Event Detection from Social Media for Epidemic Prediction

Tanmay Parekh[†] **Anh Mac**[†] **Jiarui Yu**[†] **Yuxuan Dong**[†]
Syed Shahriar[†] **Bonnie Liu**[†] **Eric Yang**[†] **Kuan-Hao Huang**[§]
Wei Wang[†] **Nanyun Peng**[†] **Kai-Wei Chang**[†]
[†]Computer Science Department, University of California, Los Angeles
[§]Department of Computer Science, University of Illinois Urbana-Champaign
{tparekh, weiwang, violetpeng, kwchang}@cs.ucla.edu

Abstract

Social media is an easy-to-access platform providing timely updates about societal trends and events. Discussions regarding epidemic-related events such as infections, symptoms, and social interactions can be crucial for informing policymaking during epidemic outbreaks. In our work, we pioneer exploiting Event Detection (ED) for better preparedness and early warnings of any upcoming epidemic by developing a framework to extract and analyze epidemic-related events from social media posts. To this end, we curate an epidemic event ontology comprising seven disease-agnostic event types and construct a Twitter dataset SPEED with human-annotated events focused on the COVID-19 pandemic. Experimentation reveals how ED models trained on COVID-based SPEED can effectively detect epidemic events for three unseen epidemics of Monkeypox, Zika, and Dengue; while models trained on existing ED datasets fail miserably. Furthermore, we show that reporting sharp increases in the extracted events by our framework can provide warnings 4-9 weeks earlier than the WHO epidemic declaration for Monkeypox. This utility of our framework lays the foundations for better preparedness against emerging epidemics.

1 Introduction

Early warnings and effective control measures are among the most important tools for policymakers to be prepared against the threat of any epidemic (Collier et al., 2008). World Health Organization (WHO) reports suggest that 65% of the first reports about infectious diseases and outbreaks originate from informal sources and the internet (Heymann et al., 2001). Social media is an important information source here, as it is more timely than other alternatives like news and public health (Lamb et al., 2013), more publicly accessible than clinical notes (Lybarger et al., 2021), and possesses a huge volume of content.[1] This underscores the need for an automated system monitoring social media to provide early and effective epidemic prediction.

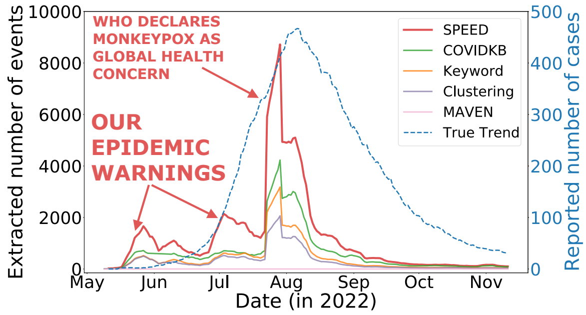

Figure 1: Number of reported Monkeypox cases and extracted events by our trained ED model from May 11 to Nov 11, 2022. Arrows indicate how our system could provide early epidemic warnings about 4-9 weeks before the WHO declared Monkeypox as a concern. We also show other baseline models for comparison.

To this end, we pioneer to leverage the task of Event Detection (ED) for epidemic prediction. ED involves identifying and categorizing significant events based on a pre-defined ontology (Sundheim, 1992; Doddington et al., 2004). Compared to existing epidemiological keyword and sentence-classification approaches (Lejeune et al., 2015; Lybarger et al., 2021), ED requires a deeper semantic understanding. This enhanced understanding aids in more effective disease-agnostic extraction of epidemic events from social media. By reporting sharp increases in extracted epidemic-related events, we can provide early epidemic warnings, as shown for Monkeypox in Figure 1 - highlighting the applicability of ED for epidemic prediction.

Existing ED datasets are unsuitable for establishing a framework to extract epidemic-related events from social media, as they focus on general-purpose events in news and wikipedia domains,

[1]A daily average of 20 million tweets were posted about COVID-19 from May 15 – May 31, 2020.

while other epidemiological works are disease-specific and too fine-grained (§ 6). Thus, we construct our own epidemic ED ontology and dataset for social media. Our created ontology comprises seven event types - *infect*, *spread*, *symptom*, *prevent*, *cure*, *control*, *death* - chosen based on their relevance for epidemics, frequency in social media, and their applicability to various diseases. We further validate our ontology through clinical sources and public health experts. For the dataset, we choose Twitter as the social media platform and focus on the COVID-19 pandemic. Using our curated ontology and expert annotation, we create our dataset **SPEED** (**S**ocial **P**latform based **E**pidemic **E**vent **D**etection) comprising 1,975 tweets and 2,217 event mentions. We complete our ED framework by training ED models (Du and Cardie, 2020; Hsu et al., 2022) on SPEED. Overall, SPEED provides disease-agnostic coverage of epidemic events for social media; thus, serving as a valuable dataset for epidemic prediction.

To validate the utility of our ED framework for disease-agnostic epidemic prediction, we perform two evaluations for three unseen diseases Monkeypox, Zika, and Dengue. First, we evaluate if our framework trained on our COVID-only SPEED dataset can detect epidemic events for the unseen diseases. Experiments reveal that our framework can successfully extract epidemic events, providing gains up to 29% F1 over the best few-shot model and 10% F1 gain over supervised models trained on limited target disease data.

Our second evaluation validates if aggregation of our extracted events can provide early epidemic warnings. Comparing our extracted events with the actual reported cases, we show that our framework can provide warnings up to 4-9 weeks earlier than the WHO declaration for the Monkeypox epidemic (Figure 1). Such early warnings aided with timely action can potentially lead to 2-4x reduction in the number of infections and deaths (Kamalrathne et al., 2023). These results underscore the strong utility of our dataset and framework for upcoming epidemic prediction and preparedness.

The contribution of this work is threefold, first, we pioneer to utilize Event Detection to develop an effective framework capable of extracting events from social media and providing early warnings for any unforeseeable epidemic. To support the proposed framework, our second contribution is the design of a disease-agnostic social-media tailored ontology and dataset SPEED. Our final

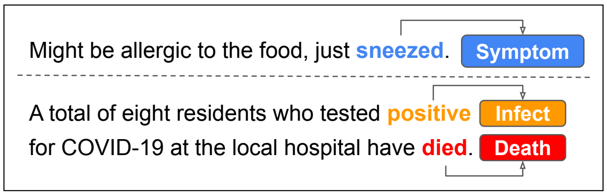

Figure 2: Illustration for the task of Event Detection. Event mentions: Event *symptom* and trigger *sneezed* (1st sentence), Event *infect* and trigger *positive* (2nd sentence), Event *death* and trigger *died* (2nd sentence).

contribution is extensive experiments to demonstrate the inadequacy of existing methods and the substantial improvements achieved by models trained on SPEED. This signifies the pivotal role of our dataset and framework in enhancing the efficacy of epidemic prediction. We release the data, code, and trained models at `https://github.com/PlusLabNLP/SPEED`.

2 From Event Detection to Epidemic Prediction

Given a social media post, Event Detection (ED) (Sundheim, 1992; Doddington et al., 2004) extracts and classifies significant events of interest. By designing disease-agnostic epidemic-based events, we aim to train ED models to extract epidemic events from social media posts for any possible disease. By detecting abnormal influx in the trends of extracted epidemic events from social media, we can thus provide early epidemic warnings for any possible disease, as we show for Monkeypox in Figure 1. Existing epidemiological approaches (Lejeune et al., 2015; Lybarger et al., 2021) are simple keyword or sentence classification-based and less accurate. Other works like COVIDKB (Zong et al., 2022) and ExcavatorCovid (Min et al., 2021a) are disease-specific and utilize events for building knowledge bases. To the best of our knowledge, we are the first ones to leverage event detection to extract epidemic events from social media and provide early warnings for any possible disease.

Formal Task Definition Following ACE 2005 guidelines (Doddington et al., 2004), we define an **event** to be something that happens or describes a change of state and is labeled by a specific **event type**. An **event mention** is the sentence wherein the event is described. Each event mention comprises an **event trigger**, which is the word/phrase that most distinctly highlights the occurrence of the event. **Event Detection** is technically defined as the task of identifying event triggers from sentences and classifying them into one of the pre-defined event types (defined by an **event ontology**). The subtask of identifying event triggers is called **Trigger Identification** and classification into event types is **Trigger Classification** (Ahn, 2006). Figure 2 shows examples for three event mentions for the events *symptom*, *infect*, and *death*.

3 Ontology Creation and Data Collection

We choose social media as our document source as it provides faster and more timely worldly information than news and public health (Lamb et al., 2013) and is more publicly accessible than clinical notes (Lybarger et al., 2021). Owing to its public access and huge content volume, we consider **Twitter**[2] as the social media platform and consider the recent **COVID-19 pandemic** as the primary disease.

Existing epidemiological ontologies are typically disease-specific, too fine-grained, or limited in coverage (§ 6 and Table 8). Similarly, standard ED datasets don't comprise epidemiological events and mostly focus on news or Wikipedia domains (§ 6). Due to these limitations, we create our own event ontology and dataset SPEED for detecting disease-agnostic epidemics from social media. Figure 3 provides a brief overview of our data creation process, with further details discussed below.

3.1 Ontology Creation

Taking inspiration from medical sources like BCEO (Collier et al., 2008), IDO (Babcock et al., 2021), and the ExcavatorCovid (Min et al., 2021b), we curate a wide range of epidemic-related event types. Next, we merge similar event types across these different ontologies (e.g. *Outbreak* event type). To create a disease-agnostic ontology, we filter out event types biased for specific diseases (e.g. *Mask Wearing* for COVID-19) and create disease-agnostic definitions using aid from public-health experts. Finally, we categorize these events into three abstractions: personal (individual-oriented events), social (large population events), and medical (medically focused events) types. We report our initial ontology comprising 18 event types in Table 9 and share additional specifications in § A.1.

Social Media Relevance To tailor our curated ontology for social media, we conduct a deeper analysis of the event types based on their frequency and specificity. Our goal is to filter and merge event

[2] https://www.twitter.com/

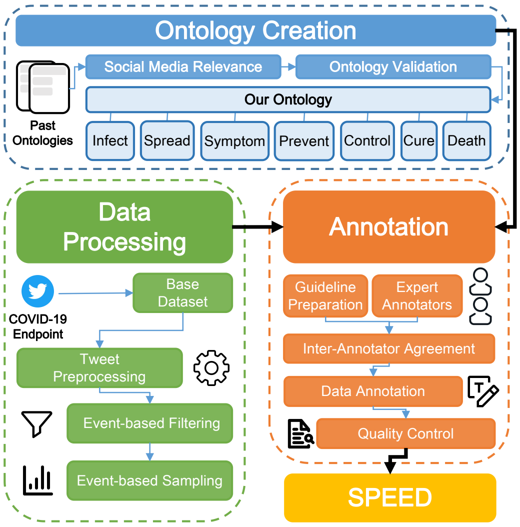

Figure 3: Overview of our dataset creation process with three major steps: Ontology Creation, Data Processing, and Data Annotation.

types that occur less frequently and less distintively in social media. To this end, using human expertise and external tools like Thesaurus,[3] we first associate each event type with specific keywords. Then we rank the event types based on the specificity and frequency of their keywords in social media posts. Based on this ranking, we merge and discard the lower ranked event types (e.g. *Respond* and *Prefigure*). Furthermore, we conduct human studies and merge event types to ensure better pairwise distinction (e.g. *Treatment* is merged with *Cure*). We provide the keywords associated with each event type and the example social media posts in Table 10 along with additional details in § A.2.

Ontology Validation and Coverage Elemental medical soundness is ensured for our ontology since it is derived from established epidemiological ontologies. To further certify this soundness, two public health experts (epidemiologists working in the Department of Public Health) validate the sufficiency and comprehensiveness of our ontology and event definitions. To verify if our ontology is characteristic of any disease, we assess our ontology coverage for four diverse diseases by estimating the percentage of event occurrence in disease-related tweets. Notably, we observe a coverage of 50% for COVID-19, 44% for Monkeypox, 70% for Dengue, and 73% for Zika (details in § A.4). Our coverage is better than the coverages of standard ED datasets ACE (Doddington et al., 2004) (27%) and ERE (Song et al., 2015) (43%), in turn, confirming the

[3] https://www.thesaurus.com/

Event Type	Event Definition	Example Event Mentions
Infect	The process of a disease/pathogen invading host(s)	1. Children can also **catch** COVID-19 ... 2. If you have antibodies, you **had** the virus. Period.
Spread	The process of a disease spreading or prevailing massively at a large scale	1. #COVID-19 CASES **RISE** TO 85,940 IN INDIA ... 2. ... the **prevalence** of asymptomatic COVID - 19 cases ...
Symptom	Individuals displaying physiological features indicating the abnormality of organisms	1. (user) (user) Still **coughing** two months after being infected by this stupid virus ... 2. If a person nearby is **sick**, the wind will scatter the virus ...
Prevent	Individuals trying to prevent the infection of a disease	1. ... wearing mask is the way to **prevent** COVID-19 2. ... an #antibody that has been succssful at **blocking** the virus
Control	Collective efforts trying to impede the spread of an epidemic	1. Social Distancing **reduces** the spread of covid ... 2. (user) COVID is still among us! Wearing masks **saves** lives!
Cure	Stopping infection and relieving individuals from infections/symptoms	1. ... **recovered** corona virus patients cant get it again 2. ... patients are **treated** separately at most places
Death	End of life of individuals due to an infectious disease	1. More than 80,000 Americans have **died** of COVID ... 2. The virus is going to get people **killed**. Stay home. Stay safe.

Table 1: Event ontology comprising seven event types promoting epidemic preparedness along with their definitions and two example event mentions. The trigger words are marked in **bold**.

robust disease coverage of our ontology.

Our final ontology comprises seven primary event types tailored for social media, disease-agnostic, and encompassing crucial aspects of an epidemic. We present our ontology in Table 1 along with event definitions and example event mentions.

3.2 Data Processing

To access a wide range of tweets related to COVID-19, we utilized the Twitter COVID-19 Endpoint released in April 2020. We used a randomized selection of **331 million tweets** between May 15 – May 31 2020, as our base dataset. For preprocessing tweets, we follow Pota et al. (2021): (1) we anonymize personal information like phone numbers, emails, and handles, (2) we normalize any retweets and URLs, (3) we remove emojis and split hashtags, (4) we filter out tweets only in English.

Event-based Filtering Most tweets in our base dataset expressed subjective sentiments, while only 3% comprised mentions aligned with our event ontology.[4] To reduce annotation costs, we further filter these tweets using a simple *sentence embedding* similarity technique. Specifically, each event type is linked to a seed repository of 5-10 diverse tweets. Example seed tweets per event are shown in Table 11 in § A.3. Query tweets are filtered based on their sentence-level similarity (Reimers and Gurevych, 2019) with this event-based seed repository.[5] This step filters about 95% tweets from our base dataset, leading to 20x reduction in the annotation cost.

Event-based Sampling Random sampling of tweets would yield an uneven and COVID-biased distribution of event types for our dataset. We instead perform a uniform sampling - wherein we over-sample tweets linked to less frequent types (e.g. *prevent*) and under-sample the more frequent ones (e.g. *death*). Such a uniform sampling has proven to ensure model robustness (Parekh et al., 2023) - as also validated by our experiments (§ B) - and in turn, would make SPEED generalizable to a wider range of diseases. In total, we sample 1,975 tweets which are utilized for ED annotation.

3.3 Data Annotation

For ED annotation, annotators are tasked with identifying whether a given tweet mentions any event outlined in our ontology. If an event is present, annotators are required to identify the specific event trigger. We design our annotation guidelines following the standard ACE dataset (Doddington et al., 2004) and amend them through several rounds of preliminary annotations to ensure annotator consistency. Additional details are provided in § C.

Annotator Details To ensure high annotation quality and consistency, we chose six experts instead of crowdsourced workers. These experts are computer science students studying NLP and well-versed for ED. They were further trained through multiple rounds of annotations and feedback.

Inter-annotator agreement (IAA) We used Fleiss' Kappa (Fleiss, 1971) for measuring IAA. We conduct two phases of IAA studies: (1) *Guideline*

[4] Based on keyword-based study conducted on 1,000 tweets
[5] We use a filtering threshold of 0.9.

Dataset	# Event Types	# Sent	# EM	Avg. EM per Event	Domain
ACE	33	18,927	5,055	153	News
ERE	38	17,108	7,284	192	News
M²E²	8	6,013	1,105	138	News
MLEE	29	286	6,575	227	Biomedical
FewEvent	100	12,573	12,573	126	General
MAVEN	168	49,873	118,732	**707**	Wikipedia
SPEED	7	1,975	2,217	**317**	Social Media

Table 2: Data Statistics for SPEED dataset and comparison with other standard ED datasets. # = "number of", Avg. = average, Sent = sentences, EM = event mentions.

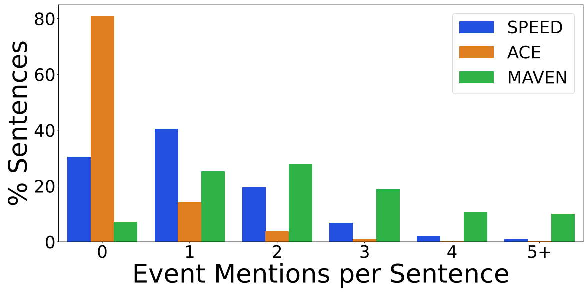

Figure 4: Distribution of number event mentions per sentence. Here % indicates percentage.

Dataset	#Event Mentions	# Unique Triggers	Avg. Triggers per Mention
ACE	5,055	1,229	0.24
MAVEN	118,732	7,074	0.06
SPEED	1,975	555	**0.25**

Table 3: Comparison of SPEED with ACE and MAVEN in terms of unique trigger words and average number of triggers per event mention. Avg = Average.

Improvement: Three annotators participated in three annotation rounds to improve the guidelines through collaborative discussion of disagreements. IAA score rose from 0.44 in the first round to 0.59 (70 samples) in the final round. (2) *Agreement Improvement:* All annotators participated in three rounds of annotations to boost consistency. IAA score improved from 0.56 in the first round to a strong 0.65 (50 samples) in the final round.

Quality Control We further ensure high annotation quality through: (1) *Multi-Annotation:* each tweet is annotated by two annotators, disagreements resolved by a third, and (2) *Flagging:* annotators flag ambiguous annotations, resolved by a third annotator via discussion. These, coupled with good IAA scores, ensure the high quality of our annotations.

3.4 Data Analysis

Our dataset SPEED comprises seven event types with 2,217 event mentions annotated over 1,975 tweets. We compare SPEED with other ED datasets like ACE (Doddington et al., 2004), ERE (Song et al., 2015), M²E² (Li et al., 2020), MLEE (Pyysalo et al., 2012), FewEvent (Deng et al., 2020), and MAVEN (Wang et al., 2020) in Table 2. We show how other datasets focus on the news, biomedical, general, and Wikipedia domains, while SPEED is the first-ever ED dataset for social media, specifically Twitter. Furthermore, none of the previous datasets comprise any of the epidemiological event types present in SPEED (§ D.1).

Comparable Datasize Since we only focus on 7 event types, SPEED has relatively lesser number of sentences and event mentions. However, SPEED has a high 316 average mentions per event type (column 5 in Table 2), more than most other standard datasets. We compare the distribution of event mentions per sentence with other ED datasets like ACE and MAVEN in Figure 4. We observe that the event density of our dataset is less than MAVEN but better than ACE. This shows that SPEED is fairly dense and reasonably sized ED dataset.

Diverse and Challenging We show the diversity of trigger words in SPEED and compare it with other datasets in Table 3. We note that SPEED has a strong average number of triggers per event mention. This demonstrates how SPEED is a diverse and challenging ED dataset.

4 Epidemic Prediction

For our ED framework, we utilize our curated dataset SPEED to train various ED models (§ 4.1). To validate the utility of models for the application of epidemic prediction, we perform evaluations using two tasks: (1) Epidemic event detection and (2) Early warning prediction. Epidemic event detection performs a formal ED evaluation of the models for detecting epidemic-based events. On the other hand, early warning prediction practically evaluates if the extracted events by the model can be aggregated to provide any early epidemic warnings.

Since SPEED focuses solely on COVID-19, we conduct these epidemic prediction evaluations for three unseen epidemics of *Monkeypox* (2022), *Zika* (2017), and *Dengue* (2018). These diseases are fairly distinct too, as Monkeypox causes rashes and rarely fatal, Zika causes birth defects, and Dengue causes high fever and can be fatal. For our evalua-

	Disease	# Sent	# EM
Train	COVID	1,601	1,746
Dev	COVID	374	471
Test	Monkeypox	286	398
	Zika + Dengue	300	274

Table 4: Statistics for data splits for epidemic event detection evaluation. # = "number of", Sent = sentences, EM = event mentions.

tions, we utilize and modify the raw Twitter dumps provided by Thakur (2022) for Monkeypox and Dias (2020) for Zika and Dengue.

4.1 Epidemic Event Detection

To validate if our SPEED-trained models can extract events for any epidemic, we perform traditional ED evaluation of these models for unseen diseases of Monkeypox, Zika, and Dengue. Following Ahn (2006), we report the F1-score for trigger identification (**Tri-I**) and classification (**Tri-C**).

Data Setup To train our ED models, we split the SPEED into 80-20 split for training and development sets. For testing, we sample tweets from the Twitter dumps of Monkeypox, Zika, and Dengue. Since the original data doesn't comprise event-based annotations, we utilize the same human experts who annotated SPEED to annotate the raw tweets for ED and create the evaluation dataset. We provide statistics for our data setup in Table 4. We release the training COVID data along with this evaluation data for Zika, Monkeypox, and Dengue for future benchmarking.

ED Models For training models using SPEED for our ED framework, we consider the following supervised models: (1) DyGIE++ (Wadden et al., 2019), (2) BERT-QA (Du and Cardie, 2020), (3) DEGREE (Hsu et al., 2022), (4) TagPrime (Hsu et al., 2023). We utilized the TextEE framework (Huang et al., 2023) to implement these models and provide more details in § E.

Baseline Models As baselines, we consider zero-shot ED models (**ZERO-SHOT**) that do not train on any supervised data and solely utilize the event definitions. We consider the following zero-shot models: (1) TE (Lyu et al., 2021), (2) WSD (Yao et al., 2021), (3) ETypeClus (Shen et al., 2021). Additional model implementation details is provided in § E. We also consider transferring from existing datasets (**TRANSFER FROM EXISTING DATASETS**) by training models on standard ED datasets like ACE (Doddington et al., 2004) and MAVEN (Wang et al., 2020) without fine-tuning on epidemic ED data.

Model	**Monkeypox Tri-I**	**Monkeypox Tri-C**	**Zika + Dengue Tri-I**	**Zika + Dengue Tri-C**
ZERO-SHOT				
TE	16.70	12.11	12.69	9.06
WSD	22.04	4.35	27.93	5.85
ETypeClus	18.31	6.78	13.99	5.33
TRANSFER FROM EXISTING DATASETS				
ACE - TagPrime	4.80	0	23.64	0
ACE - DEGREE	12.15	5.14	14.47	0
MAVEN - TagPrime	29.16	0	33.97	0
MAVEN - DEGREE	27.94	0	32.04	0
NO TRAINING				
Keyword	36.40	25.09	25.93	21.69
GPT-3.5	42.23	35.33	53.22	14.27
TRAINED ON TARGET EPIDEMIC				
BERT-QA	59.8	54.08	94.92	80.89
DEGREE	59.58	54.12	86.21	78.76
TagPrime	55.57	49.65	96.67	**84.43**
DyGIE++	55.83	50.31	73.24	65.65
TRAINED ON SPEED (OUR FRAMEWORK)				
BERT-QA	**67.38**	**64.17**	**96.77**	81.97
DEGREE	62.95	61.45	88.52	77.69
TagPrime	64.71	61.92	95.24	75.54
DyGIE++	62.76	59.82	91.8	80.34

Table 5: Evaluating ED models trained on SPEED for detecting events for new epidemics of Monkeypox, Zika, and Dengue in terms of F1 scores.

As stronger baselines, we also consider models utilizing epidemic ED data. Here, we consider models using few-shot target disease data without any model training (**NO TRAINING**) like: (1) Keyword (Lejeune et al., 2015), an epidemiological model utilizing curated event-specific keywords to detect events, and (2) GPT-3.5 (Brown et al., 2020), a large-language model (LLM) using GPT-3.5-turbo with seven target disease in-context ED examples. Finally, we consider super-strong baselines training ED models on limited target epidemic data (**TRAINED ON TARGET EPIDEMIC**). Specifically, we sample 300 event mentions for the target disease (exclusive from the test data) and use human experts to annotate these tweets (as done for COVID annotations) to create the target epidemic data. Note that these models are added for comparison, but they are practically infeasible for epidemic prediction, as it takes 3-6 weeks after the first infection to collect such target disease data.

Results We present our results in Table 5. Firstly, none of the existing data transfer, zero-shot, or no training-based models perform well for our task, mainly owing to the domain shift of social media and unseen epidemic events. Overall, ED models trained on SPEED perform the best, thus **demonstrating the capability of our ED framework to detect epidemic events for new diseases**. Compared with models trained on the target epidemic, SPEED-trained models provide a gain of 10 F1 points for Monkeypox and at par performance for Zika and Dengue. This outcome is particularly encouraging, as it **demonstrates the resilience of our framework, making it highly applicable during the early stages of an epidemic, when minimal to no epidemic-specific data is accessible**.

4.2 Early Warning Prediction

As a practical validation of the utility of our framework, we evaluate if SPEED-trained ED models are capable of providing early warnings for an unknown epidemic. More specifically, we aggregate the extracted event mentions by our framework over a time period and report any sharp increase in the rolling average of detected events as an epidemic warning. For evaluation, we compare it with the actual number of disease infections reported in the same period. Naturally, the earlier we provide an epidemic warning, the better the framework is deemed. For this evaluation, we choose Monkeypox as the unseen disease and its outbreak from May 11 to Nov 11, 2022, as the unknown period. We utilize the Twitter dump by Thakur (2022) as the datasource. We also include four other baselines for comparison: (1) MAVEN, an ED model trained on existing ED dataset MAVEN (Wang et al., 2020), (2) Clustering, a zero-shot classification baseline based on ETypeClus (Shen et al., 2021), (3) Keyword, a no-training baseline based on a previous epidemiological model (Lejeune et al., 2015), (4) COVIDKB, a BERT-based classification model trained on a previous COVIDKB (Zong et al., 2022) dataset.

Results We report the number of epidemic events extracted by the BERT-QA trained on SPEED along with the actual number of Monkeypox cases reported in the US[6] from May 11 to Nov 11, 2022, in Figure 1. As indicated by the arrows, our model could potentially provide two sets of early warnings around May 23 (9 weeks earlier, when first cases were detected) and June 29 (4 weeks earlier, when cases started rising) before the outbreak reached its peak around July 30. Comparatively, MAVEN-trained model fails completely, while the clustering, keyword, and COVIDKB models have super weak trends to provide any kind of warnings. In fact, all ED models trained on our SPEED data are capable of providing these early signals as shown in Figure 5 (further event-wise analysis in Appendix F). To further validate if the early warnings are raised owing to some actual Monkeypox epidemic discussion, we show some tweets classified as epidemic events by our system in Table 6. These tweets demonstrate that our system is picking the right signal to provide early epidemic warnings. Overall, this robust outcome underscores the **practical utility of our framework to provide early epidemic warnings and ensure better preparedness for any potential epidemic.**

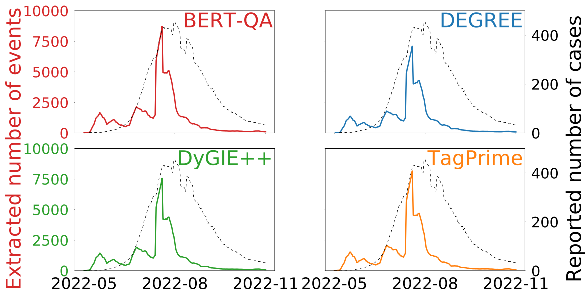

Figure 5: Number of reported Monkeypox cases and the number of extracted events from our four trained models from May 11 to Nov 11, 2022.

The first known **case** *[infect]* of human to dog #(monkey pox) **transmission** *[spread]* has been documented in medical journal, the Lancet. The dog **caught** *[infect]* it from its owners, a non-monogamous gay couple in Paris. The greyhound **developed** *[symptom]* an anal ulceration & mucocutaneous lesions. (url)
(user) Release Topxx for **treatment** *[cure]* of monkey pox people are suffering
(user) Wait what? I thought we were going to get Monkey Pox **vaccine** [prevent] in October.
Monkey Pox is not an STI. It is not an STI. #(monkey pox is not an sti) It can be **transferred** [spread] via sex the same way Covid can.
(user) Most of those people **have** [infect] Monkey Pox now.
(user) More **dead** [death] in Chicago over a weekend than the world wide health emergency monkey pox

Table 6: Sample illustrations of various tweets (anonymized) predicted as Monkeypox epidemic-related by our ED framework. The triggers are highlighted in **bold** with the respective event names in [].

[6]As reported by CDC at https://www.cdc.gov/poxvirus/mpox/response/2022/mpx-trends.html

5 Analysis and Discussion

In this section, we provide additional analyses to support the utility of our ED framework and also provide intuitions for why our framework works better than previous works.

5.1 Event-based Disease Profiling

Our ED framework offers the additional utility of generating event-based disease profiles using public sentiments. These disease profiles can be generated by plotting the percentage of mentions per event type extracted by our framework. Using 500k tweets, we depict the profiles for COVID, Monkeypox, and Zika+Dengue in Figure 6.

Distinctive profiles emerge for each disease; COVID majorly comprises *control*, Monkeypox exhibits a bias toward *infect* and *spread*, while Zika+Dengue emphasizes *control* and *death*. These trends align with the higher fatality rate of Zika and Dengue (Paixao et al., 2022), recent discoveries of transmission routes of Monkeypox (Kozlov et al., 2022), and the need for mass public control measures for the COVID pandemic (Güner et al., 2020). Relatively, Monkeypox also shows low mentions for *death*, *cure* - which aligns with low fatality and no available cure for Monkeypox (Kmiec and Kirchhoff, 2022). Overall, these profiles can provide policymakers with valuable insights about new unknown outbreaks to implement more informed and effective interventions.

5.2 Why does SPEED generalize?

We provide a qualitative analysis of why COVID-based SPEED helps detect epidemic events for other unseen diseases compared to previous epidemiological works (Collier et al., 2008; Lejeune et al., 2015) and attribute it to the difference in the task formulation and annotation schema. We demonstrate this difference (highlighted in **bold**) through illustrative examples for *Infect* and *Symptom* events in Table 7. As evident, keyword-based modeling requires annotating highly precise but disease-specific keywords like *COVID-19*, *fever*, etc. On the other hand, our ED annotation formulation emphasizes the annotation of disease-agnostic triggers like *infected*, *symptoms*, etc. This provides SPEED and our framework superior generalizability without new annotation to unseen diseases.

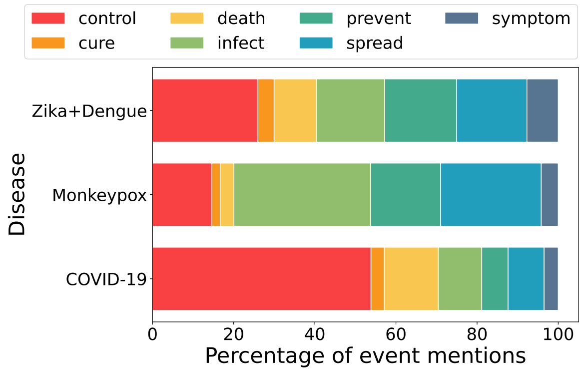

Figure 6: Disease profiles of public opinions generated by plotting the percentage of extracted event mentions for COVID-19, Monkeypox and Zika.

6 Related Work

Event Extraction Datasets Event Extraction (EE) is the task of detecting events (Event Detection) and extracting structured information about specific roles linked to the event (Event Argument Extraction) from natural text. Earliest works for this task can be dated back to MUC (Sundheim, 1992; Grishman and Sundheim, 1996) and the more standard ACE (Doddington et al., 2004). Over the years, ACE was extended to various datasets like ERE (Song et al., 2015) and TAC KBP (Ellis et al., 2015). Recent progress has been the creation of massive datasets and huge event ontologies with datasets like MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), DocEE (Tong et al., 2022), GENEVA (Parekh et al., 2023) and GLEN (Zhan et al., 2023). These ontologies and datasets cater to general-purpose events and do not comprise epidemiological event types.

Epidemiological Ontologies Earliest works (Lindberg et al., 1993; Rector et al., 1996) defined highly rich taxonomies for describing technical concepts used by biomedical experts. Further developments led to the creation of SNOMED CT (Stearns et al., 2001) and PHSkb (Doyle et al., 2005) that define a list of reportable events used for communication between public health experts. BioCaster (Collier et al., 2008) and PULS (Du et al., 2011) extended ontologies for the news domain. Recent works of NCBI (Dogan et al., 2014), IDO (Babcock et al., 2021) and DO (Schriml et al., 2022) focus on comprehensively organizing human diseases. In light of the recent COVID-19 pandemic, CIDO (He et al., 2020) define a technical taxonomy for coronavirus, while ExcavatorCovid

	Disease	Infect Event Example	Symptom Event Example
Keyword-based	COVID-19	Three students infected with **COVID-19**	COVID-19 symptoms include **fever**, **cough**, ...
Keyword-based	Monkeypox	How do you catch **Monkeypox**?	Monkeypox may cause **rashes** and **itching** ...
SPEED (Ours)	COVID-19	Three students **infected** with COVID-19	COVID-19 **symptoms** include fever, cough, ...
SPEED (Ours)	Monkeypox	How do you **catch** Monkeypox?	Monkeypox may **cause** rashes and itching ...

Table 7: Qualitative analysis for annotation difference between previous keyword-based epidemiological datasets (Collier et al., 2008; Lejeune et al., 2015) and SPEED's Event Detection based annotation schema. Our annotation schema is less disease-specific and thus, better generalizable to a wide range of diseases.

Dataset	Source	Sent-Level	Trig.	Social Eve.	Per. Eve.	SMG
SPEED (Ours)	Twitter	✓	✓	✓	✓	✓
COVIDKB	Twitter	✓	✗	✗	✓	✓
CACT	Clinical	✗	✓	✗	∼	✓
ExcavatorCovid	News	✗	✓	✓	✓	✗
BioCaster	News	✗	✗	✓	✓	✗
DANIEL	News	✗	∼	✗	∼	✓

Table 8: Objective comparison of various epidemiological datasets COVIDKB (Zong et al., 2022), CACT (Lybarger et al., 2021), ExcavatorCovid (Min et al., 2021a), BioCaster (Collier et al., 2008), and DANIEL (Lejeune et al., 2015) with our dataset SPEED. We objectify the source of data (Data Source), the level of annotation granularity (Sentence Level), the presence of trigger information (Trigger Present), the presence of social and personal events (Social Events and Personal Events), and the suitability of ontology for social media (SMG – Social Media Granular). ∼ indicates partial presence.

(Min et al., 2021a) automatically extract COVID-19 events and relations between them. Most of these ontologies are too fine-grained or limited to specific events, and can't be directly used for ED from social media, as also shown in Table 8.

Epidemiological Information Extraction Early works utilized search-engine queries and click-through rates for predicting influenza trends (Eysenbach, 2006; Ginsberg et al., 2009). Information extraction from Twitter has also been quite successful for predicting influenza trends (Signorini et al., 2011; Lamb et al., 2013; Paul et al., 2014). Over the years, various biomedical monitoring systems have been developed like BioCaster (Collier et al., 2008; Meng et al., 2022), HeathMap (Freifeld et al., 2008), DANIEL (Lejeune et al., 2015), EpiCore (Olsen, 2017). Extensions to support multilingual systems has also been explored (Lejeune et al., 2015; Mutuvi et al., 2020; Sahnoun and Lejeune, 2021). For the COVID-19 pandemic, several frameworks like CACT (Lybarger et al., 2021) and COVIDKB (Zong et al., 2022; Mendes et al., 2023) were developed for extracting symptoms, infection and treatment diagnosis. Most of these systems are disease-specific, focus on news and clinical domains, and use keyword/rule-based or simple BERT-based models, as shown in Table 8. In our work, we explore exploiting ED while focusing specifically on the social media domain.

7 Conclusion and Future Work

In this work, we develop an Event Detection (ED) framework to extract events from social media to provide early epidemic warnings. To facilitate this, we create our Twitter-based dataset SPEED comprising seven event types. Through experimentation, we show how existing models fail; while models trained on SPEED can effectively extract events and provide early warnings for unseen emerging epidemics. More broadly, our work demonstrates how event extraction can exploit social media to aid policy-making for better epidemic preparedness.

Acknowledgements

We thank I-Hung Hsu, Jiao Sun, Alex Spangher, Yu Zhou, Hritik Bansal, Yufei Tian, Chujie Zheng, and Rohan Wadhawan for their valuable insights, paper reviews, and constructive comments. We thank the anonymous reviewers and the area editors across multiple rounds of reviews for their feedback. This work was partially supported by NSF 2106859, 2200274, AFOSR MURI via Grant #FA9550- 22-1-0380, Defense Advanced Research Project Agency (DARPA) grant #HR00112290103/HR0011260656, and a Cisco Sponsored Research Award.

Limitations

Our work focuses majorly on a single source of social media - Twitter. We haven't explored other social media platforms and how ED would work on those platforms in our work. We leave that for future work, but are optimistic that our models should be able to generalize across platforms. Secondly, our work mainly only focuses on ED as the primary task, while its sister task Event Argument Extraction (EAE) is not explored. We hope to extend our work for EAE as part of our future work. Finally, we would like to show the generalization of our models on a vast range of diseases. However owing to budget constraints and the lack of publically available Twitter data for other diseases, we couldn't perform such a study. However, we believe showing results on three diseases lays the foundation for generalizability of our model.

Ethical Considerations

One strong assumption in our work is the availability of internet and social media for discussions about epidemics. Since not everyone has equal access to these platforms, our dataset, models, and results do not represent the whole world uniformly. Thus, our work can be biased and should be considered with other sources for better representation.

Our dataset SPEED is based on actual tweets posted by people all over the world. We attempted our best to anonymize any kind of private information in the tweets, but we can never be completely thorough, and there might be some private information embedded still in our dataset. Furthermore, these tweets were sentimental and may possess stark emotional, racial, and political viewpoints and biases. We do not attempt to clean any of such extreme data in our work (as our focus was on ED only) and these biases should be considered if being used for other applications.

Since our ED models are trained on SPEED, they may possess some of the dataset-based social biases. Since we don't focus on bias mitigation, these models should be used with due consideration.

Lastly, we do not claim that our models can be used off-the-shelf for epidemic prediction as it hasn't been thoroughly tested and can have false positives and negatives too. Furthermore, our results are shown on academic datasets and do not utilize all possible Twitter data. We majorly throw light to show these model capabilities and motivate future work in this direction. The usage of these systems for practical purposes should be appropriately considered.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Shane Babcock, John Beverley, Lindsay G. Cowell, and Barry Smith. 2021. The infectious disease ontology in the age of COVID-19. *J. Biomed. Semant.*, 12(1):13.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

YunSeok Choi, Hyojun Kim, and Jee-Hyong Lee. 2022. TABS: Efficient textual adversarial attack for pre-trained NL code model using semantic beam search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5490–5498, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Hung Quoc Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinform.*, 24(24):2940–2941.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 151–159. ACM.

Guilherme Dias. 2020. Tweets dataset on Zika virus.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10.

Timothy J. Doyle, Haobo Ma, Samuel L. Groseclose, and Richard S. Hopkins. 2005. Phskb: A knowledge-base to support notifiable disease surveillance. *BMC Medical Informatics Decis. Mak.*, 5:27.

Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, and Roman Yangarber. 2011. Building support tools for russian-language information extraction. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 380–387. Springer.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST.

Gunther Eysenbach. 2006. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *AMIA 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006*. AMIA.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. 2008. Model formulation: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Medical Informatics Assoc.*, 15(2):150–157.

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.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Hatice Rahmet Güner, İmran Hasanoğlu, and Firdevs Aktaş. 2020. Covid-19: Prevention and control measures in community. *Turkish Journal of medical sciences*, 50(9):571–577.

Yongqun He, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, Anthony Huffman, Hsin-Hui Huang, John Beverley, Asiyah Yu Lin, William D. Duncan, Sivaram Arabandi, Jiangan Xie, Junguk Hur, Xiaolin Yang, Luonan Chen, Gilbert S. Omenn, Brian D. Athey, and Barry Smith. 2020. CIDO: the community-based coronavirus infectious disease ontology. In *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO) joint with the 10th Workshop on Ontologies and Data in Life Sciences (ODLS) and part of the Bolzano Summer of Knowledge (BoSK 2020), Virtual conference hosted in Bolzano, Italy, September 17, 2020*, volume 2807 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org.

David L Heymann, Guénaël R Rodier, et al. 2001. Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. *The Lancet infectious diseases*, 1(5):345–353.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. 2023. A simple and unified tagging model with priming for relational structure predictions. In *Proceedings of the 61st Conference of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2023. A reevaluation of event extraction: Past, present, and future challenges. *CoRR*, abs/2311.09562.

Thushara Kamalrathne, Dilanthi Amaratunga, Richard Haigh, and Lahiru Kodituwakku. 2023. Need for effective detection and early warnings for epidemic and pandemic preparedness planning in the context of multi-hazards: Lessons from the covid-19 pandemic. *International Journal of Disaster Risk Reduction*, 92:103724.

Dorota Kmiec and Frank Kirchhoff. 2022. Monkeypox: a new threat? *International journal of molecular sciences*, 23(14):7866.

Max Kozlov et al. 2022. How deadly is monkeypox? what scientists know. *Nature*, 609(7928):663.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia. Association for Computational Linguistics.

Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artif. Intell. Medicine*, 65(2):131–143.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2022. Open relation and event type discovery with type abstraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6864–6877, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

D. A. Lindberg, B. L. Humphreys, and A. T. McCray. 1993. The unified medical language system. *Methods of information in medicine*, 32(4):281—-291.

Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. 2021. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *J. Biomed. Informatics*, 117:103761.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.

Zaiqiao Meng, Anya Okhmatovskaia, Maxime Polleri, Yannan Shen, Guido Powell, Zihao Fu, Iris Ganser, Meiru Zhang, Nicholas B. King, David L. Buckeridge, and Nigel Collier. 2022. Biocaster in 2021: automatic disease outbreaks detection from global news media. *Bioinform.*, 38(18):4446–4448.

Bonan Min, Benjamin Rozonoyer, Haoling Qiu, Alexander Zamanian, Nianwen Xue, and Jessica MacBride. 2021a. ExcavatorCovid: Extracting events and relations from text corpora for temporal and causal analysis for COVID-19. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 63–71, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bonan Min, Benjamin Rozonoyer, Haoling Qiu, Alexander Zamanian, Nianwen Xue, and Jessica MacBride. 2021b. Excavatorcovid: Extracting events and relations from text corpora for temporal and causal analysis for COVID-19. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 63–71. Association for Computational Linguistics.

Stephen Mutuvi, Antoine Doucet, Gaël Lejeune, and Moses Odeo. 2020. A dataset for multi-lingual epidemiological event extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France. European Language Resources Association.

Jennifer M Olsen. 2017. Epicore: crowdsourcing health professionals to verify disease outbreaks. *Online Journal of Public Health Informatics*, 9(1).

OpenAI. 2021. ChatGPT: Large-scale language model. Accessed: June 17, 2023.

Amir P B Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Enny S Paixao, Luciana L Cardim, Maria CN Costa, Elizabeth B Brickley, Rita CO de Carvalho-Sauer, Eduardo H Carmo, Roberto FS Andrade, Moreno S Rodrigues, Rafael V Veiga, Larissa C Costa, et al. 2022. Mortality from congenital zika syndrome—nationwide cohort study in brazil. *New England Journal of Medicine*, 386(8):757–767.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Conference of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLoS currents*, 6.

Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. 2021. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Syst. Appl.*, 181:115119.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Junichi Tsujii, and Sophia Ananiadou.

2012. Event extraction across multiple levels of biological organization. *Bioinform.*, 28(18):575–581.

Alan L Rector, Jeremy E Rogers, and Pam Pole. 1996. The galen high level ontology. In *Medical Informatics Europe'96*, pages 174–178. IOS Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sihem Sahnoun and Gaël Lejeune. 2021. Multilingual epidemic event extraction : From simple classification methods to open information extraction (OIE) and ontology. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1227–1233, Held Online. INCOMA Ltd.

Lynn M. Schriml, James B. Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J. Allen Baron, Rebecca C. Jackson, Susan M. Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion, Michelle G. Giglio, and Carol Greene. 2022. The human disease ontology 2022 update. *Nucleic Acids Res.*, 50(D1):1255–1261.

Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. 2021. Corpus-based open-domain event type induction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5427–5440, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. SNOMED clinical terms: overview of the development process and project status. In *AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*. AMIA.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Nirmalya Thakur. 2022. Monkeypox2022tweets: A large-scale twitter dataset on the 2022 monkeypox outbreak, findings from analysis of tweets, and open research questions. *Infectious Disease Reports*, 14(6):855–883.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. 2021. Connect-the-dots: Bridging semantics between words and definitions via aligning word sense inventories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7741–7751, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qiusi Zhan, Sha Li, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. 2023. GLEN: general-purpose event detection for thousands of types. *CoRR*, abs/2303.09093.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2022. Extracting a knowledge base of COVID-19 events from social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3810–3823, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Ontology Creation - Additional Details

A.1 Complete ontology

Here, we first describe the selection steps for event types for our ontology as follows:

1. *Curation of event types*: We scan through existing medical ontologies like BCEO (Collier et al., 2008), IDO (Babcock et al., 2021), and the ExcavatorCovid (Min et al., 2021b) and curate a large list of event types for infectious and epidemic-related diseases.

2. *Merge event types across ontologies*: Since these existing ontologies may have repetitive event types, we perform a merging step. Specifically, two human experts manually examine and merge event types that are exactly similar in our curated list of event types.

3. *Filter out disease-specific event types*: Some event types in our curated list are specific to certain diseases. We identify and filter out such event types (e.g. Mask Wearing for COVID-19 which may not be observed for other diseases). We utilize opinions from public health experts to aid this step ensuring our event types are disease-agnostic.

4. *Definition Correction*: Utilizing aid from public health experts, we add and refine definitions for the curated set of event types and ensure they are disease-agnostic.

5. Organization - Following ExcavatorCovid (Min et al., 2021b), we organize our curated list of event types into three larger categories: social (events involving larger populations), personal (individual-oriented events), and medical (medically focused events) types.

Our complete initial event ontology comprises 18 event types along with their event definitions organized into three abstract categories as shown in Table 9.

A.2 Initial analysis of events

Our initial ontology (§ A.1) was constructed using previous ontologies and human knowledge. But the relevance of each event type for social media (specifically Twitter) remains unknown. To evaluate this relevance, we first associate each event type with event-specific keywords. Then we utilize frequency and specificity as two guiding heuristics for further filtering/merging of event types in our curated ontology. We utilize the base Twitter dataset for SPEED for conducting this analysis. We describe each of these steps in more detail here:

Keyword Association In order to objectively conduct this analysis, we associate each event type with a set of keywords.[7] This association involves two simple steps:

1. *Human expert curation*: For each event type, a human expert curates 2-3 simple yet specific keywords for each event based on commonsense knowledge. For example, for the *Cure* event, the set of curated keywords were [cure, recovery].

2. *Thesaurus-based expansion*: For each human-expert curated list, we utilize an external resource - Thesaurus[8] to further find event-relevant keywords. Human experts manually curate keywords from this thesaurus list such that the curated keyword is not generic (e.g. *display* is filtered out for event *Symptom* since it has other meanings as well).

We provide some example keywords for each event type along with some social media posts in which they appear in Table 10. We can also see how some example social media posts with the keywords can have different meanings as well. But we ignore them as this is a preliminary analysis.

Frequency-based filtering Using frequency, we aim to filter out event types that are less mentioned in social media. To approximately estimate the frequency of each event type in social media, we count the number of social media posts containing any of the curated keywords for each event type. We show the keyword-count based frequency for each event type in Figure 7. We observe that most events under the medical abstraction occur much lesser than others. Furthermore, the variance in frequency is large as the most frequent event type *control* is 180 times more likely to occur than the least frequent event type *variant*. Since such low-frequency events (e.g. *Variant*, *Cause*, *Prefigure*, etc.) are less likely to be mentioned in a smaller sample of data, we discard or merge such events for our final ontology.

Specificity-based filtering Specificity ensures that each event type is uniquely identifiable with a

[7] We release these keywords as part of our final code.
[8] https://www.thesaurus.com/

Event name	Event Definition	Action for Final Ontology
	SOCIAL SCALE EVENTS	
Prefigure	The signal that precedes the occurrence of a potential epidemic.	Discarded
Outbreak	The process of disease spreading among a certain amount of the population at a massive scale.	Merged into *Spread*
Spread	The process of disease spreading among a certain amount of the population but at a local scale.	Final Event
Control	Collective efforts trying to impede the spread of a epidemic.	Final Event
Promote	The relationship of a disease driver leading to the breakout of a disease.	Discarded
	PERSONAL SCALE EVENTS	
Prevent	Individuals trying to prevent the infection of disease.	Final Event
Infect	The process of a disease/pathogen invading host(s).	Final Event
Symptom	Individuals displaying physiological features indicating the abnormality of organisms.	Final Event
Treatment	The process that a patient is going through with the aim of recovering from symptoms.	Merged into *Cure*
Cure	Stopping infection and relieving individuals from infections/symptoms.	Final Event
Immunize	The process by which an organism gains immunization against an infectious agent.	Merged into *Prevent*
Death	End of life of individuals due to infectious disease.	Final Event
	MEDICAL SCALE EVENTS	
Cause	The causal relationship of a pathogen and a disease.	Discarded
Variant	An alternation of a disease with genetic code-carrying mutations.	Discarded
Intrude	The process of an infectious agent intruding on its host.	Merged into *Infect*
Respond	The process of a host responding to an infection.	Discarded
Regulate	The process of suppressing and slowing down the infection of a virus.	Merged into *Cure*
Transmission route	The process of a pathogen entering another host from a source.	Discarded

Table 9: Complete initial epidemic event ontology comprising 18 event types along with their event definitions organized into 3 higher-level abstract categories. We also present details about the action taken for each event type in the final ontology.

good confidence and mainly aims to reduce ambiguity and make the event types more distinct. To estimate specificity, for each curated keyword of an event type, we randomly sample a small number of non-duplicate social media posts. Human experts then manually evaluate the keyword specificity based on the percentage of posts wherein the semantic meaning of the keyword matches the definition of its event and is specific only to this event type. This specificity and distinctivity classifies keywords as high, medium, or low.

For example, the *Control* event comprises high specificity keywords such as *quarantine*, *protocol*, *guidelines*; medium specificity keywords such as *restrict*, *postpone*, *investigate*; and low specificity keywords such as *battle*, *separation*, *limitation*. On the other hand, the event *Prefigure* doesn't have any high specificity keywords, but only medium specificity keywords such as *foreshadow* and low specificity keywords such as *foretell*.

Our analysis suggests that medium and low specificity keywords are more likely to give false positives relative to high specificity ones. Thus, we filter/merge event types that have a high number of low-confidence keywords (e.g. *Intrude*, *Promote*).

Final Ontology Thus, with the above filtering and merging, we shrink our ontology from 18 event types to seven event types that are distinguishable, frequent, and have a low false-positive rate. We provide details about the action taken for each event type with respect to the final ontology in Table 9.

A.3 Association of Seed Tweets with Events

For event-based filtering, we associate each event type with a set of 5-10 seed tweets. In Table 11, we provide a couple of the seed tweets for each event type for reference. We also release these seed tweets as part of the main code.

Event name	Associated Keywords	Example Social Media Post
	Social Scale Events	
Prefigure	alert, foreshadow, warn, indicate, ...	(user) Alabama hospitals with no ICU beds **foreshadow** rural coronavirus crisis - Business Insider (url)
Outbreak	epidemic, pandemic, outbreak, ...	(user) If Boris goes, who is Nicola going to follow in the coronavirus **epidemic**?
Spread	outbreak, spread, contagious, ...	COVID-19 **outbreak** detected in immigration detention centres ...
Control	quarantine, isolate, restriction, protocol, ...	The doctors in my family are losing their minds seeing these videos of people flagrantly violating social distancing **protocols**.
Promote	boost, lead, accelerate	(user) Covid-19 Will **Accelerate** the AI Health Care Revolution (url)
	Personal Scale Events	
Prevent	prevent, protect, avoid, vaccine, ...	(user) why can't they wear masks and god will cover them anyways then you have double the **protection**
Infect	infect, spread, outbreak, infection, ...	(user) warns of COVID-19 **spread** in group gatherings after 2 people infect dozens at church. (url)
Symptom	symptom, fever, ill, sick, ache, ...	... average person dying from coronavirus has serious **illness** and is about to die anyway.
Treatment	doctor, drug, hospitalize, nurse, ...	Hyperbaric oxygen **therapy** can help fight one of the major issues(lack of oxygen) caused by covid/corona virus?
Cure	recover, cure, heal, ...	Patients who've **recovered** from Covid-19 and test positive again aren't contagious, a study says (url)
Immunize	immunize, antibody, vaccine, ...	We need to hope the number of cases go up in order to Herd **immunize**, and continually watch the # of hospitalizations.
Death	death, decease, kill, ...	Highest **death** toll in Europe. Economy on the floor. Quarantine for international travellers months after the horse bolted.
	Medical Scale Events	
Cause	cause, lead	Once again, 1-2 infections **lead** to a large number of new infections.
Variant	lineage, variant, mutate	#DeltaVariant surging in U.S. New data show Delta much more contagious than previous **versions** of #COVID19.
Intrude	attack, intrude, harm	(user) Peru found 100% cure zero **harm** ivermectin. Pinned tweet medical paper. Plagues over. Retweet.
Respond	respond, react	Aid agencies face huge challenges **responding** to #COVID19 in countries affected by conflict ...
Regulate	regulate, limit, decelerate, slow, ...	How can you **regulate** a child in school after lockdown?
Transmission route	airborne, contact, air, water, ...	(user) Very unfortunate. You can't workout with a mask on, is Corona **airborne**? Just asking..

Table 10: All the events in our ontology along with their associated keywords and example social media posts.

A.4 Coverage analysis of ontology

To quantitatively verify the coverage of our ontology, we conduct an analysis on four diseases with very different characteristics - COVID-19, Monkeypox, Dengue, and Zika. For each disease, we randomly sample 300 tweets and then filter them if they are related to the disease or not. Next, we annotate the filtered disease-related tweets based on our ontology and evaluate the proportion of event occurrences relative to the number of disease-related tweets. We find that our ontology has high coverage of 50% for COVID-19, 44% for Monkeypox, 70% for Dengue, and 73% for Zika. This in turn assures that our ontology can be used to detect epidemic events for various different kinds of diseases.

Event Type Distribution As part of our analysis, we also study our ontology's event type distribution for each disease and its correlation with the disease properties and outbreak stage. We show this event distribution in Figure 8 for each of the diseases. We note that distributions for Dengue and Monkeypox exhibit a strong focus on *spread* and *infect* events. This makes sense as the data for these diseases was collected at earlier stages of the outbreak when mitigation measures were not being discussed yet. On the other hand, for COVID-19, the distribution is vastly dominated by *control* and *death* events. Our

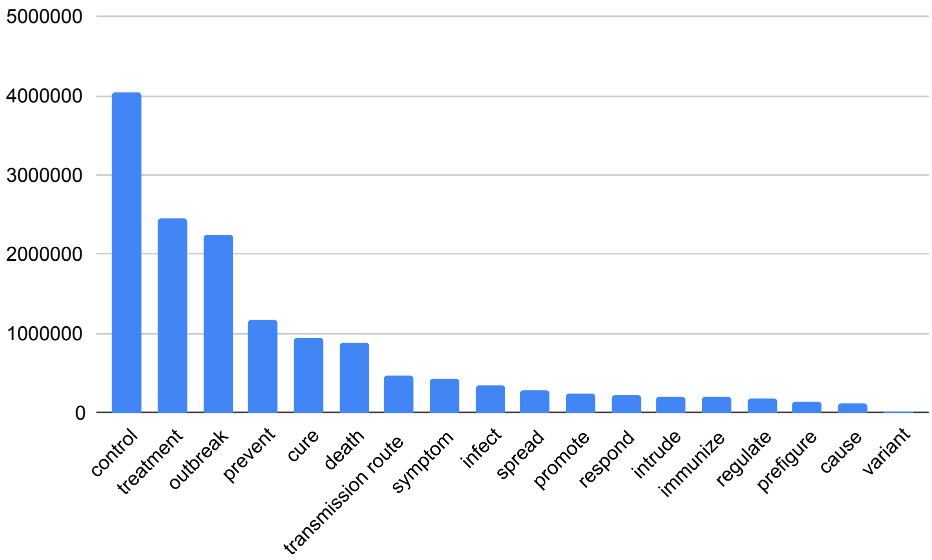

Figure 7: Frequency of occurrence based on keyword search for all event types in the initial complete ontology.

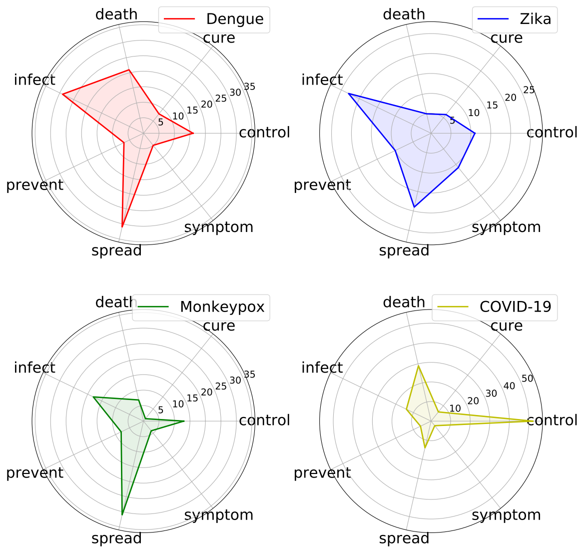

Figure 8: Event type distribution of the disease-related tweets for each disease. Numbers on the axis represent count of mentions for a given event type.

COVID-19 data was collected in May 2020 when the outbreak had vastly spread in America. Thus our distribution reflects more notions of lockdowns and control measures as well reflects the deadly nature of the disease.

B Uniform Sampling v/s Random Sampling for Data Selection

Previously Parekh et al. (2023) had shown how uniform sampling of data for events can yield more robust model performance. To validate the same for our ontology and data, we conduct additional experiments comparing uniform sampling with random sampling. More specifically, we annotate 200 tweets that conform to a 'real distribution'[9] based on random sampling and compare the trained models on this data with models trained on 200 tweets of uniform-sampling data. We further annotated 300 tweets based on the 'real-distribution' which was used for the evaluation of these two sampling techniques.

We present our results in Table 12 averaged over three model runs. We show that in terms of best model performance, uniform sampling is better by 5.5 F1 points compared to random sampling. On average, uniform-sampling trained models outperform the random-sampling trained models by up to 11 points. Both these results prove how despite train-test distribution differences, uniform sampling leads to better training of downstream models.

Generalizability to Other Diseases We also evaluate the models trained on the uniform and random-sampled data for generalizability to other diseases of Monkeypox, Zika, and Dengue. We show the results in Table 13. Clearly, we can see superior generalizability of uniform-sampling trained models as they outperform random-sampling trained models by 37 F1 points for Monkeypox and 28 F1 points for Zika + Dengue. Overall, this result strongly highlights the impact of uniform sampling

[9]Event-based filtering was still applied before sampling.

Event Name	Sample Seed Tweets
Infect	The pandemic will infect many older people My brother tested positive for COVID-19
Spread	The COVID-19 has spread throughout Europe Number of positive cases is rising
Symptom	Fever is a key symptom of COVID-19 I became incredibly sick after catching the COVID
Prevent	Officials hope the vaccine will prevent the spread of COVID-19 in high-risk populations Medical experts encourage young kids to wash their hands
Control	Many countries are requiring mask wearing to reduce the spread of the pandemic Government officials have imposed a lockdown on certain districts
Cure	There is no effective treatment for COVID so far Even patients already recovered from COVID remain coughing for a while
Death	My friend has passed away because of the pandemic Millions of people are dead because of COVID

Table 11: Sample seed tweets for the different event types in our ontology.

for robust and generalizable model training.

C Annotation Guidelines and Interface

C.1 Annotation Guidelines

Inspired by Doddington et al. (2004), we develop an extensive set of instructions with tricky cases and examples that have been developed through multiple rounds of expert annotation studies. We present the task summary with the major instructions in Figure 14. To reduce ambiguity in trigger selection, we present extensive examples and tricky cases with priority orders as shown in Figure 15. Finally, we also provide a wide range of annotated positive and negative examples as part of the guidelines and show those in Figure 16.

C.2 Annotation Interface

We utilize Amazon Mechanical Turk[10] as the interface for quick annotation. To annotate, annotators can select any word and label it into one of the seven pre-defined event types. Event definitions and examples are provided alongside for reference. Each batch (also known as HIT) comprises five tweets for flexibility in annotations. We show the interface and various utilities in Figure 17, 18, and 19 respectively.

[10] https://www.mturk.com/

Model	Tri-I	Tri-C
TRAINED ON UNIFORM DISTRIBUTION		
BERT-QA	**58.19**	52.30
DEGREE	55.83	**52.88**
TagPrime	55.48	50.51
DyGIE++	53.22	47.64
Average	55.68	50.83
TRAINED ON RANDOM DISTRIBUTION		
BERT-QA	46.11	43.76
DEGREE	46.11	45.23
TagPrime	25.03	24.15
DyGIE++	**51.10**	**47.35**
Average	42.09	40.12

Table 12: Benchmarking ED models trained on uniformly-sampled and randomly-sampled SPEED data on real-distribution based test data of 300 samples.

Model	Monkeypox		Zika + Dengue	
	Tri-I	Tri-C	Tri-I	Tri-C
TRAINED ON UNIFORM SAMPLED DATA				
BERT-QA	56.56	49.30	56.35	46.19
DEGREE	58.35	53.39	**58.37**	**51.27**
TagPrime	**58.36**	**53.56**	57.05	48.53
DyGIE++	55.73	48.30	56.90	47.10
TRAINED ON REAL SAMPLED DATA				
BERT-QA	9.48	7.97	21.68	20.43
DEGREE	10.76	10.53	19.33	19.00
TagPrime	10.37	8.57	12.78	12.28
DyGIE++	19.59	16.62	26.43	23.40

Table 13: Generalizability benchmarking of ED models trained on 200 samples of uniformly-sampled and randomly-sampled COVID data on other diseases of Monkeypox, Zika, and Dengue.

D Data Analysis for SPEED

D.1 Event Coverage for previous datasets

To show the distinction of the event types covered in SPEED compared to other previous datasets, we calculate the percentage event types from SPEED present in various diverse previous dataset ontologies. We show the results of this analysis in terms of partial coverage (similar events present) and exact coverage (exact event present) in Table 14.

Overall, from the table, we can note that there is no dataset with exact matches with our ontology. This proves the distinctive coverage of our dataset.

Dataset	Partial Match	Exact Match
ACE (Doddington et al., 2004)	14%	0%
ERE (Song et al., 2015)	14%	0%
MAVEN (Wang et al., 2020)	42%	0%
MEE (P B Veyseh et al., 2022)	14%	0%
M^2E^2 (Li et al., 2020)	14%	0%
MLEE (Pyysalo et al., 2012)	0%	0%
FewEvent (Deng et al., 2020)	28%	0%

Table 14: Comparison of SPEED with ACE and MAVEN in terms of unique trigger words and average number of triggers per event mention. Avg = Average.

D.2 Event Distribution Analysis

As part of data processing, we attempt to sample tweets in a more uniform distribution between the event types (§ 3.2). In Figure 9, we show the distribution of our dataset in terms of event types. In contrast to tail-ending distributions of other standard datasets like ACE (Doddington et al., 2004) and MAVEN (Wang et al., 2020) as shown in Figures 10 and 11 respectively, our distribution of event mentions is more uniform.

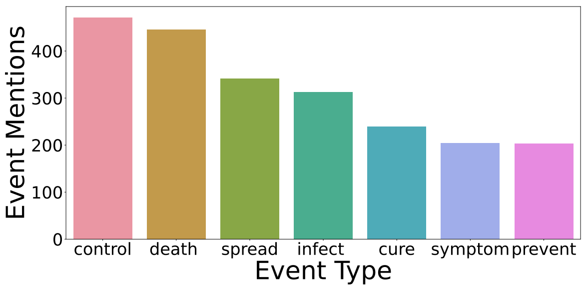

Figure 9: Distribution of event mentions per event type for our dataset SPEED.

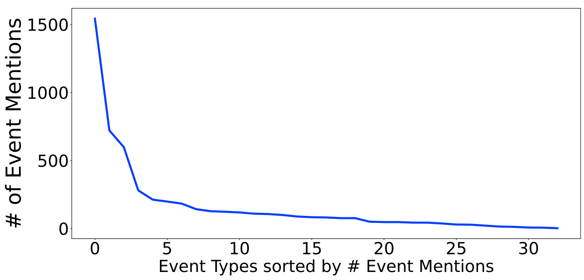

Figure 10: Distribution of event mentions for the event types in the ACE dataset.

D.3 Monkeypox Test Data Statistics

We share the data statistics of the evaluation dataset used for Monkeypox in Table 15 split according to each event type. We observe that there is a disparity

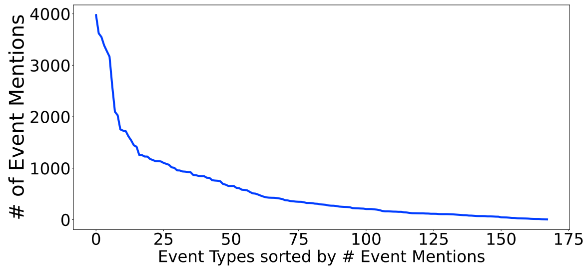

Figure 11: Distribution of event mentions for the event types in the MAVEN dataset.

in distribution across different event types, with *spread* mostly discussed and *cure* and *death* are least discussed.

Event Type	# Event Mentions
infect	78
spread	119
symptom	43
prevent	70
control	62
cure	13
death	13
Total	**389**

Table 15: Data Statistics for the evaluation dataset for Monkeypox Event Detection categorized by event types.

D.4 Zika + Dengue Test Data Statistics

We share the data statistics of the evaluation dataset used for Zika + Dengue in Table 16 split according to each event type. We observe a more even distribution of event types with more focus on *infect*, *spread*, and *death* well-discussed.

Event Type	# Event Mentions
infect	57
spread	53
symptom	34
prevent	22
control	28
cure	20
death	60
Total	**274**

Table 16: Data Statistics for the evaluation dataset for Zika+Dengue Event Detection categorized by event types.

E ED models and Implementation Details

We present details about each ED model that we benchmark along with the extensive set of hyperparameters and other implementation details.

E.1 TE

TE (Lyu et al., 2021) is a pre-trained model that formulates ED as a textual entailment and question-answering task. We run our experiments for TE on an NVIDIA 1080Ti machine with support for 8 GPUs. Our hyperparameters are as listed in the original paper.

E.2 WSD

WSD (Yao et al., 2021) is a classification model using on the joint encoding of the contextualized trigger and event definitions. We run our experiments for WSD on an NVIDIA A100 machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 17.

Pre-trained LM	RoBERTa-Large
Training Batch Size	64
Eval Batch Size	8
Learning Rate	0.00001
Weight Decay	0.01
# Training Epochs	7
Max Sentence Length	128
Max gradient norm	1

Table 17: Hyperparameter details for WSD model.

E.3 TABS

TABS (Li et al., 2022) is an event type induction model, wherein the goal is to discover new event types without a pre-defined event ontology. It utilizes two complementary trigger embedding spaces (mask view and token view) for classification. To adapt this for ED, we follow the end-to-end event discovery setting in (Choi et al., 2022) while making the following modifications: (1) **Dataset Composition**: We utilize ACE (Doddington et al., 2004) dataset for training and development and our SPEED dataset for testing. Our training data comprises 26 known event types from ACE, the validation set comprises 7 ACE event types, while our test set comprises 7 event types from SPEED. (2) **Candidate Trigger Extraction**: To improve trigger coverage, we extract all nouns and non-auxiliary verbs as candidate trigger mentions. (3) **Evaluation Setup**: Trigger identification (**Tri-I**) F1 score is evaluated using the extracted candidate triggers. For trigger classification (**Tri-C**), we first find the best cluster assignment of the predicted event clusters to the gold event types and then evaluate the F1 score.

We run our experiments for TABS on an NVIDIA RTX 2080 Ti machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 18.

Pre-trained LM	BERT-Base
Training Batch Size	8
Eval Batch Size	8
Gradient Accumulation Steps	2
Learning Rate	0.00005
Gradient Clipping	1
# Pretrain Epochs	10
# Training Epochs	30
Consistency Loss Weight	0.2
# Target Unknown Event Types	30

Table 18: Hyperparameter details for TABS model.

E.4 ETypeClus

ETypeClus (Shen et al., 2021) extracts salient predicate-object pairs and clusters their embeddings in a spherical latent space. For consistency across our evaluations, we follow the re-implementation of the ETypeClus model in (Choi et al., 2022), which consists of the latent space clustering stage of the ETypeClus pipeline and uses the embeddings of trigger mentions to be the input features. We utilize the contextualized embeddings of the candidate triggers extracted from SPEED for unsupervised training. The candidate trigger extraction process and the evaluation setup are the same as described in § E.3.

We run our experiments for ETypeClus on an NVIDIA RTX 2080 Ti machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 19.

Pre-trained LM	BERT-Base
Training Batch Size	64
Eval Batch Size	64
Learning Rate	0.0001
Gradient Clipping	1
# Pretrain Epochs	10
# Training Epochs	50
KL Loss Weight	5
Temperature	0.1
# Target Unknown Event Types	30

Table 19: Hyperparameter details for ETypeClus model.

E.5 BERT-QA

BERT-QA (Du and Cardie, 2020) is a classification model utilizing label semantics by formulating event detection as a question-answering task. We run our experiments for BERT-QA on an NVIDIA RTX A6000 machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 20.

Pre-trained LM	RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	30
Warmup Epochs	5
Max Sequence Length	175
Linear Layer Dropout	0.2

Table 20: Hyperparameter details for BERT_QA model.

E.6 DEGREE

DEGREE (Hsu et al., 2022) is a generation-based model prompting using natural language templates. We run our experiments for DEGREE on an NVIDIA RTX A6000 machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 21.

Pre-trained LM	BART-Large
Training Batch Size	32
Eval Batch Size	32
Learning Rate	0.00001
Weight Decay	0.00001
Gradient Clipping	5
Training Epochs	45
Warmup Epochs	5
Max Sequence Length	250
Max Output Length	20
Negative Samples	15
Beam Size	1

Table 21: Hyperparameter details for DEGREE model.

E.7 TagPrime

TagPrime (Hsu et al., 2023) is a sequence tagger priming words to input text to convey more task-specific information. We run our experiments for TagPrime on an NVIDIA RTX A6000 machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 22.

Pre-trained LM	RoBERTa-Large
Training Batch Size	64
Eval Batch Size	8
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	100
Warmup Epochs	5
Max Sequence Length	175
Linear Layer Dropout	0.2

Table 22: Hyperparameter details for TagPrime model.

E.8 DyGIE++

DyGIE++ (Wadden et al., 2019) is a multi-task classification-based model utilizing local and global context via span graph propagation. We run our experiments for DyGIE++ on an NVIDIA RTX A6000 machine with support for 8 GPUs. The major hyperparameters for this model are listed in Table 23.

Pre-trained LM	RoBERTa-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	0.001
Weight Decay	0.001
Gradient Clipping	5
Training Epochs	60
Warmup Epochs	5
Max Sequence Length	200
Linear Layer Dropout	0.4

Table 23: Hyperparameter details for DyGIE++ model.

E.9 Keyword

This baseline model basically curates a list of keywords specific to each event and predicts a trigger for a particular event if it matches one of the curated event keywords. Event keywords are curated by expert annotators based on the gold triggers appearing in the SPEED dataset and classified as high confidence, medium confidence, and low confidence based on their occurrence counts and false positive rates (as described in § A.2.[11] Although this baseline accesses gold test data, it is meant to be a baseline to provide the upper cap for models of this family.

E.10 GPT-3

We use the GPT-3.5 turbo model as the base GPT model. We experiment with ChatGPT (OpenAI, 2021) for tuning our prompts that ensure output consistency. Our final prompt (as shown in Figure 12) comprises a task definition, ontology details, 1 example for each event type, and the final test query. We conducted a looser evaluation for GPT and only match if the predicted trigger text matches the gold trigger text (we didn't check the exact span match basically).

[11] We will release the set of keywords with our final code.

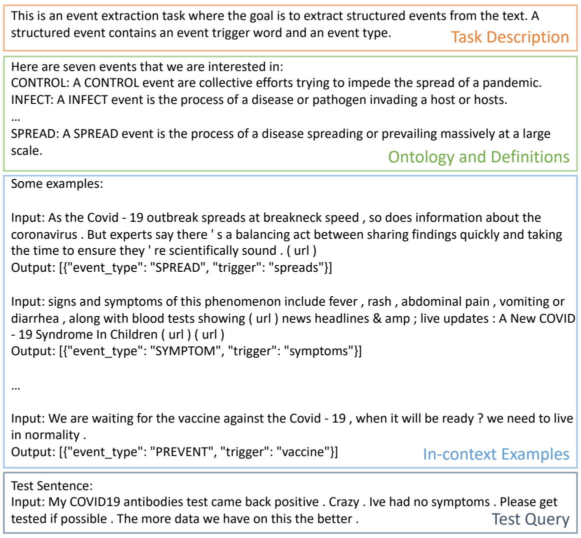

Figure 12: Illustration of the prompt used for GPT-3 model. It includes a task description, followed by ontology details of event types and their definitions. Next, we show some in-context examples for each event type and finally, provide the test sentence.

F Predicting Early Warnings for Monkeypox

F.1 Event-wise Analysis

As BERT-QA yields the strongest early warning signal (shown in Figure 5), we conduct an analysis at a more granular level on the contribution of each event type to the early warning signal based on the trained BERT-QA output. We present the results in Figure 13, which leads to the following observations: (1) **Strength of indication varies among event types**: As indicated in Figure 13, event type *infect* and *spread* are strong indicators of the incoming surge in reported cases, while event type *prevent* and *control* can serve as indicators of medium strength. Event type *symptom*, *cure*, and *death* are weak indicators that barely contribute to the early warning signal. (2) **Distribution across event types can potentially reveal high-level disease characteristics**: We can infer some properties of diseases based on the frequency of mentions about particular events. For example, *death* is less mentioned, which can indicate that *Monkeypox* is less fatal compared to other epidemics like COVID. We would like to mention that these are hypothetical properties based on predictions of our best model (which can be imperfect) and should be taken with a pinch of salt.

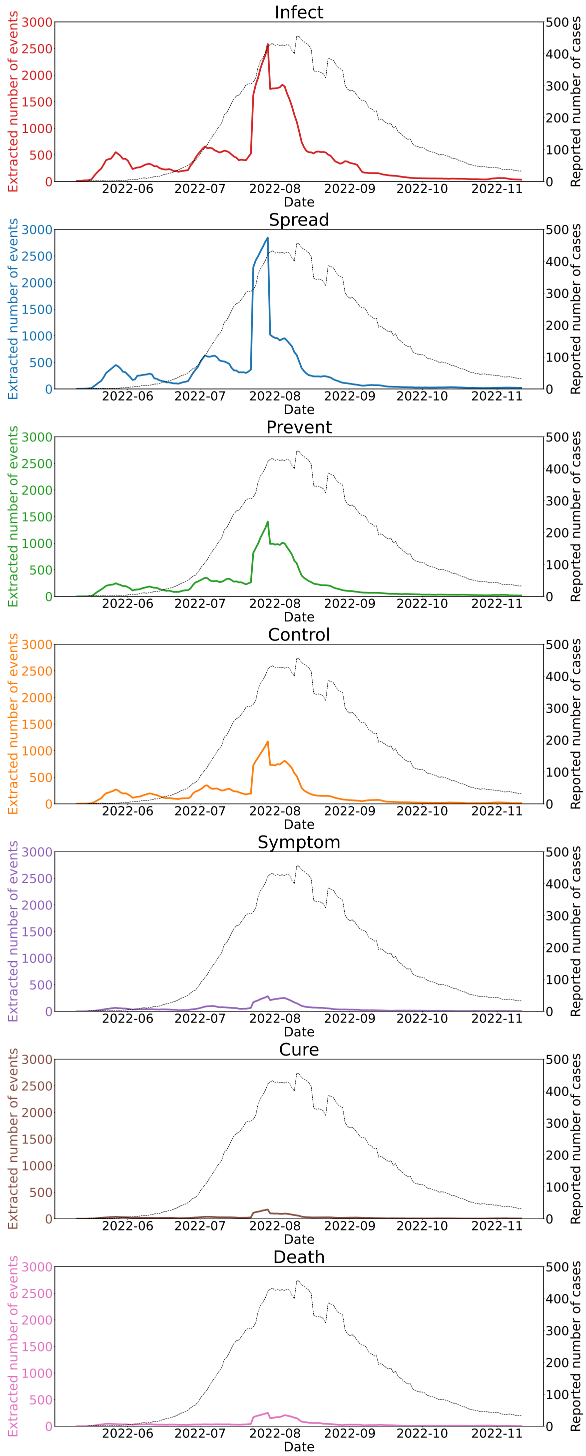

Figure 13: Number of reported Monkeypox cases and the number of extracted events for each SPEED event type from our trained BERT-QA model from XX to XX

An Event is defined as something happens in a sentence. In this task, we are trying to identity whether one or more of the following events exist in a given string: *infect, spread, symptom,prevent,control, cure, and death*. And if an event exist, what is the major **triggering word** that mostly manifest its occurrence.	
Event	Definition
infect	The process of a disease/pathogen invading host(s).
spread	The process of a disease spreading/pervailing massively at a large scale.
symptom	Individuals displaying physiological features indicating the abnormality of organisms.
prevent	Individuals trying to prevent the infection of a disease.
control	Collective efforts trying to impede the spread of a pandemic.
cure	Stopping infection and relieving individuals from infections/symptoms.
death	End of life of individuals due to infectious disease.
If there exist any explicit negation of an Event, we say that Event does NOT exist and do not mark it.	
Important Notes:	
There can be sentences without any events. No need to annotate anything for such sentences.	
A trigger word can be linked to one or more events. Choose all possible events in such cases.	
Multiple events can be presented in a given sentence. Mark all such events.	
The same event can occur multiple times (at different parts) in the same sentence. Mark all occurrences of the event.	
You will be able to submit the HIT at the last sentence once you finish annotating all the sentences.	
Select "flag" event if you see multiple triggering words or any other tricking situations that needs revisiting, but do not abuse this function.	

Figure 14: Task summary and the major annotation guidelines.

Here are more detailed instructions for how to choose the most appropriate triggering word.

Goal: Look for the one word that MOST LIKELY manifests the event's occurrence. You can use the following priority order for annotation:

1. Most of the times, the trigger of the event will be the main verb in the sentence.
2. If the verb is ambigous/vague, the trigger would be a noun semantocally related to the event.
3. (Rare case) If no such noun exist, the trigger would be any adjective/adverb that is realated to the event.
4. If still confused, use your best judgement to select the trigger.

In the following illustrations, correct trigger words are marked **blue**.

CASE I : main verb

Example Sentence: "I was coughing and got a fever yesterday and today confirmed I did not get COVID"

Annotation: There are 2 events of symptom

a. ...**got** a fever...-->Event symptom.

b. ...was **coughing**... -->Event symptom.

c. Note 1: "fever" and "COVID" are Not marked as triggering word of the events since the main verbas indicate the event.

Note 2: Here, due to the presence of "and", we have two occurrences of the event symptom.

d. Although "get COVID" appears, "not" is the negation emphasizing no infection happens, so event infect does NOT occur

e. More examples of main verbs as triggering word:

Example	**Event**
fight against the pandemic	control
caught a flu	infect
recover from COVID	cure
COVID **takes** lives	death
prevent infection	prevent
stomach **hurts**	symptom
number of infection **increases**	spread

CASE II : nouns

Example Sentence: "Fever, cough, and headache are the most common symptoms of COVID"

Annotation: Here we have 1 event of symptom event:

a. ...**symptoms** -->Event symptom.

b. Note: "fever","cough", and "headache" manifest the symptom event but they are NOT triggering words because "symptom" better manifests the Event.

c. More examples of nouns as triggering word:

Example	Event
death rate	death
therapy for COVID	cure
infection **prevention**	prevent
control of spread	control
signs of infection	symptom
spreading of COVID	spread
infection rate	infect

CASE III : adjective

Example Sentence: "I am feverish since 2 days ago"

Annotation: Here we have 1 event of the symptom event

a. ...**feverish** -->Event symptom.

b. Note: Here, we do not have a strong verb/noun for marking the trigger. Thus we mark "feverish".

c. More examples of nouns as triggering word:

Example	Event
get **rid** of disease	cure
stay **cautious** against virus	prevent
contagious virus	infect

Figure 15: Guidelines to choose the proper triggering word.

Good Examples

Example 1 : "3000+ people are dead due to COVID, so every one please remember to wear a mask and follow the rules to prevent infection and protect our nation from the virus."

Annotation:

a. **prevent** --> evemt prevent

b. **protect** --> event control

c. **dead**-->event death

Note1: Although "infection" is mentioned, it is prevented, meaning no infection is happening in the sentence, so event infect does NOT exist

Note2: Do not mark negation of an event.

Note3: intuitively, people die of COVID must have been infected, but event infect DOES NOT edist here because
An event must be triggered via triggering word and cannot be infered from another event.

Example 2: "if you ever have a fever, or cough, or have a sore throat, or feel difficult breathing, get tested immediately since you may have been infected."

Annotation:

a. **...have** a fever --> event symptom

b. ...been **infected** --> event infect

Note1: if have more than two parallel phrases triggering an event, only mark the first one instead of all of them.

Note2: event infect has no explicit negation, so event infect exists here.

Bad Examples

Example 1: "Wear a mask""

Wrong annotation:

a. **wear**-->event prevent

Note1: we may link the action of wearing a mask with pandemic prevention directly, but here it is just an action similar to "read a book" or "eat my lunch".

Note2: if the sentence is instead "wear a mask to prevent COVID." we mark prevent as a triggering word for event prevent instead of "wear"
Look for Events themselves instead of actions/policies related to Events.

Example 2:"Two weeks of quarantine is killing me! May God cure my mind and stop my crazy thoughts."

Wrong annotation:

a. killing-->event death

b. cure--> event cure

Note1: killing does not indicate any body is dying, and cure does not indicate a therapy against a disease.

Note2: Do NOT mark hyperbole or rhetorics as Events

Figure 16: Positive and Negative examples in the annotation guideline.

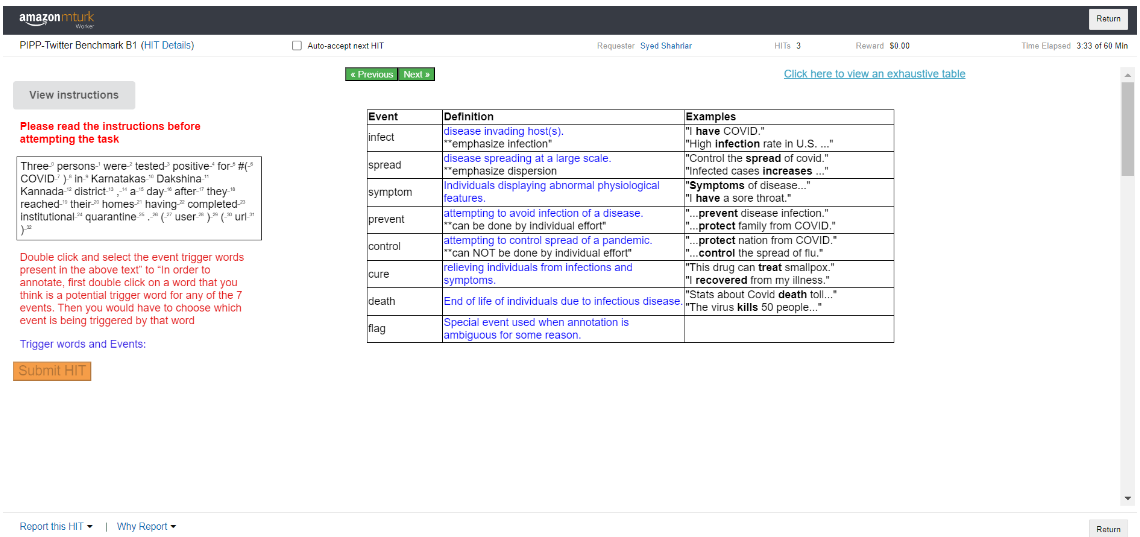

Figure 17: Illustration of the default annotation interface on Amazon Mechanical Turk.

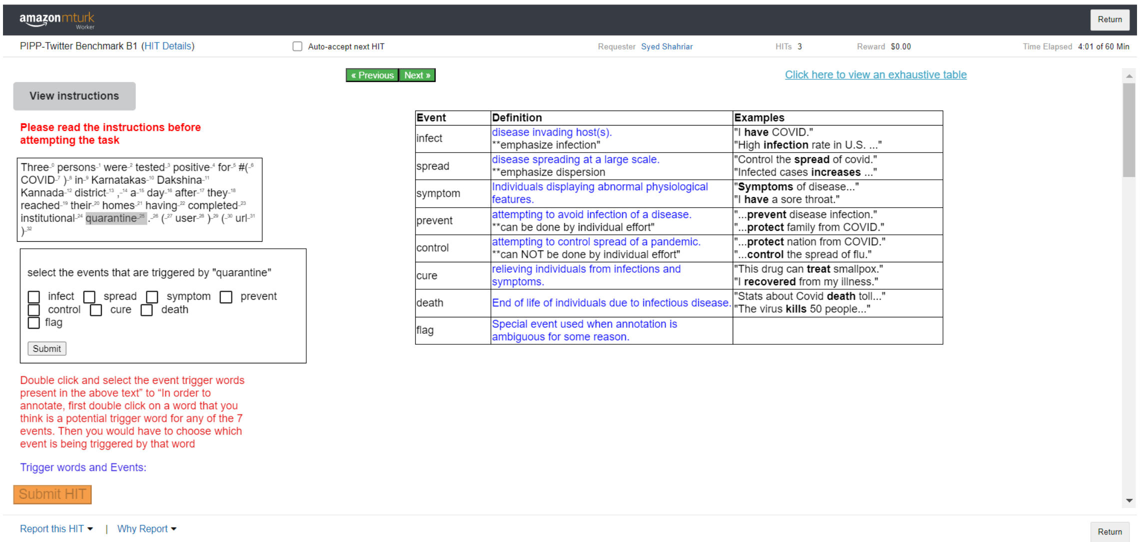

Figure 18: Illustration of selection of a word within a tweet for annotation in the interface.

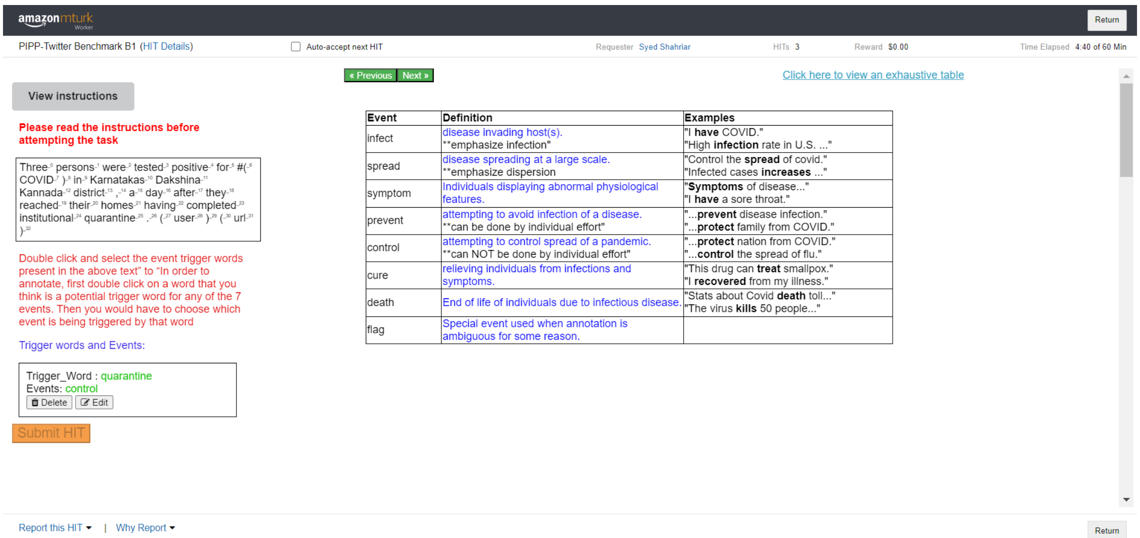

Figure 19: Illustration of the format and options available for a completed annotation in the interface.