A Biomedically oriented automatically annotated Twitter COVID-19 Dataset

Luis Alberto Robles Hernandez, Tiffany J. Callahan, Juan M. Banda

Background

The use of social media data for biomedical research has gradually increased over the years. With the COVID-19 pandemic, researchers have turned to more non-traditional sources of clinical data to characterize the disease in near-real time (1), spread of misinformation about facemasks (2), study health-related societal implications (3,4), characterization of what unproven therapies are people discussing (5,6), as well as the sequelae that recovered COVID-19 cases present (7). However, manually curated social media datasets are difficult to come by due to the expensive costs of manual annotation and the efforts needed to identify the correct texts. When datasets are available, they are usually very small and their annotations do not generalize well over time or to larger sets of documents. As part of the 2021 virtual Biomedical Linked Annotation Hackathon, we created and released (8), a dataset of over 120 million automatically annotated tweets for biomedical research purposes. Incorporating best-practices, we identified tweets with potentially clinical relevance, allowing biomedical researchers unfamiliar with processing and annotating social media data to directly focus on relevant tweets. To bridge the gap between using social media data for health research with OHDSI researchers, we have annotated this dataset using concept identifiers from the OHDSI Vocabulary, producing a total of 751,245,366 annotations.

Methods

To prepare the dataset released in this work, we looked for named entity recognition (NER) pipelines to identify biomedical entities in text. We opted to evaluate: MedSpaCy (9), MedaCy (10), and ScispaCy (11), alongside a traditional text tagging pipeline from Social Media Mining Toolkit (SMMT)(12). The main reason for selecting these text processing pipelines is the fact that they are all based on SpaCy (13), a widely adopted open-source library for Natural Language Processing (NLP) in Python. Several preprocessing steps like URL and emoji removal were performed on all tweets. Please note that the selected NER pipelines are usually tuned and developed to annotate specific types of clinical/scientific text, from either electronic health records, clinical notes, or scientific literature. The only general-purpose tagger is the Social Media Mining Toolkit, which does not perform any specialized tasks other than tagging or annotating text.

As the data source for this work, we used one of the largest COVID-19 Twitter chatter datasets available (14). We used version 44 of the dataset (14), which contains 903,223,501 unique tweets. To improve the quality and relevance of the annotations, we used the clean version of this dataset, which has all retweets removed. Leaving us with a total of 226,582,903 unique tweets to annotate. From this subset, we selected only English tweets, as all the systems evaluated were created to extract/annotate biomedical concepts in this language.

For transparency and reproducibility, all the code used for this work can be at: https://github.com/thepanacealab/annotated_twitter_covid19_dataset

Results

For the evaluation of the annotations from each NER system and the SMMT tagger, we will use as a gold standard, a manually annotated dataset created for symptoms, conditions, prescriptions, and measurement procedures identification in patients with long COVID phenotypes [21]. This dataset consists of 10,315 manually annotated tweets, by multiple clinicians. Currently, the dataset is not publicly available but will be released at a later date. To determine which system to use for the large-scale annotation of the Twitter COVID-19 chatter dataset, we evaluated all systems against the manually annotated gold-standard. Here, while we grouped the annotations into three categories: drugs, conditions/symptoms, and measurements. We did not use the systems' annotation categories, but rather their annotated terms and spans. This was done to accommodate the custom entity categories that systems like MedSpaCy and MedaCy have in their default settings and the fact that we are using only the first UMLS concepts identified by ScispaCy. Table 1 shows the annotation overlap analysis.

	Drugs	Conditions / symptoms	Measurements	Average
SMMT Tagger	*69.31%*	*71.91%*	*39.83%*	60.35%
MedSpaCy	19.98%	13.49%	7.45%	13.64%
MedaCy	47.04%	27.14%	12.56%	28.91%
ScispaCy	59.71%	44.65%	26.98%	43.78%

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

We would like to stress again that MedSpaCy and MedaCy are at a disadvantage as their models are trained on considerably different data that does not work well with Twitter data. ScispaCy, however, performs fairly decently (in comparison) as the larger models capture relevant annotations when the tweet's text is clean and well-formed. While it is clear that regular text annotation performed the best in replicating the annotations that our clinicians made, we still annotated all 226,582,903 dataset tweets and evaluated the overlap of annotations made by the different systems. Table 2 shows the comparison between counts of produced annotations, processing time, and overlaps in annotations between the systems.

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

Table 2. Annotation overlap evaluation for complete dataset.

Figure 1. presents an example of a tweet that has been annotated with OMOP concept identifiers (magenta highlighting). Considering just this part of the figure first, you can see that aside from creating a link between the words in the tweet and concepts in OMOP vocabularies, this work also includes rich context about what these words mean. For example, the chain of “hospital” + “positive” + “COVID-19” . In addition to the OMOP concept identifiers, we have also included NER person (red highlighting), date (orange highlighting), organization (blue highlighting), and location (grey highlighting) tags. While these tags are not provided as part of the current release, we include them here as an example of how this work can easily be extended using current NLP tools (like those examined in this work).

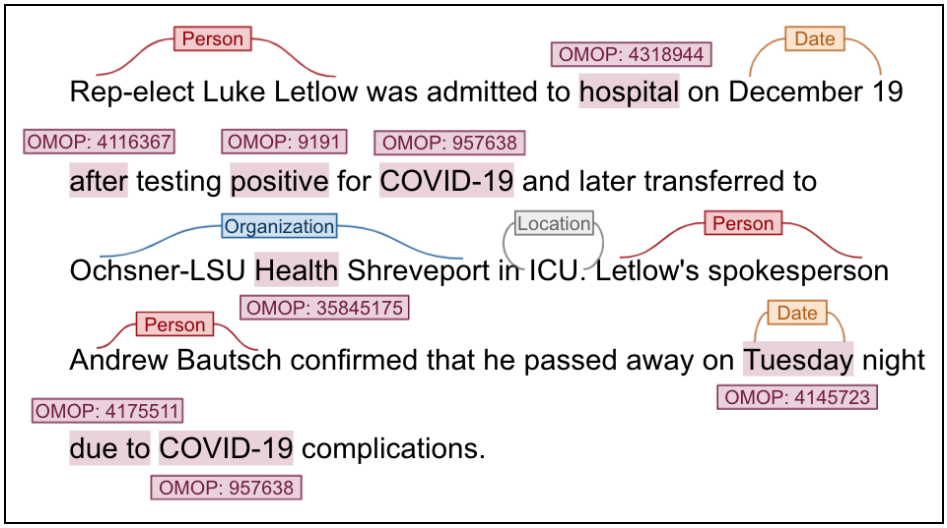

Figure 1. Illustration of a tweet annotated with OMOP concept identifiers (magenta) and enriched with NER person, date, location, and organization tags.

Conclusion

In this work we release a biomedically oriented automatically annotated dataset of COVID-19 chatter tweets. We believe that such a dataset will bridge some of the gaps between the researchers of the OHDSI community into using social media data for large-scale epidemiological, social, or even clinical analyses. 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, nicely standardized using the OHDSI vocabulary.

References/Citations

1. Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health [Internet]. 2021 Mar;3(3):e175–94. Available from: http://dx.doi.org/10.1016/S2589-7500(20)30315-0

2. Ayers JW, Chu B, Zhu Z, Leas EC, Smith DM, Dredze M, et al. Spread of Misinformation About Face Masks and COVID-19 by Automated Software on Facebook. JAMA Intern Med [Internet]. 2021 Jun 7; Available from: http://dx.doi.org/10.1001/jamainternmed.2021.2498

3. Ayers JW, Leas EC, Johnson DC, Poliak A, Althouse BM, Dredze M, et al. Internet Searches for Acute

Anxiety During the Early Stages of the COVID-19 Pandemic. JAMA Intern Med [Internet]. 2020 Dec 1;180(12):1706–7. Available from: http://dx.doi.org/10.1001/jamainternmed.2020.3305

4. Jamison AM, Broniatowski DA, Dredze M, Sangraula A, Smith MC, Quinn SC. Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 “infodemic” on Twitter. HKS Misinfo Review [Internet]. 2020 Sep 9; Available from: https://misinforeview.hks.harvard.edu/?p=2462

5. Liu M, Caputi TL, Dredze M, Kesselheim AS, Ayers JW. Internet Searches for Unproven COVID-19 Therapies in the United States. JAMA Intern Med [Internet]. 2020 Aug 1;180(8):1116–8. Available from: http://dx.doi.org/10.1001/jamainternmed.2020.1764

6. Tekumalla R, Banda JM. Characterizing drug mentions in COVID-19 Twitter Chatter. 2020 Dec; Available from: https://www.aclweb.org/anthology/2020.nlpcovid19-2.25/

7. Banda JM, Singh GV SR, Alser OH, Prieto-Alhambra D. Long-term patient-reported symptoms of COVID-19: an analysis of social media data [Internet]. bioRxiv. medRxiv; 2020. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.07.29.20164418

8. Robles Hernandez LA, Callahan TJ, Banda JM. A Biomedically oriented automatically annotated Twitter COVID-19 Dataset [Internet]. 2021. Available from: https://zenodo.org/record/4606734

9. medspacy [Internet]. Github; [cited 2021 Mar 9]. Available from: https://github.com/medspacy/medspacy

10. Mulyar A, Mahendran D, Maffey L, Olex A, Matteo G. TAC SRIE 2018: Extracting systematic review information with medacy. In: National Institute of Standards and Technology (NIST) 2018 Systematic Review Information Extraction (SRIE) - Text Analysis Conference [Internet]. researchgate.net; 2018. Available from: https://www.researchgate.net/profile/Darshini_Mahendran/publication/340870892_TAC_SRIE_2018_Extracting_Systematic_Review_Information_with_MedaCy/links/5ea1add5a6fdcc88fc381e4c/TAC-SRIE-2018-Extracting-Systematic-Review-Information-with-MedaCy.pdf

11. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task [Internet]. Florence, Italy: Association for Computational Linguistics; 2019. p. 319–27. Available from: https://www.aclweb.org/anthology/W19-5034

12. Tekumalla R, Banda JM. Social Media Mining Toolkit (SMMT). Genomics Inform [Internet]. 2020 Jun;18(2):e16. Available from: http://dx.doi.org/10.5808/GI.2020.18.2.e16

13. Explosion AI. spaCy-Industrial-strength Natural Language Processing in Python. URL: https://spacy io [Internet]. 2017; Available from: https://spacy.io/

14. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration [Internet]. 2020. Available from: https://zenodo.org/record/4263444