

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

7/21/2020

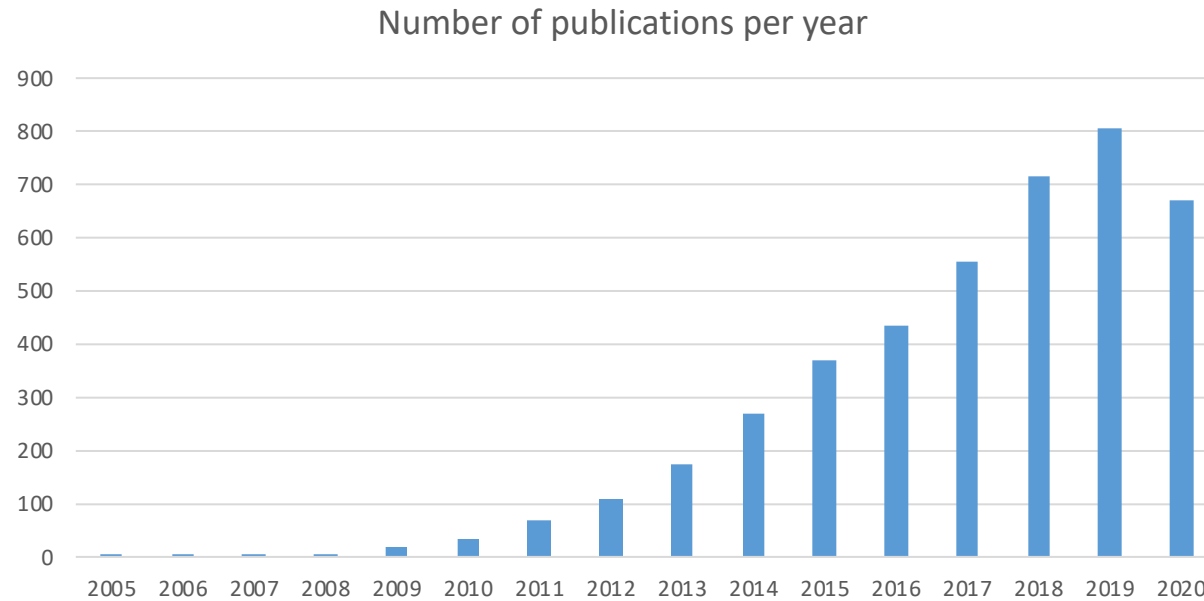
Juan M. Banda

www.panacealab.org

jbanda@gsu.edu

@drjmbanda

Preface: Twitter is gaining attention for health-related research since 2009



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)



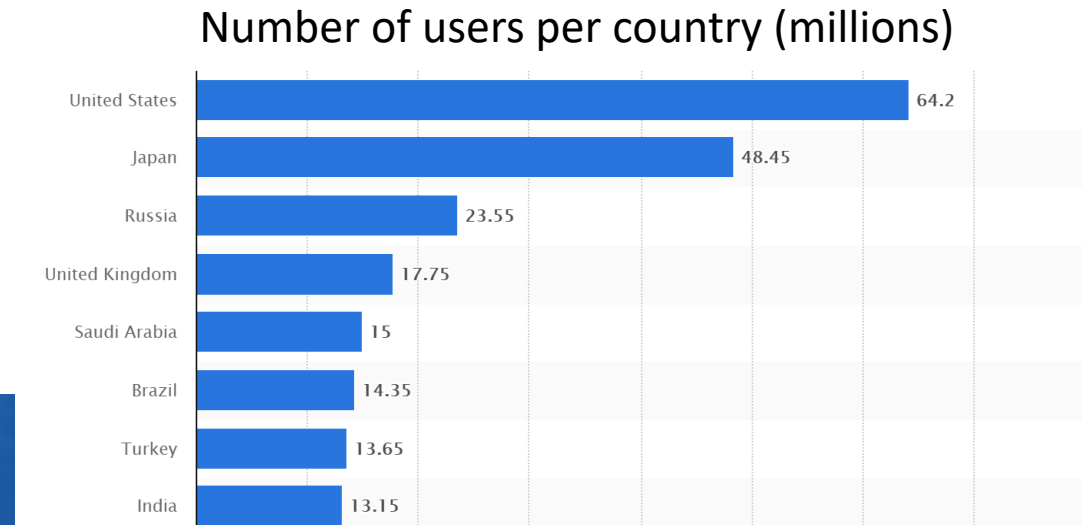
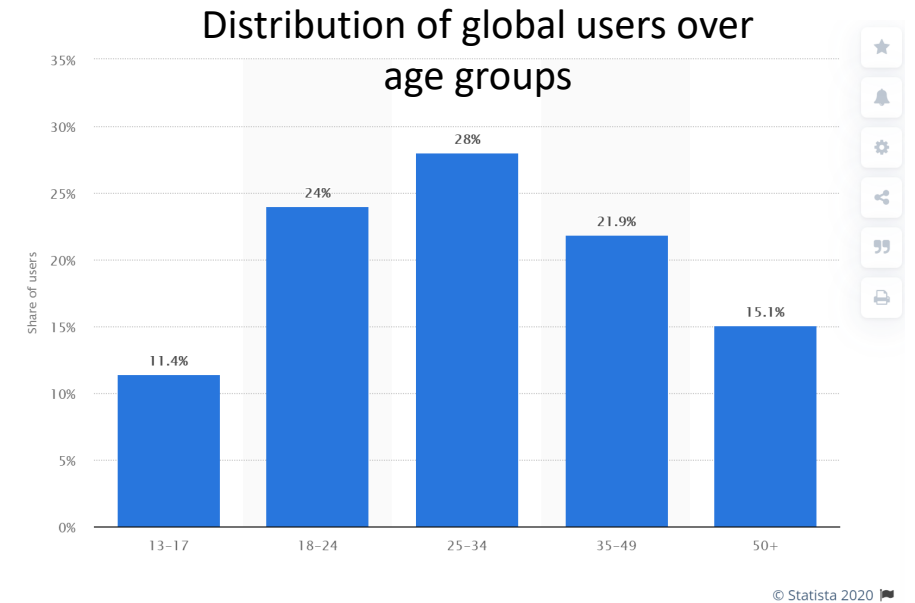
- OHDSI collaborator on probabilistic phenotyping using weak supervision and NLP Workgroup member standardizing resources to OHDSI vocabulary



- 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)



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 (🤔)

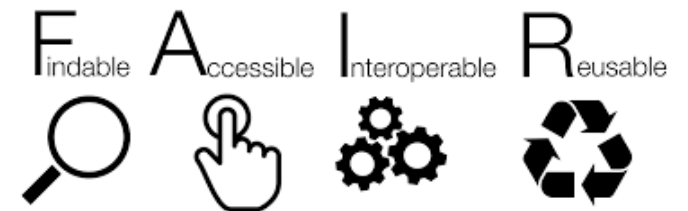
Collecting the data

- 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

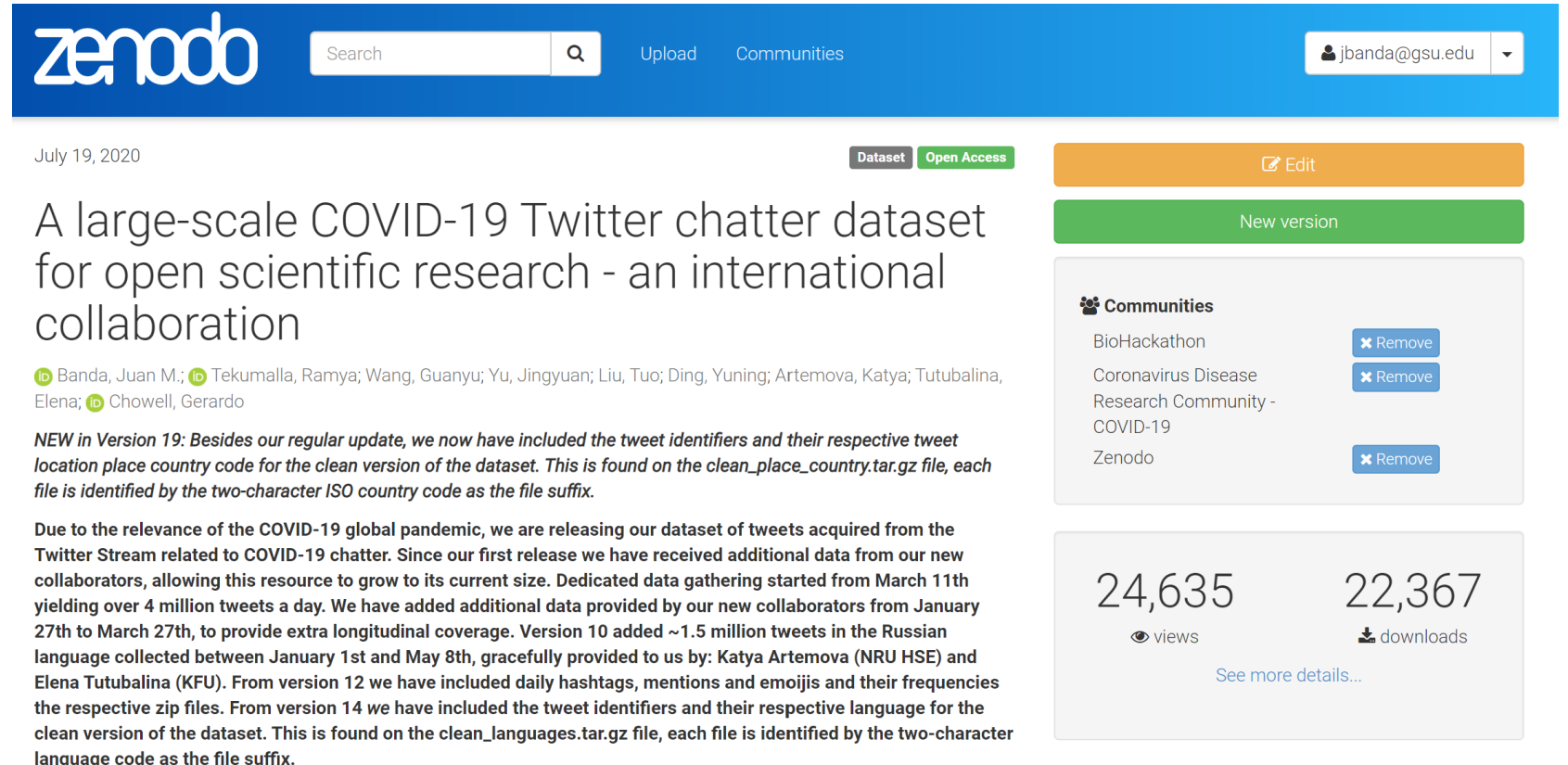


..... within all legal and ethical responsibilities of the data being shared

The dataset:

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

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



The screenshot shows the Zenodo dataset page for a COVID-19 Twitter chatter dataset. The header includes the Zenodo logo, a search bar, and links for Upload and Communities. The user profile 'jbanda@gsu.edu' is visible in the top right. The dataset is dated July 19, 2020, and is marked as 'Dataset' and 'Open Access'. The title is 'A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration'. The authors listed are Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; and Chowell, Gerardo. A note about Version 19 states: 'Besides our regular update, we now have included the tweet identifiers and their respective tweet location place country code for the clean version of the dataset. This is found on the clean_place_country.tar.gz file, each file is identified by the two-character ISO country code as the file suffix.' A detailed paragraph explains the dataset's growth and updates: 'Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and Elena Tutubalina (KFU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. This is found on the clean_languages.tar.gz file, each file is identified by the two-character language code as the file suffix.' On the right, there are buttons for 'Edit' and 'New version'. Below these, a 'Communities' section lists 'BioHackathon', 'Coronavirus Disease Research Community - COVID-19', and 'Zenodo', each with a 'Remove' button. At the bottom right, statistics show 24,635 views and 22,367 downloads, with a link to 'See more details...'.

zenodo Search Upload Communities jbanda@gsu.edu

July 19, 2020 Dataset Open Access Edit New version

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo

NEW in Version 19: Besides our regular update, we now have included the tweet identifiers and their respective tweet location place country code for the clean version of the dataset. This is found on the clean_place_country.tar.gz file, each file is identified by the two-character ISO country code as the file suffix.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and Elena Tutubalina (KFU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. This is found on the clean_languages.tar.gz file, each file is identified by the two-character language code as the file suffix.

Communities BioHackathon Coronavirus Disease Research Community - COVID-19 Zenodo Remove Remove Remove

24,635 views 22,367 downloads See more details...

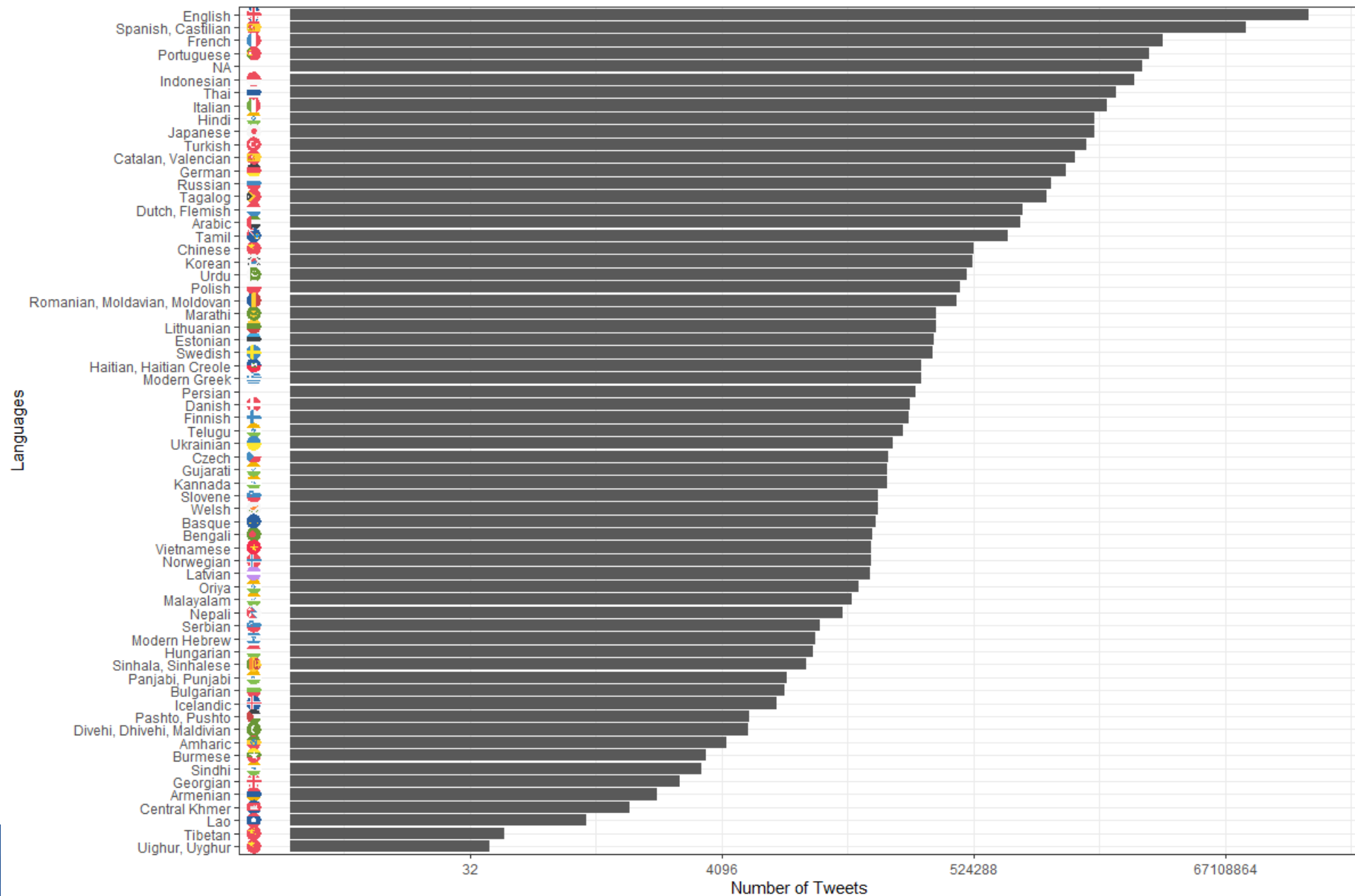
Dataset: <https://doi.org/10.5281/zenodo.3723939>

Pre-print: <https://arxiv.org/abs/2004.03688>

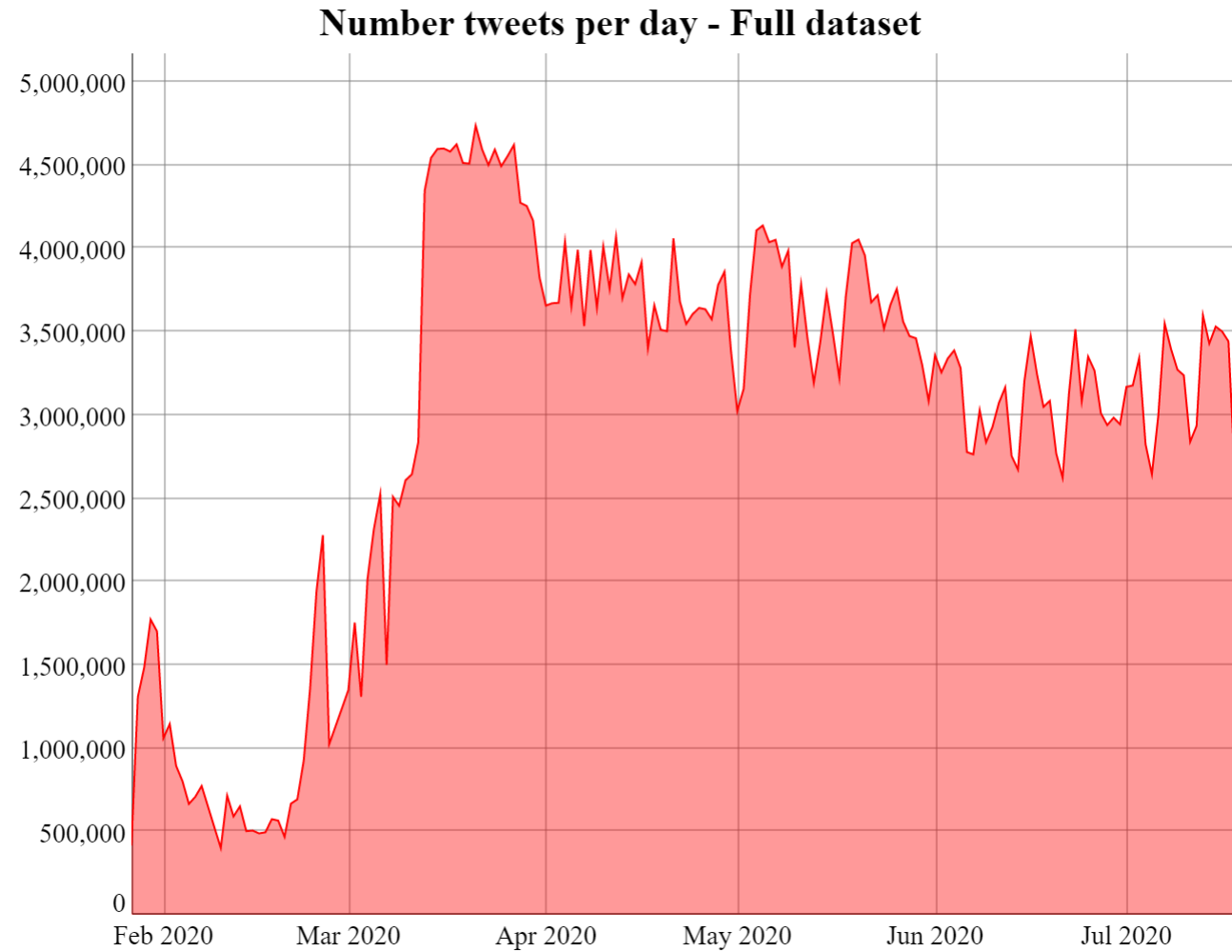
Recent additions: https://github.com/thepanacealab/covid19_twitter

Languages
available:

69 in total

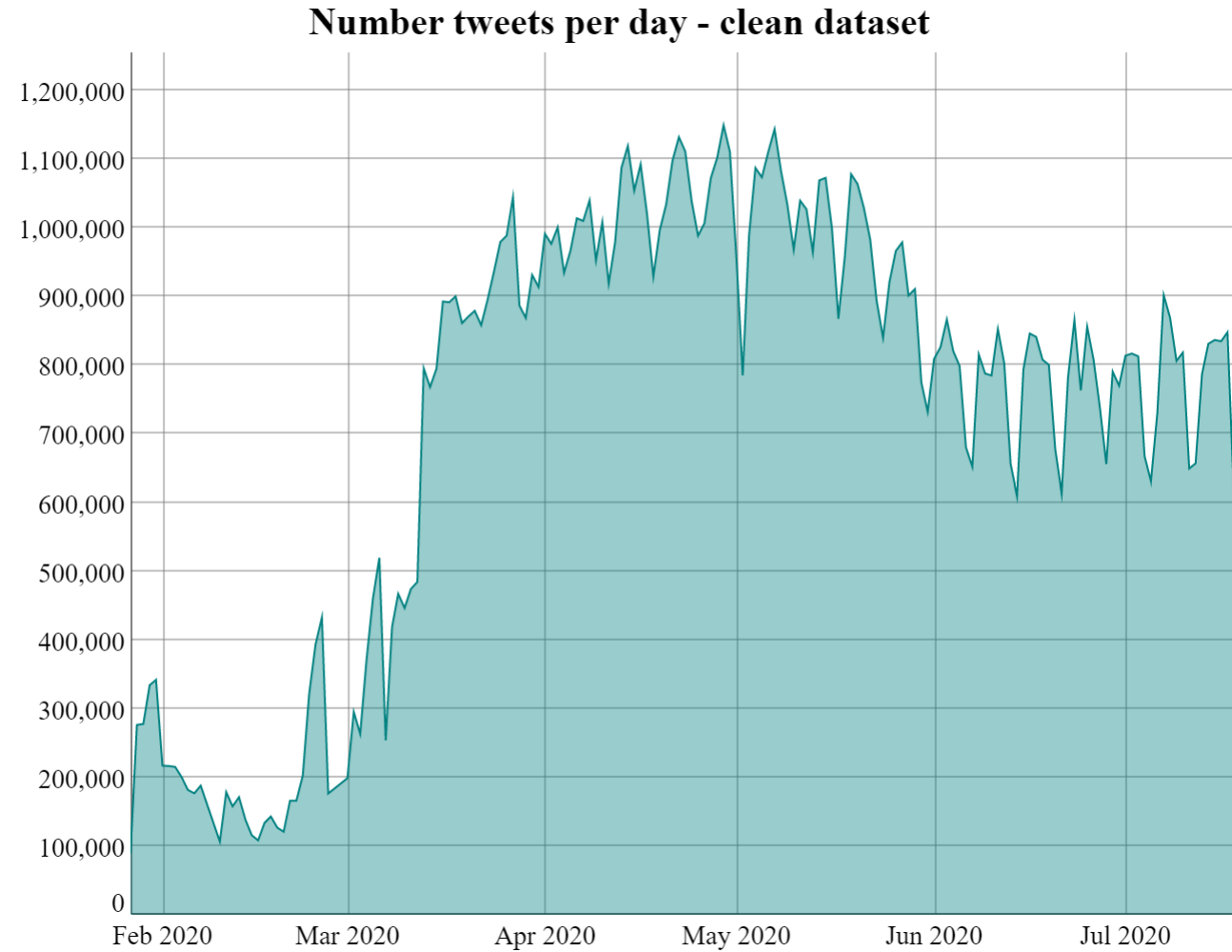


Tweets per day (all):



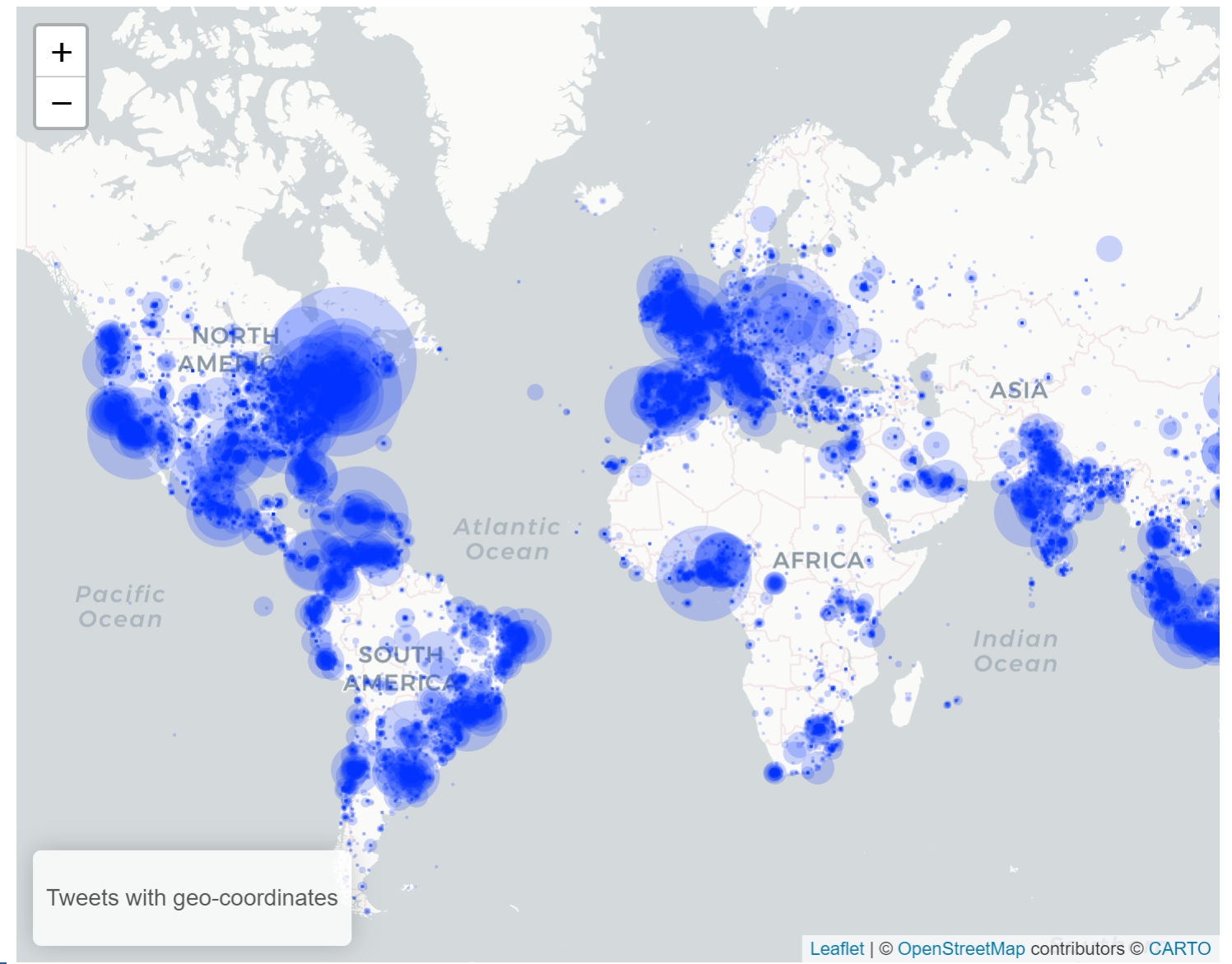
<http://www.panacealab.org/covid19/>

Tweets per day (clean):



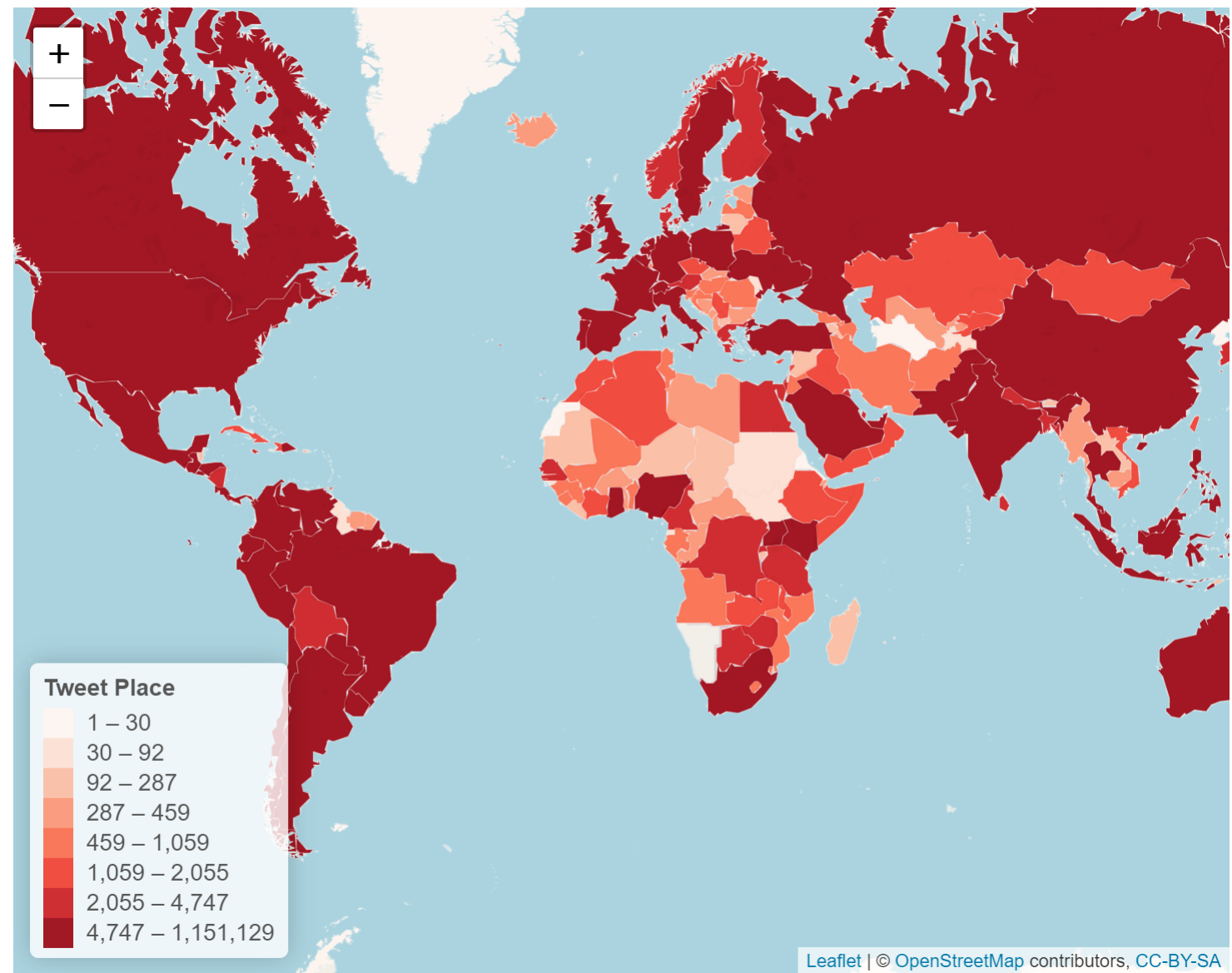
<http://www.panacealab.org/covid19/>

Geolocations?
... on some
less than 1%



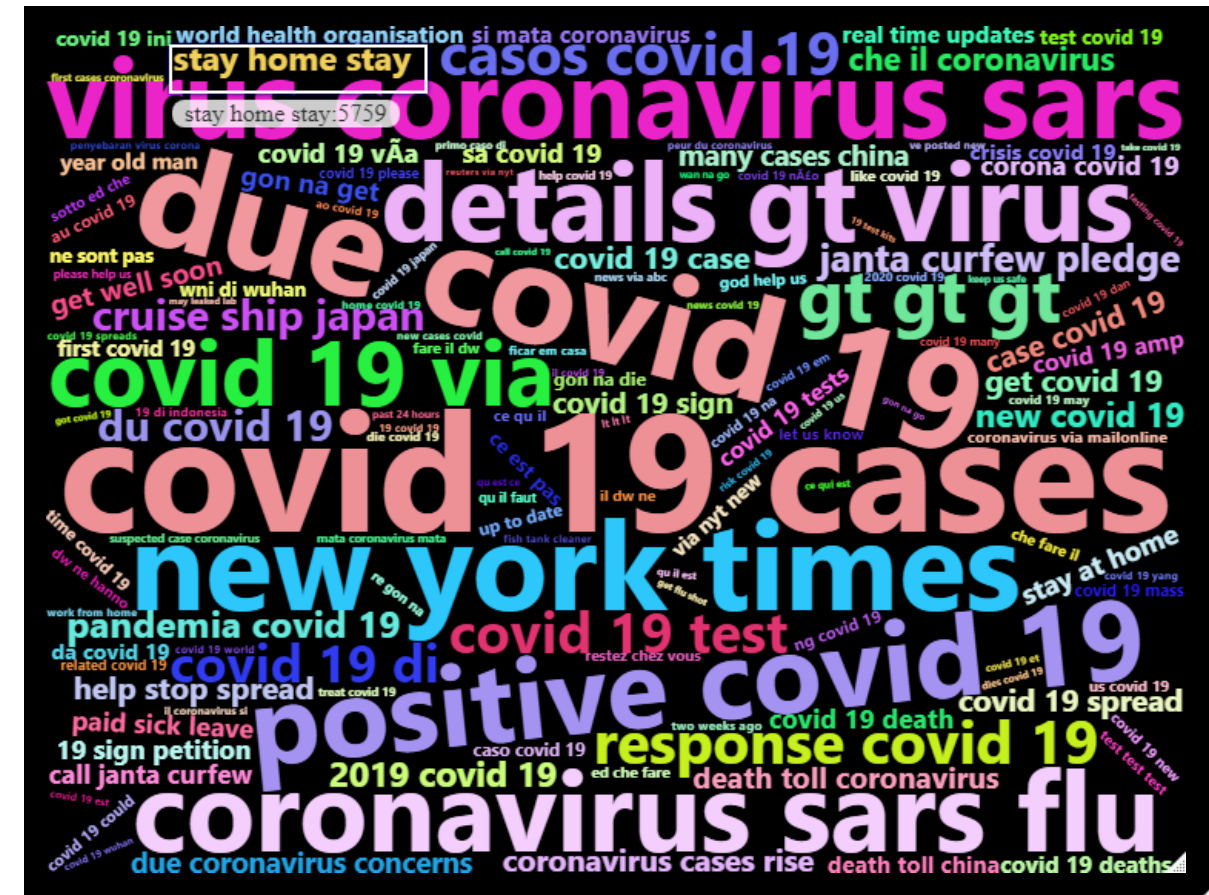
Tweets with place
location enabled?
... some

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

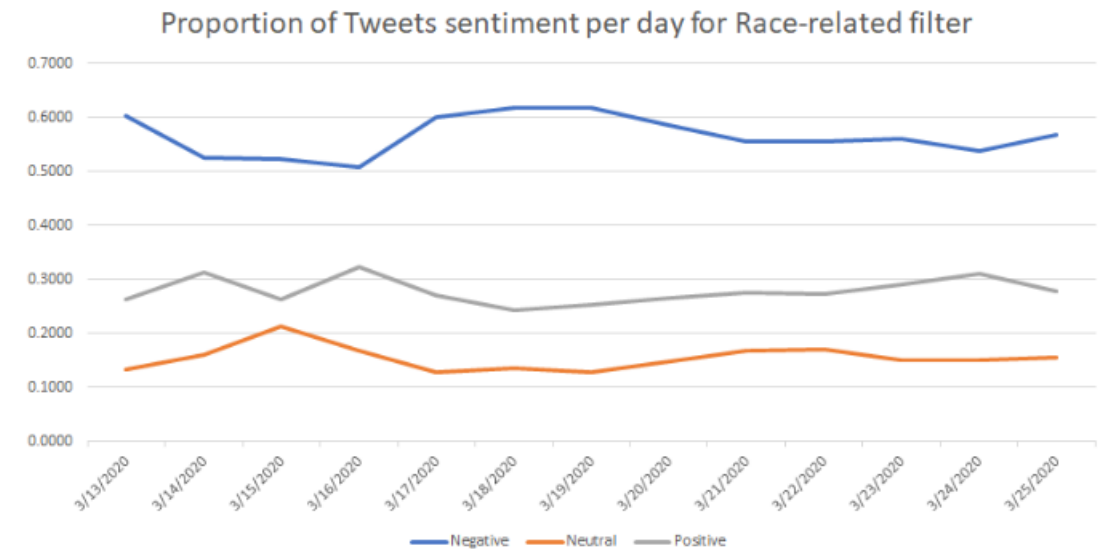
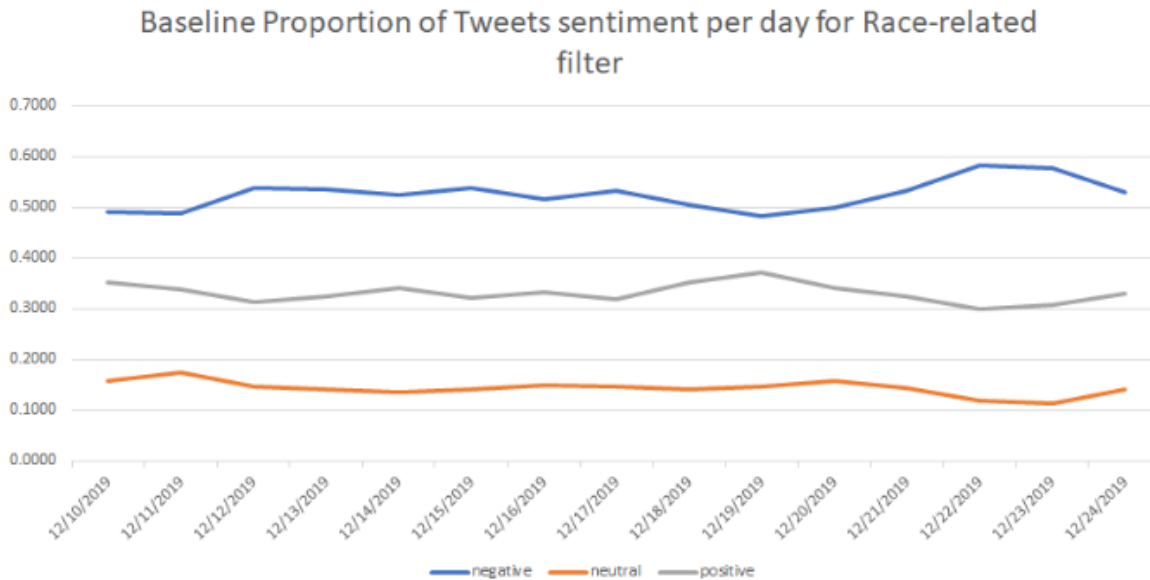
- Being a fellow at Stanford SAGE Research Center one of our first questions was to identify the perception of the Twitter users to ethnic minorities and elderly populations
- Turns out we can!
- Turns out is not as expected!

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)

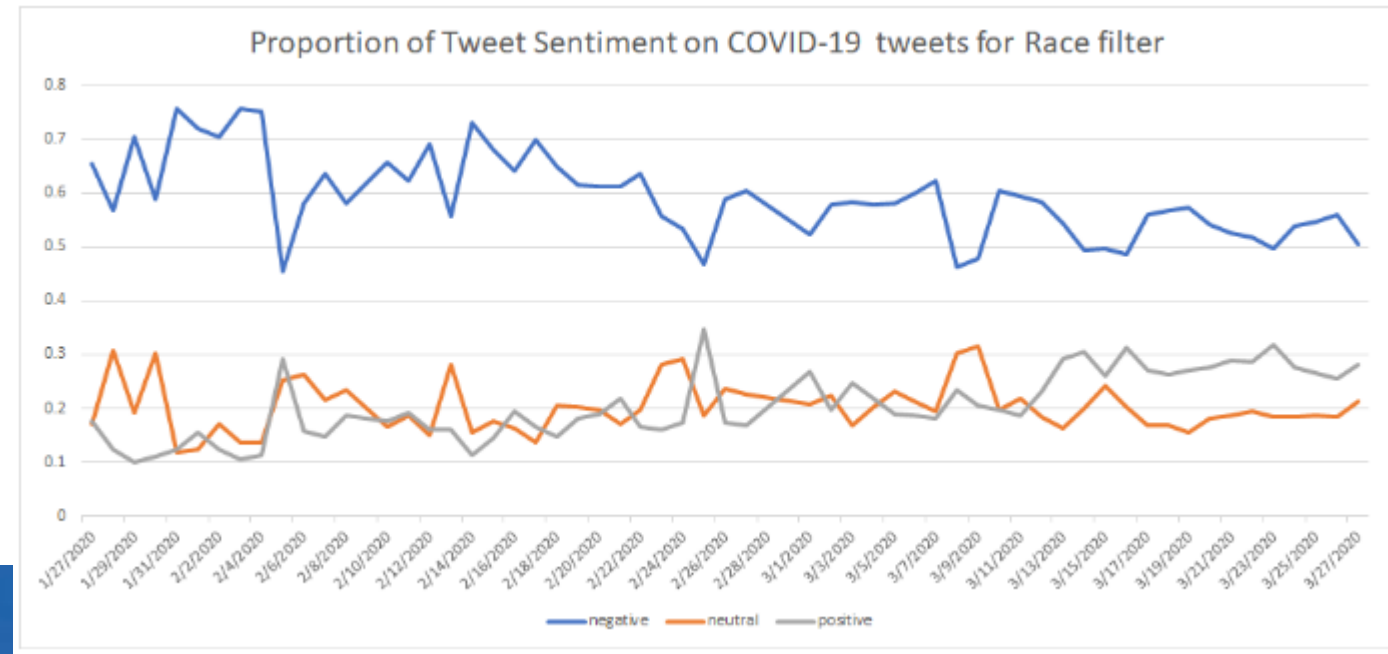
Perception towards the elderly and ethnic minorities (2)

- For ethnic minorities:

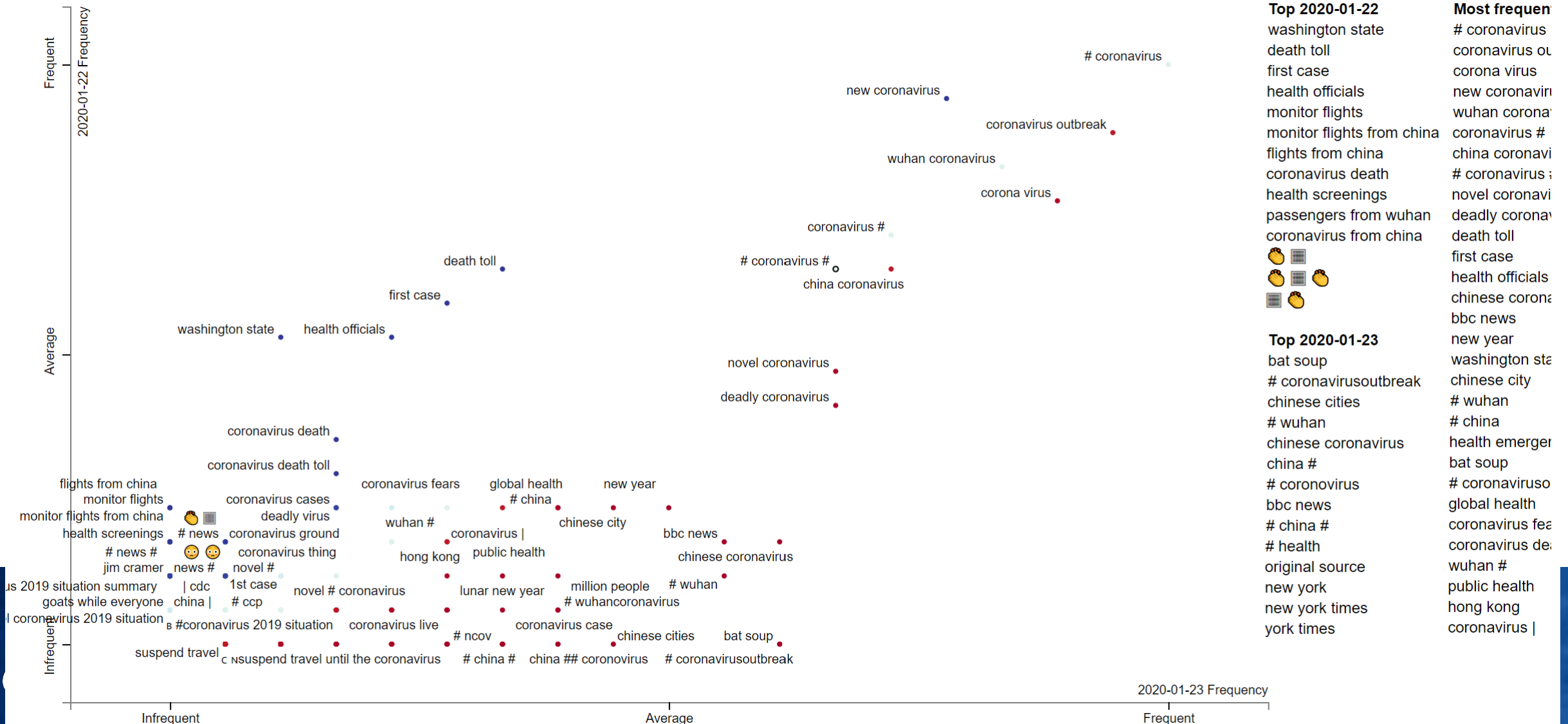


Perception towards the elderly and ethnic minorities (3)

- 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.

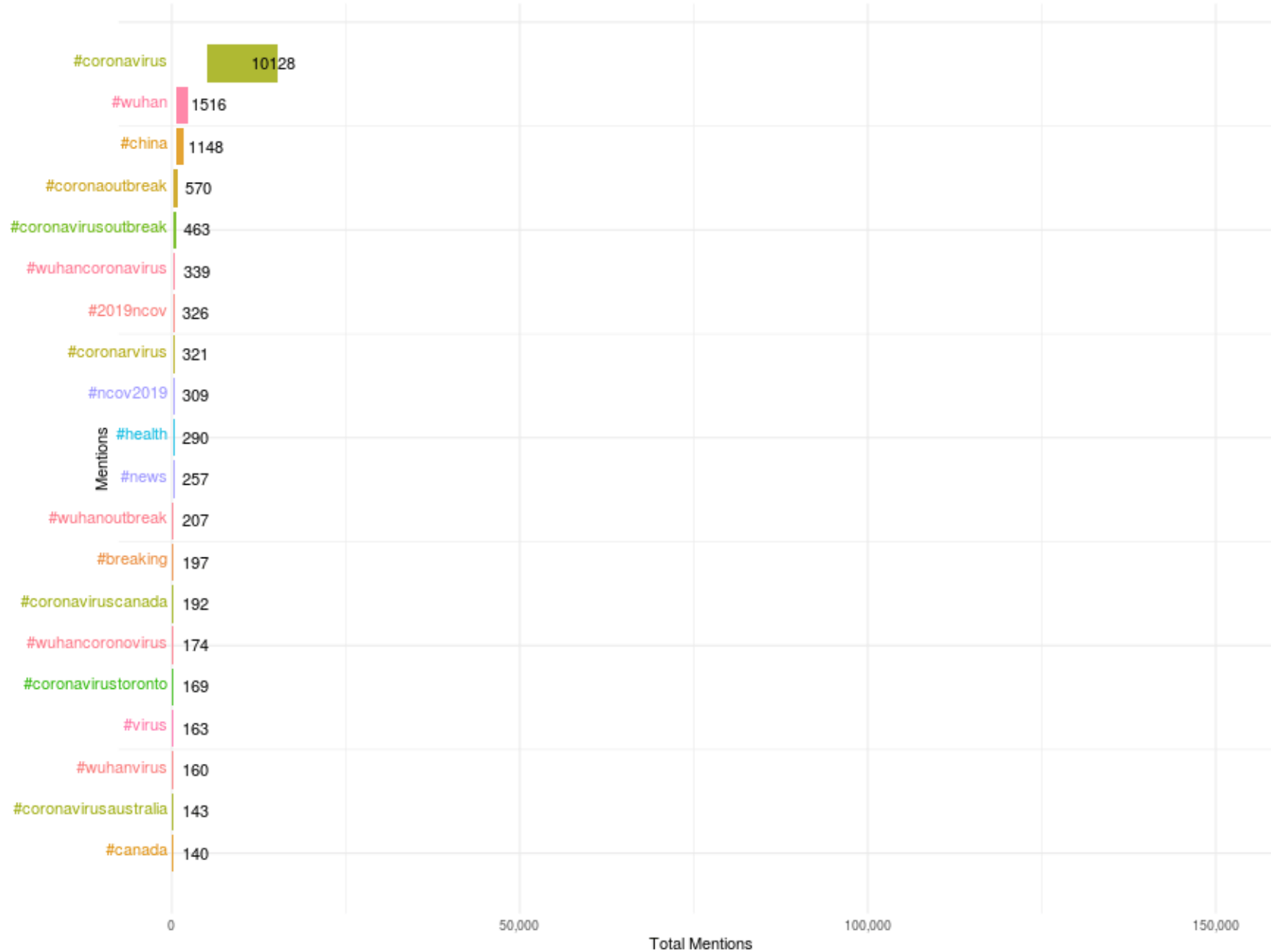


Hashtag Evolution over time

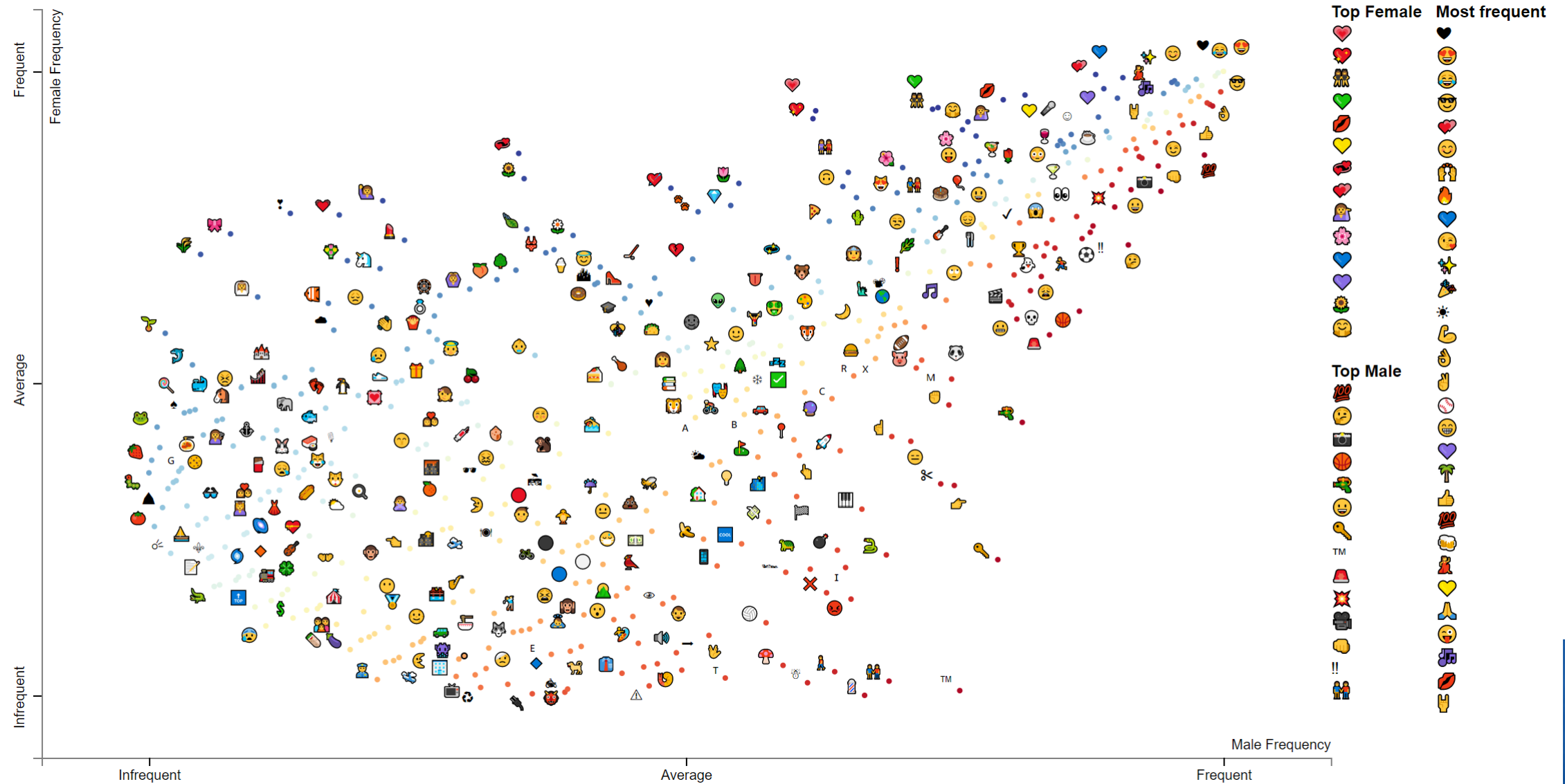


Total hashtag mentions in : 2020-01-27

Top 10 Hashtags



Ge

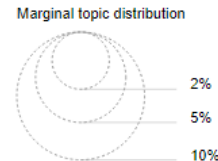
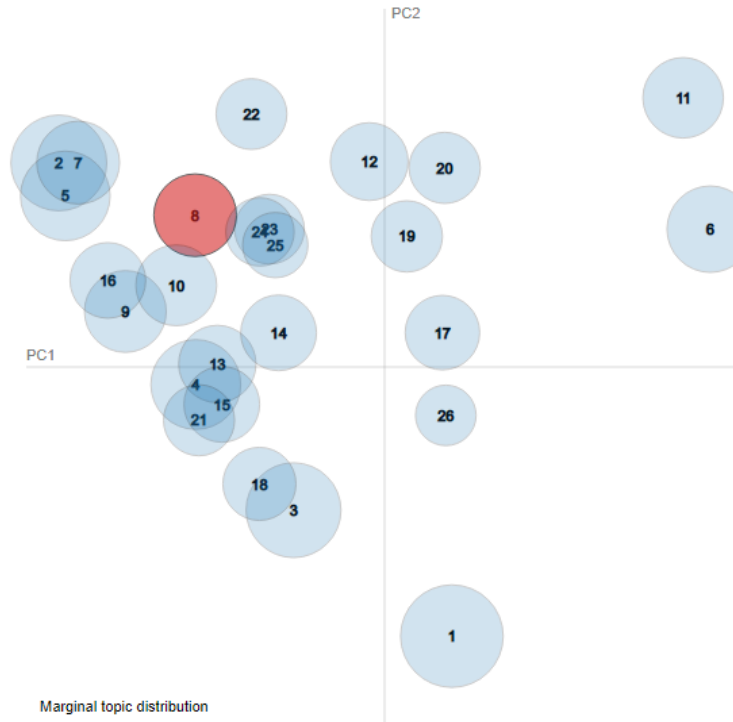


Topic modeling?

- Quite tricky with Twitter data!
- Too many topics, not enough clarity on our initial analyses with LDA and Dynamic LDA

Selected Topic:

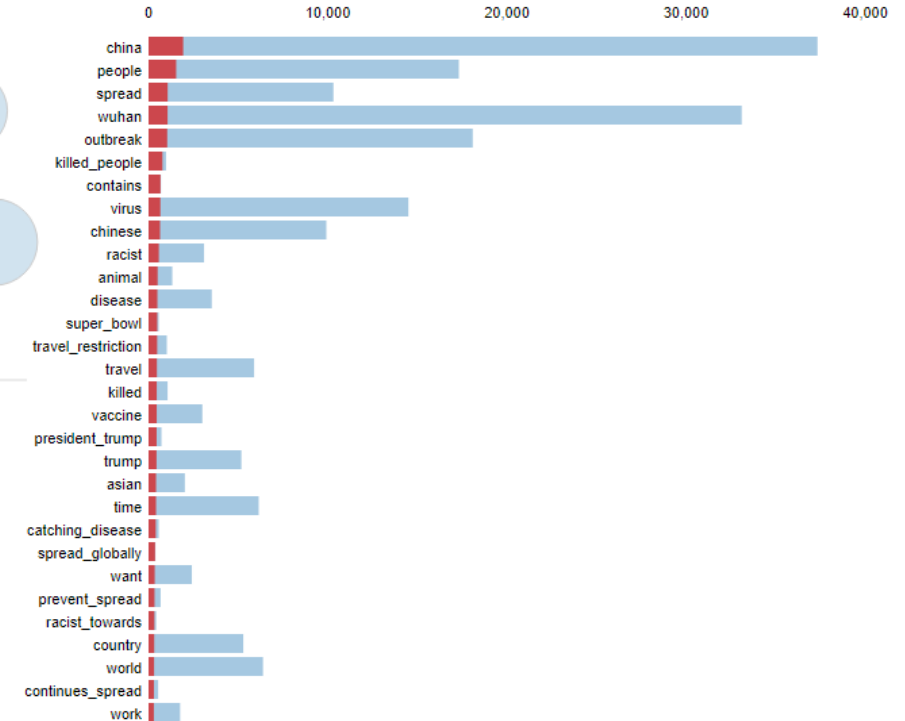
Intertopic Distance Map (via multidimensional scaling)



Slide to adjust relevance metric:⁽²⁾

$\lambda = 1$ 0.0 0.2 0.4 0.6 0.8 1.0

Top-30 Most Relevant Terms for Topic 8 (4.2% of tokens)



Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

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?

OHDSI Dictionary for NLP

- 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.

* <https://genominfo.org/journal/view.php?number=605>

What we got:

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

Is this useful this way?

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

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

What the vocabulary gets us:

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.

Cool story... show me some RWE

Can we characterize drug mentions?

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

Timeline of Tweet mentions of COVID-19 potential drug treatments

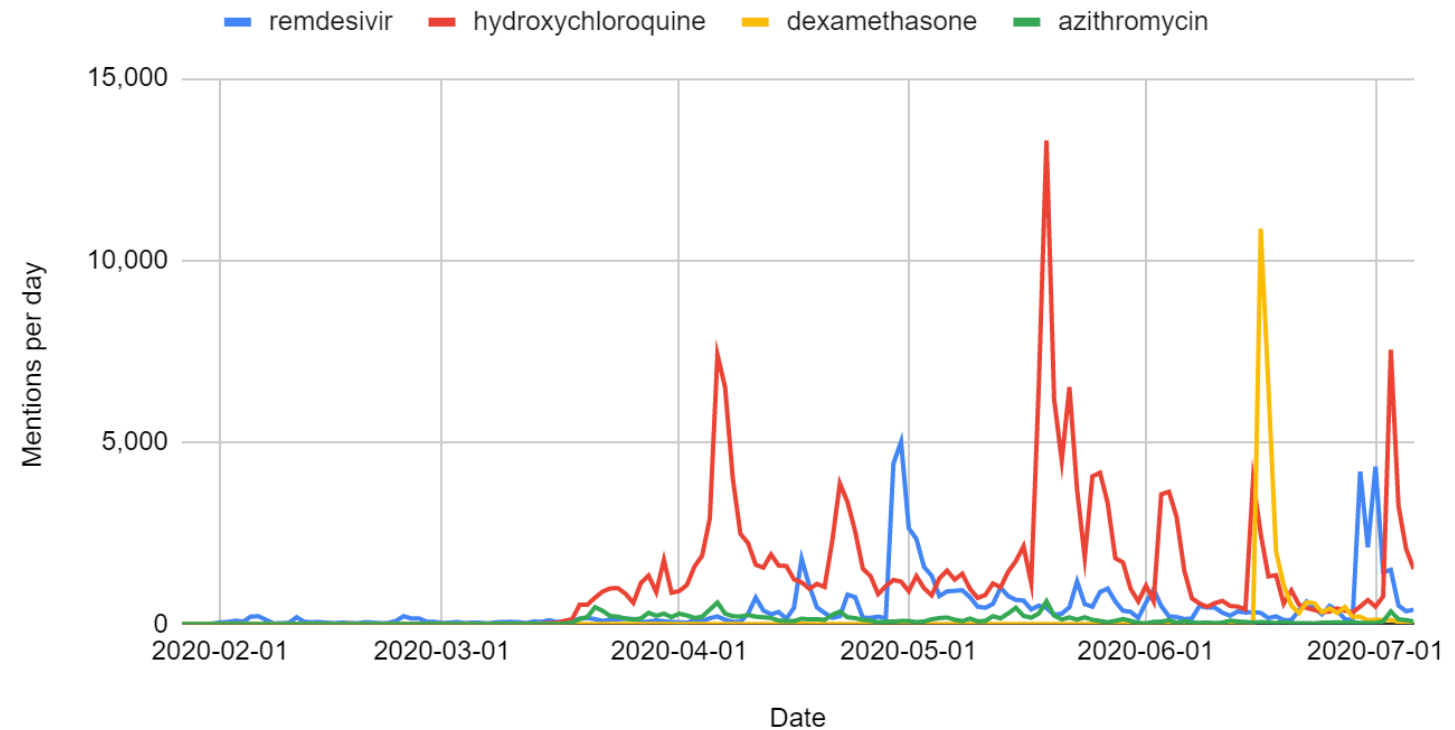
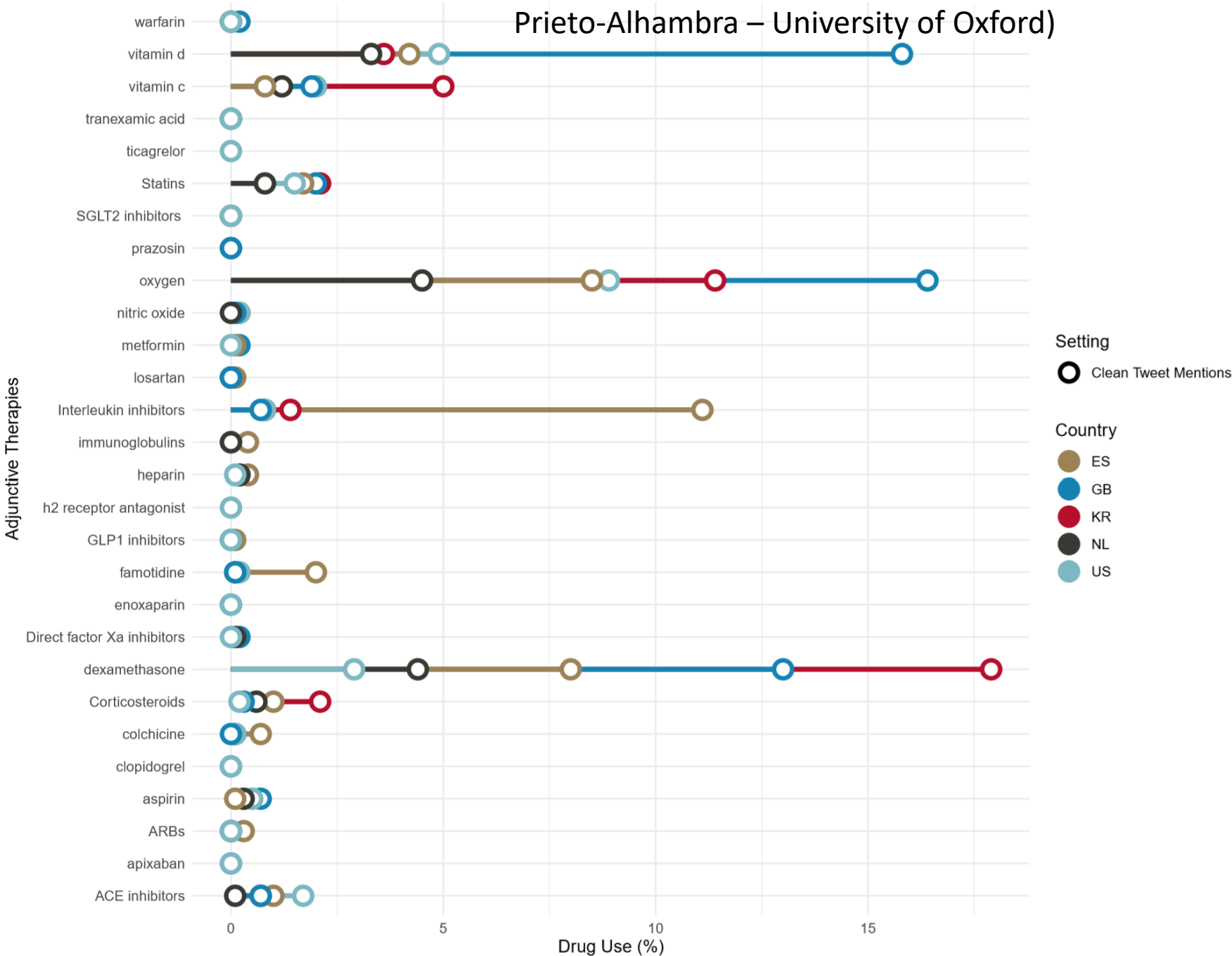


Figure 1. Timeline of Tweets with potential drug treatment mentions.

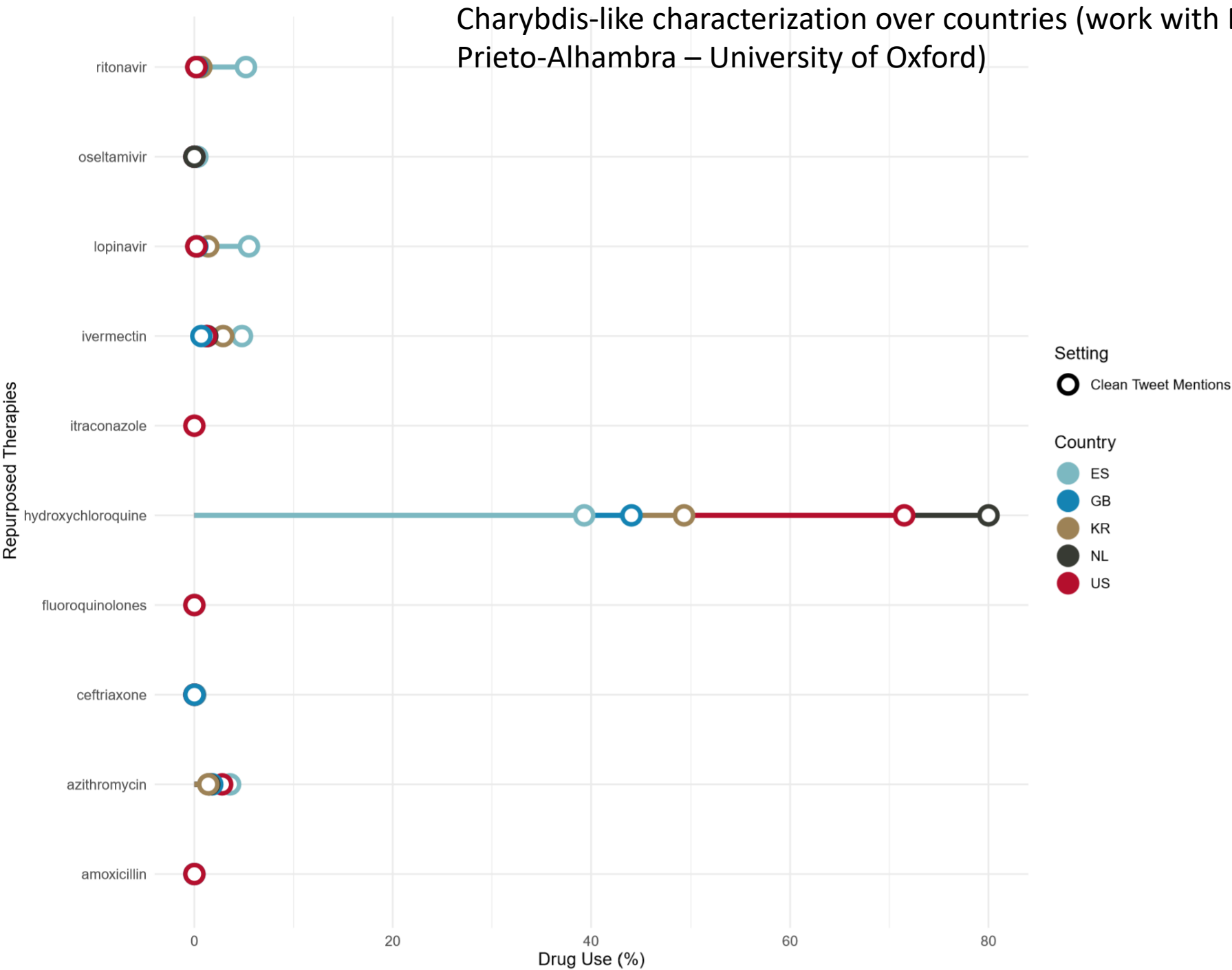
Charybdis style?

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



Charybdis style?

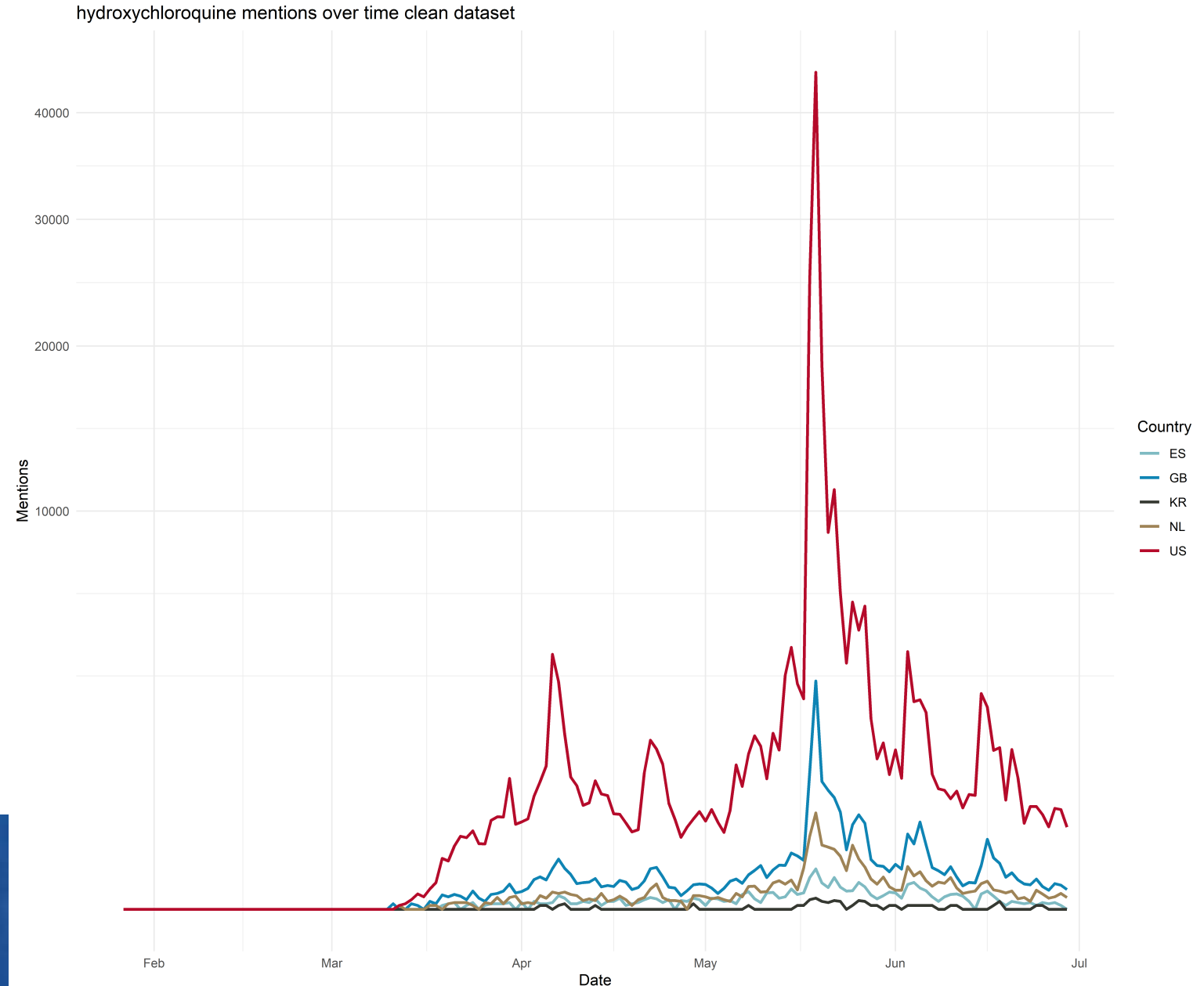
Charybdis-like characterization over countries (work with Dani Prieto-Alhambra – University of Oxford)



Charybdis
style?

... over
time?

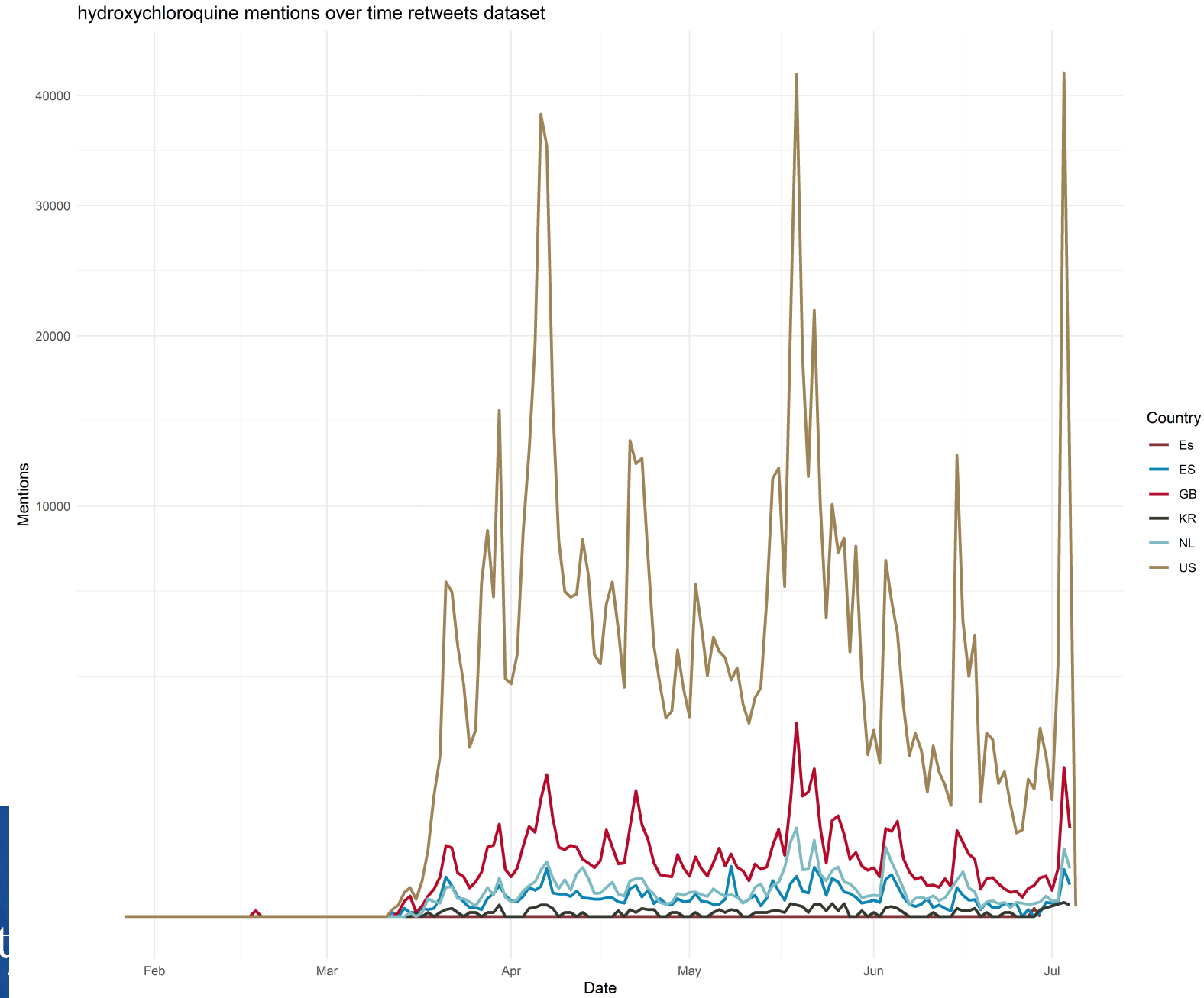
Charybdis-like characterization over countries (work with Dani Prieto-Alhambra – University of Oxford)



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>) but on Twitter?

Term	Frequency
pneumonia	110124
infection	71882
influenza	36390
cough	35753
anxiety	34658
pain	12773
depression	12189
asthma	8307

* https://github.com/thepanacealab/covid19_biohackathon/tree/master/user_symptoms

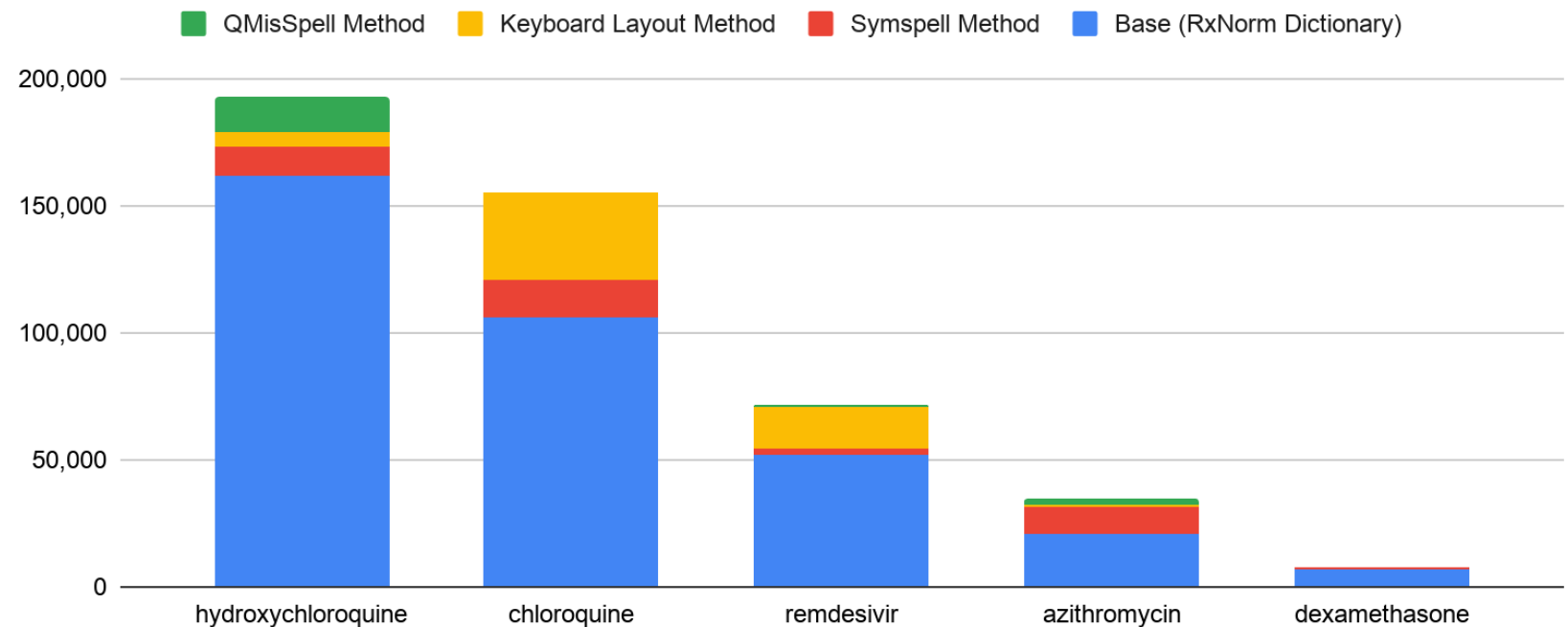
Is this real RWE?..... Not quite.... Yet!

How do we get there?

Data is very messy!

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

Frequency vs Drug Name



Attribution is critical

- What about attributions?
 - Others have done work on this:
 - Klein, Ari, Arjun Magge, Karen O'Connor, Haitao Cai, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. "A Chronological and Geographical Analysis of Personal Reports of COVID-19 on Twitter." Health Informatics. medRxiv. <https://doi.org/10.1101/2020.04.19.20069948>
 - Sarker, Abeed, Sahithi Lakamana, Whitney Hogg, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. "Self-Reported COVID-19 Symptoms on Twitter: An Analysis and a Research Resource." Health Informatics. medRxiv. <https://doi.org/10.1101/2020.04.16.20067421>
 - 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

Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold

Ramya Tekumalla

Georgia State University

Javad Rafiei Asl

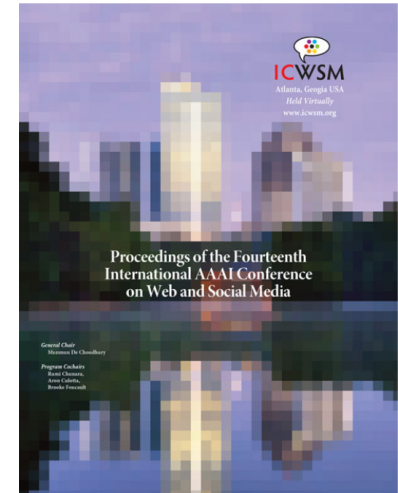
Georgia State University

Juan M. Banda

Georgia State University

Abstract

In the last few years, Twitter has become an important resource for the identification of Adverse Drug Reactions (ADRs), monitoring flu trends, and other pharmacovigilance and general research applications. Most researchers spend their time crawling Twitter, buying expensive pre-mined



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!

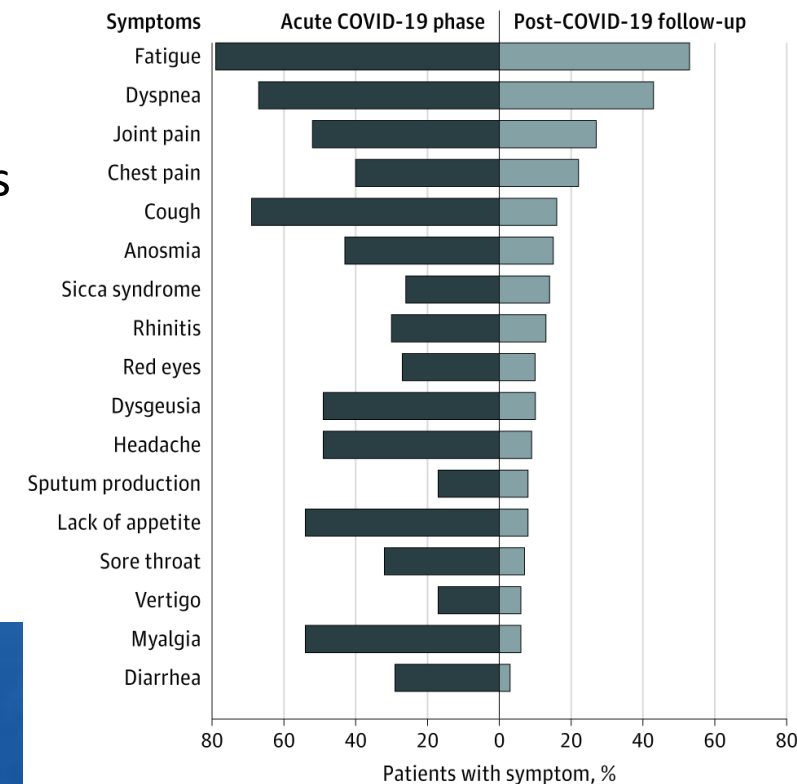
JAMA doi:10.1001/jama.2020.12603

- 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/>