UNIVERSITAT
ROVIRA i VIRGILI

Big Data Analytics

Report Final Project

Student:

Giorgio Rossi

January 6, 2021

Contents

1 Introduction

I have collected all the materials related to this assignment (code, report, python notebooks and dataset) on a Github repository [1], also available at the following link: `jeorjebot/covid_vaccine_tweets_italy`.

1.1 Description of the problem

With this project I want to analyse how public opinion varies on hot themes of this period, the **vaccine** against Covid-19 and the start of the **vaccination campaign** in Italy.

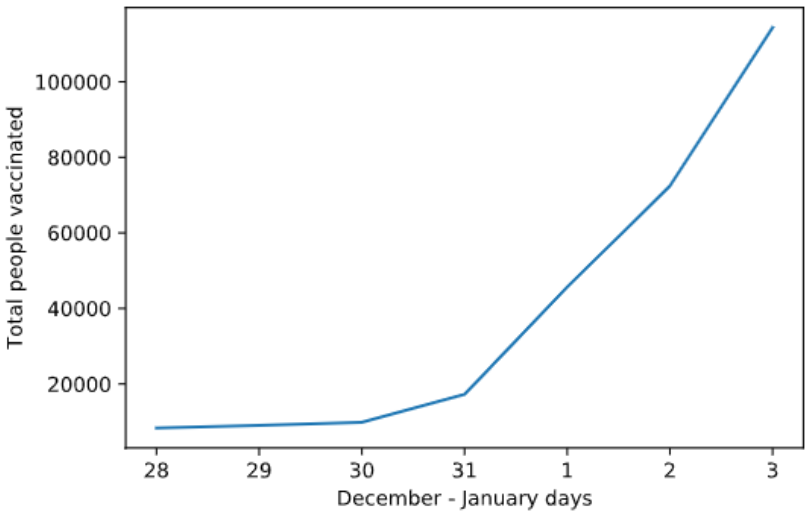

Figure 1: Prediction of available doses on Italy according to the Strategic Plan and the agreements with the European Commission (millions of doses)

1.2 Temporal Events

I have summarised briefly the events regarding the starting of the vaccination campaign as follow:

- **21 December 2020**: the EMA [2] approved Pfizer-Biontech vaccine.
- **26 December 2020**: Pfizer delivered first symbolic doses of the vaccine to all the European countries. Italy received 9750 doses.
- **27 December 2020**: European Vaccine Day
- **31 December 2020**: the Italian Ministry of Health made available a dashboard [3] with updated data and statistics on the ongoing vaccinal campaign.

- **1 January 2021**: Pfizer delivered in Italy the first tranche of vaccines, which amount on 469950 doses.

1.3 Vaccination Campaign informations

All the information about the vaccinal campaign are extracted from the Strategic Plan approved by the Italian Government [4]. The week analyzed fall inside the plan for first quarter of the year.

VACCINE	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	TOTAL
Astra Zeneca	16,155	24,225	-	-	-	-	40,38
PF/BT	8,749	8,076	10,095	-	-	-	26,92
J&J *	-	14,806	32,304	6,73	-	-	53,84
Sanofi/GSK**	-	-	-	-	20,19	20,19	40,38
Curevac	2,019	5,384	6,73	8,076	8,076	-	30,285
Moderna	1,346	4,711	4,711	-	-	-	10,768
TOTAL	**28,269**	**57,202**	**53,84**	**14,806**	**28,266**	**20,19**	**202,573**
AVG x month	9,421	19,065	17,947	4,935	9,422	6,73	

Figure 2: Prediction of available doses on Italy according to the Strategic Plan and the agreements with the European Commission (millions of doses)

1.4 Data sources

I have used the dashboard [3] mentioned before to gain informations about the total amount of vaccines administred per day. Since the dashboard was available only from December 31th, on late night, I have used as sources the data provided by the splendid work of the **Luigi Einaudi Foundation** [5][6] for the days not covered by the dashboard (Dec 27th - 28th - 29th - 30th).

Furthermore the official dashboard provided only total data and not day by day data, so every night at 23:59 I have manually annotated the count of the vaccines administred that day.

The dataset with the tweets collected by my script is available here [7].

1.5 Language and Tools

I have coded the assignment in **Python**, taking advantage of the library tweetpy for interacting with the Twitter API as suggested by the professor. I have used a **shell script** to allow the script to run endlessy. I have coded the analysis of the data on **Ipython Notebook**. The tweets were stored on a **MongoDB** database running inside a **Docker Container**.
Finally, this reports is written in LaTeX

2 Data Gathering

The tweets were collected thanks to the python script `stream.py`, where, after a Oauth Authentication [8] start a `Stream` object capable of gathering all the tweets submitted in real time by the users to the Twitter Platform thart match some filters. In particular, I have setted the language filter on `"it"`, so Italian, and the keywords (which are about the word *vaccine* in Italian and the names of some pharmaceutical enterprises) as follow:

```
WORDS = [
    'vaccino dosi',
    'vaccino',
    'vaccini',
    'vaccinazione',
    'vaccinato',
    'vaccinati',
    'pfizer',
    'biontech',
    'moderna',
    'astrazeneca',
    'curevac'
]
```

It is intersting to notice that for example the first set of keywords, `'vaccino dosi'` (*vaccine* and *doses*), is a superset of the second keyword, `'vaccino'`. I have repeated the keyword `'vaccino'` because the keyword `'dosi'` alone is a generic word that can also not be related to the Covid-19 pandemic.

This is also the case of each keywords, because a tweet containing a keyword like *'vaccine'* can be related both to the pandemic or the vaccination of a stray cat rescued by volunteers. Due to the media-bombing with pandemic news, I have assumed all the tweets to be related to the vaccine against Covid-19, and I have chosen proper keywords to be sure of this assumption.

To resolve the problem of disconnection from the API due to the exceed of a rate limit of tweets streamed, I have coded a very simple shell script (`stream.sh`) that in case of disconnection wait some seconds and then relaunch the python script.

So at the end I have collected **415432 tweets** in a span time of **7 days**.

3 Data Description

As described very well in the official documentation, the Tweet object [9] has a long list of 'root-level' attributes, including fundamental attributes such as `id`, the unique identifier of the tweet, `created_at`, the UTC time when the tweet

was created, and `text`, that contains the first 140 characters of the content of tweet. This limit of 140 characters was a key feature of the old Twitter, now the length is doubled to 280.

Tweet objects are also the 'parent' object to several child objects. Tweet child objects include `user` [10], `entities` [11], and `extended_entities` [12]. Tweets that are geo-tagged will have a `place` child object [13].

As I have writed in the Notebook, the structure of the Tweet object is affected by some legacy unused fields, for example `geo`, that is deprecated, and by the transformations that the Twitter Platform has undergone in this years. A clear example is the field `text` can hold only 140 characters, the text that exceed that limit is stored in the `full_text` field of the `extended_tweet` substructure, with a flag field, `truncated`, that distinguish from these two lenghts. This legacy can tell the story of metamorphosis of the tweet length: from 140 to 280 characters.

In my analysis I have mainly used the `'created_at'` field because I was interested only in the frequency of the tweets, but with the purpose of showing statistics of the tweets gathered, i have used also:

- `lang` field, that store the machine-detected language of the text content of the tweet.
- `hashtags` field, in the `entities` child object, which as the name suggest store the hashtags of the tweet.
- `retweeted_status` field, that contains a representation of the original Tweet that was retweeted. It is useful to understand if a tweet was a retweet or not, in the latter case the content is `None`.
- `is_quote_status` is a boolean field that indicates whether this is a quoted tweet or not.
- `in_reply_to_status_id` field contain the integer representation of the original tweet's ID, it the tweet is a reply, otherwise `None`.
- `location` field, in the `user` child object, is the user-defined location for this account's profile. Not necessarily a location, nor machine-parseable. I have tried to analyze this field in my analysis, but I have concluded that contains too many missing values and fantasy values to be helpful.
- `screen_name` field, in the `user` child object, handle the name used by the user on twitter. When a user tag another user on a tweet, this tag mechanism involve this field.
- `followers_count` field, in the `user` child object,
- `verified` boolean field, in the `user` child object, indicates whether the user has a verified account [14] (the little blue badge near the name) or

not. The blue verified badge on Twitter lets people know that an account of public interest is authentic. To receive the blue badge, the user account must be authentic, notable, and active and respect a lot of measurable criteria established by Twitter.

4 Data Cleaning

The initial intention was to analyse the frequency of tweets about the vaccine against Covid-19 for each Italian region, or at least areas (North-West, North-East, Center, South Italy and Islands), but as I said before in the Data Description section, the location field contains too many missing or invented values, so I have decided to drop the informations on location, also because the geo-tagged tweets were a insignificant amount of the total.

The time field of the tweet I have used in my analysis haven't missing values. All the others fields that can contain missing values, or as the Twitter official documentation state, that are *Nullable* [9], contains as missing value `null`, that Python convert in the corrisponding `None`.

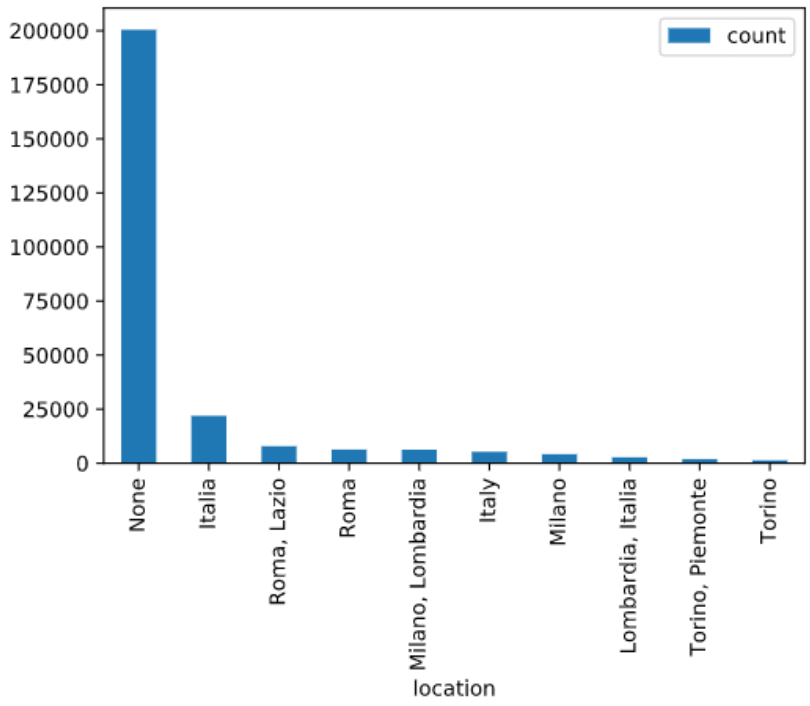

Figure 3: Statistics about the `location` user-defined attribute: about half of tweets lack this information

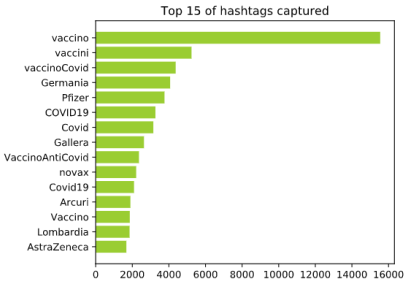

Figure 4: Top 15 hashtags

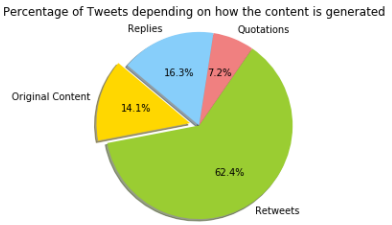

Figure 5: Percentage of retweets, replies, quotations or original tweets

5 Data Analysis

5.1 Statistics

In the first part of my analysis I have computed some statistics of the dataset.

- **Language**: I have verified that all the 415432 are classified with the Italian language by Twitter. Assuming that only Italians and Swiss citizens of the southern cantons speaks Italians, I have assumed that the majority of the tweets come from Italian citizens, due to the fact also that the Italian population was approx 60 millions in 2019 according to the Italian Institute of Statistics [15] while Swiss speaking Italian were approx half million in the observation period of 2016-2018, according to the Swiss Bureau of Statistics [16].

- **Most frequent hashtags**: using existing code I have showed the top 15 of the hashtag captured. As expected the word *vaccino* (vaccine), singolar and plural, were on the first and second position, but some hashtag surprised me, for example *Germania* (Germany), due to the eternal confront between what our country does vs what Germany does, some names as *Gallera*, a controversial regional political man, *Arcuri*, the Emergency Covid Commissioner (the public super-manager on who depends the emergency management of the pandemic), and last but not least *novax*, that refers to the anti-vax movements.

- **Percentage of retweets, replies, quotations or original tweets**: same as the previous, I have used existing code to show the percentage in a pie chart.

- **Locations**: I have also analysed the locations of the users, but without achieving a significant result, as explained in the previous sections.

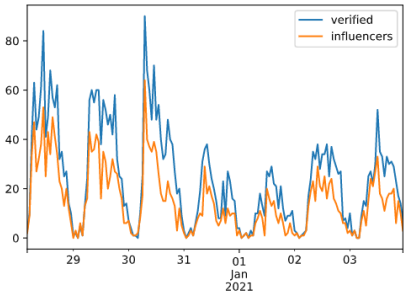

Figure 6: Trend of the frequency of influencers and verified users tweets with vaccine-related keywords

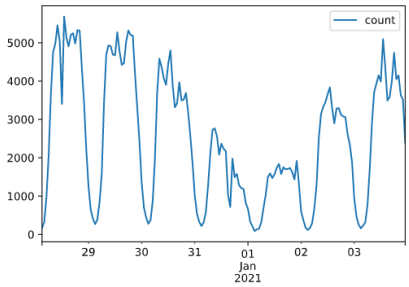

Figure 7: Trend of the frequency of all tweets with vaccine-related keywords

5.2 Influencers and Verified tweets trend

I have computed the top 200 of the people with the highest number of followers in order to compare the trend of the "influencers" activity over the week, compared to the activity of people with the verified profile. A lot of "influencers" are also verified users. The trend is similar as expected, but compared with the general trend computed with all the tweets, shows a high activity before the new year, but without the increase trend showed in the general trend. Some hypothesis which I have made are the following: people for sure enjoyed the New Years Eve without caring about problems or solutions, like Covid-19 and vaccine, and the lack of the reprise can be linked with the publications of vaccine data by the government in the dashboard mentioned in the section before. These hypotheses need to be confirmed with future research that I have not addressed in this project.

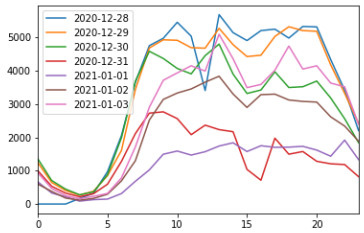

Figure 8: Hourly Analysis of the frequency of all tweets: we can clearly identify the launch break

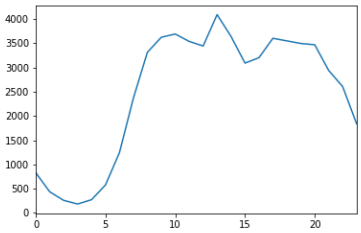

Figure 9: Hourly Analysis of the frequency of all tweets, average of the week

5.3 Hourly Analysis of the frequency of all tweets

The frequency of the tweets are very similar in all the days, with the exceptions of the 1st of January and the last hours of the 31 December, near the New Years Eve. It is interesting to notice the the downhill of the tweets near 15:00. At first I didn't understand why this hour, but then reading on the docs I have discovered that the `created_at` field that I have used for the analysis provides time in UTC standard, and so due to the fact that Italy lies on UTC+1 timezone, I can clear state that people tweet less during the launch break, at least about the vaccine!

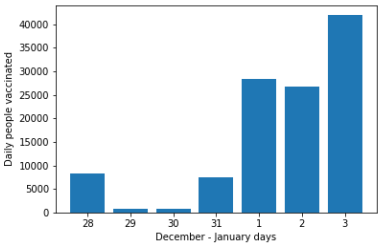

Figure 10: Vaccine administred durigt the week Dec 28th - 3rd Jan

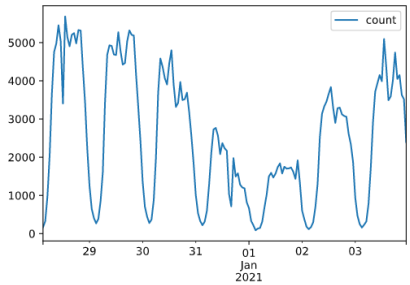

Figure 11: Frequency of the tweets about the vaccine during the week Dec 28th - 3rd Jan

6 Conclusions

I want to briefly conclude the analysis of the data with the observation of the graphics of the vaccine administered during the first week vs the impact on social of the vaccinal campaign. We can clearly see that the first three day, despite the little doses of vaccine administered, the vaccine has become a top trend on Twitter, and this was quite understandable due to the high simbolic meaning of the vaccine. We can notice also that after the 1st of January, the trend is more correlate with the data of the vaccinal campaign.

7 Future improvements

Future improvements surely will involve a better data gathering script, that I have already started to write in the `improvements` folder on the repository. A good approach can be to store the tweets no directly in the MongoDB database, that can be overwhelmed if the number of keywords were much greater, but in a infinite buffer, and and in the meantime have some thread workers that each time remove one item from the buffer and insert into the database.

Another improvement will be a longer observation period in order to gather more useful informations, not polluted by the festivity period.

References

[1] *BDA Final Project repository on Github*. URL: https://github.com/jeorjebot/covid_vaccine_tweets_italy.

[2] *European Medicines Agency*. URL: https://www.ema.europa.eu/en/about-us/who-we-are.

[3] Italian Ministry of Health. *Report Vaccinal Campagin against Covid-19*. URL: http://bit.ly/38gOFpw.

[4] The Italian Governement. *Strateic Plan for the vaccination against SARS-CoV-2/COVID-19*. NOTE: ITA language only. 2020. URL: http://www.salute.gov.it/imgs/C_17_pubblicazioni_2986_allegato.pdf.

[5] Fondazione Luigi Einaudi ONLUS. *Vaccine Data - Quntivaccini Project*. URL: http://bit.ly/3okNT0p.

[6] *Luigi Einaudi Foundation website*. URL: https://www.fondazioneeinaudi.it.

[7] Giorgio Rossi. *Full tweet dataset - JSON format*. URL: http://bit.ly/3hKZu6o.

[8] Twitter Docs. *Twitter Oauth Authentication*. URL: https://developer.twitter.com/en/docs/authentication/oauth-1-0a.

[9] Twitter Docs. *Tweet Object: Standard v1.1*. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.

[10] Twitter Docs. *User Object: Standard v1.1*. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user.

[11] Twitter Docs. *Entities Object: Standard v1.1*. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/entities.

[12] Twitter Docs. *Extended Entities Object: Standard v1.1*. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/extended-entities.

[13] Twitter Docs. *Geo Object: Standard v1.1*. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo.

[14] Twitter Docs. *Verified Account*. URL: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

[15] ISTAT - Italian Institute of Statistics. *National Demographic Balance - year 2019*. URL: https://www.istat.it/en/archivio/245600.

[16] BFS - Federal Bureau of Statistics Swiss. *Languages usually spoken at home - year 2016-2018*. URL: https://www.bfs.admin.ch/bfs/en/home/statistics/population/languages-religions.assetdetail.12228973.html.