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

Peilin Zhou, Zeqiang Wang, Dading Chong, Zhijiang Guo, Yining Hua, Zichang Su, Zhiyang Teng, Jiageng Wu, Jie Yang

Zhejiang University, Peking University, University of Cambridge, Harvard Medical School, and Westlake University

Introduction

Motivation:

- There is a large body of social media-based public health studies, especially during the pandemic when clinical and survey studies are difficult to conduct.
- Existing natural language processing (NLP) tools struggle to fulfill the surging demand for accurate **social media-based public healthcare analysis** due to the lack of relevant datasets.
- Named entity recognition (NER) and targeted sentiment analysis (TSA) are two important tasks for studying user foci and attitudes (Figure 1).

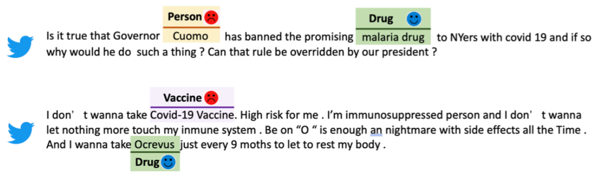

Figure 1: Examples of medical entities and targeted sentiments in tweets.

Contribution:

- We released **METS-CoV** (Medical Entities and Targeted Sentiments on CoVid-19-related tweets), a dataset annotated with:
  - ***10,000 tweets***
  - ***7 types of entities***: *Disease, Drug, Symptom, Vaccine, Person, Location,* and *Organization.*
  - ***sentiments of 4 types of entities***: *Person, Organization, Drug,* and *Vaccine*.
- We designed detailed **guidelines** for annotating medical entities (*Disease, Drug, Symptom, Vaccine*) on tweets.
- We **benchmarked** the performance of classical machine learning models and state-of-the-art deep learning models including pre-training language models on NER and TSA tasks of METS-CoV.

The METS-CoV Dataset

Data Collection:

We collected COVID-19-related tweets ranging from February 1, 2020 to September 30, 2021. We removed non-English tweets, retweets, and tweets with URLs. We used a list of symptom keywords to match medical-related tweets. **2,208,676** tweets remained in our final dataset.

GitHub

Data Annotation:

We defined 7 entity types based on public health research needs, including 3 general types and 4 medical types. Then we selected 4 entity types for sentiment annotation using 3 sentiment labels: *positive*, *negative*, and *neutral*.

Paper

Data Statistics:

Figure 2 , Table1 and 2. Read the paper for more:

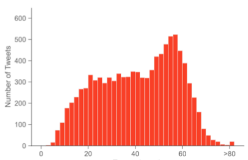

Figure 2: The distribution of tweets length o METS-CoV.

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

Model Benchmarking

We evaluated the performance of (a) statistical machine learning models, (b) neural networks, (c) general domain large-scale pre-trained language models (PLM), and (d) COVID-19-related PLM for the NER task and the TSA task on METS-CoV. In addition, we selected the best model from each group for in-depth analysis and discussion.

Named Entity Recognition (NER)

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

	Results (F1 value ± std)	Person	Location	Organization	Disease	Drug	Symptom	Vaccine	Overall
(a)	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
(b)	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
(c)	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
(d)	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

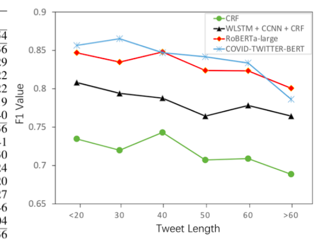

Figure 3: F1 values of different NER models against the tweet length.

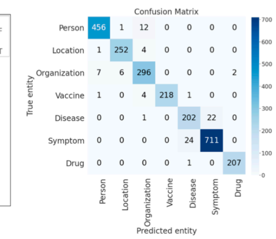

Figure 4: The confusion matrix of COVID-TWITTER-BERT on NER test set.

Observations:

- **Overall performance** (Table 3): COVID-TWITTER-BERT is a strong baseline on the NER subset, but F1 values of *Disease* and *Organization* entities are suboptimal.
- **Effect of tweet lengths** (Figure 3): All models' performance decreases as tweet lengths increase. The performance is generally better when tweet lengths ≤ 40 tokens.
- **In-depth study** (Figure 4): COVID-TWITTER-BERT extracts entities correctly in most cases. But it tends to confuse *Symptom* and *Disease*.

Targeted Sentiment Analysis (TSA)

Table 4: 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
(a)	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
(b)	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
(c)	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
(d)	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

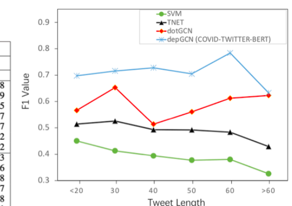

Figure 5: F1 values of different TSA models against the tweet length.

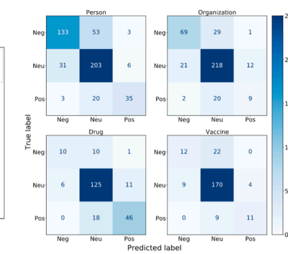

Figure 6: The confusion matrix of depGCN (COVID-TWITTER-BERT) on TSA test set.

Observations:

- **Overall performance** (Table 4): Pre-trained models on COVID-19 tweets, such as COVID-TWITTERBERT, has better performance on TSA subset.
- **Impact of tweet lengths** (Figure 5): For SVM and TNET, their F1 values gradually decrease as tweet lengths increase. But for depGCN (COVID-TWITTER-BERT), its F1 value remains stable when the tweet length ≤ 50. The performance increases to 0.8 when the tweet length = 60, and decreases to about 0.6 when the length further increases to > 60.
- **In-depth study** (Figure 6): the current best TSA model has moderate performance. More robust models are needed to accurately distinguish sentiment polarities.

Conclusions & Future Work

Public health researchers can use METS-CoV to mine valuable medical information from tweets. For example, The dataset can be used as a training dataset for examining public attitudes toward COVID-19 vaccines and drugs, tracking the public's mental status change during different COVID-19 phases, etc. Our experiments also show that current models have not fully exploited the dataset's potential. We call for more efforts on developing models for social media-based public health studies.

For ethics discussions, code, data, and guidelines, please refer to our paper and GitHub page.

YLab