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

7/21/2020

Juan M. Banda
www.panacealab.org
jbanda@gsu.edu
@drjmbanda
Preface: Twitter is gaining attention for health-related research since 2009

![Graph showing the number of publications per year from 2005 to 2020, with a significant increase starting around 2009.](image-url)

Results of PubMed Query for Twitter and Health
My path to working on COVID-19 research

• Ph.D in CS – Data Mining (Image data)

• Postdoc and Research Scientist at Stanford University – Shah Lab on Biomedical Informatics (Medical text data – structured and unstructured)

• OHDSI collaborator on probabilistic phenotyping using weak supervision and NLP Workgroup member standardizing resources to OHDSI vocabulary

• Assistant Professor at Georgia State University – PI of Panacea lab
  • Working with epidemiologists on identifying mobility patterns during natural disasters using social media data
  • Extracting drug use from social media data
Benefits of using Twitter:

1) Good population representation
2) Everybody can post and have an account
3) Anonymity = unfiltered opinions
4) Data is freely available*
5) Tons of data generated each day (hundreds of millions of tweets get posted every day)
6) Easy filtering (hashtag usage, people mentions)

* Data is freely available when not protected by copyright or other legal restrictions.

Traditional disadvantages of using Twitter:

- Messy data (plenty of misspellings, shorthand, emojis, etc.)
  - There are at least 25 different ways people misspell hydroxychloroquine

- Attribution is an issue – are people just mentioning something or did it happen to them?

- Freely available data is only a 1% sample of whole set

- Collection is hard and needs to be ongoing for days/weeks before getting considerable mass

- Very unique challenges
  - Short form text (up to 280 chars in Twitter)
  - More colloquial, ambiguous, and expressive in different ways (✍️️)
Collecting the data

• Thanks to our work with Dr. Chowell (GSU – Public Health) we started collecting COVID-19 data early!

• We also received contributed data for January and February from (new) collaborators – after we first shared our dataset publicly

........ wait, the data was released before any publications/work has been completed on it???
Short answer......Yes!

But it needs to be done right!
and for the benefit of everybody

...... within all legal and ethical responsibilities of the data being shared
The dataset:

- 513+ Million Tweets
- ONLY COVID related chatter is included

Longitudinal – January 27th to today... and growing

Dataset: https://doi.org/10.5281/zenodo.3723939
Recent additions: https://github.com/thepanacealab/covid19_twitter
Languages available: 69 in total
Tweets per day (all):

http://www.panacealab.org/covid19/
Tweets per day (clean):

http://www.panacealab.org/covid19/
Geolocations? ... on some less than 1%
Tweets with place location enabled? ...

... some

less than 2%
For instant NLP uses

- We include:
  - Top 1000 frequent terms
  - Top 1000 frequent bigrams
  - Top 1000 frequent trigrams
Is the dataset being used?

22K downloads since we started

Over 15 pre-prints cite us..... 3 datasets ‘aggregated’ us without attribution or even asking

Has been used on multiple hackathons: MIT-COVID19 challenge, Lumiata COVID hackathon, COVID-19 biohackathon

We have been invited to participate on:
- Bay Area Summer Institutes in Computational Social Science
- Harvard The Coronavirus Visualization Team – Xenophobia Project

Started new collaborations with over six academic institutions
So what can we do with this type of data?
Perception towards the ethnic minorities and elderly populations

• Being a fellow at Stanford SAGE Research Center one of our first questions was to identify the perception of the Twitter users to ethnic minorities and elderly populations

• Turns out we can!

• Turns out is not as expected!
Perception towards the elderly and ethnic minorities (1)

• What we did:
  • Curated terms to identify tweets for elderly populations and for ethnic minorities
  • Identify Tweets with mentions. Build machine learning models to disambiguate ambiguous terms after manual curation.
  • Sentiment Analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner)
  • Manual Evaluation of identified Tweets for correctness of polarity
  • Comparison of Baseline tweets (2019) with epidemic related tweets (March)
Perception towards the elderly and ethnic minorities (2)

• For ethnic minorities:
Perception towards the elderly and ethnic minorities (3)

- The calculated proportion of negative versus positive tweets has increased to 28% more negative sentiment on racial minorities. This leads to an almost 10% increase in negative sentiments in the time periods we analyze.
Hashtag Evolution over time
Emoji distribution by gender
Topic modeling?

• Quite tricky with Twitter data!

• Too many topics, not enough clarity on our initial analyses with LDA and Dynamic LDA
Wait..... Isn’t this the OHDSI call?

• Yes! And the previous and the following work was 100 times easier by leveraging the OHDSI vocabulary!

.... How?
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.
Annotation process

• The tweet data pre-processing and automatic annotation was performed by using the Social Media Mining Toolkit (SMMT)*, and Spacy.

• NOTE: 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.

What we got:

- We found a total of 1,147,782,412* terms on 115M tweets

Is this useful this way?

<table>
<thead>
<tr>
<th>concept_name</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronavirus</td>
<td>25,766,403</td>
</tr>
<tr>
<td>People</td>
<td>5,304,705</td>
</tr>
<tr>
<td>China</td>
<td>2,407,265</td>
</tr>
<tr>
<td>Virus</td>
<td>2,323,879</td>
</tr>
<tr>
<td>Time</td>
<td>2,078,697</td>
</tr>
<tr>
<td>Home</td>
<td>1,532,694</td>
</tr>
<tr>
<td>Is a</td>
<td>1,521,450</td>
</tr>
<tr>
<td>death</td>
<td>1,467,443</td>
</tr>
<tr>
<td>Due to</td>
<td>1,379,014</td>
</tr>
<tr>
<td>Spread</td>
<td>1,335,751</td>
</tr>
<tr>
<td>Today</td>
<td>1,318,637</td>
</tr>
<tr>
<td>Crisis</td>
<td>1,301,976</td>
</tr>
<tr>
<td>State</td>
<td>1,299,045</td>
</tr>
<tr>
<td>Outbreak</td>
<td>1,225,292</td>
</tr>
<tr>
<td>Country</td>
<td>1,192,392</td>
</tr>
<tr>
<td>Support</td>
<td>1,013,937</td>
</tr>
<tr>
<td>India</td>
<td>897,104</td>
</tr>
<tr>
<td>Vaccine</td>
<td>846,643</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>concept_name</th>
<th>frequency</th>
<th>concept_class_id</th>
<th>vocabulary_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronavirus</td>
<td>25,766,403</td>
<td>LOINC Component</td>
<td>LOINC</td>
</tr>
<tr>
<td>People</td>
<td>5,304,705</td>
<td>Social Context</td>
<td>SNOMED</td>
</tr>
<tr>
<td>China</td>
<td>2,407,265</td>
<td>2nd level</td>
<td>OSM</td>
</tr>
<tr>
<td>Virus</td>
<td>2,323,879</td>
<td>LOINC Component</td>
<td>LOINC</td>
</tr>
<tr>
<td>Time</td>
<td>2,078,697</td>
<td>LOINC Component</td>
<td>LOINC</td>
</tr>
<tr>
<td>Home</td>
<td>1,532,694</td>
<td>Visit</td>
<td>CMS Place of Service</td>
</tr>
<tr>
<td>Is a</td>
<td>1,521,450</td>
<td>Relationship</td>
<td>Relationship</td>
</tr>
<tr>
<td>death</td>
<td>1,467,443</td>
<td>Table</td>
<td>CDM</td>
</tr>
<tr>
<td>Due to</td>
<td>1,379,014</td>
<td>Attribute</td>
<td>SNOMED</td>
</tr>
<tr>
<td>Spread</td>
<td>1,335,751</td>
<td>Attribute</td>
<td>SNOMED</td>
</tr>
<tr>
<td>Today</td>
<td>1,318,637</td>
<td>Brand Name</td>
<td>RxNorm</td>
</tr>
<tr>
<td>Crisis</td>
<td>1,301,976</td>
<td>Clinical Finding</td>
<td>SNOMED</td>
</tr>
<tr>
<td>State</td>
<td>1,299,045</td>
<td>Concept Class</td>
<td>Concept Class</td>
</tr>
<tr>
<td>Outbreak</td>
<td>1,225,292</td>
<td>Context-dependent</td>
<td>SNOMED</td>
</tr>
<tr>
<td>Country</td>
<td>1,192,392</td>
<td>LOINC Component</td>
<td>LOINC</td>
</tr>
<tr>
<td>Support</td>
<td>1,013,937</td>
<td>LOINC Component</td>
<td>LOINC</td>
</tr>
<tr>
<td>India</td>
<td>897,104</td>
<td>Answer</td>
<td>LOINC</td>
</tr>
</tbody>
</table>
What the vocabulary gets us:

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

*Table 1. Number of top five unique concepts captured by domain, concept class, and vocabulary identifiers.*
Cool story… show me some RWE
Can we characterize drug mentions?

Table 2. Drug ingredient mentions found

<table>
<thead>
<tr>
<th>Drug Ingredient</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydroxychloroquine</td>
<td>204,879</td>
</tr>
<tr>
<td>remdesivir</td>
<td>72,841</td>
</tr>
<tr>
<td>chloroquine</td>
<td>49,915</td>
</tr>
<tr>
<td>oxygen</td>
<td>37,961</td>
</tr>
<tr>
<td>vitamin D</td>
<td>25,445</td>
</tr>
<tr>
<td>dexamethasone</td>
<td>25,142</td>
</tr>
<tr>
<td>zinc</td>
<td>24,843</td>
</tr>
<tr>
<td>azithromycin</td>
<td>16,079</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>8,469</td>
</tr>
<tr>
<td>ivermectin</td>
<td>6,390</td>
</tr>
</tbody>
</table>

Figure 1. Timeline of Tweets with potential drug treatment mentions.
Charybdis style?

Charybdis-like characterization over countries (work with Dani Prieto-Alhambra – University of Oxford)
Charybdis style?

Charybdis-like characterization over countries (work with Dani Prieto-Alhambra – University of Oxford)
Charybdis style?

... over time?
Charybdis style?

... over time?

Very different pattern for retweets
What about symptom/condition characterization?

- Self-reported symptoms on Twitter vs EHR lists *
  - Can we find related symptoms both found on EHR’s (Callahan, A., Steinberg, E., Fries, J.A. et al. Estimating the efficacy of symptom-based screening for COVID-19. npj Digit. Med. 3, 95 (2020). [https://doi.org/10.1038/s41746-020-0300-0](https://doi.org/10.1038/s41746-020-0300-0)) but on Twitter?

* [https://github.com/thepanacealab/covid19_biohackathon/tree/master/user_symptoms](https://github.com/thepanacealab/covid19_biohackathon/tree/master/user_symptoms)
Is this real RWE?..... Not quite.... Yet!

How do we get there?
Data is very messy!

- Misspellings are the norm – ignore them: lose 15% of data!

Attribution is critical

• What about attributions?
  • Others have done work on this:
  • However, needs tons and tons of manual review

• Can this be done easier and maybe semi-supervised (like APHRODITE)?
  • Yes! We have done this for drugs!
• Mined 9 Billion tweets from public domain

• Using heuristic we found ~6M tweets with drugs

• Trained models on subsets of them

• Used models to predict already existing labeled sets (~93% accuracy on them)

• Subset of 3 million tweets gave us these results! No manual review on our side
So we have many pieces now:
On going super exciting work
Tracking Self-Reported symptoms after infection recovery

• Since we can find symptoms and drugs, we can also find people that had COVID and their symptoms after infection!

• On-going work with Dani Prieto-Alhambra and others
  • Incorporates methods shown before + manual review by clinicians

Some very preliminary findings:

fatigue = 789
shortness of breath = dyspnea = 701
chest pain = 687
palpitations = 674
anxiety = 212
post-exertional malaise = 36
Tired = fatigue = 36
muscle pain = myalgia = 35

UNDER REVIEW!!!
Identification of drug-safety signals for COVID19 drug treatments

- We have found the drug mentions
- We have found the adverse side effects

- In process: generating PRR and OR from them
- Next up: Proper attribution of signals
The future

• We have users
• We have user timelines
• We have self-reported conditions
• We have self-reported drug usage
• We have self-reported lab tests (and with results sometimes!)
• We have (some) methods to attribute these things

We have ‘patient’ timelines that can go into CDM
We have hundreds/thousands of them
The obstacles

• Getting funding for Twitter research (in the Health sciences space) is super hard!

• There is considerable noise on this data
  • Cleaning/extracting stuff from clinical notes is a cake walk in comparison

• Attribution is very hard
The gory details:

• Technical stuff:
  • “Building tools and frameworks for large-scale social media mining: Creating data infrastructure for COVID-19 research” dair.ai meetup 7/22: https://www.meetup.com/dair-ai/events/271690722/
Acknowledgments

- All this work would not be possible without the help of my Ph.D student: Ramya Tekumalla

- Collaborations with: Dani Prieto-Alhambra, Gurdas Viguruji Singh, Osaid H. Alser

- Additional GSU collaborators: Dr. Gerardo Chowell

- Extra data provided by: Guanyu Wang², Jingyuan Yu³, Tuo Liu⁴, Yuning Ding⁵.

- Dr. VJ Periyakoil at Stanford University

- Funding by: National Institute of Aging through Stanford University’s Stanford Aging & Ethnogeriatrics Transdisciplinary Collaborative Center (SAGE) center (award 3P30AG059307-02S1)
Want to get involved?

• Do you have interesting questions?
• Do you have funding? 😊

• Feel free to get in touch: jbanda@gsu.edu or @drjmbanda

• Access the data and related items: http://www.panacealab.org/covid19/