Tracking Twitter Conversations on the COVID-19 Pandemic: An Analysis for Canada[1]

Pearl Herrero, Nicole Poynarova, Luis Quiñones and Renato Vassallo

March 5, 2023

Abstract

The objective of this Technical Note is to understand how the conversation on Twitter about COVID-19 evolved during the beginning of the pandemic. In particular, we are interested in tracking only MD users. To do this, the first step was to collect a large-scale dataset of COVID-19-specific geotagged tweets using the Twarc library, then filter and clean the tweets to obtain some descriptive statistics that could serve as input for further research.

1 Streaming

The first step would be to obtain a list of user names and user ids for Medical Doctors. We implement this with the stream endpoint, making use of the bio description of the users.

```
twarc2 stream-rules add "(bio:\"Medical Doctor\"
OR bio:\"M.D.\" OR bio:\"MD\"
OR bio:\"M.D\" OR bio:\"PhD
in Medicine\" OR bio:\"Phd
in medicine\" OR bio:\"PHD
in Medicine\" OR bio:\"Phd
in Medicine\" OR bio:\"PHD
in medicine\" OR bio:\"PhD
in medicine\" OR bio:\"#MD\"
OR bio:\"Doctor of Medicine\" OR bio_name:\"MD\"
OR bio_name:\"M.D.\"
OR bio_name:\'’M.D\’’
(bio_location:Canada
OR bio_location:canada
OR bio_location:Toronto
OR bio_location:Montreal) lang:en"

twarc2 stream doctors.jsonl
```

[1]This technical note is part of the solution to Problem Set 2 of the Text Mining and Natural Language Processing course, corresponding to the Data Science 2022-2023 masters program at the Barcelona School of Economics.

2 Getting followers from the smaller side

```
unique_ids = set()
# Iterate through the set of usernames
for index, row in df.iterrows():
    username = row["username"]
    followers_count = row["followers_count"]
    following_count = row["following_count"]
    # check which list is smaller
    if followers_count <= following_count:
        # Retrieve the followers of the user
        followers = list(t.follower_ids(username))
        # add unique ids to the set
        unique_ids.update(followers)
    else:
        # Retrieve the followings of the user
        following = list(t.friend_ids(username))
        # add unique ids to the set
        unique_ids.update(following)

# write the set of unique ids to the followers.txt file
with open("followers.txt","w+") as h:
    for id in unique_ids:
        h.write(str(id)+"\n")
```

Given we have 117 user's we will get the follower's or following from, the time estimate is around 1 hour and 57 minutes. Based on the rate limit of 15 requests per 15 minutes.

3 Filtering

Now, we filter, among the followers, those that are medical doctors using the same text matching criteria that we used at the beginning.

```
twarc users followers.txt > users.jsonl

twarc2 csv users.jsonl users.csv


df=df_2[df_2['description'].str.contains('|'.join(keywords), case=False)
& df_2['location'].str.contains('|'.join(location_keywords), case=False)]
```

See code for more detail.

4 Users activity

The next step is to obtain all the users' activity around the period spanning 2019-10-01 to 2020-10-01 (included).

```
twarc2 timelines userids.txt
--start-time "2019-10-01" --end-time "2020-10-01" > useractivity.jsonl
```

Count estimate:

```
df['public_metrics.tweet_count'].sum()
```

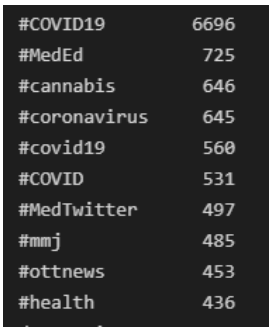
9598620

5 Descriptive statistics

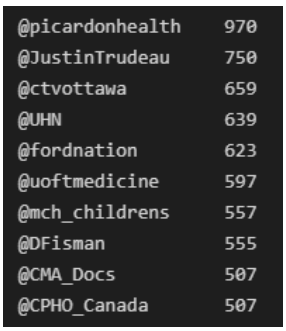
Now, we will focus on getting some descriptive statistics, in particular we are interested on analyzing the top 10 hashtags, mentions and emojis. Then we are going to plot the timeline evolution for the total number of retweets and the evolution of tweets that contain any word related to the pandemic.

5.1 Top 10

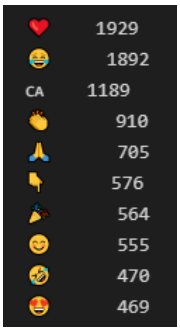
From users activity data around the period spanning 2019-10-01 to 2020-10-01, we extract relevant English-language tweets and filter for information on incidence of mentions, hashtags, and emojis. Figure 1 presents the ranking of the 10 most used hashtags, mentions and emojis.



(a) Top 10 hashtags



(b) Top 10 mentions



(c) Top 10 emojis

Figure 1: Descriptive statistics: top 10.

5.2 Evolution of retweets

Figure 2 shows the evolution of the total number of retweets from our user activity data in Canada. The series shows peaks in March and July 2020, associated with the outbreak of the COVID-19 pandemic, the start of quarantines and the different outbreaks.

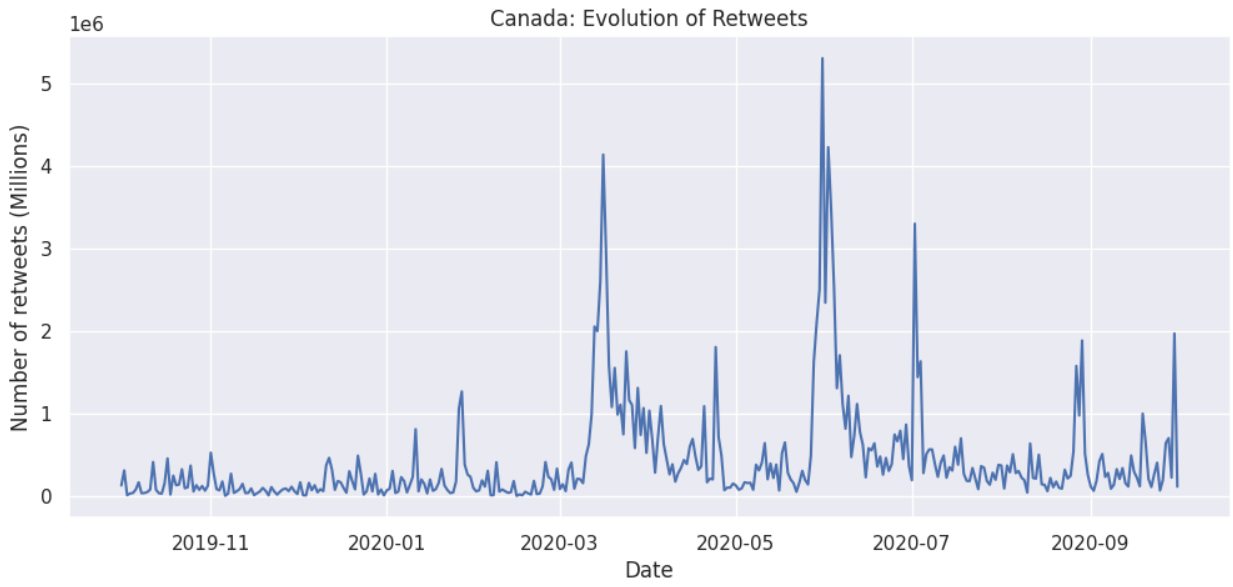


Figure 2: Time evolution of total number of retweets.

5.3 Evolution of COVID-19 indicator

We construct a pandemic indicator using the following list of **keywords**: pandemic, covid, asymptomatic, coronavirus, incubation, pathogen, distancing, quarantine, isolation, antibody, outbreak, epidemic, mask, immunity, and respirator.
The indicator, shown in Figure 3, is consistent with the dynamics of COVID-19: a violent increase is observed at the beginning of March 2020 and then gradually falls, showing a certain rebound towards the end of the year (as a consequence of the second wave of infections). A certain erratic behavior of the series is observed, which would be associated with a seasonal weekend component.

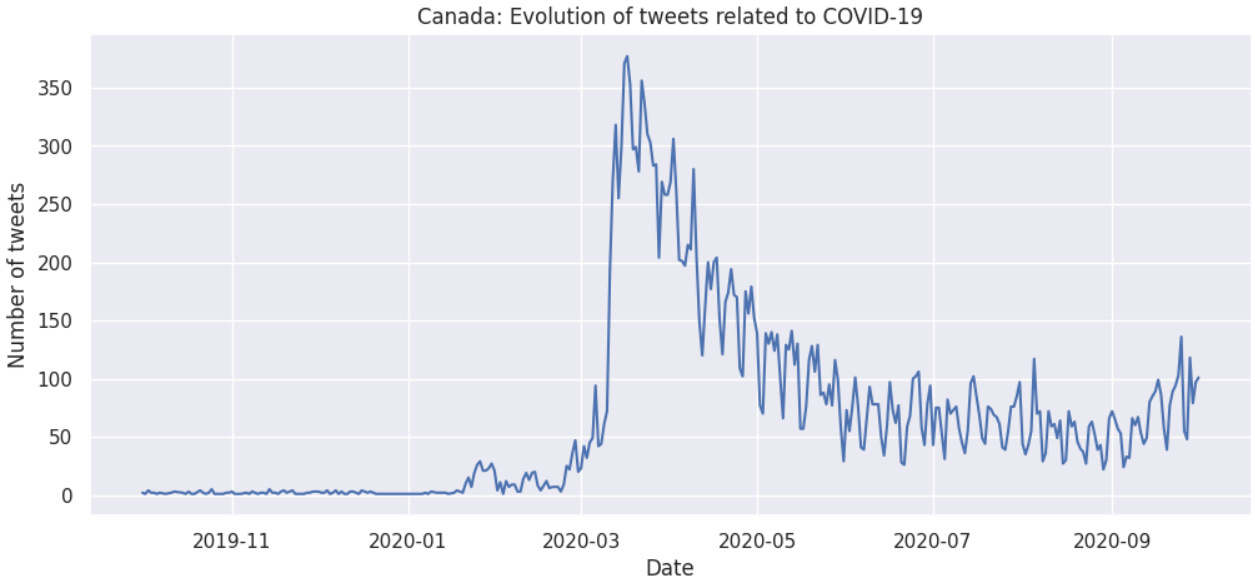


Figure 3: Pandemic indicator using words related to COVID-19.

6 Research project

An analysis of how sentiment and emotion are expressed on Twitter. An examination of the emotional and sentimental expressions on the platform. Additional tweet metadata, such as sentiments (positive, negative, neutral) and emotions (happy, sad, angry, etc.), would be collected as part of the project (the tweets). With regard to the most popular emojis, hashtags, and mentions, it would analyze the sentiment and emotion distributions. For each type of sentiment in a tweet, a different dataset could be created. Different patterns would be investigated, e.g: Are certain mentions more likely to be used in tweets with a certain narrative? Are certain emojis or hashtags more likely to be used in tweets with positive or negative sentiment? The project would look at how sentiment and emotion affect the use of emojis, hashtags, and mentions in tweets. Finally, is there a connection, for instance, between the tone of a tweet and the emojis used therein? By analysing these connections and patterns, the project would provide insight on how sentiment and emotion are expressed on social media platforms like Twitter.