Event Detection from Social Media for Epidemic Preparedness

Anonymous ACL submission

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 locally deployed measures can be crucial for policy making during epidemic outbreaks. In this work, we exploit Event Detection (ED) for extracting and capturing relevant events from social media posts to provide better preparedness for any upcoming epidemic. To facilitate this task, we curate an epidemic event ontology comprising seven generic event types such as *infect*, *symptom*, *prevent*, etc. Using our event ontology and human expert annotation, we construct our epidemic preparedness Twitter dataset SPEED comprising 1,975 tweets and 2,217 event mentions for the COVID-19 pandemic. Experiments reveal that existing ED models and datasets cannot transfer well for our task, highlighting the challenging nature of our dataset. Finally, we provide empirical evidence highlighting the utility and generalizability of our dataset by showing that ED models trained on our COVID-only dataset SPEED, can effectively identify epidemic events and offer timely warnings for three unseen epidemics of Monkeypox, Zika, and Dengue. This generalizability of SPEED lays the foundations for better preparedness against emerging epidemics.[1]

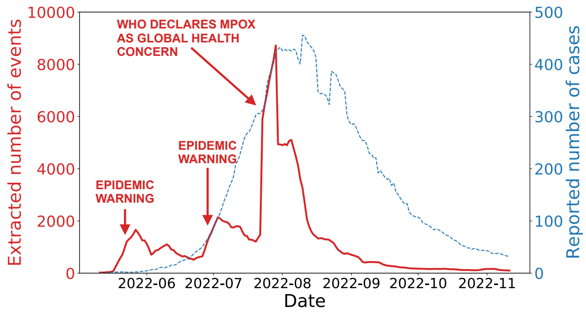

Figure 1: Number of reported Monkeypox cases and the number of extracted events from our trained BERT-QA model from May 11 to Nov 11, 2022. Indicated arrows show how our system can potentially provide early epidemic warnings almost 4-8 weeks before the WHO declared Monkeypox as a pandemic.

1 Introduction

Early epidemic 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 becomes an important information source here, as it's 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.[2] In our work, we explore the social application of information extraction towards building an automated system to efficiently extract epidemic events from social media to provide better epidemic preparedness.

The process of identifying and categorizing significant events based on a pre-defined ontology is a well-established task in NLP, known as *Event Detection* (ED) (Sundheim, 1992; Doddington et al., 2004). However, standard ED datasets mostly focus on general-purpose events for news or Wikipedia domains, and can't be transferred to the epidemic domain (§ 5). Furthermore, most prior epidemiological ED ontologies restrict themselves only to certain diseases or are too fine-grained and specific in nature; while the corresponding datasets majorly focus on news or clinical domains (§ 8). Thus, existing ED ontologies and datasets are not sufficient and models trained on them cannot be readily utilized for extracting events from social

[1] Code and data will be released upon acceptance.

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

media for emerging epidemics.

To this end, we construct our own epidemic ED ontology and dataset for social media. Our ontology comprises seven event types - *infect*, *spread*, *symptom*, *prevent*, *cure*, *control*, *death* - chosen based on their relevance for epidemic preparedness, frequency of mentions in social media, and their applicability to various diseases. Our ontology and event definitions are derived from clinical sources (Collier et al., 2008; Babcock et al., 2021) and its sufficiency and coverage are validated by public health experts and quantitative analyses. For our dataset, we choose Twitter as the social media platform and focus on the recent COVID-19 pandemic. Since our task requires domain expertise, we hire six expert annotators to ensure high annotation quality for our dataset. Using our curated ontology and expert annotation, we create our epidemic preparedness dataset **SPEED** (**S**ocial **P**latform based **E**pidemic **E**vent **D**etection) comprising 1,975 tweets and 2,217 event mentions. SPEED provides good coverage of events characteristic of any disease and is granular for social media; thus, serves as a valuable ED benchmark for epidemic preparedness from social media.

We benchmark various existing models including four zero-shot models (Shen et al., 2021; Lyu et al., 2021) and two supervised models (Hsu et al., 2022) pre-trained on existing ED datasets of ACE (Doddington et al., 2004) and MAVEN (Wang et al., 2020) on our SPEED benchmark. Experiments reveal that none of the existing models perform well on our dataset mainly owing to the domain-shift and noise in social media as well as unseen epidemic-based event types. Furthermore, training on limited in-domain SPEED data provides significant gains compared to the existing models, highlighting the importance of domain-specific training. Overall, these results reveal how SPEED is a challenging ED dataset.

Tying back to our original motivation of epidemic preparedness, we evaluate the utility and generalizability of our COVID-only dataset SPEED to detect events for any emerging epidemics. More specifically, we evaluate models trained only on SPEED to detect events for three unforeseen epidemics of *Monkeypox*, *Zika*, and *Dengue*. Experiments reveal that SPEED-trained models can successfully detect events for all these epidemics while providing improvements of 29% F1 over zero-shot models and 10% F1 over supervised models trained on small samples of target epidemic data. Furthermore, by comparing the trends of our extracted events with the actual reported cases, we show that our model can provide early preparedness warnings for the Monkeypox epidemic (Figure 1). These results underscore the strong generalizability and applicability of our dataset SPEED for general epidemic preparedness.

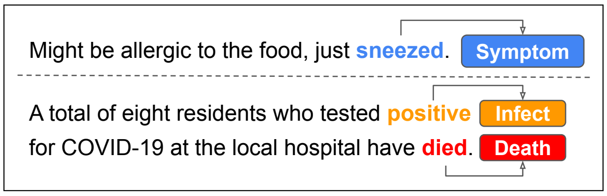

Figure 2: Illustration for the task of Event Detection. Event mentions: Event *symptom* and trigger *sneezed* (1st sentence), Event *infect* and trigger *positive* (2nd sentence), Event *death* and trigger *died* (2nd sentence).

Overall, we make the following contributions: (1) We create an ED ontology and dataset SPEED tailored for predicting epidemic events charactersitc of any disease from social media, (2) We show that existing zero-shot models and datasets cannot transfer well to our dataset, highlighting the significance of our dataset, (3) We validate the generalizability of our framework by demonstrating how SPEED-trained ED models using only COVID-tweets can successfully detect events and provide early warnings for three unforeseen epidemics.

2 Task Definition

We employ the task of Event Detection (ED) (Sundheim, 1992; Grishman and Sundheim, 1996) for identifying epidemic events from social media. We define ED based on the ACE 2005 guidelines (Doddington et al., 2004). An **event** is 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 the task of identifying event triggers from sentences and classifying them into one of the pre-defined event types. The subtask of identifying event triggers is called **Trigger Identification** and classification into event types is **Trigger Classification** (Ahn, 2006). The event types of interest are pre-defined by an **event ontology**. Figure 2 shows examples for three event mentions for the events *symptom*, *infect*, and *death*.

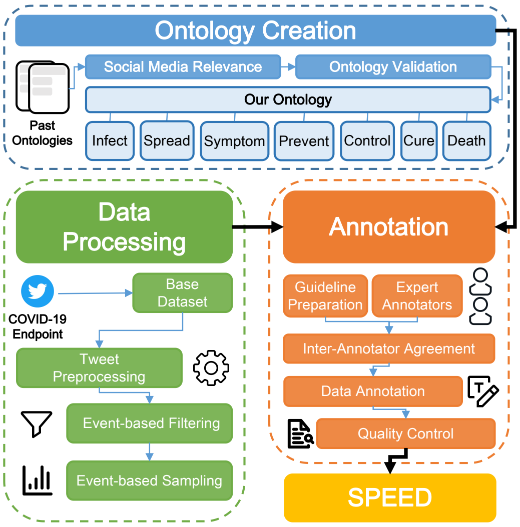

Figure 3: Overview of our dataset creation process.

3 Ontology Creation and Data Collection

We choose social media as our document source since it provides faster and more timely worldly information than other alternatives like 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**[3] as the social media platform and consider the recent **COVID-19 pandemic** as the primary disease for our dataset.

Previous epidemiological ontologies are typically specific to a particular disease, too fine-grained, or cover only a few event types (§ 8 and Table 6) and cannot be readily utilized for ED from social media. Similarly, standard ED datasets don't comprise epidemiological events and are mostly confined to news or wikipedia domains (§ 8). Owing to these reasons, we create our own event ontology and dataset SPEED specific to detecting epidemics from social media. We provide a brief overview of our data creation process in Figure 3 and discuss these steps in more detail 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., 2021a), we curate a wide range of epidemic events while ensuring that they are not biased for specific diseases. We categorize these events into three abstractions of social (events involving larger populations), personal (individual-oriented events), and medical (medically focused events) types and create our initial ontology comprising 18 event types as reported in Table 19 (§ A.1).

Social Media Relevance To adapt our curated ontology better for social media, we conduct a deeper analysis of the event types based on their frequency and relevance. Majorly, we associate each event type with a certain set of keywords and rank them based on the confidence and frequency of the occurrence of their keywords in social media posts (more details in § A.2). Based on this relevance-based ranking, we merge and discard some event types. Furthermore, we conduct human studies and merge event types to ensure better pairwise distinction.

Ontology Validation and Coverage Drawing upon established epidemiological ontologies serves to guarantee the medical soundness of our ontology. In addition, we assess the sufficiency and comprehensiveness of our ontology and definitions through evaluation by two public health experts. We also quantify our ontology coverage for four diverse diseases by evaluating the percentage of event occurrence in disease-related tweets. We observe a high coverage of 50% for COVID-19, 44% for Monkeypox, 70% for Dengue and 73% for Zika (more details in § A.3), ensuring strong disease coverage of our ontology.

Our final SPEED ontology comprises seven major event types that are better suited for social media and cover important 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 Despite COVID-based filtering, most tweets in our base dataset expressed subjective public sentiments, while only 3% comprised mentions adhering to our curated event ontology.[4] To reduce annotation costs, we further filter

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

[4] Based on keyword-based study conducted on 1,000 tweets

Event Type	Event Definition	Example Event Mention
Infect	The process of a disease/pathogen invading host(s)	Children can also **catch** COVID-19 ...
Spread	The process of a disease spreading/prevailing massively at a large scale	#COVID-19 CASES **RISE** TO 85,940 IN INDIA ...
Symptom	Individuals displaying physiological features indicating the abnormality of organisms	(user) (user) Still **coughing** two months after being infected by this stupid virus ...
Prevent	Individuals trying to prevent the infection of a disease	... wearing mask is the way to **prevent** COVID-19
Control	Collective efforts trying to impede the spread of epidemic	Social Distancing is our responsibility to **reduce** spread of COVID-19 ...
Cure	Stopping infection and relieving individuals from infections/symptoms	... **recovered** corona virus patients cant get it again
Death	End of life of individuals due to infectious disease.	More than 80,000 Americans have **died** of COVID ...

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

these tweets based on our curated ontology using a simple *sentence embedding* similarity technique. Specifically, we associate each event type with a seed repository of 5-10 diverse tweets. Query tweets are filtered out based on their sentence-level similarity (measured using the BERT sentence embedding model (Reimers and Gurevych, 2019)) with this event-based seed repository. This step filters about 95% tweets from our base dataset significantly reducing 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 an 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. We sample a total of 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 of the events outlined in our ontology. If an event is indeed mentioned, annotators are required to identify the specific event trigger. Following the standard ACE dataset (Doddington et al., 2004), we design our annotation guidelines and amend them through several rounds of preliminary annotations to ensure consistency amongst the annotators. Additional details and illustrations of the annotation guidelines and interface are provided in Appendix C.

Annotator Details To ensure high annotation quality and enforce consistency, we choose six experts instead of crowdsourced workers for our annotation. These experts are computer science students studying NLP and are well-versed with the task of ED. They were further trained on our task through multiple loops 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 Improvement:* Three annotators participated in three annotation rounds with a focus on improving the guidelines. 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. IAA score improved from 0.56 in the first round to a strong 0.65 (50 samples) in the final round.

Quality Control Apart from extensive IAA studies, we deploy two mechanisms to ensure the high annotation quality: (1) *Multi-Annotation:* Each tweet is annotated by two annotators and disagreements are resolved by a third annotator. (2) *Flagging:* Annotators can "flag" ambiguous annotations, which are then resolved and annotated by a third annotator through collective discussion. Both these mechanisms along with a good IAA score ensure that the annotations have high quality.

4 Data Analysis

In this section we present quantitative analyses of our dataset for comparison with other standard ED datasets. Comparison with other epidemiological datasets is discussed in § 8 along with an objective comparison in Table 6.

Dataset	# Event Types	# Sent	# EM	Avg. EM per Event	Domain
ACE	33	18,927	5,055	153.2	News
ERE	38	17,108	7,284	191.7	News
MAVEN	168	49,873	118,732	706.7	Wikipedia
SPEED	7	1,975	2,217	316.7	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.

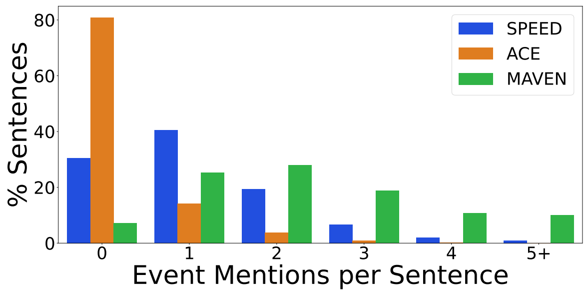

Figure 4: Distribution of number event mentions per sentence. Here % indicates percentage.

Dataset	# Unique Triggers	Avg. Triggers per Mention
ACE	1,229	0.24
MAVEN	7,074	0.06
SPEED	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.

Data Statistics Our dataset SPEED comprises seven event types with 2,217 event mentions annotated over 1,975 tweets. We compare SPEED with other standard ED datasets like ACE (Doddington et al., 2004), ERE (Song et al., 2015), and MAVEN (Wang et al., 2020) in Table 2. Despite the lesser number of sentences and event mentions (since we focus only on 7 event types), SPEED has a reasonable size of 316 average event mentions per event, which is more than the standard ACE and ERE datasets. We also note the differences of the domain of data sources as ACE/ERE focus on News, MAVEN on Wikipedia, while SPEED is based on social media, specifically Twitter data.

Event Mention Density Analysis 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 despite having just seven event types, SPEED is a fairly dense dataset.

Trigger Word Analysis 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.

5 Transfer Existing Methods

Since existing ED datasets and models are based on general-purpose event ontologies and news/wikipedia domains, they may not transfer well to our social-media-based epidemic detection task. In order to verify this hypothesis, we benchmark the transfer capabilities of these existing methods to our dataset SPEED. For this experimentation, we assume no access to any annotated social media data for epidemic events. We majorly consider the following two families of models:
Zero-shot models do not train on any supervised data and utilize names and definitions of the events for ED. For this, we consider (1) **TE** (Lyu et al., 2021), a pre-trained model that uses event definitions to formulate ED as a textual entailment and question-answering task, (2) **WSD** (Yao et al., 2021) which encodes the contextualized trigger and event definitions jointly and uses a classification head atop for event detection. (3) **TABS** (Li et al., 2022), a model that utilizes two complementary embedding spaces ("mask view" and "token view") to classify examples of new event types. (4) **ETypeClus** (Shen et al., 2021), that extracts salient predicate-object pairs and clusters the embeddings of these pairs in a spherical latent space.
Data transfer models are supervised models pre-trained on other standard ED datasets like ACE (Doddington et al., 2004) and MAVEN (Wang et al., 2020) and transfer to SPEED in a zero-shot manner. For this, we consider (5) **DEGREE** (Hsu et al., 2022), a generation-based model prompting using natural language templates, (6) **TagPrime** (Hsu et al., 2023), a sequence tagging approach that utilizes priming words to input text to convey more task-specific information.

5.1 Evaluation

We evaluate the above models on the 1,683 tweets from the SPEED dataset. Following previous works (Ahn, 2006), we report the F1-score for the two tasks of Trigger identification (**Tri-I**) and trig-

Model	Tri-I	Tri-C
DATA-TRANSFER		
ACE - TagPrime	0	0
ACE - DEGREE	1.82	1.71
MAVEN - TagPrime	**27.65**	0
MAVEN - DEGREE	26.72	0
ZERO-SHOT		
TE	9.64	5.54
WSD	17.68	3.65
TABS	3.70	1.61
ETypeClus	17.56	**7.66**

Table 4: Benchmarking existing zero-shot and data-transfer models on SPEED in terms of Tri-I and Tri-C F1 scores.

ger classification (**Tri-C**) respectively. The results are shown in Table 4. We observe that models pre-trained on the news dataset ACE absolutely fail, while Wikipedia dataset MAVEN pre-training helps to improve Tri-I scores, but still has a nil Tri-C score. The zero-shot models using event definitions perform slightly better, while the best performance is provided by ETypeClus which is an unsupervised clustering model. Overall, **all existing zero-shot and data-transfer models fail to detect epidemic events, mainly owing to the domain shift of social media data and the finer granularity of epidemic events. In turn, this renders SPEED as a challenging ED dataset.**

6 Training with Limited SPEED Data

To improve model performance for SPEED, we conduct experiments trained ED models using limited amounts of in-domain SPEED training data. Majorly, we consider two training paradigms: (1) *Few-shot (FS)*: Models are provided access to n mentions per event (n-shot) for training. We explore 2-shot and 5-shot with three splits of data. (2) *Low Resource (LR)*: Models have access to a limited 100-300 event mentions for training. (Data Statistics in Table 9 in Appendix § D.1).

For training, we consider the following ED models: (1) **DyGIE++** (Wadden et al., 2019), a multi-task classification-based model utilizing local and global context via span graph propagation, (2) **BERT-QA** (Du and Cardie, 2020), a classification model utilizing label semantics by formulating event detection as a question-answering task. We also consider (3) **DEGREE** and (4) **TagPrime** models (as described before in § 5). Other baselines also include (5) **Keyword** (Lejeune et al., 2015),

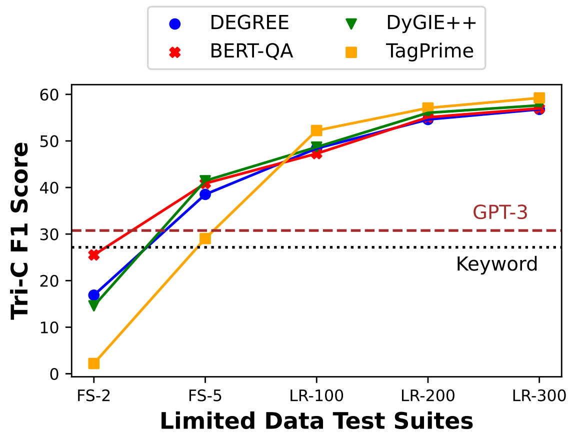

Figure 5: Model performances on the Few-Shot (FS) and Low Resource (LR) test suites in terms of Tri-C F1 scores. Here, LR-XX represents low resource with XX training event mentions and FS-Y represents few-shot with Y training mentions per event.

a popular epidemiological model, that predicts an event if any of the event-specific *curated* keywords are present in the sentence, (6) **GPT-3** (Brown et al., 2020), a large-language model (LLM) baseline using GPT-3.5-turbo as the base model with seven *in-context* examples.

6.1 Evaluation

We follow the same evaluation setup described in § 5.1. Figure 5 presents the model performances for the few-shot and low-resource settings. Majorly, we observe how training on in-domain data can yield performance gains upto 50 F1 points compared to zero-shot and data-transfer methods. We also note that GPT-3 and Keyword baselines are easily outperformed by models trained with just 30 event mentions. Furthermore, these gains are highly consistent for the different ED models. Overall, we note that **small amounts of in-domain training data can provide significant gains in model performance compared to the existing zero-shot and data-transfer models**.

7 Generalization for New Epidemics

Since SPEED focuses solely on COVID-19, its transferability for detecting events for new epidemics remains unknown. To effectively evaluate this generalization, we test if models trained only using our in-domain COVID dataset can detect events for unseen epidemics without any further fine-tuning on the new epidemic data. Specifically, we consider the outbreaks of three diverse diseases

Model	Monkeypox Tri-I	Monkeypox Tri-C	Zika + Dengue Tri-I	Zika + Dengue Tri-C
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 + 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
Keyword	36.40	25.09	25.93	21.69
GPT-3*	42.23	35.33	53.22	14.27
Trained for 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
Transfer from SPEED				
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: Benchmarking ED models trained on COVID-only SPEED for generalizability to new epidemics of Monkeypox, Zika and Dengue in terms of F1 scores.

of *Monkeypox* (2022), *Zika* (2017), and *Dengue* (2018) as the unseen epidemics.

Experimental Setup For creating datasets, we utilize the Twitter datasets of Thakur (2022) for Monkeypox and Dias (2020) for Zika and Dengue. Using expert annotation for a sample of the tweets, our final evaluation dataset comprises 286 tweets with 398 event mentions for Monkeypox while 300 tweets with 274 event mentions for Zika and Dengue (statistics in § D.3 and § D.4).

For model training, we use a 80-20 split of our COVID-only SPEED dataset to train various ED models (Transfer from SPEED). For comparison, we benchmark models trained on existing datasets (Transfer from Existing Datasets) and models requiring no training data (No Training). As strong baselines, we also consider supervised models trained on a small sample of 300 tweets for the target epidemic (Trained on Target Epidemic).

Results We present our results in Table 5. None of the existing data transfer methods or zero-shot methods perform well. Overall, we observe that ED models transferring from SPEED perform the best with model performance ranging from 60-65 F1 points, thus **demonstrating the generalizability of our SPEED dataset to new epidemics**. Furthermore, we observe SPEED-trained models even outperform models trained on for Monkeypox by 10 F1 points and are at par for Zika. This outcome is particularly encouraging, as it **demonstrates the resilience of SPEED-trained models, making them highly applicable during the early stages of an unfamiliar epidemic, when minimal to no epidemic-specific data is accessible**.

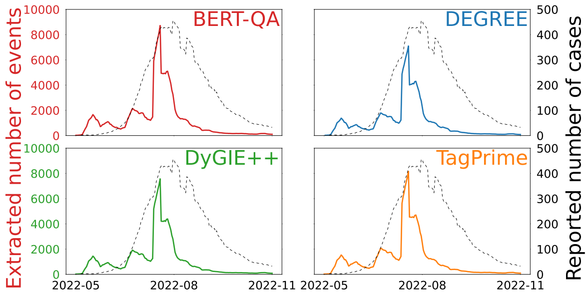

Figure 6: Number of reported Monkeypox cases and the number of extracted events from four trained models from May 11 to Nov 11, 2022.

7.1 Providing Early Epidemic Warnings

As a challenging yet practical evaluation, we evaluate our SPEED-trained models in their ability to provide early warnings for an unknown epidemic. We choose the Monkeypox as the unknown epidemic. 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[5] from May 11 to Nov 11, 2022, in Figure 1. As shown by the arrows in the figure, our model could potentially provide two sets of early warnings around May 23 and June 29 before the outbreak reached its peak around July 30. In fact, all our trained ED models are capable of providing these early signals as shown in Figure 6 (further analysis in Appendix F). This robust outcome underscores the **real-time practicality to provide early warnings and broad applicability to unknown epidemics of our SPEED dataset.**

8 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

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

Dataset	Data Source	Sentence Level	Trigger Present	Social Events	Personal Events	Social Media Granular
SPEED (Ours)	Twitter	✓	✓	✓	✓	✓
COVIDKB (Zong et al., 2022)	Twitter	✓	✗	✗	✓	✓
CACT (Lybarger et al., 2021)	Clinical	✗	✗	✗	∼	✓
ExcavatorCovid (Min et al., 2021b)	News	✗	✓	✓	✓	✗
BioCaster (Collier et al., 2008)	News	✗	✗	✓	✓	✗
DANIEL (Lejeune et al., 2015)	News	✗	∼	✗	∼	✓

Table 6: Objective comparison of various epidemiological datasets with our dataset SPEED. We objectify the source of base data (Data Source), the level of annotation granularity (Sentence Level), the presence of trigger information (Trigger Present), the presence of social and personal events in ontology (Social Events and Personal Events), and the suitability of ontology for social media (Social Media Granular). ∼ indicates partial presence.

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 (Min et al., 2021b) 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 6.

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) were developed for extracting symptoms and infection statistics respectively. Most of these systems focus on the domains of news and clinical notes and use keyword/rule-based or simple BERT-based models, as shown in Table 6. In our work, we explore more recent ED models while focusing specifically on the social media domain.

9 Conclusion and Future Work

In this work, we leverage the framework of Event Detection (ED) to extract epidemic events from social media to promote better epidemic preparedness. To facilitate this, we create our Twitter-based dataset SPEED comprising seven major epidemic event types. Through experimentation, we show how existing datasets and models fail to transfer for our task. Contrastingly, we show how models trained on SPEED can generalize and provide early warnings for unseen emerging epidemics. More broadly, our work demonstrates how event extraction and in general, information extraction can exploit social media to aid policy-making for better epidemic preparedness.

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 social biases embedded in SPEED. Since our work didn'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. 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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

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*.

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.

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.

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.

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, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 63–71. 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*, pages 63–71, Online and Punta Cana, Dominican Republic. 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.

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 Systems with Applications*, 181:115119.

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. *arXiv preprint arXiv: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

We present our complete initial event ontology comprising 18 event types organized into 3 abstract categories in Table 19. We also describe each event type by its definition and also present details about the action taken for its role in the final event ontology.

A.2 Initial analysis of events

Our initial ontology (§ A.1) was constructed using previous ontologies and human knowledge. But the suitability of each event type for social media (specifically Twitter) remains unknown. To evaluate this suitability, we use frequency and confidence as two guiding heuristics and use them for final filtering/merging. We utilize the base Twitter dataset for SPEED for conducting this analysis. We describe each of these heuristics here:

Frequency To approximately estimate the frequency of events, we curate a list of keywords for each event type and count the number of posts containing any of these keywords. Keyword curation involves creating a seed list using human expert knowledge and expanding that list using synonyms from external sources like Thesaurus.[6] We show the results in Figure 7. We observe that most events under the medical abstraction occur much lesser than others. Furthermore, the count variance is large as the most frequent event type *control* is 180 times more likely to occur than the least frequent event type *variant*. Since low-frequency events are less likely to be mentioned in a smaller sample of data, we discard or merge such events for our final ontology.

Confidence For each keyword, we randomly sample a small number of non-duplicate tweets and manually rate the keyword confidence based on the percentage of tweets wherein the semantic meaning of the keyword matches the definition of its event. We mainly categorize this confidence as high, medium, or low.[7] Take event *control* as an example, it has high confidence keywords such as "quarantine", "protocol", and "distancing"; medium confidence keywords such as "restrict", "postpone", and "investigate"; low confidence keywords such as "battle", "limitation", and "separation". On the other hand, event *prefigure* does not have high confidence keywords, but only medium confidence keywords such as "foreshadow" and low confidence keywords such as "foretell". Our heuristic suggests that low-confidence keywords are more likely to give false positives relative to high-confidence ones. Thus, we filter/merge event types that have a high number of low-confidence keywords.

Eventually, our final ontology comprises seven events that are distinguishable, frequent, and have a low false-positive rate.

A.3 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 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

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

[7] We release these keywords as part of our final code.

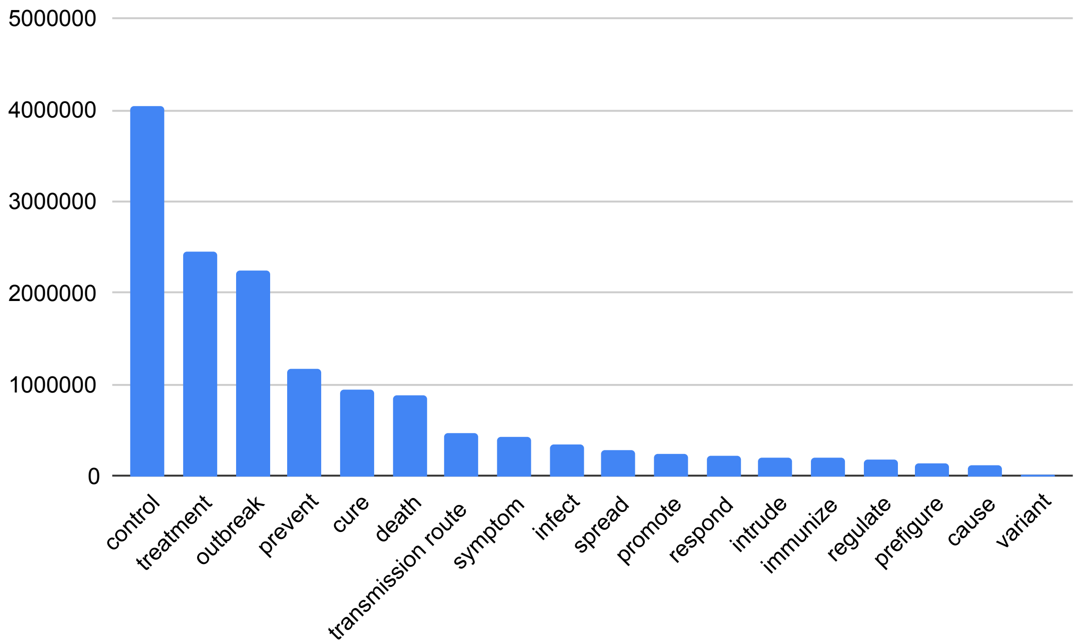

Figure 7: Frequency of occurrence based on keyword search for all event types in the initial complete ontology.

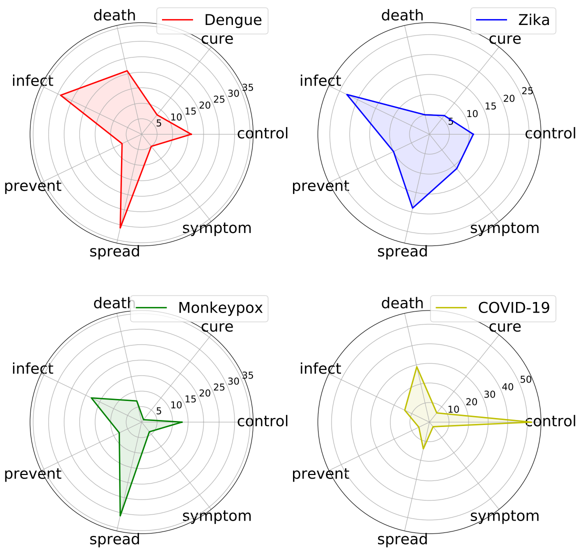

Figure 8: Event type distribution of the disease-related tweets for each disease. Numbers on the axis represent count of mentions for a given event type.

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’[8] 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.

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

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

We present our results in Table 7 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 8. 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 for robust and generalizable model training.

Model	Monkeypox Tri-I	Monkeypox Tri-C	Zika + Dengue Tri-I	Zika + Dengue 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 8: 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.

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. For our interface, we utilize Amazon Mechanical Turk.[9] 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.

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

D Data Analysis for SPEED

D.1 Benchmarking Test Suites Statistics

We provide the statistics in terms of number of event mentions and tweets for the various benchmarking test suites based on SPEED in Table 9.

	Test Suite	# Mentions	# Tweets
Train	FS-2	14	11
	FS-5	35	24.33
	LR-100	99	67
	LR-200	198	139
	LR-300	306	211
Dev	LR/FS	101	81
Test	All	1,810	1,683

Table 9: Data Statistics for the various benchmarking test suites in terms of number of event mentions and number of tweets. Here, LR-XX represents low resource with XX training event mentions and FS-YY represents few-shot with YY training mentions per event. For FS, we take the average over three different splits of data.

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.

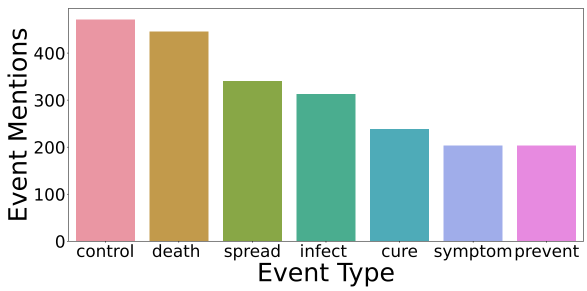

Figure 9: Distribution of event mentions per event type for our dataset SPEED.

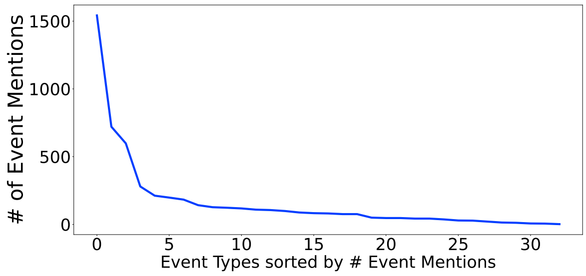

Figure 10: Distribution of event mentions for the event types in the ACE dataset.

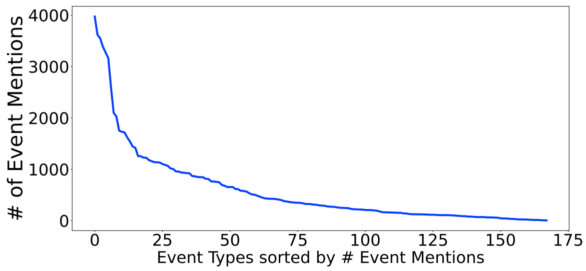

Figure 11: Distribution of event mentions for the event types in the MAVEN dataset.

D.3 Monkeypox Test Data Statistics

We share the data statistics of the evaluation dataset used for Monkeypox in Table 10 split according to each event type. We observe that there is a disparity 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 10: 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 11 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 11: Data Statistics for the evaluation dataset for Zika+Dengue Event Detection categorized by event types.

E Implementation Details for models

We present the extensive set of hyperparameters and other implementation details about the various ED models we benchmarked in our work.

E.1 BERT-QA

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 12.

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 12: Hyperparameter details for BERT_QA model.

E.2 DEGREE

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 13.

E.3 TagPrime

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 14.

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 13: Hyperparameter details for DEGREE model.

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 14: Hyperparameter details for TagPrime model.

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 15: Hyperparameter details for DyGIE++ model.

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 16: Hyperparameter details for WSD model.

E.4 DyGIE++

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 15.

E.5 TE

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 (Lyu et al., 2021).

E.6 WSD

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 16.

E.7 TABS

TABS is an event type induction model, wherein the goal is to discover new event types without a pre-defined event ontology. 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 17.

E.8 ETypeClus

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

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 17: Hyperparameter details for TABS model.

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.7.

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 18.

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 18: Hyperparameter details for ETypeClus 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.

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

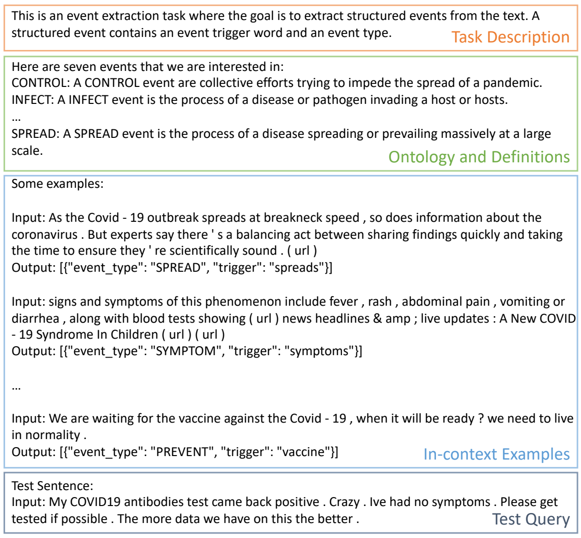

Figure 12: Illustration of the prompt used for GPT-3 model. It includes a task description, followed by ontology details of event types and their definitions. Next, we show some in-context examples for each event type and finally, provide the test sentence.

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).

F Predicting Early Warnings for Monkeypox

F.1 Event-wise Analysis

As BERT-QA yields the strongest early warning signal (shown in Figure 6), 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.

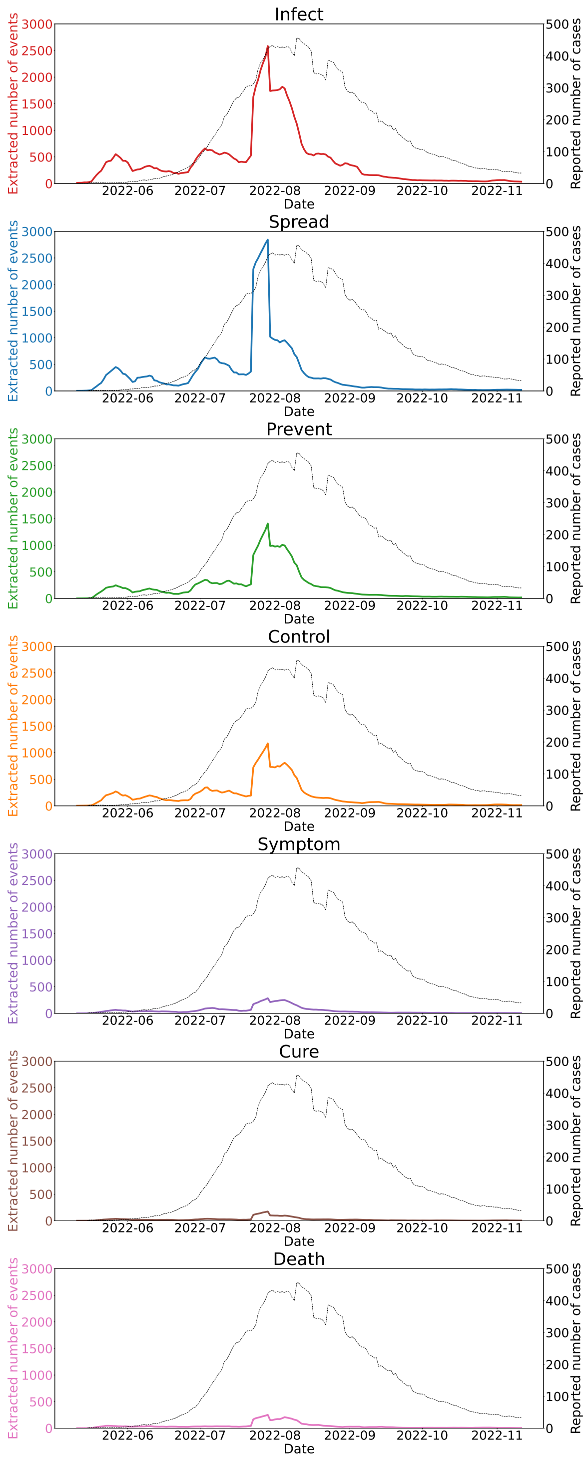

Figure 13: Number of reported Monkeypox cases and the number of extracted events for each SPEED event type from our trained BERT-QA model from XX to XX

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.

Event name	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 19: Complete initial epidemic event ontology comprising 18 event types organized into 3 higher-level abstract categories. We also present details about the event definitions and the action taken for each event type in the final ontology.

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.

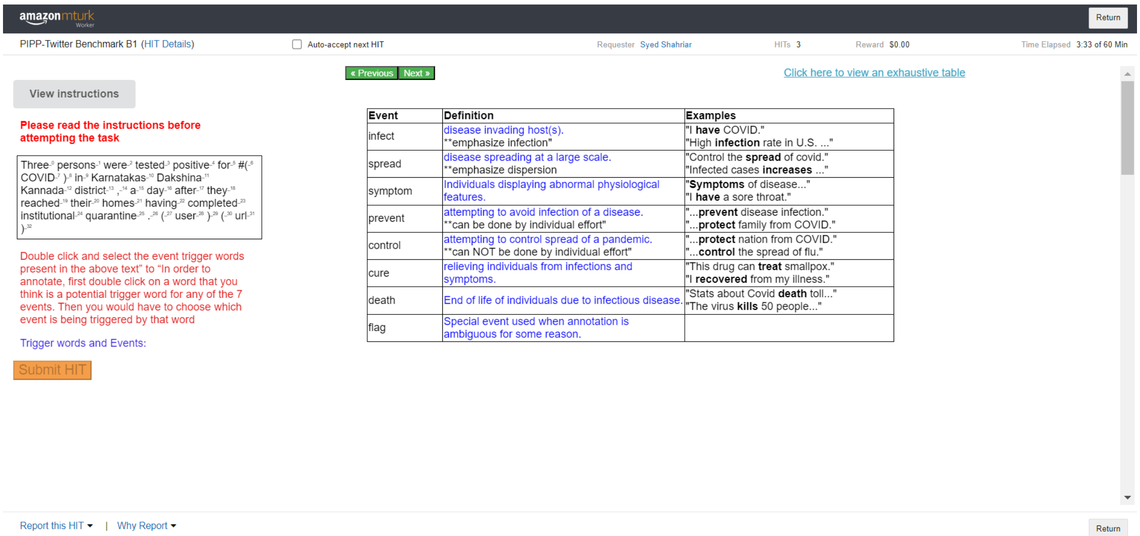

Figure 17: Illustration of the default annotation interface on Amazon Mechanical Turk.

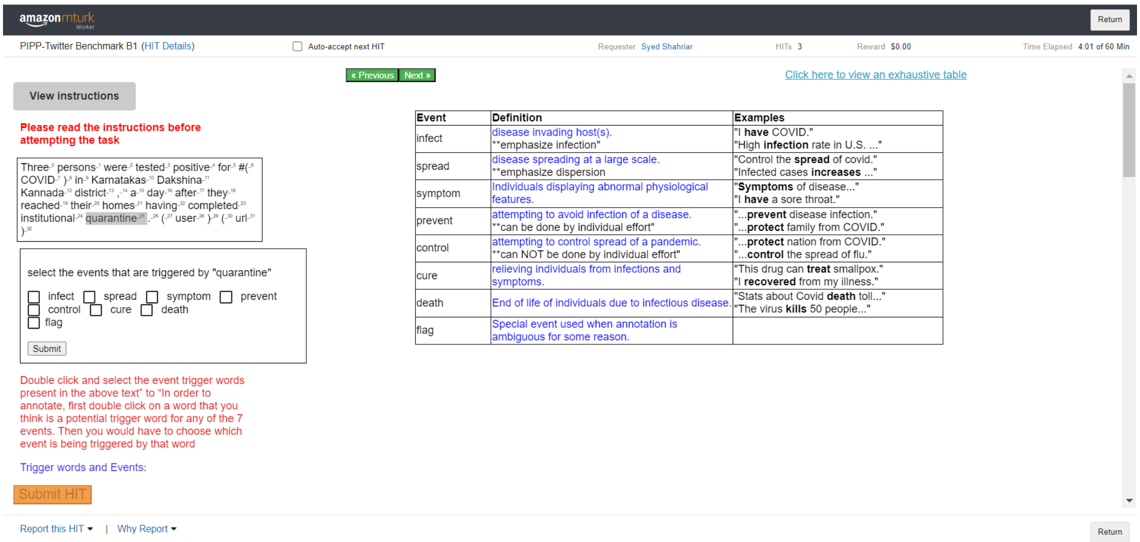

Figure 18: Illustration of selection of a word within a tweet for annotation in the interface.

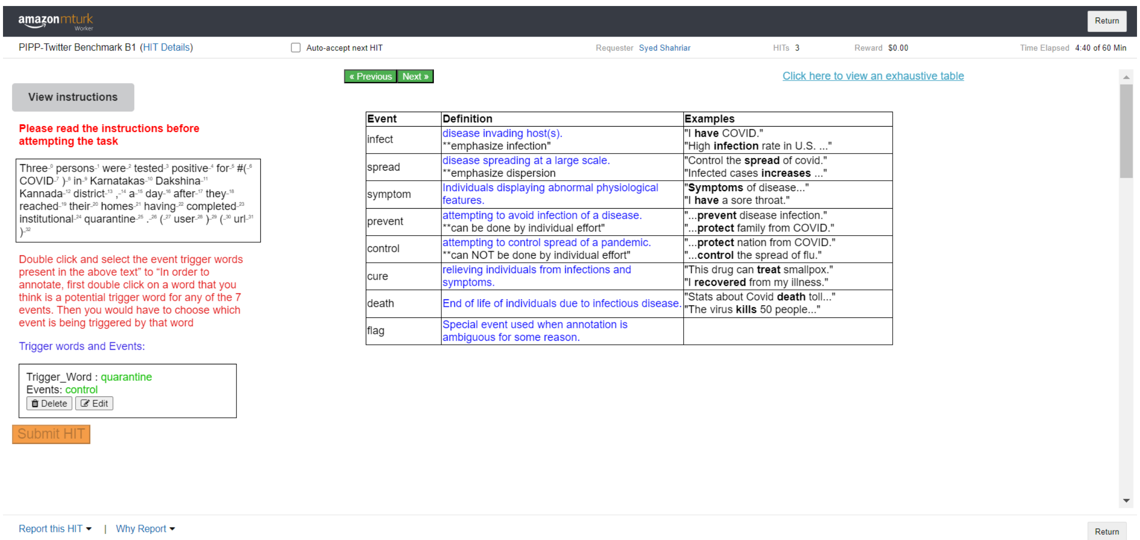

Figure 19: Illustration of the format and options available for a completed annotation in the interface.