METS-CoV: A Dataset of Medical Entity and Targeted Sentiment on COVID-19 Related Tweets

Peilin Zhou[1,2] **Zeqiang Wang**[1,2] **Dading Chong**[3] **Zhijiang Guo**[4]
Yining Hua[5] **Zichang Su**[1,6] **Zhiyang Teng**[7] **Jiageng Wu**[1,2] **Jie Yang**[1,2*]

[1]School of Public Health and the Second Affiliated Hospital, Zhejiang University, China
[2]The Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, China
[3]School of Electronic and Computer Engineering, Peking University, China
[4]Department of Computer Science and Technology, University of Cambridge, UK
[5]Department of Biomedical Informatics, Harvard Medical School, USA
[6]Chu Kochen Honors College, Zhejiang University, China
[7]School of Engineering, Westlake University, China

{zhoupalin, jieynlp}@gmail.com, {wzq99, suzc,jiagengwu}@zju.edu.cn
1601213984@pku.edu.cn, tengzhiyang@westlake.edu.cn
zg283@cam.ac.uk, yining_hua@hms.harvard.edu

Abstract

The COVID-19 pandemic continues to bring up various topics discussed or debated on social media. In order to explore the impact of pandemics on people's lives, it is crucial to understand the public's concerns and attitudes towards pandemic-related entities (e.g., drugs, vaccines) on social media. However, models trained on existing named entity recognition (NER) or targeted sentiment analysis (TSA) datasets have limited ability to understand COVID-19-related social media texts because these datasets are not designed or annotated from a medical perspective. This paper releases METS-CoV, a dataset containing medical entities and targeted sentiments from COVID-19-related tweets. METS-CoV contains 10,000 tweets with 7 types of entities, including 4 medical entity types (*Disease*, *Drug*, *Symptom*, and *Vaccine*) and 3 general entity types (*Person*, *Location*, and *Organization*). To further investigate tweet users' attitudes toward specific entities, 4 types of entities (*Person*, *Organization*, *Drug*, and *Vaccine*) are selected and annotated with user sentiments, resulting in a targeted sentiment dataset with 9,101 entities (in 5,278 tweets). To the best of our knowledge, METS-CoV is the first dataset to collect medical entities and corresponding sentiments of COVID-19-related tweets. We benchmark the performance of classical machine learning models and state-of-the-art deep learning models on NER and TSA tasks with extensive experiments. Results show that the dataset has vast room for improvement for both NER and TSA tasks. With rich annotations and comprehensive benchmark results, we believe METS-CoV is a fundamental resource for building better medical social media understanding tools and facilitating computational social science research, especially on epidemiological topics. Our data, annotation guidelines, benchmark models, and source code are publicly available (`https://github.com/YLab-Open/METS-CoV`) to ensure reproducibility.

*Corresponding Author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

1 Introduction

The outbreak of the COVID-19 pandemic has had severe implications for people's lives and global health (Khan et al., 2020). To assess the impact of this pandemic on the public, epidemiologists and medical researchers conducted research through various methods such as clinical follow-up (Huang et al., 2021), questionnaires (Kumari et al., 2020), and app tracking (Zens et al., 2020). Social media is one of the most popular media types in the world. With its large user base, rapid information dissemination, and active participation, social media has become an important channel for the public to express their feelings and opinions on COVID-19-related topics, which provides researchers with large-scale and low-cost materials to track the impact of COVID-19. For example, as shown in Fig.1, Twitter users express their attitudes toward medical entities (e.g., drugs and vaccines) through their Twitter posts. Large-scale analyses of such tweets are vital to understanding the public's opinions towards medical topics, which can further help medical research or public health management. Therefore, many studies have been proposed to track and analyze people through social media platforms such as Twitter, one of the most widely used social media platforms (Banda et al., 2021; Shahi et al., 2021; Xue et al., 2020; Boon-Itt et al., 2020; Hua et al., 2022).

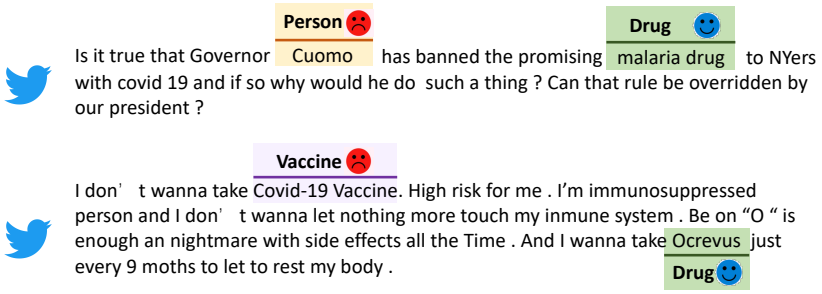

Figure 1: Examples of medical entities and targeted sentiments in tweets.

However, existing natural language processing (NLP) tools (Bird, 2006; Schmitt et al., 2019; Cao et al., 2022) struggle to fulfill the surging demand for accurate COVID-19 tweet analysis due to several reasons:

- Most NLP tools are not explicitly designed for social media texts, leading to dramatic performance degradation when applied to noisy texts such as tweets (Duarte et al., 2018).
- Most current NLP tools are developed for general applications, so domain adaptation can be poor without adding external knowledge from medical studies.
- Existing NLP tools are not designed from the medical or public health research perspective, making it challenging to meet the needs of epidemiologists to analyze medical topics.

In short, the very underlying reason for the poor applicability of these tools in COVID-19 social media studies is the lack of COVID-19-related social media datasets designed and annotated from a medical perspective. Furthermore, one of the most critical analysis goals for COVID-19-related studies conducted on tweets is to find the entities (both general and medical entities) discussed by users and their attitudes or feelings towards them. Such goals correspond to two basic NLP tasks: named entity recognition (NER) and targeted sentiment analysis (TSA). NER aims at entity extraction from unstructured texts, and TSA aims at user sentiment prediction towards the targeted entities.

In this work, we release **METS-CoV** (**M**edical **E**ntities and **T**argeted **S**entiments on **CoV**id-19-related tweets), a dataset that contains 10,000 tweets annotated with 7 types of entities, including 4 medical entity types (*Disease*, *Drug*, *Symptom*, and *Vaccine*) and 3 general entity types (*Person*, *Location*, and *Organization*). In addition, 4 types of entities (*Person*, *Organization*, *Drug*, and *Vaccine*) are selected and annotated with user sentiments to explore the attitudes of tweet users toward specific entities. Unlike other general NER and TSA datasets, METS-CoV is built from a public health research perspective and contributes to developing tailored natural language processing tools for the medical domain to mine valuable information from social media rather than only relying on keyword search. For example, based on the NER and TSA model trained on METS-CoV dataset, researchers can track public attitudes toward COVID-19 vaccination for more efficient vaccination policies. The models could also track the public's mental status in different COVID-phases, providing potential solutions for addressing the global mental health crisis. In addition, we design detailed annotation guidelines

for medical entities on tweets. The guidelines have been applied to our annotation process as strict supervision and control to ensure quality. We benchmark the performance of classical machine learning models and state-of-the-art deep learning models (including pretraining language models) on NER and TSA tasks of METS-CoV. According to the benchmark results, the NER and TSA tasks on the dataset have much space for performance improvement.

2 Related Work

METS-CoV supports two basic NLP tasks from a medical perspective: 1) named entity recognition, i.e., identifying general and medical entities, and 2) targeted sentiment analysis, i.e., predicting the attitudes of Twitter users towards specific entities (including *Drug* and *Vaccine*). This section reviews several commonly used open-source datasets for these two tasks and compares them with the proposed dataset.

2.1 Named Entity Recognition Datasets

CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) is one of the most widely used NER datasets with its newswire texts collected from the Reuters Corpus. The dataset consists of 4 general entity types: PER (Person), LOC (Location), ORG (Organization), and MISC (Miscellaneous), which are also adopted in the SciTech News dataset (Jia et al., 2019). WNUT NER (Strauss et al., 2016) is a benchmark NER dataset for the social media domain that consists of manually annotated tweets with 10 entity types. Nevertheless, none of the entity types is medical-related. Similarly, the recently release Tweebank-NER dataset is neither medical-related (Jiang et al., 2022). In the medical domain, NER is often used to extract medical terminologies from clinical case reports (CCRs) or electronic medical records (EMRs). A representative medical NER dataset is i2b2-2010 dataset (Uzuner et al., 2011), which includes discharge summaries and progress notes provided by well-known medical centers with 3 entity types: test, problem, and treatment. Besides, one of the SMM4H shared tasks (Klein et al., 2020; Weissenbacher et al., 2019) released a dataset for extracting tweet text spans with adverse drug reactions (ADR). This dataset is not COVID-19-related. On the other hand, the CORD-NER dataset (Wang et al., 2020b) has 75 fine-grained types of entities from scientific papers about COVID-19 and historical coronavirus research. But since social media texts have way more colloquial forms than scientific papers, models trained on WNUT or CORD-NER are unsuitable for social media analyses.

2.2 Targeted Sentiment Analysis Datasets

Most TSA studies typically experiment on 3 datasets: LAPTOP (Pontiki et al., 2014), TWITTER (Mitchell et al., 2013), and REST (Pontiki et al., 2015, 2016). Specifically, LAPTOP and REST are user review datasets collected from the laptop and restaurant domains. The TWITTER dataset has tweets but only with general types of entities (*Person* and *Organization*). At the same time, the data might be outdated for the ever-involving social media languages. There are several recent open-domain TSA datasets. For example, YASO (Orbach et al., 2021) is an open-domain TSA dataset containing more than 2,000 English user comments extracted from YELP [2], AMAZON (Keung et al., 2020), SST (Socher et al., 2013), and OPINOSIS (Ganesan et al., 2010), covering a variety of topics in multiple domains. COVIDSenti (Naseem et al., 2021) includes 90,000 COVID-19-related tweets annotated with overall sentiment polarity.

Despite the existing open-domain and in-domain datasets, NER and TSA on clinical social media texts remain an under-explored area. There is a pressing need for such datasets to facilitate social media-based public health studies. To fill in this gap, we release METS-CoV, a COVID-19 tweets-based NER and TSA dataset with 3 general entity labels (*Person*, *Location*, *Organization*,) 4 medical entity labels (*Disease*, *Drug*, *Symptom*, and *Vaccine*) as well as sentiment labels for *Person*, *Organization*, *Drug*, and *Vaccine* entities.

3 METS-CoV

In this section, we provide a detailed description of the collection methodology, annotation process, and statistics of the NER and TSA subsets of METS-CoV.

[2]https://www.yelp.com/dataset

3.1 Data Collection Methodology

We collect COVID-19 related tweets ranging from February 1, 2020, to September 30, 2021, whose unique Tweet Identifier (Tweet ID) came from an open-source database (Chen et al., 2020a). All the tweets are downloaded following Twitter's automation rules and data security policy. For data filtering, we first remove non-English tweets and retweets, resulting in 368,816,761 tweets. Then we filter out the tweets containing URLs because they are often restatements of third-party messages and do not directly reflect the users' intentions and attitudes. Finally, we use a list of symptoms (including symptoms of COVID-19 as well as common diseases) as keywords to match the tweets to extract medical-related tweets (Wang et al., 2020a; Goss et al., 2018; Sarker et al., 2020; Lopez-Leon et al., 2021; Mao et al., 2020). 2,208,676 tweets remain after the pre-processing step.

3.2 Data Annotation Process

We define 7 entity types based on public health research needs (Tsao et al., 2021; Xu et al., 2022), including 3 general entity types and 4 medical entity types for annotation. In particular, we select 4 entity types for additional sentiment annotation with 3 types of sentiment labels: positive, negative, and neutral. All the annotation work is done using the YEDDA annotation platform by Yang et al. (2018b). We first randomly sample 6,000 tweets from the pre-processed tweets for NER annotation. Then we use these 6,000 annotated NER data to train a BERT-based NER tagger and annotate the rest of the tweets. In order to include more medical entities in the dataset, we select additional 4,000 tweets from the model labeled data (with higher drug and vaccine entity ratios) and manually validate the entities to extend the dataset to a total number of 10,000 tweets.

Here we describe detailed annotation guidelines and processes for the METS-CoV-NER dataset and the METS-CoV-TSA dataset in detail. Note that all our annotators are from medical domains, including medicine, public health and pharmaceutical sciences.

The annotation process of METS-CoV-NER. The annotation includes 3 phases:

1. In the pre-annotation phase, all annotators are requested to conduct 3 rounds of annotation (with training). F1 value is used as the metric of inter-annotator agreement. All the annotators are assigned the same corpus and required to annotate following the guidelines. After annotation, the project leader compares all the labels and determined the final gold labels, which are used to calculate the inter-annotator agreement. Annotators with F1 greater than 80% are selected to enter the formal annotation process. The annotation guidelines are also iteratively updated throughout this process.
2. In the formal annotation phase, annotators label the tweets in pairs (3 pairs in total) to ensure each tweet is annotated twice. When an inconsistency occurs, another annotator steps in to determine the final annotation of the tweet.
3. After the formal annotation phase, the project team conducts a quality control check on the labeled results to ensure that the annotated tweets meet the annotation guidelines' requirements. The final inter-annotator agreement is 85.0% in the F1 value.

The annotation process of METS-CoV-TSA. Similar to the NER dataset's annotation process, METS-CoV-TSA's annotation has 3 phases.

1. First, we conduct 6 rounds of sentiment pre-labeling and update the labeling guidelines in each iteration. We use accuracy as the metric of inter-annotator agreement. The procedure to determine the gold labels is the same as for METS-CoV-NER.The annotators who meet the consistency criteria are selected to participate in the subsequent annotation process.
2. We randomly pair up the annotators (4 pairs in total) and assign the same tweets to each group. A third annotator determines the final annotation when inconsistency occurs.
3. The project team conducts a secondary check to ensure that the annotation meets the guidelines. The final inter-annotator agreement is 78.4% in accuracy.

Our annotation guidelines are customized to understand medical-related entities and sentiments better. NER guidelines include rules such as "when the manufacturer name refers to a vaccine in the tweet context, the name should be annotated as a vaccine rather than an organization." Sentiment guidelines include rules such as "when irony appears in an annotation entity, its sentiment should be annotated from the standpoint of the person who posted the tweet." More details can be found in our guidelines (see supplementary data).

3.3 Dataset Statistics

Fig.2 shows the distribution of tweet lengths in METS-CoV. Most tweets have lengths shorter than 80 (split with white spaces). Among them, The largest proportion of tweets have a length of around 50. Table 1 shows the statistics of the METS-CoV-NER dataset. 10,000 tweets contain 19,057 entities in total, resulting in 1.91 entities per tweet in this dataset. We observe that *Symptom* entities have the highest frequency. This is an expected result of pre-filtering tweets using symptom-related keywords[3]. Other than *symptoms*, all other 6 entity types have relatively balanced proportions. Table 2 presents the statistical information of the METS-CoV-TSA dataset, where we observe that neutral sentiment accounts for the highest proportion across all 4 target entities. For the *drug* entities, users have significantly more positive sentiments than negative sentiments, whereas for the *vaccine* entities, users have similar positive and negative sentiments.

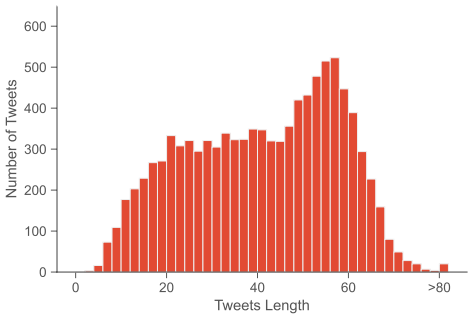

Figure 2: The distribution of tweets length of METS-CoV.

Table 1: Statistics of METS-CoV-NER dataset.

Number	Train	Dev	Test
Tweets	7,000	1,500	1,500
Tokens	285k	59k	60k
All Entities	13,570	2,749	2,738
Person	2,253	472	487
Location	1,371	294	279
Organization	1,934	396	381
Disease	1,555	301	258
Drug	1,077	262	227
Symptom	4,223	806	869
Vaccine	1,157	218	237

Table 2: Statistics of METS-CoV-TSA dataset.

Number		Train	Dev	Test
Person	POS	260	64	58
	NEU	1293	256	240
	NEG	700	152	189
Organization	POS	126	24	31
	NEU	1346	284	251
	NEG	462	88	99
Drug	POS	234	85	64
	NEU	730	147	142
	NEG	113	30	21
Vaccine	POS	112	25	20
	NEU	913	173	183
	NEG	132	20	34

4 Model Benchmarking

In this section, we evaluate the performance of statistical machine learning models, neural networks, general domain large-scale pre-trained language models (PLM), and COVID-19-related PLM for the NER task and the TSA task on METS-CoV, respectively. In addition, we select the best model from each group for in-depth analysis and discussion.

4.1 Named Entity Recognition

Models. The NER models that we benchmark on the NER dataset can be divided into 4 branches: a traditional statistical machine learning model: Conditional Random Field (CRF) (Lafferty et al., 2001); 6 neural network models from a combination of (1) a BiLSTM for word-level feature extraction (denoted as WLSTM in Table 3) and (2) CRF or non-CRF as the inference layer or (3) without character information or encoding character feature with structure CNN, LSTM (denoted as CCNN and CLSTM respectively) (Yang et al., 2017; Lample et al., 2016; Huang et al., 2015; Ma and Hovy, 2016); 3 general domain PLM including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020); two COVID-19-related PLM: BERTweet-covid19 (Nguyen et al., 2020) and COVID-TWITTER-BERT (Müller et al., 2020). BERTweet is RoBERTa-base further trained on

[3]Some matched symptom keywords are not real symptom entities in tweets due to the ambiguity of words, that is why the total number of Symptom entities is less than 10,000.

Table 3: Model performance on METS-CoV-NER dataset. (* means uncased model)

Results (F1 value ± std)	Person	Location	Organization	Disease	Drug	Symptom	Vaccine	Overall
CRF	64.43±1.59	76.37±0.62	54.64±2.08	73.61±0.44	77.34±1.60	74.05±0.56	84.85±0.82	71.58±0.54
WLSTM	72.05±0.79	79.82±0.61	60.79±0.77	73.52±1.26	79.63±1.36	76.72±0.83	86.03±0.84	75.02±0.36
WLSTM + CCNN	80.63±0.62	81.47±0.89	61.30±0.91	74.52±0.75	80.46±0.28	76.63±0.91	85.91±1.17	76.78±0.29
WLSTM + CLSTM	81.16±0.29	81.37±0.48	62.28±1.41	74.80±1.49	79.50±1.01	76.70±0.25	85.59±1.46	76.91±0.22
WLSTM + CRF	72.41±0.44	79.25±0.59	62.39±0.97	74.89±1.28	79.60±0.71	**79.14**±0.51	88.72±0.62	76.38±0.22
WLSTM + CCNN + CRF	**81.38**±0.44	**82.15**±0.44	62.79±0.91	**76.12**±0.76	80.41±0.58	78.12±0.51	**89.11**±0.36	**78.10**±0.19
WLSTM + CLSTM + CRF	77.49±1.67	81.26±1.19	**63.21**±0.93	75.61±0.76	**81.27**±0.65	**79.14**±0.53	87.85±0.43	77.63±0.40
BERT-base*	86.99±1.27	84.47±0.84	71.01±0.52	76.18±1.49	84.78±1.02	80.26±0.43	89.83±0.56	81.49±0.36
BERT-base	86.71±0.73	84.09±1.67	71.90±0.91	**76.93**±1.11	84.54±1.05	79.70±1.22	89.48±0.99	81.34±0.41
BERT-large*	87.47±1.22	84.43±1.05	71.21±0.59	76.34±0.96	85.53±1.52	**81.44**±0.18	89.33±1.25	81.98±0.30
BERT-large	**88.25**±0.52	84.63±1.38	73.30±1.38	76.52±1.11	86.05±1.06	80.12±0.55	89.16±1.58	82.05±0.24
RoBERTa-base	85.58±0.73	85.46±0.86	72.21±1.11	76.49±1.63	85.38±1.15	79.81±0.67	89.89±0.28	81.43±0.20
RoBERTa-large	86.79±0.44	**85.85**±2.12	**73.78**±0.72	76.84±0.57	**86.79**±0.78	81.32±0.67	**90.42**±1.12	**82.55**±0.27
BART-base	84.24±0.59	82.85±0.71	70.60±1.46	75.01±1.80	83.39±1.03	79.03±0.48	90.22±1.58	80.17±0.46
BART-large	81.60±4.93	80.04±4.74	64.66±8.86	71.24±1.90	80.61±2.90	74.27±4.45	81.21±6.20	75.56±5.04
BERTweet-covid19-base*	**91.63**±0.79	85.79±0.75	**77.07**±0.51	77.09±1.61	83.57±0.65	81.16±0.45	**91.16**±1.54	83.63±0.36
BERTweet-covid19-base	91.50±0.81	**86.26**±0.97	76.47±0.46	**77.80**±0.57	84.16±1.22	80.89±0.52	89.98±1.04	83.49±0.18
COVID-TWITTER-BERT*	91.29±0.42	85.68±0.92	76.27±0.64	77.48±0.81	**86.35**±0.96	**81.85**±0.53	90.44±0.94	**83.88**±0.20

Table 4: Span F1 of NER models.

Model (F1 value ± std)	Person	Location	Organization	Disease	Drug	Symptom	Vaccine	Overall
CRF	64.43±1.59	76.37±0.62	54.64±2.08	73.61±0.44	77.34±1.60	74.05±0.56	84.85±0.82	74.52±0.77
WLSTM + CCNN + CRF	81.38±0.44	82.15±0.44	62.79±0.91	76.12±0.76	80.41±0.58	78.12±0.51	89.11±0.36	82.15±0.33
RoBERTa-large	86.79±0.44	**85.85**±2.12	73.78±0.72	76.84±0.57	**86.79**±0.78	81.32±0.67	**90.42**±1.12	85.91±0.43
COVID-TWITTER-BERT	**90.51**±0.67	85.37±0.30	**76.31**±0.55	**77.14**±1.77	86.64±0.65	**81.36**±0.38	89.71±2.35	**86.70**±0.45

850 million general English tweets and 23 million COVID-19-related tweets. COVID-TWITTER-BERT is BERT-large further trained on 97 million COVID-19-related tweets. All the experiments of NER models are conducted using NCRF++ (Yang and Zhang, 2018).

Training and Test Sets. To compare the model performance, we perform a train-dev-test splitting of our dataset with a ratio of 70:15:15. Statistics of the train-dev-test splits are presented in Table 1. Hyperparameters of the models are default settings from Yang et al. (2018a).

Results and Discussion. Table 3 shows the performance of NER models on the NER test set evaluated by micro-F1. We list both the mean values and standard deviations. We can observe that COVID-TWITTER-BERT achieves the best performance with an overall micro-F1 value of 83.88, outperforming both general domain PLM and classical NER models based on CRF or BiLSTM (and their variants). Specifically, for the 3 general entity types (*Person*, *Organization* and *Location*), the language models pre-trained on COVID-19-related tweets (COVID-TWITTER-BERT) outperform the best general PLM (RoBERTa-large). The absolute F1 improvements are 3.38, 0.41, and 3.29 for *Person*, *Organization*, and *Location*, respectively. For the 4 medical entity types, COVID-TWITTER-BERT outperforms RoBERTa-large except for the Drug entity type, for which a slightly worse performance is observed.

To compare various models in detail, we select the best models from each branch, i.e., CRF, WLSTM + CCNN + CRF, RoBERTa-large, and COVID-TWITTER-BERT, to represent the state-of-the-art statistical machine learning model, traditional neural network model, general domain PLM, and COVID-19 related PLM, respectively. Following Liu et al. (2021), we adopt the Span F1 and the Type Accuracy (Type Acc) as the two metrics to evaluate the 4 models. Span F1 indicates the correctness of entity spans in NER, while Type Acc refers to the proportion of the predicted entities that have both correct spans and types out of predicted entities with the correct spans.

As shown in Table 4 and Table 5, we observe that COVID-TWITTER-BERT achieves the best overall performance in both metrics, followed by RoBERTa-large. Specifically, COVID-TWITTER-BERT achieves the best performance on *Person* and *Organization* entities compared to the RoBERTa-large, with 3.72% and 2.53% improvement in terms of Span F1, respectively. For Type Acc, COVID-TWITTER-BERT performs the best on four entity types including *Person, Location, Organization* and *Drug*. These results demonstrate the effectiveness of incremental pre-training of language models on COVID-19-related tweets. In addition, we also conduct an experiment to investigate the effect of tweet lengths on the model performance. Experimental results are presented in Fig 3. We observe that all models perform better when the tweet length is short (less than 40 tokens). Their performance

Table 5: Type Acc of NER models.

Model (Acc ± std)	Person	Location	Organization	Disease	Drug	Symptom	Vaccine	Overall
CRF	97.25±0.85	93.75±0.66	93.38±0.57	**92.61**±2.07	98.00±0.30	97.13±0.14	97.63±0.53	96.06±0.32
WLSTM + CCNN + CRF	95.88±0.61	93.89±0.15	88.08±1.32	90.12±1.47	97.64±0.90	97.42±0.22	97.97±0.41	95.06±0.20
RoBERTa-large	97.38±0.66	95.72±0.89	93.87±1.03	87.46±0.44	98.75±0.86	**97.58**±0.34	**99.27**±0.42	96.18±0.21
COVID-TWITTER-BERT	**97.66**±0.34	**96.86**±0.83	**94.34**±0.77	89.38±0.77	**99.42**±0.21	97.20±0.59	97.65±0.57	**96.39**±0.16

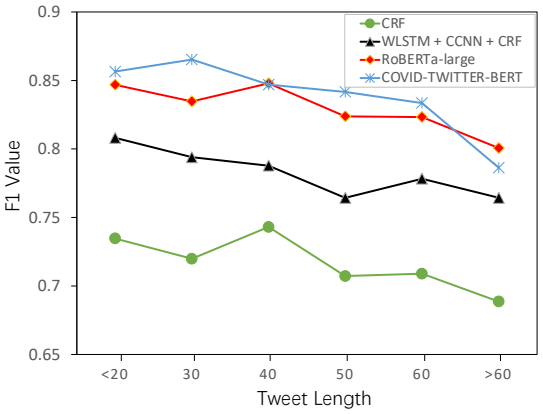

Figure 3: F1 values of different NER models against the tweet length.

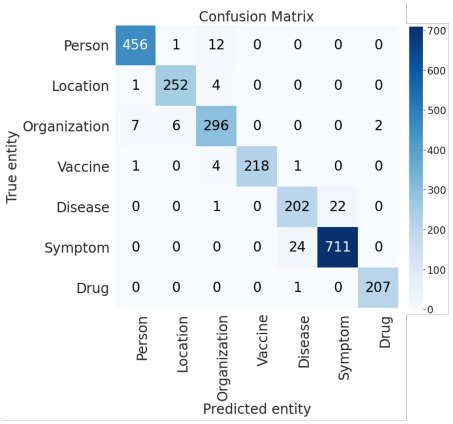

Figure 4: The confusion matrix of COVID-TWITTER-BERT on NER test set.

gradually decreases dealing with longer tweets. We also visualize the confusion matrix of the best performing model, COVID-TWITTER-BERT, on the test set in Fig 4. We find that in most cases, COVID-TWITTER-BERT extracts the entities correctly. However, it tends to get confused when labeling *Symptom* and *Disease* entities. This confusion is expected because diseases and symptoms oftentimes have similar expressions and contexts. Besides, we also notice that all models performed worse on *Organization* entities. This problem has also been encountered and discussed in previous Twitter NER dataset papers such as (Jiang et al., 2022).

Through the above experiments and analysis, we conclude that COVID-TWITTER-BERT can be regarded as a strong baseline on the NER dataset of METS-CoV, and there is still vast room for improvement. Specifically, the F1 value of *Disease* and *Organization* entities is suboptimal. More effective NER models are needed for this challenging dataset.

4.2 Targeted Sentiment Analysis

Models. The TSA models that we benchmark on METS-CoV-TSA could be classified into 4 categories: 1 statistical machine learning model: SVM (Vo and Zhang, 2015); 7 traditional neural network models: ASGCN (Zhang et al., 2019), LSTM (Hochreiter and Schmidhuber, 1997), TD-LSTM (Tang et al., 2016b), MemNet (Tang et al., 2016a), IAN (Ma et al., 2017), MGAN (Fan et al., 2018) and TNet-LF (Li et al., 2018); 6 general domain PLM (BERT-base-uncased): AEN (Song et al., 2019), LCF (Zeng et al., 2019), BERT-SPC (Devlin et al., 2019), depGCN (Zhang et al., 2019), kumaGCN (Chen et al., 2020b) and dotGCN (Chen et al., 2022); and 4 models (BERT-SPC, depGCN , kumaGCN and dotGCN) with COVID-19 related PLM (COVID-TWITTER-BERT).

Train and Test Sets. The TSA dataset's train-dev-test splitting (with a ratio of 70:15:15) corresponds to that of the NER dataset mentioned above. In this setting, the TSA training dataset is the subset of the NER training dataset, where only the tweets containing targeted entities are kept. A similar procedure is also adopted to obtain the dev and test sets of TSA. The details of TSA split statistics can be found in Table 2. Still, we set the hyperparameters of the models according to the original papers.

Results and Discussion. The experiment results of TSA are listed in Table 6. We can observe that models incorporating COVID-TWITTER-BERT as feature extractor significantly outperform other types of models. Specifically, depGCN (COVID-TWITTER-BERT) achieves the best performance on

[4] https://github.com/duytinvo/ijcai2015

Table 6: Model performance on METS-CoV-TSA dataset.
(† means our implementation. The standard deviation of SVM model is not reported because the prediction of LibSVM[4] is not affected by random seeds when the dataset splitting is fixed.)

Model (mean ± std)	Person		Organization		Drug		Vaccine		Overall	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SVM (Vo and Zhang, 2015)	50.72	36.99	64.57	42.02	58.15	30.17	70.89	46.09	59.53	38.73
LSTM (Hochreiter and Schmidhuber, 1997)	58.56±1.79	50.41±2.48	61.00±0.95	45.64±0.47	56.39±2.18	41.53±1.92	65.99±2.41	40.03±3.62	60.21±1.53	49.08±1.58
TD-LSTM (Tang et al., 2016b)	59.26±0.98	49.54±1.81	63.90±2.19	41.57±2.76	59.91±1.72	41.04±2.22	73.08±0.77	38.14±2.73	63.16±0.65	48.26±1.09
MemNet (Tang et al., 2016a)	**59.79**±1.57	43.97±3.30	64.79±2.48	37.98±1.92	59.21±1.86	40.24±1.19	74.43±1.21	36.98±2.79	**63.73**±0.85	45.04±1.65
IAN (Ma et al., 2017)	52.81±1.72	32.75±2.65	**67.68**±0.45	36.70±1.94	59.73±0.45	25.22±0.10	**77.22**±0.00	30.88±0.00	62.59±0.55	34.62±1.77
MGAN (Fan et al., 2018)	57.17±2.00	43.84±4.70	63.84±2.68	40.09±1.20	58.33±2.22	35.32±5.01	72.49±0.98	37.05±5.10	62.00±1.10	42.55±3.17
TNet-LF (Li et al., 2018)	58.07±1.19	**51.17**±2.10	63.16±1.52	**47.68**±1.59	**60.00**±2.00	**46.30**±2.51	68.52±3.15	**41.57**±2.17	61.71±1.01	**50.80**±1.22
ASGCN(Zhang et al., 2019)	58.89±0.63	42.48±2.61	63.89±0.84	39.73±2.62	58.41±1.90	31.12±3.91	72.66±2.02	38.24±2.95	62.69±0.23	41.32±2.22
AEN (Song et al., 2019)	56.84±2.54	47.91±4.12	60.63±4.22	45.68±3.58	52.51±3.12	37.08±4.31	69.12±4.59	41.46±3.31	59.37±2.43	46.28±3.73
LCF (Zeng et al., 2019)	60.29±2.10	52.43±2.31	68.58±1.06	50.15±4.17	58.42±3.16	44.01±1.64	71.14±2.73	41.75±1.90	64.27±1.61	51.29±2.06
BERT-SPC (Devlin et al., 2019) †	64.39±0.91	60.06±1.01	73.28±0.93	58.48±1.50	62.38±1.33	49.11±2.90	77.22±0.96	**50.28**±5.09	68.87±0.34	59.31±1.18
depGCN (Zhang et al., 2019)†	**67.39**±1.16	62.35±2.08	**74.02**±0.71	**58.62**±1.06	**63.61**±1.07	49.32±0.74	77.13±1.54	47.50±2.73	**70.38**±0.40	59.96±0.77
kumaGCN (Chen et al., 2020b)†	66.28±1.33	61.84±2.46	72.86±0.59	58.01±2.04	**63.61**±0.77	49.67±3.14	76.88±0.73	50.17±4.56	69.59±0.73	59.91±2.48
dotGCN (Chen et al., 2022)†	67.06±1.63	**62.56**±1.54	73.33±0.68	58.34±1.52	63.35±2.13	**50.20**±1.98	**77.55**±0.77	45.98±2.90	70.09±0.75	**60.05**±1.41
BERT-SPC(COVID-TWITTER-BERT)†	73.72±1.47	70.25±2.10	78.53±0.61	66.24±1.59	75.07±1.06	62.67±3.10	**79.15**±1.02	**61.68**±3.70	76.29±0.57	70.03±0.92
depGCN (COVID-TWITTER-BERT) †	**75.85**±1.22	**72.70**±1.78	**79.42**±1.25	**66.94**±2.66	**76.92**±1.49	**67.35**±2.45	77.89±2.14	59.85±4.45	**77.42**±0.77	**71.39**±1.65
kumaGCN (COVID-TWITTER-BERT) †	74.37±1.39	71.46±1.43	78.48±1.46	64.56±1.61	76.30±2.38	62.96±7.41	78.73±1.50	58.87±2.63	76.65±1.20	70.20±1.95
dotGCN (COVID-TWITTER-BERT) †	74.95±1.60	72.53±1.54	79.11±0.89	65.21±2.06	74.10±1.90	61.26±2.14	78.65±1.72	59.41±2.92	76.65±0.49	70.32±0.96

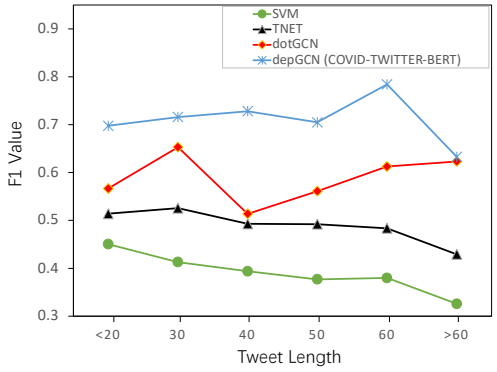

Figure 5: F1 values of different TSA models against the tweet length.

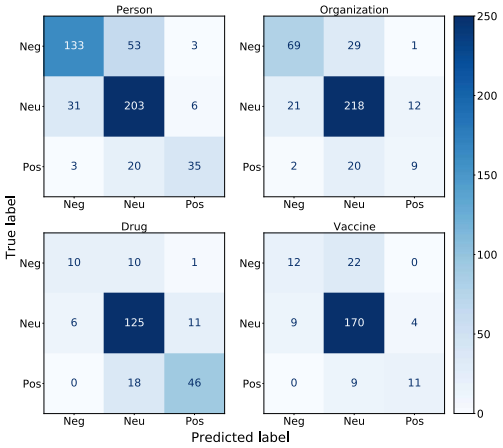

Figure 6: The confusion matrix of depGCN (COVID-TWITTER-BERT) on TSA test set.

the *Person* entities compared to the BERT-based depGCN, with 8.46% and 10.35% improvement in terms of ACC and F1, respectively. For *Organization* entities, depGCN (COVID-TWITTER-BERT) still performs the best, with at least 5.4% and 8.32% improvement on ACC and F1, respectively. For *Drug* entities, depGCN (COVID-TWITTER-BERT) outperforms other models by a large margin and achieves 13.31% and 18.03% absolute improvements in terms of ACC and F1 compared with BERT-based depGCN. For *Vaccine* entities, BERT-SPC (COVID-TWITTER-BERT) outperforms all other models, with at least 1.6% and 11.4% improvement in terms of ACC and F1.

To investigate the impacts of tweet lengths on different TSA models, we select the best models from each branch, i.e., SVM, MemNet, depGCN (BERT-base), and depGCN (COVID-TWITTER-BERT). As shown in Fig 5, we observe that the impacts of tweet length varies for different TSA models. For SVM and TNET, the F1 value gradually decrease when the length of tweets gets longer. For dotGCN, the F1 value fluctuates sharply when the tweet length is between 20 and 40. Afterwards, it increases moderately as the tweet length grows from 40. For depGCN (COVID-TWITTER-BERT), the F1 value remains stable when the tweet length is less than 50 and then increases to 0.8 and finally decreases to about 0.6.

We choose depGCN (COVID-TWITTER-BERT) for an in-depth analysis since it achieves the best overall performance. Fig 6 shows the confusion matrix of applying depGCN (COVID-TWITTER-BERT) to the TSA test set. The results show that, for all the targeted entities, most confusion is caused by the misclassification between positive (negative) and neutral.

To summarize, we find that the pre-trained models on COVID-19 tweets, such as COVID-TWITTER-BERT, can be utilized to further improve the performance of previous TSA models on the TSA dataset of METS-CoV. Moreover, the in-depth study shows that the performance of the current best TSA model remains insufficient, and more robust TSA models are needed to distinguish sentiment polarities sufficiently.

5 Ethics

General Ethical Conduct. METS-CoV is the first dataset to include medical entities and targeted sentiments on COVID-19-related tweets. These human-derived data are openly displayed on the Twitter platform and are allowed to be used for research purposes following Twitter's Developer Agreement and Policy. Following the guidelines, we release only the Tweet IDs but not the original content. Meanwhile, the download script we provide for promoting replicable research could be configured not to store user information. Readers can request desensitized tweets if the tweets can no longer be accessed. All tweets used in this study were retrieved in an entirely random manner. No new bias should have been introduced except for potentially associated issues with label imbalance. We expect future studies to investigate such issues and provide solutions. We also encourage users to conduct sanity checks to avoid potential bias for specific sub-populations regarding social economics status, cultural groups, etc. Note that we did not filter offensive content in our study because directly removing such words can alter user sentiments expressed in the original tweet. However, we encourage the users to consider doing so in tailored ways for developing fair models.

Potential Negative Societal Impacts. Readers should not use the dataset to assess the reputation of public figures or accounts because it was collected and annotated only to develop models for NER and TSA rather than direct causal referential analyses. It is also vital to note that TSA models trained on METS-CoV-TSA should not be used to analyze public attitudes toward people or organizations for non-medical purposes, given that METS-CoV derives from COVID-19-related tweets. Even using such models in a non-COVID-19 setting should be evaluated because public attitudes towards specific topics may change in different settings, and the model may have learned information associated explicitly with COVID-19. Moreover, sentiment analysis for vaccines and drugs can only reveal the user's overall viewpoint and attitude, not their willingness to take drugs or get vaccinated. In addition, readers should be aware that the TSA dataset only reflects user attitudes in the provided contexts and does not explain underlying reasons. Although our work does not directly result in negative social impacts, it is necessary to take appropriate precautions to avoid such impacts.

6 Limitations

We acknowledge the following limitations: First, METS-CoV has imbalanced entity distribution in that the medical-related tweets are matched using a symptom lexicon to reduce the sparsity of medical entities. To mitigate this problem, we have provided performance evaluation for each entity in the NER benchmark analysis to help readers understand how the models perform on different entities. Second, we do not filter tweets in MES-CoV-TSA because we want to show the actual distribution of sentiments. This again results in label imbalance, reflected by the high proportion of neutral labels. Readers should be aware of this imbalance when using the dataset. Third, the TSA annotations unavoidably contain subjectivity. To relieve this problem, we have made strict guidelines, conducted multi-rounds of pre-annotation training, and had annotators work in pairs with third-party validation.

7 Conclusion & Future Work

In this work, we introduce METS-CoV, the first dataset to include medical entities and targeted sentiments on COVID-19-related tweets. Based on this dataset, we evaluate the performance of both classical and state-of-the-art models for NER and TSA tasks. Results show that existing models can not fully exploit the potential of METS-CoV. The METS-CoV dataset is built from a medical research perspective. It fully considers the characteristics of the medical field and can therefore be used to help researchers use natural language processing models to mine valuable medical information from tweets. Much COVID-19 research could leverage this dataset. For example, investigating the public attitudes toward COVID-19 vaccines and drugs, tracking the public's mental status change during different COVID-19 phases, etc. Besides the data, we hope our released annotation guidelines,

benchmark models, and source code could facilitate and encourage the curation of more datasets and novel models for medical social medial research.

Acknowledgments and Disclosure of Funding

We thank Minghui Li for making the tweet length distribution figure. We appreciate the annotation work of Shixu Lin, Minghui Li, Wanxin Li, Yujie Zhang, Junjie Wang, Subatijiang, and Bingtao Guan. This research received no grant from any funding agency.

References

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia* 2, 3 (2021), 315–324.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle (Eds.). The Association for Computer Linguistics. https://doi.org/10.3115/1225403.1225421

Sakun Boon-Itt, Yukolpat Skunkan, et al. 2020. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance* 6, 4 (2020), e21978.

Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. LocVTP: Video-Text Pre-training for Temporal Localization. *arXiv preprint arXiv:2207.10362* (2022).

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete Opinion Tree Induction for Aspect-based Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2051–2064. https://doi.org/10.18653/v1/2022.acl-long.145

Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020b. Inducing target-specific latent structures for aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5596–5607.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020a. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* 6, 2 (29 May 2020), e19273. https://doi.org/10.2196/19273

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Natasha Duarte, Emma Llansó, and Anna C. Loup. 2018. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 106. http://proceedings.mlr.press/v81/duarte18a.html

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained Attention Network for Aspect-Level Sentiment Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3433–3442. https://doi.org/10.18653/v1/D18-1380

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Grapased Approach to Abstractive Summarization of Highly Redundant Opinions. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, Chu-Ren Huang and Dan Jurafsky (Eds.). Tsinghua University Press, 340–348. https://aclanthology.org/C10-1039/

Foster R. Goss, Kenneth H. Lai, Maxim Topaz, Warren W. Acker, Leigh Kowalski, Joseph M. Plasek, Kimberly G. Blumenthal, Diane L. Seger, Sarah P. Slight, Kin Wah Fung, Frank Y. Chang, David W. Bates, and Li Zhou. 2018. A value set for documenting adverse reactions in electronic health records. *J. Am. Medical Informatics Assoc.* 25, 6 (2018), 661–669. https://doi.org/10.1093/jamia/ocx139

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

Yining Hua, Hang Jiang, Shixu Lin, Jie Yang, Joseph M Plasek, David W Bates, and Li Zhou. 2022. Using Twitter data to understand public perceptions of approved versus off-label use for COVID-19-related medications. *Journal of the American Medical Informatics Association* (07 2022). https://doi.org/10.1093/jamia/ocac114 arXiv:https://academic.oup.com/jamia/advance-article-pdf/doi/10.1093/jamia/ocac114/45019593/ocac114.pdf ocac114.

C. Huang, L. Huang, Y. Wang, X. Li, L. Ren, X. Gu, L. Kang, L. Guo, M. Liu, X. Zhou, J. Luo, Z. Huang, S. Tu, Y. Zhao, L. Chen, D. Xu, Y. Li, C. Li, L. Peng, Y. Li, W. Xie, D. Cui, L. Shang, G. Fan, J. Xu, G. Wang, Y. Wang, J. Zhong, C. Wang, J. Wang, D. Zhang, and B. Cao. 2021. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* 397, 10270 (01 2021), 220–232.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). arXiv:1508.01991 http://arxiv.org/abs/1508.01991

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-Domain NER using Cross-Domain Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2464–2474. https://doi.org/10.18653/v1/P19-1236

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 7199–7208. https://aclanthology.org/2022.lrec-1.780

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4563–4568. https://doi.org/10.18653/v1/2020.emnlp-main.369

M. Khan, S. F. Adil, H. Z. Alkhathlan, M. N. Tahir, S. Saif, M. Khan, and S. T. Khan. 2020. COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far. *Molecules* 26, 1 (Dec 2020).

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, Barcelona, Spain (Online), 27–36. https://aclanthology.org/2020.smm4h-1.4

Archana Kumari, Piyush Ranjan, Naval K Vikram, Divjyot Kaur, Anamika Sahu, Sada Nand Dwivedi, Upendra Baitha, and Aastha Goel. 2020. A short questionnaire to assess changes in lifestyle-related behaviour during COVID 19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 6 (2020), 1697–1701.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 260–270. https://doi.org/10.18653/v1/N16-1030

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation Networks for Target-Oriented Sentiment Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 946–956. https://doi.org/10.18653/v1/P18-1087

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5847–5858. https://doi.org/10.18653/v1/2021.acl-long.454

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

Sandra Lopez-Leon, Talia Wegman-Ostrosky, Carol Perelman, Rosalinda Sepulveda, Paulina A Rebolledo, Angelica Cuapio, and Sonia Villapol. 2021. More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Scientific reports* 11, 1 (2021), 1–12.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 4068–4074. https://doi.org/10.24963/ijcai.2017/568

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1064–1074. https://doi.org/10.18653/v1/P16-1101

Ling Mao, Huijuan Jin, Mengdie Wang, Yu Hu, Shengcai Chen, Quanwei He, Jiang Chang, Candong Hong, Yifan Zhou, David Wang, et al. 2020. Neurologic manifestations of hospitalized patients with coronavirus disease 2019 in Wuhan, China. *JAMA neurology* 77, 6 (2020), 683–690.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1643–1654. https://aclanthology.org/D13-1171

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *CoRR* abs/2005.07503 (2020). arXiv:2005.07503 https://arxiv.org/abs/2005.07503

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W Eklund, and Jinman Kim. 2021. COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 1003–1015.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 9–14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. YASO: A Targeted Sentiment Analysis Evaluation Dataset for Open-Domain Reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9154–9173. https://doi.org/10.18653/v1/2021.emnlp-main.721

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, 19–30. https://doi.org/10.18653/v1/S16-1002

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, 486–495. https://doi.org/10.18653/v1/S15-2082

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, 27–35. https://doi.org/10.3115/v1/S14-2004

Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *J. Am. Medical Informatics Assoc.* 27, 8 (2020), 1310–1315. https://doi.org/10.1093/jamia/ocaa116

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves Le Traon. 2019. A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS 2019, Granada, Spain, October 22-25, 2019*, Mohammad A. Alsmirat and Yaser Jararweh (Eds.). IEEE, 338–343. https://doi.org/10.1109/SNAMS.2019.8931850

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media* 22 (2021), 100104.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 1631–1642. https://aclanthology.org/D13-1170/

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional Encoder Network for Targeted Sentiment Classification. *ArXiv* abs/1902.09314 (2019).

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan, 138–144. https://aclanthology.org/W16-3919

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016b. Effective LSTMs for Target-Dependent Sentiment Classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 3298–3307. https://aclanthology.org/C16-1311

Duyu Tang, Bing Qin, and Ting Liu. 2016a. Aspect Level Sentiment Classification with Deep Memory Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 214–224. https://doi.org/10.18653/v1/D16-1021

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://aclanthology.org/W03-0419

Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A Butt. 2021. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health* 3, 3 (2021), e175–e194.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18, 5 (2011), 552–556.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence*.

Liqin Wang, Suzanne V. Blackley, Kimberly G. Blumenthal, Sharmitha Yerneni, Foster R. Goss, Ying-Chih Lo, Sonam N. Shah, Carlos A. Ortega, Zfania Tom Korach, Diane L. Seger, and Li Zhou. 2020a. A dynamic reaction picklist for improving allergy reaction documentation in the electronic health record. *J. Am. Medical Informatics Assoc.* 27, 6 (2020), 917–923. https://doi.org/10.1093/jamia/ocaa042

Xuan Wang, Xiangchen Song, Bangzheng Li, Yingjun Guan, and Jiawei Han. 2020b. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218* (2020).

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. Association for Computational Linguistics, Florence, Italy, 21–30. https://doi.org/10.18653/v1/W19-3203

Hua Xu, David L Buckeridge, Fei Wang, and Peter Tarczy-Hornoch. 2022. Novel Informatics Approaches to COVID-19 Research: from methods to applications. *Journal of Biomedical Informatics* (2022).

Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, Tingshao Zhu, et al. 2020. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of medical Internet research* 22, 11 (2020), e20550.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018a. Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3879–3889. https://aclanthology.org/C18-1327

Jie Yang and Yue Zhang. 2018. NCRF++: An Open-source Neural Sequence Labeling Toolkit. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, 74–79. https://doi.org/10.18653/v1/P18-4013

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018b. YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, 31–36. https://doi.org/10.18653/v1/P18-4006

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=ByxpMd9lx

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences* 9, 16 (2019), 3389.

Martin Zens, Arne Brammertz, Juliane Herpich, Norbert Südkamp, Martin Hinterseer, et al. 2020. App-based tracking of self-reported COVID-19 symptoms: analysis of questionnaire data. *Journal of medical Internet research* 22, 9 (2020), e21956.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. 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)*. Association for Computational Linguistics, Hong Kong, China, 4568–4578. https://doi.org/10.18653/v1/D19-1464

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the datasets, benchmarks and code are available at `https://github.com/YLab-Open/METS-CoV`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Those details were listed in Section 4. For the hyperparameter selecting, we used the default hyperparameters of the benchmark models.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we reported the results based on experiments on 5 different random seeds. Mean ± std were reported in this paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We didn't include the consumption of resources as we are releasing a new dataset rather than proposing new architecture.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used code from several models in our benchmarks, all the sources were properly cited in this paper.
   (b) Did you mention the license of the assets? [No] The code we used are all open available, they were used to evaluate model performance in our new dataset. We do not claim any copyright from the code.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 5. This work was conducted on public available data, so this study is waived from the participant's consent. We follow the privacy policy of Twitter platform when sharing this dataset.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 5
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] This work was conducted on public available data, it doesn't have participants.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] This dataset is based on public available tweet text, it doesn't have potential participant risks.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] This dataset was voluntarily annotated by the authors and members of Prof. Jie Yang's group.