A Biomedically oriented automatically annotated Twitter COVID-19 Dataset

PRESENTER: **Luis Alberto Robles Hernandez**

INTRODUCTION:

Twitter has been used extensively during the COVID-19 outbreak [8], providing insight into everything from monitoring communication between public health officials and world leaders [9], tracking emergency symptoms [10] and access to testing facilities [11], to understanding the public's top fears and concerns about infection rates and vaccination [12].

Researchers face a myriad of challenges when trying to utilize Twitter data. Firstly, it can be difficult to obtain access to these data and hard to keep it up with real-time content collection [14,15]. Once the data has been collected, researchers must then perform several preprocessing steps to ensure data are sufficient for analysis. Concerning COVID-19, there are several social media repositories [16-20]. Unfortunately, most of these repositories are not frequently updated, do not provide any preprocessing or data cleaning, and either do not provide the raw data or lack appropriate metadata or provenance. The COVID-19 Twitter Chatter dataset [20] is a robust large-scale repository of tweets that is well-maintained and frequently updated, and recent work utilizing this resource has shown great promise for tracking long-term patient-reported symptoms [21] as well as highlighted mentions of drugs relevant to the treatment of COVID-19 [22].

This paper presents preliminary work achieved during the 2021 Biomedical Linked Annotation Hackathon (BLAH 7) [23], which aimed to **enhance and extend the COVID-19 Twitter Chatter dataset [20] to include biomedical entities, and we hope to improve the downstream clinical utility of these data and provide researchers with a means to clinically characterize personally-reported COVID-19 phenomena.** We envision this work as the first step towards our larger goal of deriving mechanistic insights from specific types of entities within COVID-19 tweets by integrating these data with larger and more complex sources of biomedical knowledge, like PheKnowLator [24] and the KG-COVID-19 [25] knowledge graphs.

METHODS:

1. Data were collected from one of the largest COVID-19 Twitter chatter datasets available [20], which contains around 900 million unique tweets, in which around 226 million unique tweets were annotated (excluding all retweets).
2. To prepare the previous dataset, named entity recognition (NER) pipelines were used to identify biomedical entities in text.
3. The NER pipelines used were MedSpaCy [26], MeDaCy [27], and SciSpaCy [28], alongside a traditional text annotation pipeline from Social Media Mining Toolkit [29].

Including biomedical entities in large-scale Twitter datasets will provide researchers with a means to clinically characterize personally-reported COVID-19 experiences

Take a picture to download the full paper

CONCLUSION:

In this work we release a biomedically oriented automatically annotated dataset of COVID-19 chatter tweets. We demonstrate that while there are SpaCy-based systems for NER on clinical and scientific documents, they do not generalize well when used on non-clinical sources of data like tweets.

The resulting dataset and biomedical annotations **is the first and largest of its kind**, making it a substantial contribution with respect to using large-scale Twitter data for biomedical research. As for future work, the release of this dataset **will facilitate continued development of fine-tuned resources for mining social media data for biomedical and clinical applications.**

RESULTS:

As shown in ***Table 1***, we evaluated all NER systems against a manually annotated gold-standard dataset grouped into three categories: drugs, conditions/symptoms, and measurements. From all the evaluated NER systems, SciSpaCy performed fairly decent when capturing relevant annotations from a tweet's text.

	Drugs	Conditions/ Symptoms	Measurements
SMMT Tagger	**69.31%**	**71.91%**	**39.83%**
MedSpaCy	19.98%	13.49%	7.45%
MedaCy	47.04%	27.14%	12.56%
SciSpaCy	59.71%	44.65%	26.98%

Table 1. Annotation overlap analysis between gold standard dataset and evaluated NER systems.

We can see from ***Table 2*** that regular text annotation (SMMT) performed the best in replicating the annotations that our clinicians made. As mentioned before, around 226 million tweets were annotated, as well as evaluated the overlap by the different NER systems. The table also shows the comparison between counts of produced annotations, processing time, and overlaps in annotation between the systems.

	SMMT Tagger	MedSpaCy	MedaCy	SciSpaCy
Annotations Produced	751,245,366	582,768,145	656,311,799	**775,615,621**
Processing Time (minutes)	**24,120**	159,267	26,147	325,620
Overlaps with SMMT	**100%**	53.48%	51.14%	**89.17%**
Overlaps with MedSpaCy	20.12%	**100%**	44.92%	34.77%
Overlaps with MedaCy	33.91%	42.23%	**100%**	44.17%
Overlaps with SciSpaCy	**72.28%**	55.39%	49.73%	**100%**

Table 2. Annotation overlap evaluation for complete dataset.

REFERENCES:

For the full references, visit the full paper at https://arxiv.org/abs/2107.12565

Juan M. Banda, Luis Alberto Robles Hernandez
Georgia State University, Atlanta, Georgia, USA
Tiffany J. Callahan
University of Colorado AMC, Aurora, Colorado, USA