Leveraging the OHDSI vocabulary to characterize the COVID-19 epidemic using Twitter data and NLP

Ramya Tekumalla, MS.[1], Juan M. Banda, PhD[1]
[1]Georgia State University, Atlanta, Georgia, USA

Abstract

The ongoing COVID-19 pandemic has created a data deluge unlike any other public health emergency in history. One of the growing sources of data is social media, specifically Twitter. On this platform people have been sharing everything from conspiracy theories, to personal health information related to symptomatic or asymptomatic COVID-19 infections, off-label drug intake, and current or persisting COVID-19 symptoms. Trying to harness all this publicly available data, we have, since March 2020, gathered a dataset of over 460 million COVID-19 tweets. While all this data is an excellent resource into social discourse of the pandemic, the need to efficiently and effectively characterize it is vital to generate knowledge and actionable insights. In this work we showcase how by leveraging the OHDSI vocabulary and social media mining tools, we are able to quickly uncover interesting trends and mentions of clinical concepts and drug mentions, using the vocabulary's internal structure of domains, concept classes and individual vocabulary identifiers. We present a short proof of concept analysis by tracking the mentions over time of widely discussed potential drug treatments for COVID-19.

Research Category: Observational data management

Introduction

The Coronavirus Disease 2019 (COVID-19) has spread all over the world since the beginning of 2020, being characterized as a global pandemic on March 11, 2020[1]. As of July 13, 2020, a total of 12,945,505 confirmed cases and 571,444 deaths were recorded worldwide. It is vital to use any available data resources to try to characterize this pandemic and try to generate insights that are useful for public health officials and scientists. Social media data sources, especially Twitter and Reddit, have recently been proven to be useful resources for the characterization of several disasters and outbreaks in the past [2–5]. While Twitter data contains many structured data fields, the most interesting and insightful ones are unstructured, making it a non-trivial task to standardize any findings and generate any insights from them. When trying to look at this data from the health informatics context, the need for data standardization is vital. Per design, the ODSHI[6] vocabulary is a collection of external vocabularies and ontologies like SNOMED-CT, MeSh, ICD, RxNorm among hundreds of others. Leveraging this vocabulary alongside social media mining tools, will allow us to extract and standardize[7] the unstructured data more effectively, and be able to analyze it more efficiently. While this application does not directly involve using the OMOP CDM, we demonstrate that by leveraging one of the community's most important tools, the vocabulary, we have an advantage for the characterization of the Twitter data over using standard non-OHDSI tools, as we will describe in the following section.

Methods

For this work, we used version 17 of the largest available COVID-19 Twitter dataset which consists of **468,169,539** tweets[8]. Since retweets usually amplify the signals and add bias, we used the clean version of the dataset which consists of **115,262,201** tweets with no retweets. A terms dictionary was created from the OHDSI vocabulary by selecting the uniquely distinct terms (by concept_name) with the following adjustments: a) Since Twitter has a limit of 280 characters per tweet, we removed any term string longer than 100 characters, b) all the terms less than 3 characters are also removed due to their ambiguous nature, c) stop words were removed, and d) all the terms were lower cased. The final dictionary consists of **2,938,998** unique terms. The tweet data pre-processing and automatic annotation was performed by using the Social Media Mining Toolkit (SMMT)[9], and Spacy[10]. When collapsing the vocabulary by unique terms, we lose the domain, concept class, and vocabulary identifiers of repeated strings, however, this is recovered after annotation by joining the annotations back with the original vocabulary. The annotation process resulted in a total of **1,147,782,412** terms tagged by our annotation tool. Table 1 represents the top five number of unique terms found for respective domain, concept class and vocabulary identifiers. Seemingly trivial, it is quite powerful to be able to have term groupings as it will allow researchers to focus on any given particular task they have without needing to create lists of specific terms per domain/class/vocabulary before filtering. The variety of vocabularies in the OHDSI vocabulary gives enough flexibility as well for plenty of health informatics tasks.

Domain ID	Distinct	Concept Class ID	Distinct	Vocabulary ID	Distinct
Drug	24,045	Clinical Finding	9,144	SNOMED	31,444
Condition	18,373	Brand Name	6,893	MedDRA	8,456
Observation	17,593	Substance	5,231	RxNorm Extension	5,469
Procedure	4,013	LLT	4,790	dm+d	4,598
Geography	2,415	Ingredient	3,870	RxNorm	4,102

Table 1. Number of top five unique concepts captured by domain, concept class, and vocabulary identifiers.

Characterizing potential drug treatments for COVID-19: a proof of concept

One of the most discussed topics on Twitter has been the discourse around drug treatment. With the produced text annotations, we are able to elucidate time trends when each of the most discussed drugs (Table 2) have gained traction based on news reports, politician mentions, and scientists on Twitter (Figure 1). Notice the individual trends on the drug discourse over time has changed with relevant events, showing first the wide discussion of hydroxychloroquine, touted by many as the first potentially game-changing drug around the beginning of April. This trend reached new heights when the first negative results were published from scientific studies near mid-May. Similar patterns can be seen for remdesivir and dexamethasone later on during the pandemic study period.

Table 2. Drug ingredient mentions found

Drug Ingredient	Frequency
hydroxychloroquine	204,879
remdesivir	72,841
chloroquine	49,915
oxygen	37,961
vitamin D	25,445
dexamethasone	25,142
zinc	24,843
azithromycin	16,079
ibuprofen	8,469
ivermectin	6,390

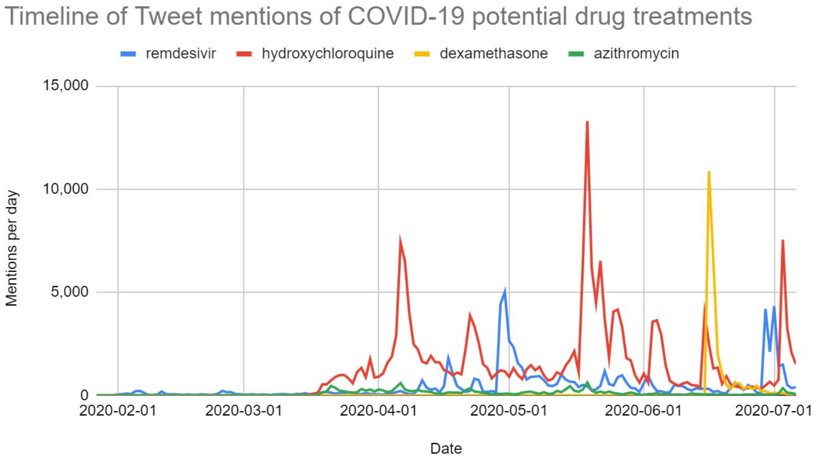

Figure 1. Timeline of Tweets with potential drug treatment mentions.

Conclusion

In this work we have shown that leveraging the OHDSI vocabulary for NLP tasks on social media data adds great value when looking into health informatics topics. We also show that people share a considerable amount of health data on Twitter, making it a valuable resource for health research. Note that we are just characterizing drug mentions and no attributions to actual intake are made, this is out of the scope of this work and currently under active research by our group.

References

1. World Health Organization. WHO characterizes COVID-19 as a pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.
2. Khatua, A., Khatua, A. & Cambria, E. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Inf. Process. Manag.* **56**, 247–257 (2019).
3. Alam, F., Ofli, F., Imran, M. & Aupetit, M. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *arXiv [cs.SI]* (2018).
4. Earle, P. Earthquake Twitter. *Nat. Geosci.* **3**, 221–222 (2010).
5. Zou, L., Lam, N. S. N., Cai, H. & Qiang, Y. Mining Twitter Data for Improved Understanding of Disaster Resilience. *Ann. Assoc. Am. Geogr.* **108**, 1422–1441 (2018).
6. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
7. Banda, J. M. Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data. *Genomics & Informatics* vol. 17 e13 (2019).
8. Banda, J. M. *et al.* A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration* (2020) doi:10.5281/zenodo.3941294 .
9. Tekumalla, R. & Banda, J. M. Social Media Mining Toolkit (SMMT). *Genomics Inform.* **18**, e16 (2020).
10. Explosion, A. I. spaCy-Industrial-strength Natural Language Processing in Python. *URL: https://spacy. io* (2017).