Jurafsky and James H Martin. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Grace E. Lee and Aixin Sun. 2019. A study on agreement in pico span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1149–1152, New York, NY, USA. Association for Computing Machinery.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*.

Bonan Min and Ralph Grishman. 2012. Compensating for annotation errors in training a relation extractor. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. GeoCOV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from Twitter. In *Proceedings of the 24th International Conference on World Wide Web*.

Evangelia Spiliopoulou, Salvador Medina Maza, Eduard Hovy, and Alexander G Hauptmann. 2020. Event-related bias removal for real-time disaster events. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3858–3868.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alane Suhr, Srinivasan Iyer, and Yoav Artzi. 2018. Learning to map context-dependent sentences to executable formal queries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*.

Deepak Venugopal, Chen Chen, Vibhav Gogate, and Vincent Ng. 2014. Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Stefan Wojcik and Adam Hughes. 2019. Sizing up twitter users. Technical report, Pew Internet and American Life Project.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the COVID-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125*.

Deyu Zhou, Xuan Zhang, and Yulan He. 2017. Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Shi Zong, Alan Ritter, Graham Mueller, and Evan Wright. 2019. Analyzing the perceived severity of cybersecurity threats reported on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Dataset

A.1 Keywords for Data Collection

We provide the keywords used for collecting data along with starting date in Table A1. Keywords in our experiments are carefully chosen to both have a wide coverage of tweets with different linguistic phenomena and have a good precision of collecting tweets that are relevant to our tasks.

Event Type	Start From	Keywords
TESTED POSITIVE	2020/01/15	(test OR tests OR tested) positive AND VIRUS
TESTED NEGATIVE	2020/02/15	(test OR tests OR tested) negative AND VIRUS
CAN NOT TEST	2020/01/15	(can't OR can not) get (tested OR test OR tests) (can't OR can not) be tested (couldn't OR could not) get (tested OR test OR tests) (couldn't OR could not) be tested
DEATH	2020/02/15	(died OR pass away OR passed away) AND VIRUS
CURE & PREVENTION	2020/03/01	(cure OR prevent) AND VIRUS

Table A1: Keywords used for each event type. We consider the following variants for VIRUS: VIRUS = (COVID19 OR COVID-19 OR corona OR coronavirus).

A.2 Data Annotation

The complete slot filling questions used for annotating COVID-19 events are listed in Table A2. We also provide the annotation interface shown to Mechanical Turk workers in Figure A1.

Event Type	Slot Name	Slot Filling Questions
TESTED POSITIVE —— TESTED NEGATIVE	who	Who tested positive (negative)?
	close contact	Who was in close contact with the person who tested positive (negative)?
	relation	Does the affected person have a relationship with the author of the tweet?
	employer	Who is the employer of the person who tested positive?
	recent travel	Where did the people who tested positive recently visit?
	when	When were positive (negative) cases reported?
	where	Where were positive (negative) cases reported?
	age	What is the age of the people who tested positive (negative)?
	duration	How long did it take to know the result of the test?
CAN NOT TEST	who	Who can not get a test?
	relation	Does the untested person have a relationship with the author of the tweet?
	when	When was the person unable to obtain a test?
	where	Where was the person unable to obtain a test?
	symptoms	Is the affected person currently experiencing any COVID-19 related symptoms?
DEATH	who	Who died from COVID-19?
	relation	Does the deceased person have a personal relationship with the author of the tweet?
	when	When was the death reported?
	where	Where was the death reported?
	age	What is the age of the person who died?
CURE & PREVENTION	opinion	Does the author of the tweet believe cure/prevention is effective?
	what	Which method of cure/prevention is mentioned?
	who	Who is promoting the cure or prevention?

Table A2: Slot filling questions used for annotating COVID-19 events.

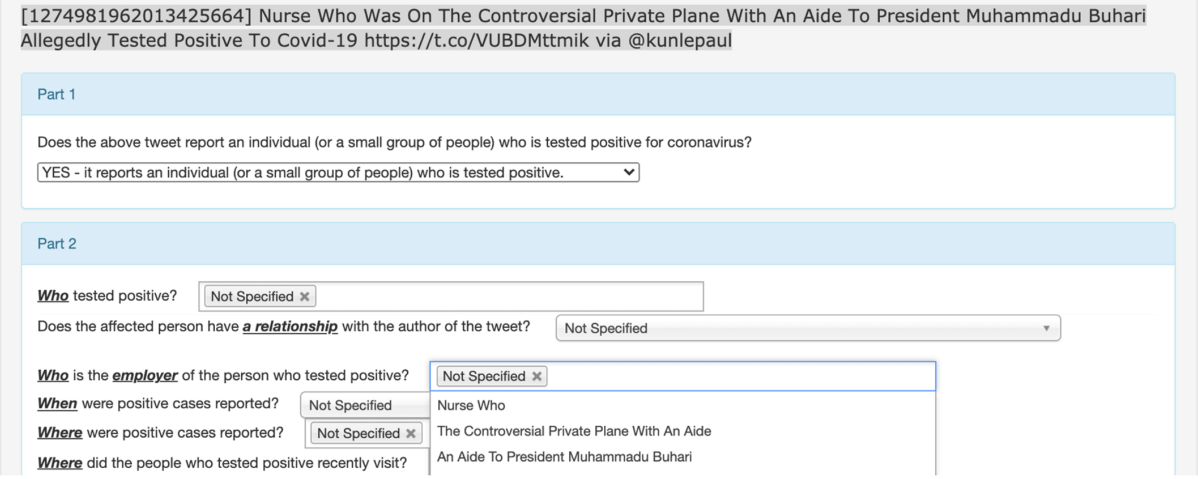


Figure A1: Main portion of the annotation interface shown to Mechanical Turk workers for annotating TESTED POSITIVE events.

A.3 Annotated Samples

Examples of our annotated tweets are presented in Table A3.

Event Type	Tweet	Annotations
POSITIVE	*#Karnataka | A 26-year-old man returning from #Greece tested positive for #COVID19, becoming the fifth positive case in the state, a health official said on Thursday. #CoronavirusPandemic #COVID #COVID19india* `[URL]`	WHO AGE WHERE RECENT V.
NEGATIVE	*Live updates: Boris Johnson tested negative for Covid-19 on leaving hospital, says Downing Street #coronavirus*	WHO
DEATH	*'#TopChef Masters' winner Floyd #Cardoz dies after #coronavirus diagnosis' "World-renowned chef Floyd Cardoz died Wednesday in New Jersey at age 59." "Cardoz admitted himself to the hospital on March 17 after feeling feverish."*	WHO AGE WHERE WHEN
CAN NOT TEST	*Nurse working in ITU couldn't get tested, & was told that the test was "very expensive", so he couldn't have a test.* `[URL]` ...	WHO

Table A3: Examples of our annotated tweets.

B COVIDKB Knowledge Base

B.1 Statistics of Our Knowledge Base

We report the number of extracted events along with the breakdown statistics for each slot in Table A4.

Event Types	# Extracted	Number of Events per Slot											
		who	relation	when	where	age	close contact	employer	recent travel	duration	symptoms	opinion	what
TESTED POS	2,354,363	2,098,964	164,126	81,053	602,552	32,361	122,952	264,275	84,157	–	–	–	–
TESTED NEG	411,071	387,354	47,325	17,044	28,447	851	7,733	–	–	9,049	–	–	–
CAN NOT TEST	30,552	26,468	17,432	94	7,637	–	–	–	–	–	14,881	–	–
DEATH	779,074	629,323	91,121	164,282	230,672	143,270	–	–	–	–	–	–	–
CURE & PREV.	665,422	319,077	–	–	–	–	–	–	–	–	–	270,493	461,290
Total	4,240,482	3,461,186	320,004	262,473	869,308	176,482	130,685	264,275	84,157	9,049	14,881	270,493	461,290

Table A4: Number of extracted events, with a breakdown for each slot in our knowledge base. Slot filling questions that are not applied to specific event types are marked with "–".

B.2 Interface of Our Knowledge Base

Our structured query interface of the knowledge base is presented in Figure A2.

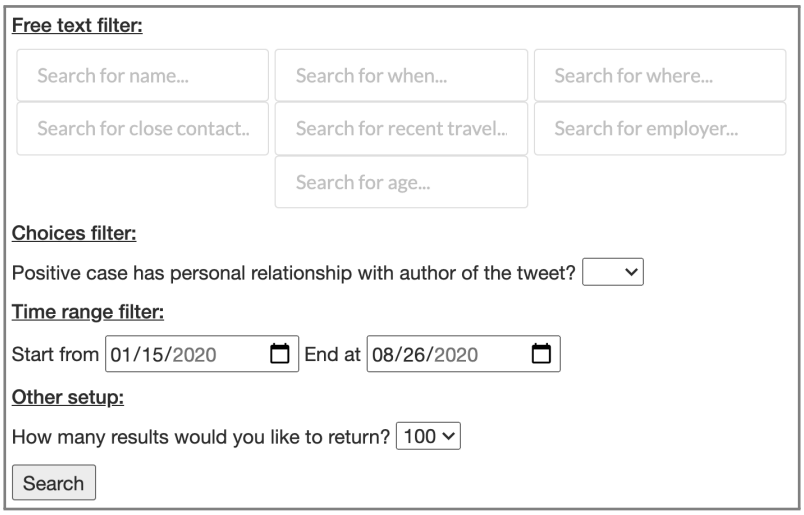


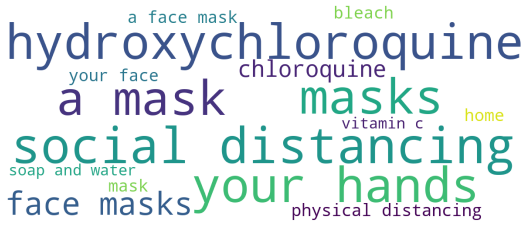
Figure A2: Structured query interface of our knowledge base.

B.3 Public Attention Shifts for Cure and Prevention Methods over Time

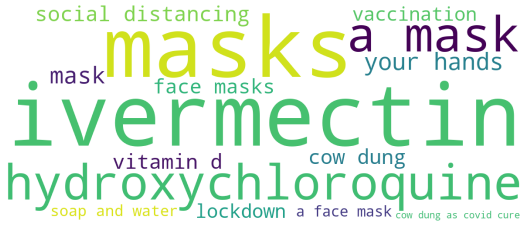
We present the top 15 frequently mentioned potential cure and prevention methods that people believe are effective within different time ranges in Table A5. Larger fonts indicate more frequent terms.

(A-3) What methods of cure and prevention do people think are effective?

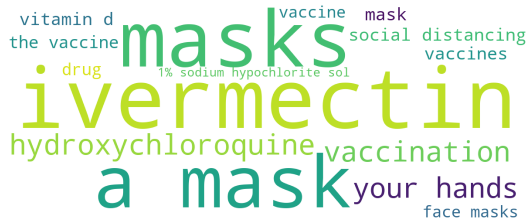
Before 2021/01/01



From 2021/02/15 to 2021/06/15 (First Wave in 2021)



From 2021/06/16 to 2021/10/15 (Second Wave in 2021)



From 2021/10/16 to 2022/04/01

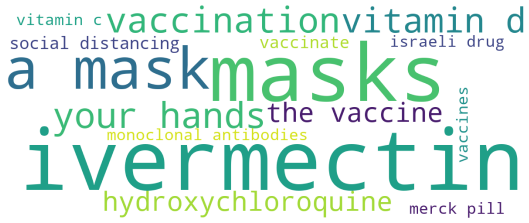


Table A5: Top 15 most frequent potential cure and prevention methods that people think are effective over different time ranges.