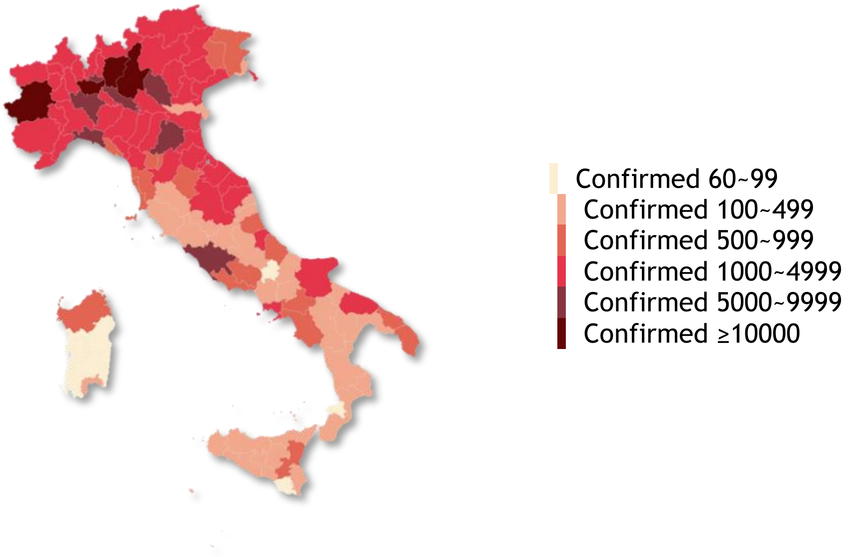
Tzvi Puchinsky
Alla Kitaieva
Or Shalit

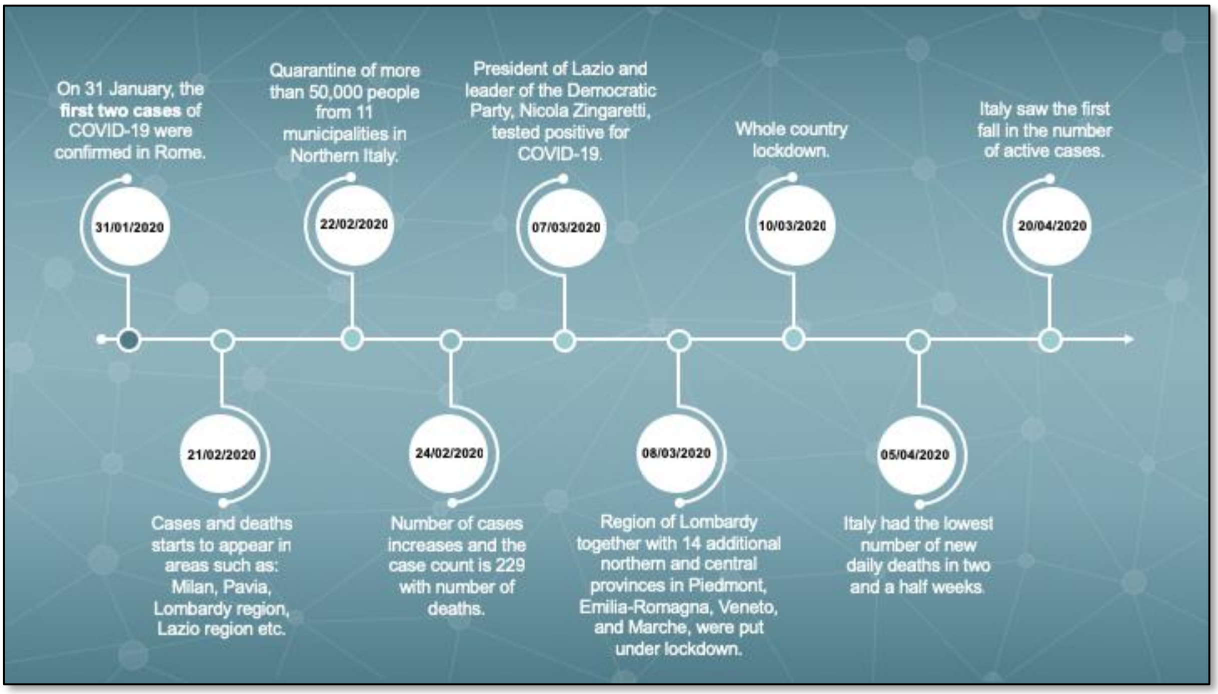
#coronavirus in Italy - how you feel about it?

The COVID-19 virus that started in China affected the world and our daily routine. One of the countries that felt the COVID-19 in the worst case scenario is Italy. With rapid infection rate and high death toll.
As part of our studies we decided to use our knowledge to see if we can analyze the overall mood of twitter users in Italy during specific events and specific areas.



To achieve this goal number of steps implemented:

1. **Create a timeline**
   The first step is to establish a timeline of events, from the first confirmed case through number of key events and the progression of cases in Italy. Using the Wikipedia article *"COVID-19 pandemic in Italy"* a rough timeline established to use as a guideline for tweets scraping.



31/01/2020
On 31 January, the **first two cases** of COVID-19 were confirmed in Rome.

21/02/2020
Cases and deaths starts to appear in areas such as: Milan, Pavia, Lombardy region, Lazio region etc.

22/02/2020
Quarantine of more than 50,000 people from 11 municipalities in Northern Italy.

24/02/2020
Number of cases increases and the case count is 229 with number of deaths.

07/03/2020
President of Lazio and leader of the Democratic Party, Nicola Zingaretti, tested positive for COVID-19.

08/03/2020
Region of Lombardy together with 14 additional northern and central provinces in Piedmont, Emilia-Romagna, Veneto, and Marche, were put under lockdown.

10/03/2020
Whole country lockdown.

05/04/2020
Italy had the lowest number of new daily deaths in two and a half weeks.

20/04/2020
Italy saw the first fall in the number of active cases.

2. **Define queries and geo locations.**
   The goal is to analyze specific regions and Italy in general, to do so specific queries needed. Using python project Twint to download tweets with geo-location where the location is the center and define radius from that location to collect tweets in time range. All the queries used the same query hashtag: *#coronavirus*

   The queries parameters table:

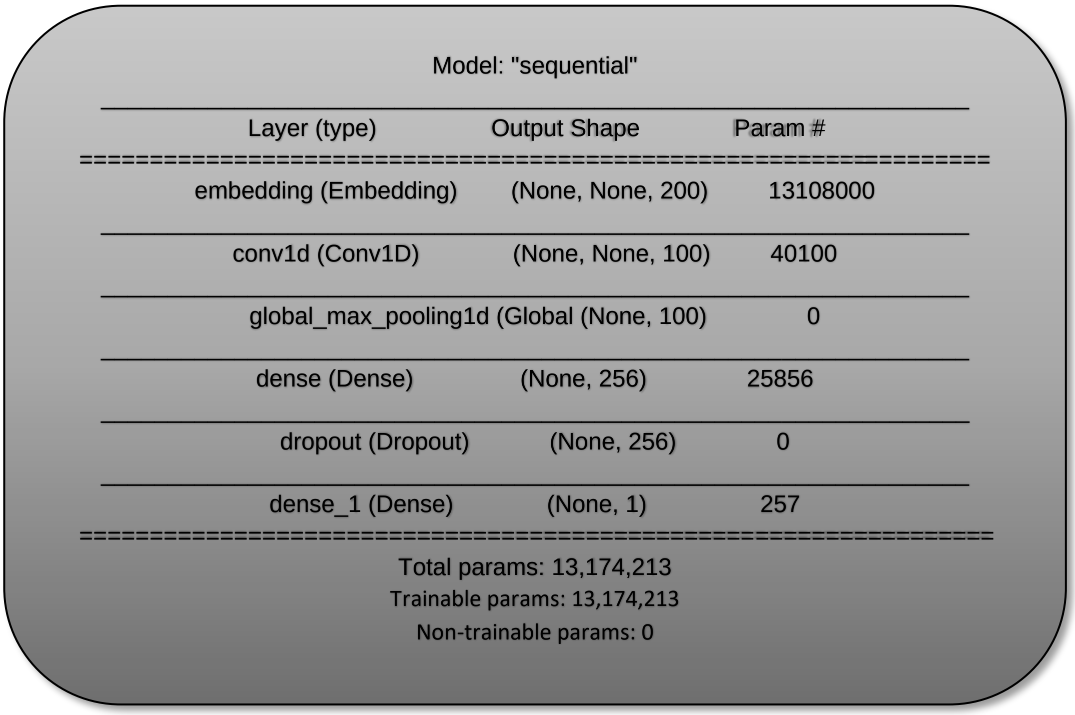
Date	Location	Geo	Radius	Description
31/01/2020	Rome	41.902782,12.496365	15km	Location
21/02/2020	Veneto	45.735802,11.861790	82km	Location
21/02/2020	Casalpusterlengo	45.177811,9.650170	15km	Location
21/02/2020	Lombardy	45.657581,9.963600	100km	Location
21/02/2020	Milan	45.464203,9.189982	15km	Location
21/02/2020	Pavia	45.191652,9.172623	15km	Location
21/02/2020	Emilia-Romagna	44.551499,10.926799	100km	Location
21/02/2020	Piedmont	45.082506,7.919908	100km	Location
22/02/2020	Rome	41.902782,12.496365	500km	Event
23/02/2020	Crema	45.378779,9.682331	15km	Location
23/02/2020	Pavia	45.191652,9.172623	15km	Location
24/02/2020	Bergamo	45.692919,9.685191	15km	Location
24/02/2020	Codongo	45.163569,9.704778	15km	Location
25/02/2020	Como	45.808555,9.095428	15km	Location
07/03/2020	Rome	41.902782,12.496365	500km	Event
08/03/2020	Rome	41.902782,12.496365	500km	Event
10/03/2020	Rome	41.902782,12.496365	500km	Event
05/04/2020	Rome	41.902782,12.496365	500km	Event
20/04/2020	Rome	41.902782,12.496365	500km	Event

3. **Download the data and arrange it**
   Using the *Twint* each query used to get as much as possible tweets with the parameters defined. Each query results saved in a separate csv file for further processing. The largest file contain over 11000 tweets where the smallest contained only 8 tweets.

4. **Pre-processing and classifying tweets**
   Using a trained convolution neural network to classify the tweets into two separated classes: *negative* and *positive* tweets.
   Each tweet is pre-processed to eliminate any unwanted signs and characters:
   - Removing @.
   - Removing URL links.
   - Keeping only letters.
   - Removing additional whitespaces.

   The final result is that we have for each location query a set of pre-processed tweets and their classification to *negative* or *positive* value.

   The CNN model structure that was used to classify the tweets:



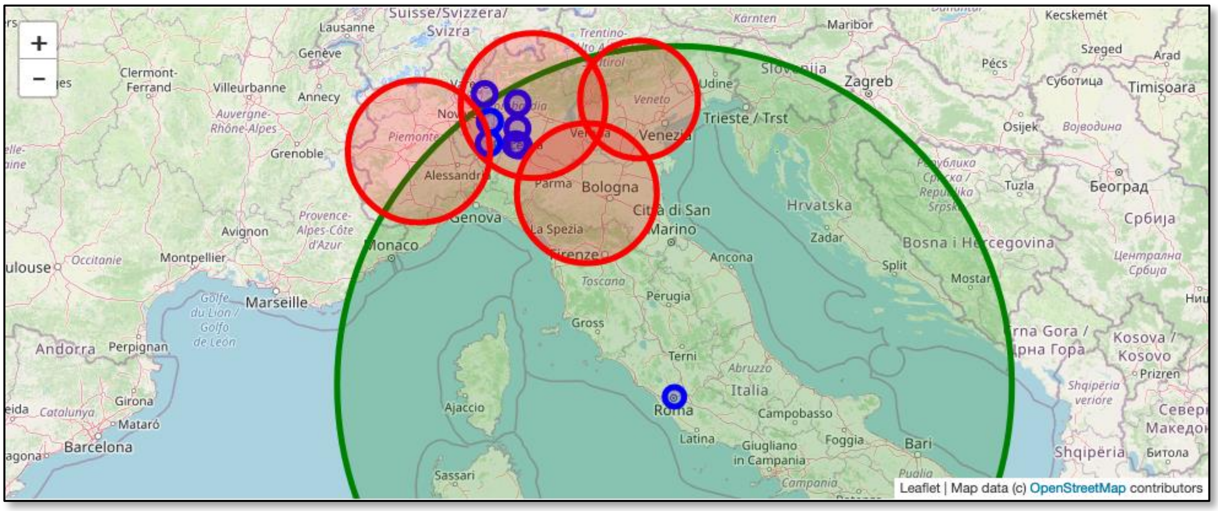
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 200)         13108000
_________________________________________________________________
conv1d (Conv1D)              (None, None, 100)         40100
_________________________________________________________________
global_max_pooling1d (Global (None, 100)               0
_________________________________________________________________
dense (Dense)                (None, 256)               25856
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 257
=================================================================
Total params: 13,174,213
Trainable params: 13,174,213
Non-trainable params: 0
```

Loss: 0.4070 - Accuracy: 0.8143

5. Visualize the data

a. Query map

To better understand the results decided to first show the areas that we searched for tweets. For the map visualization the *ipyleaflet* [1]python package used in jupyter notebook.



This is the query map, each circle is a query with a center location and different radius.
The color difference is just to distinguish between radius sizes.

[1] https://ipyleaflet.readthedocs.io/en/latest/

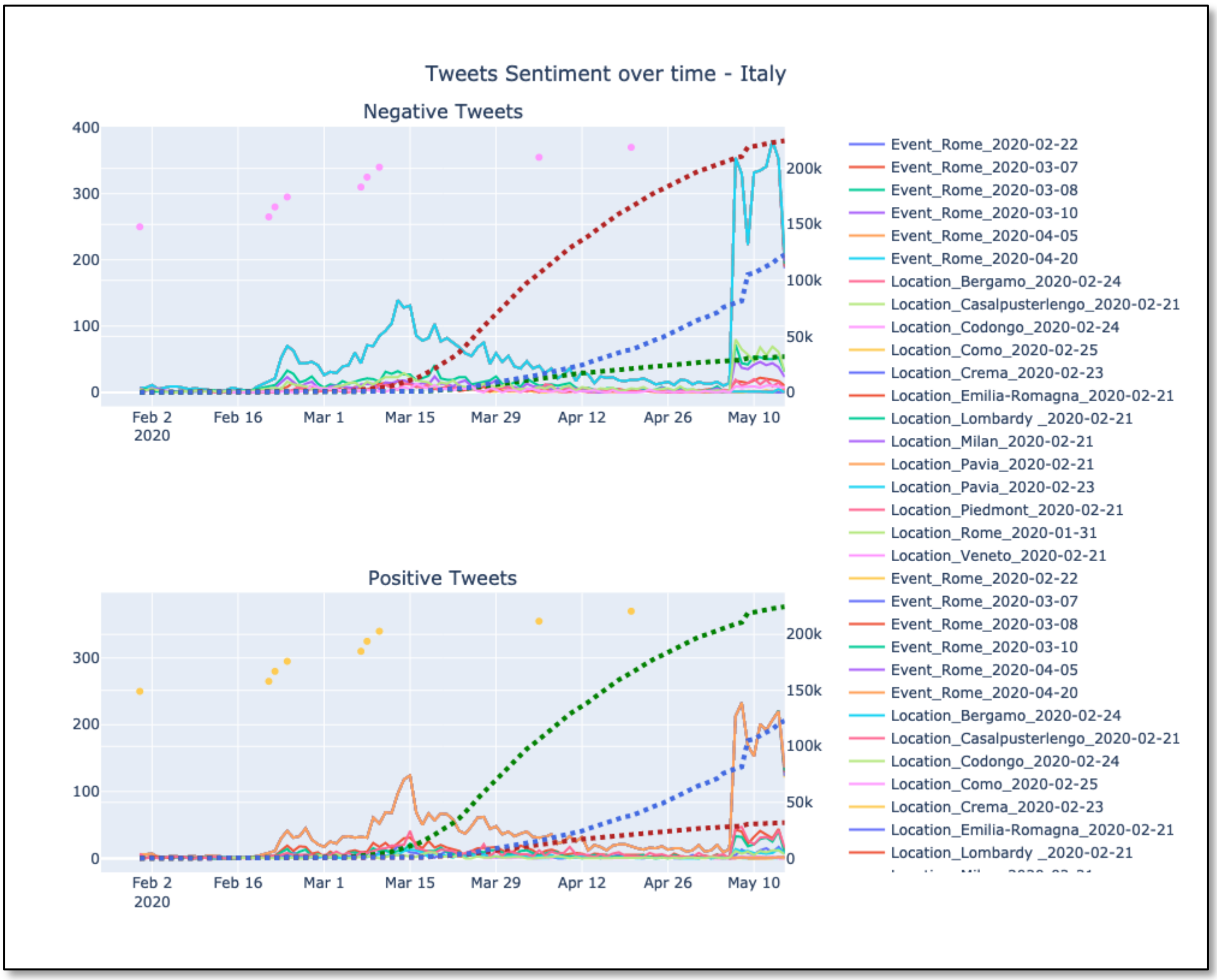
b. Tweets sentiment over time

For plotting the results over graphs the plotly[2] python package used in jupyter notebook to provide an interactive graphs to analyze the results in the best way.

The graphs separated into two groups:

- Negative tweets.
- Positive tweets.

Also, to provide a better picture a second axis added using the *NOVELCovid/API* [3] to download historical data of the COVID19 case, recovered, death status on each day in Italy.



The interactive version can be seen in the jupyter notebook html provided.

Note: The small upper dots that you see on the graph are key events during the timeline as defined in the first section.

[2] https://plotly.com/

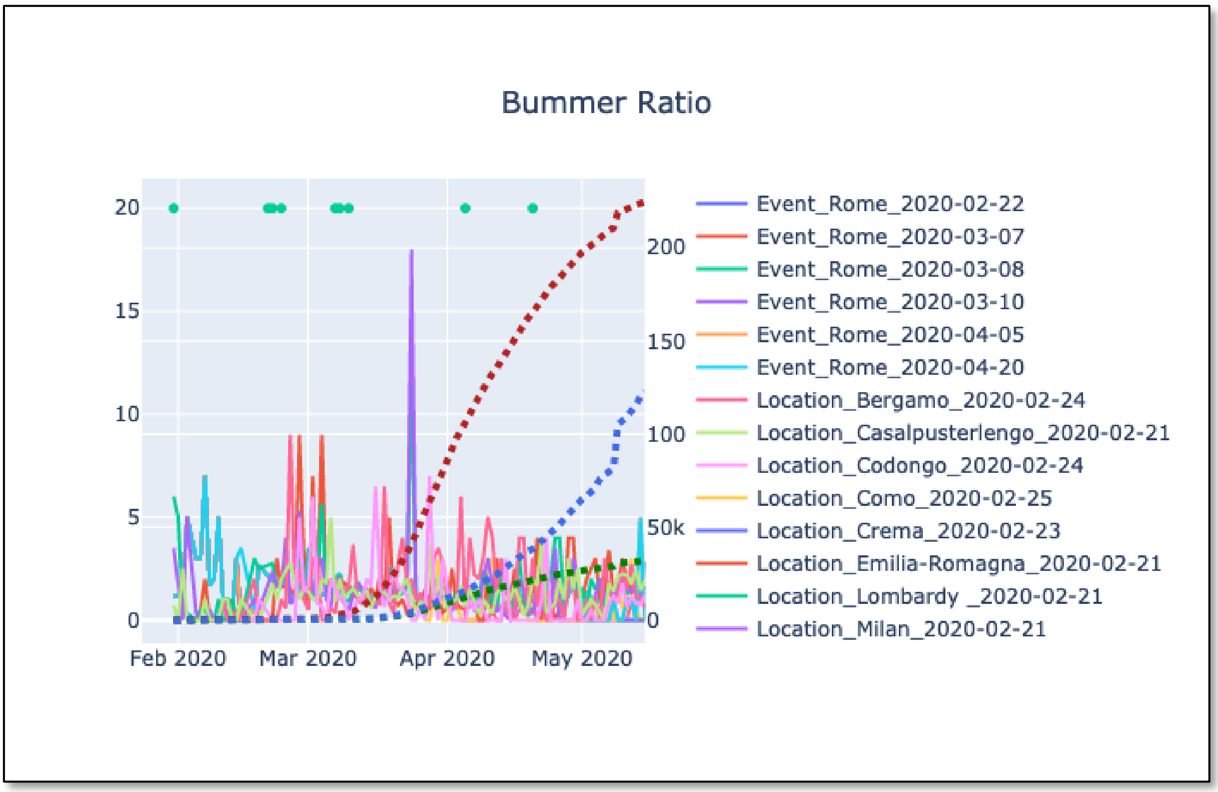
[3] https://corona.lmao.ninja/docs/

c. **Bummer ratio**

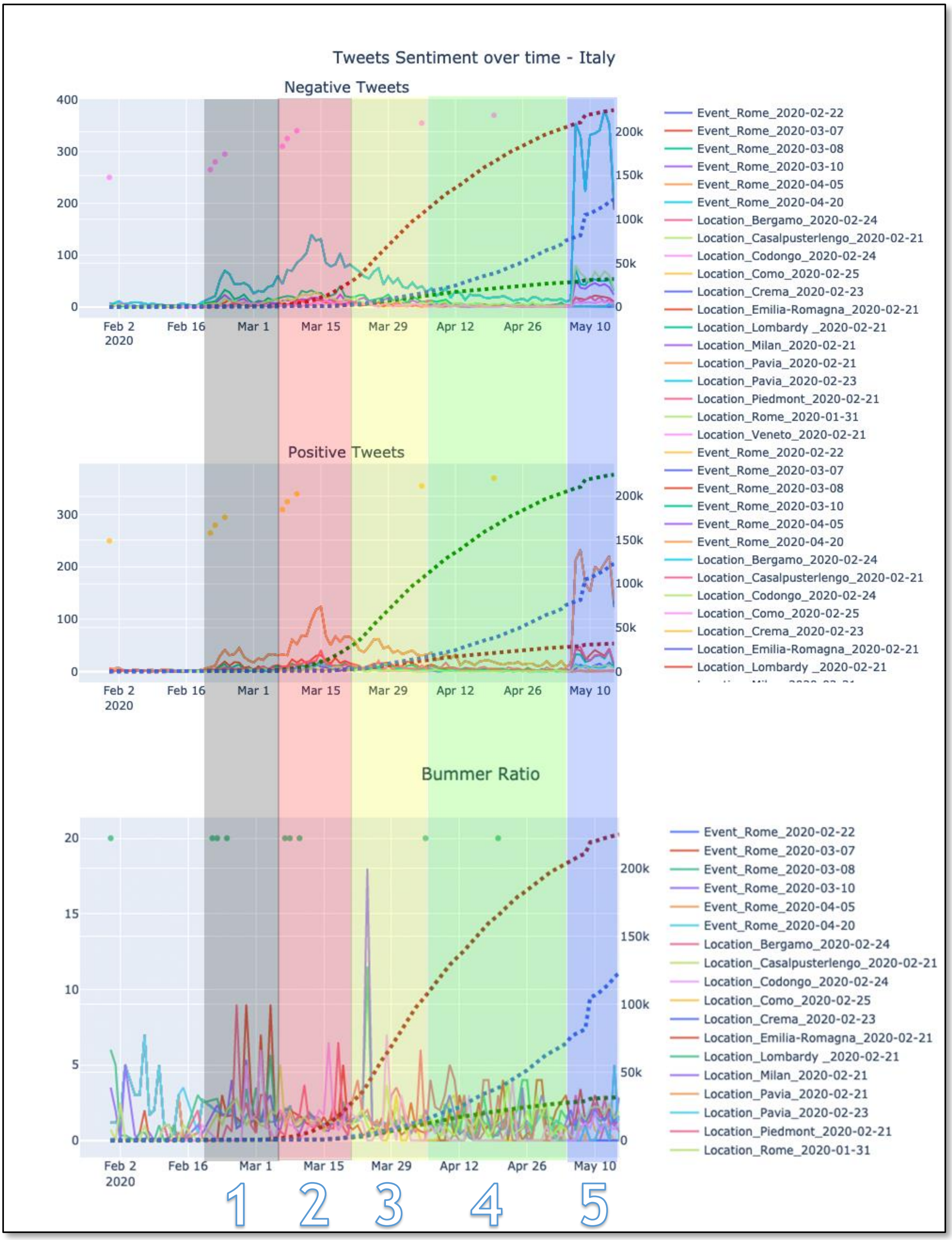
Using the results gathered additional ratio is introduced, the bummer ratio defined as:

$$bummer\ ratio = \frac{Negative\ tweets\ count}{Positive\ tweets\ count}$$

This "bummer ratio" is calculated for each query result for each day and then graphed on a separate graph with same secondary axis.



6. Conclusions



a. Part 1 – 19/02/20 – 06/03/20

Major events:

- 21/02/20 – Increase in number of cases.
- 22/02/20 – Quarantine of more than 50000 people.
- 24/02/20 – Number of cases continues to increase rapidly.
- 04/03/20 - Educational institutions were closed

In the first part of the timeline we can see the effects of those major events on the amount of tweets with #coronavirus and how the COVID-19 gaining interest in the tweeter community. In the “bummer ratio” we can see the spikes that means strong negative feelings about the events.

b. Part 2 – 07/03/20 – 22/03/20

Major events:

- 07/03/20 – On 7 March, President of Lazio and leader of the Democratic Party, Nicola Zingaretti, tested positive for COVID-19. Ten days before, he was in Milan attending public events. The following day, President of Piedmont Alberto Cirio also tested positive.
- 08/03/20 - Starting on 8 March, the region of Lombardy together with 14 additional northern and central provinces in Piedmont, Emilia-Romagna, Veneto, and Marche, were put under lockdown. Italy comes first in terms of mortality due to coronavirus. Every 20th sick person dies.
- 10/03/20 - Whole country lockdown.
- 19/03/20 – Italy surpassed China in the number of deaths due to coronavirus.

In the second part of the timeline we can see a further increase in tweets regarding the COVID-19. If we observe the “bummer ratio” we can see that the spikes are lower meaning that the number of negative/positive ratio is getting smaller.

c. Part 3 – 23/03/20 – 06/04/20

Major events:

- 23/03/20 – The first person who spent a long time in critical condition in intensive care is released.
- 24/03/20 – A press conference was held with the Prime Minister of Italy. They increased the fine for non-compliance with measures and refused to extend the state of emergency.
- 05/04/20 – Lowest number of daily deaths.

In the third part of timeline we can see a start of decrease in the amount of tweets regarding the COVID-19. The specific spike that we see in the “bummer ratio” may be connected to the press conference and the announcements of the new regulations.

d. **Part 4 – 07/04/20 – 03/05/20**

Major events:

- 20/04/20 – The first fall in the number of active cases.

During the fourth part of the timeline we can the lowest number of tweets regarding the COVID-19. We can assume that the population learned to “live” with the virus the new reality.

e. **Part 5 – 04/05/20 – 15/05/20**

Major events:

- 04/05/20 – Italy lifts quarantine restrictions.
- 06/05/20 – Record number of recovered, more than 8000 per day.

In the fifth part of the timeline we can see a significant rise in number of tweets regarding the COVID-19 largely due to this two major events that changed again the day-to-day routine. Interesting to note that the “bummer ratio” is lowest during this part of the timeline and the meaning is that the negative/positive difference between the number of tweets is much closer. Lower “bummer ratio” means that the population of Italy is much happier and this corresponds to this two events during the fifth part of the timeline.