

# Studying the emotional impact of the Covid-19 pandemic using social media

Bachelor Degree Thesis Presentation, (TeX)

---

Simone Alghisi

June 7, 2021

Università degli Studi di Trento



Project description

Dataset and tweets preprocessing

Sentiment analysis over time - by language

Sentiment analysis over time - by gender

Sentiment analysis over time - by region

Conclusions

Tools used

Bibliography

## Project description

---

The COVID-19 pandemic is having a huge impact on our lives, that goes beyond the direct effects of the virus. Besides the fear of infection, lockdown measures adopted by many countries are limiting the possibility to move, work, have contact with others, and are creating a situation of economic crisis and generalized uncertainty about the future. The psychological effects of this unprecedented situation need to be studied.

The project consisted in an **analysis of emotions as emerging from Twitter messages** during the pandemic.

**Lexicon-based sentiment analysis tools** have been employed to characterize emotions associated with content on a large scale. This could allow us to contrast the emotional reaction with the evolution of contagions and deaths, and with the different lockdown and de-escalation stages, in different areas.

## Dataset and tweets preprocessing

---



The tweets to analyze were retrieved from **echen102/COVID-19-TweetIDs** GitHub repository<sup>1</sup> and relate to the period **from January 2020 to March 2021**.

Below, some general statistics

Number of files	10 402
Number of identified languages	65
Number of tweets	1 055 843 481
Number of unique tweets (no retweets)	323 504 667
Dataset compressed size	865GB
Dataset estimated uncompressed size	6.252TB

---

<sup>1</sup>Emily Chen, Kristina Lerman, and Emilio Ferrara. "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set". In: *JMIR Public Health and Surveillance* 6.2 (2020), e19273.

# Languages with the most tweets



UNIVERSITÀ  
DI TRENTO

Language breakdown of top 10 most prevalent languages :

Language	ISO	No. tweets	% total Tweets
English	en	855,574,399	65.38%
Spanish	es	162,884,466	12.45%
Portuguese	pt	54,469,049	4.16%
French	fr	39,402,060	3.01%
Undefined	und	37,005,463	2.83%
Indonesian	in	33,470,927	2.56%
German	de	22,167,543	1.69%
Japanese	ja	15,410,569	1.18%
Italian	it	14,345,605	1.1%
Turkish	tr	13,396,516	1.02%



We retrieved tweets given their id thanks to Twarc and we kept the most relevant information.

## Final JSON object

```
{
  "id": 1307025659294674945,
  "full_text": "Here's an article that highlights the updates...",
  "lang": "en",
  "created_at": "Fri Sep 18 18:36:15 +0000 2020",
  "retweet_count": 11,
  "favorite_count": 70,
  "user": {
    "id": 2244994945,
    "id_str": "2244994945",
    "screen_name": "TwitterDev",
    "name": "Twitter Dev",
    "description": "The voice of the #TwitterDev team and your official...",
    "location": "127.0.0.1",
    "followers_count": 513958,
    "statuses_count": 3635,
    "default_profile_image": false,
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/1283786620521652229/1E0DkLTh_normal.jpg"
  }
}
```





To access directly to a specific category of tweets (e.g. the tweets written in English on the 3rd of June 2020), we grouped them

- ▶ first according to the **language**, using the `lang` field from Twitter
- ▶ secondly by **year-month**
- ▶ and finally by **day**

Later on, we decided to **aggregate them in weekly batches** for better data visualization and to average the results.

## Sentiment analysis over time - by language

---

We used the NRC Word-Emotion Association Lexicon (aka EmoLex)<sup>2</sup> for sentiment analysis. EmoLex consists of " . . . a list of English words and their associations with **eight basic emotions** (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and **two sentiments** (negative and positive). "

Sets of Categories: A treemap showing the number of words associated with \*sets\* of categories



Word-Sentiment Associations

absent	negative
absentee	negative
absenteeism	negative
absolute	positive
absolution	positive
absorbed	positive
absurd	negative
absurdity	negative
abundance	positive
abundant	positive

Word-Emotion Associations

abacus	trust
abandon	fear
abandon	sadness
abandoned	anger
abandoned	fear
abandoned	sadness

We have chosen this lexicon because it **has been translated in over one hundred languages**.

<sup>2</sup>NRC Emotion Lexicon. URL: <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>. (accessed: 2021/06/03).



We considered **Catalan, English, Italian** and **Spanish tweets** for further analysis.  
This helped to focus on a even more restricted set of data, in particular:

language	tweets number	tweets percentage
Catalan	1 377 225	0.4%
English	195 645 826	60.4%
Italian	5 256 748	1.6%
Spanish	35 533 886	10.9%

**Table 1:** Number of tweets for a specific language

However, our statistics could end up biased because of **particularly active** (or emotional) **users**.

Given that, we decided to follow one of the approaches discussed by Aiello et al.,<sup>3</sup> in particular we have considered

- ▶ **emotions in a binary way** (e.g. whether in a given week the user expressed joy or not)
- ▶ **users over tweets** (e.g. the number of unique users, instead of tweets, that expressed joy in a week)

## Definition

Given  $U_e(t)$ , the number of distinct users that expressed emotion  $e$  at time  $t$  in a tweet, and  $U(t)$ , the number of distinct users that tweeted at time  $t$ ,

$$f_e(t) = \frac{U_e(t)}{U(t)}$$

is the proportion of users that expressed emotion  $e$  at time  $t$ .

---

<sup>3</sup>Luca Maria Aiello et al. *How Epidemic Psychology Works on Social Media: Evolution of responses to the COVID-19 pandemic*. 2020. arXiv: 2007.13169 [cs.CY].

# Italian tweets per week

The graphics below shows the proportion of tweets that express a particular emotion (color) w.r.t. the total number of tweets in a week. The dashed grey line indicates the total volume of tweets normalized w.r.t. the maximum value observed in the period Jan 2020-Apr 2021.

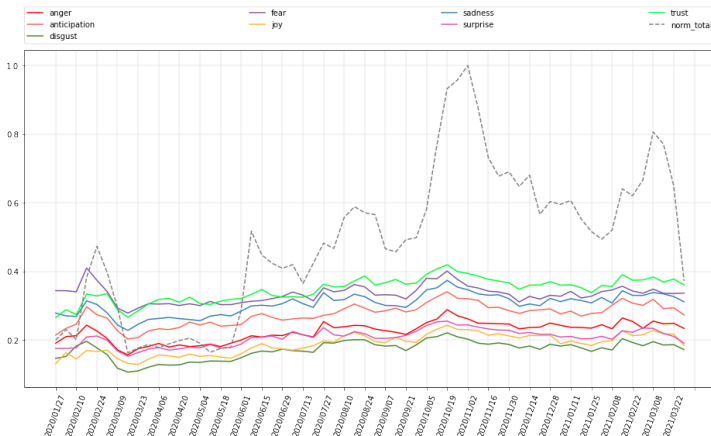


Figure 1: Emotions expressed in Italian tweets per week.

## Italian tweets per week (subplot)

The following graphics are just a variation from the one showed before, in order to have a cleaner visualization of the course of a single emotion.

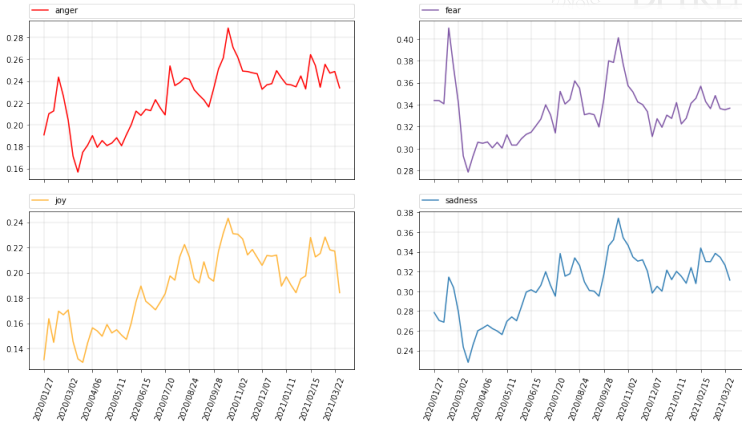


Figure 2: Emotions expressed in Italian tweets per week (single emotion).



Unfortunately,

- fig. 1 emotions seem to be all the same
- fig. 2 is easier to understand, but it is more difficult to determine which emotion has the most impact,
- there is simply no way to determine which days are globally more relevant than others

For this reason, it was decided to normalize the data using  $z_e(t)$  (i.e. z-score).

## Definition

Given  $f_e(t)$  and the period of time  $[0, T]$ ,

$$z_e(t) = \frac{f_e(t) - \mu_{[0, T]}(f_e)}{\sigma_{[0, T]}(f_e)}$$

where  $\mu_{[0, T]}(f_e) = \frac{1}{|T|} \sum_{t=0}^T f_e(t)$ , and  $\sigma_{[0, T]}(f_e) = \sqrt{\frac{1}{|T|} \sum_{t=0}^T (f_e(t) - \mu_{[0, T]}(f_e))^2}$



# Normalized Italian tweets per week

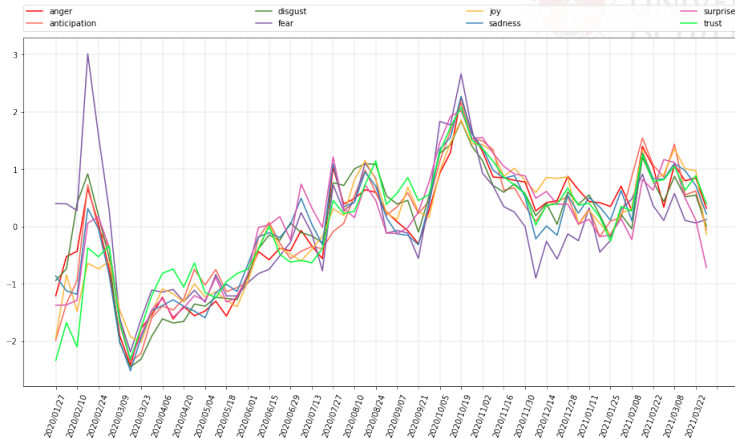


Figure 3: Emotions expressed in Italian tweets per week, normalized w.r.t. the displayed period.



# Most used Italian words 2020/02/17 - 23 (subplot)

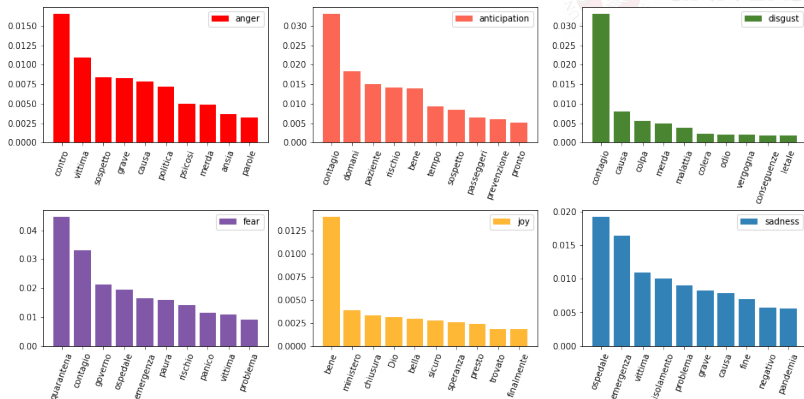


Figure 5: Most used Italian words from 2020/02/17 to 2020/02/23 per emotion #1

## Most used Italian words 2020/02/17 - 23 (subplot)

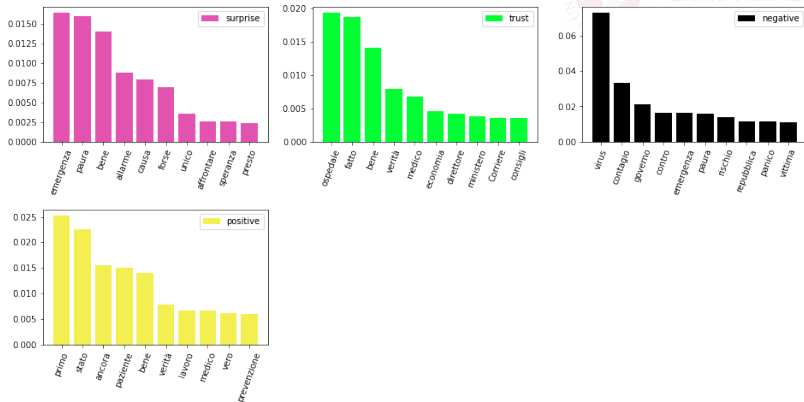
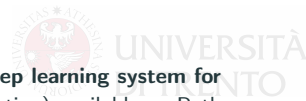


Figure 6: Most used Italian words from 2020/02/17 to 2020/02/23 per emotion #2

## Sentiment analysis over time - by gender

---



To infer data about the users, we used m3inference,<sup>4</sup> a **deep learning system for demographic inference** (gender, age, and person/organization) available on Python.

m3inference bases its results on the analysis of the user's

- description
- name
- screen name
- profile image

## m3inference prediction


```
{  
  "gender": {"male": 0.8758, "female": 0.1242},  
  "age": {"<=18": 0.0053, "19-29": 0.0363, "30-39": 0.9239, ">=40": 0.0346},  
  "org": {"non-org": 0.9965, "is-org": 0.0035}  
}
```

---

<sup>4</sup>Zijian Wang et al. "Demographic inference and representative population estimates from multilingual social media data". In: *The World Wide Web Conference*. ACM. 2019, pp. 2056–2067.



During the project I encountered several errors while using m3inference, for this reason I decided to open a Pull Request on GitHub<sup>5</sup> to contribute to the project.

## fix urllib errors while trying to fetch a profile image from twitter

 Merged computermagcyver merged 1 commit into `euagendas:master` from `Simone-Alghisi:fix-urllib-err` on 13 Apr



Simone-Alghisi commented on 13 Apr

Contributor  

Added more exceptions in preprocess.py to handle `urllib` remaining errors like `ContentTooShortError`, which occurred while I was fetching profile images from Twitter.

The same goes for `ValueError`, which I have encountered when the field `profile_image_url_https` in the twitter json was empty (i.e. `""`)

At last, I have added a line in m3twitter.py to verify if the profile image was successfully downloaded: if that's not the case, `TW_DEFAULT_PROFILE_IMG` is used to avoid crash during the infer phase.

The version of m3inference on the packet manager was updated when the same issue was notified by another user<sup>6</sup>.

---

<sup>5</sup>Pull request: fix urllib errors while trying to fetch a profile image from twitter #20

<sup>6</sup>issue: Error fetching images will fail the infer method #21



## Definition

A user  $u$  belongs to the category  $c \in C$  iif their prediction confidence  $pc$  is greater or equal than 0.95, i.e.

$$u \in c \iff pc(u) \geq 0.95$$

In particular, the following methodology was applied

- ▶ we first check if the user's account belongs to an organization
- ▶ then, if the user is male (or female)
- ▶ finally, if none of the previous constraints were satisfied, we do not consider this user



# Italian tweets per week with user categories

In the graphics below it is possible to observe the emotions course divided by week and also per category. The gray line shows the general course of the emotion (i.e. without considering the division per category)



Figure 7: Emotions expressed in Italian tweets per week and category #1

# Italian tweets per week with user categories



Figure 8: Emotions expressed in Italian tweets per week and category #2



Instead, fig. 9 and fig. 10 below, are meant to show whether a certain category  $c \in C$  expressed at time  $t$  more (or less) emotion  $e$  (e.g. sadness, anger) w.r.t the mean value for emotion  $e$  in the period of time  $[0, T]$ , regardless of the category.

### Definition

Given  $f_{e,c}(t)$ , i.e. the proportion of users belonging to category  $c \in C$  that expressed emotion  $e$  at time  $t$ , and the period of time  $[0, T]$ ,

$$v_{e,c}(t) = \frac{f_{e,c}(t) - \mu_{[0,T]}(f_e)}{\mu_{[0,T]}(f_e)}$$

$$\text{where } \mu_{[0,T]}(f_e) = \frac{1}{|T|} \sum_{t=0}^T f_e(t) = \frac{1}{|T|} \sum_{t=0}^T \sum_{c \in C} f_{e,c}(t)$$

# Italian tweets per week with user categories



Figure 9: Value of the emotions per category w.r.t the average value among all users #1

# Italian tweets per week with user categories

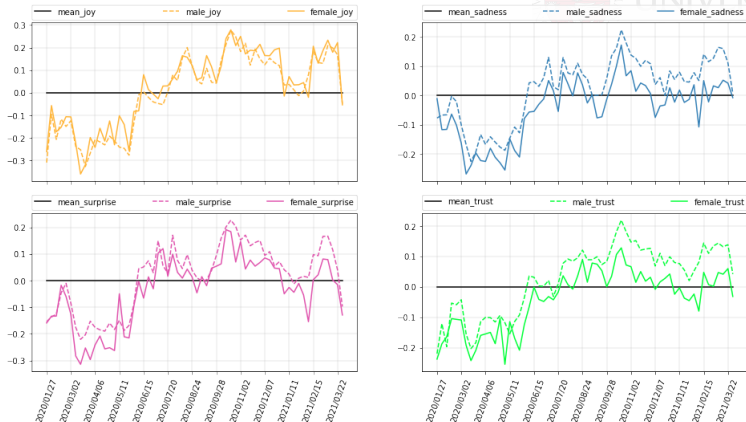


Figure 10: Value of the emotions per category w.r.t average value among all users #2

## Sentiment analysis over time - by region

---



Users on Twitter can specify their location so, for the third sentiment analysis, we thought about **analyzing the emotions of the users from a specific location** (e.g. state, country, ...).

Unfortunately, Twitter does not provide any format or restriction for the location, so

- ▶ not all the users inserted a location
- ▶ some locations could be fake or misspelled
- ▶ the same location could be specified with a different syntax

We linked the users to a specific place through **address geocoding**. Address geocoding is the process of taking a text-based description of a location and returning its geographic coordinates.



Figure 11: OSM Logo

For this task, we decided to use the data made available by **OpenStreetMap (OSM)**,<sup>7</sup> a collaborative project to create a free editable map of the world.

---

<sup>7</sup>OpenStreetMap contributors. *Planet dump* retrieved from <https://planet.osm.org>.  
<https://www.openstreetmap.org>. 2017.





In particular, given a location we used

- **geopy**<sup>8</sup> to contact the Nominatim public API
- **Nominatim**<sup>9</sup> to get the coordinates and the address

## Result obtained given "Milano" as location

```
{
  "place_id": 317098601,
  "licence": "Data \u00a9 OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
  "boundingbox": ["45.3867381", "45.5358482", "9.0408867", "9.2781103"],
  "lat": "45.4668",
  "lon": "9.1905",
  "display_name": "Milano, Lombardia, Italia",
  "address": {
    "city": "Milano",
    "county": "Milano",
    "state": "Lombardia",
    "country": "Italia",
    "country_code": "it"
  }
}
```

---

<sup>8</sup>Python client for several geocoding web services

<sup>9</sup>tool to search through OSM data by name and address

# Italian users distribution in Italy

After assigning to each user location its corresponding state, the following data were available:

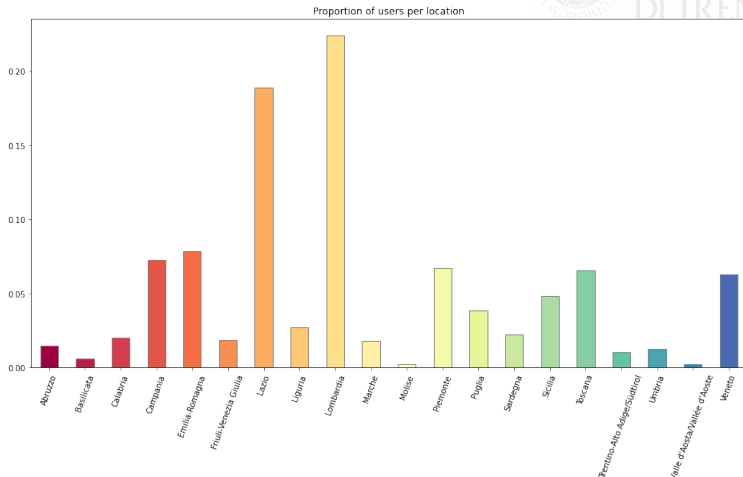
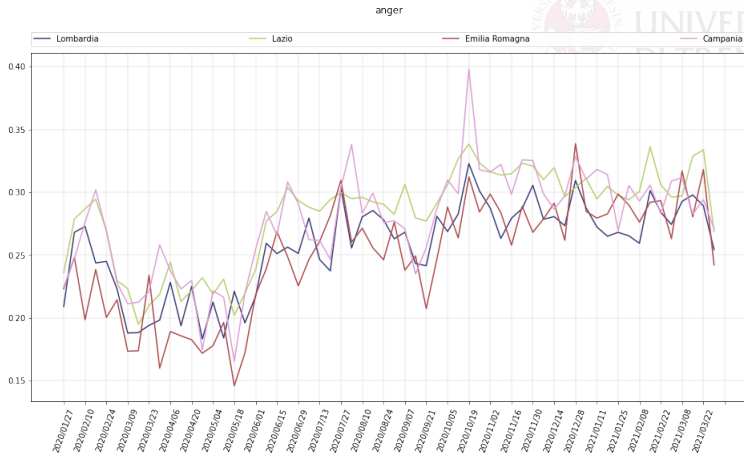


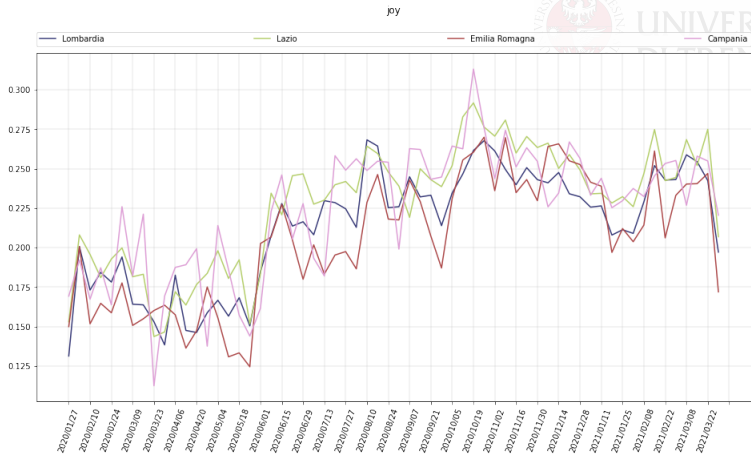
Figure 12: Italian users distribution in Italy per state

# Italian tweets expressing anger per week and state



**Figure 13:** Italian tweets expressing anger per week from Lombardia, Lazio, Emilia Romagna and Campania

# Italian tweets expressing joy per week and state



**Figure 14:** Italian tweets expressing joy per week from Lombardia, Lazio, Emilia Romagna and Campania

## Tweets from Campania expressing anger per week

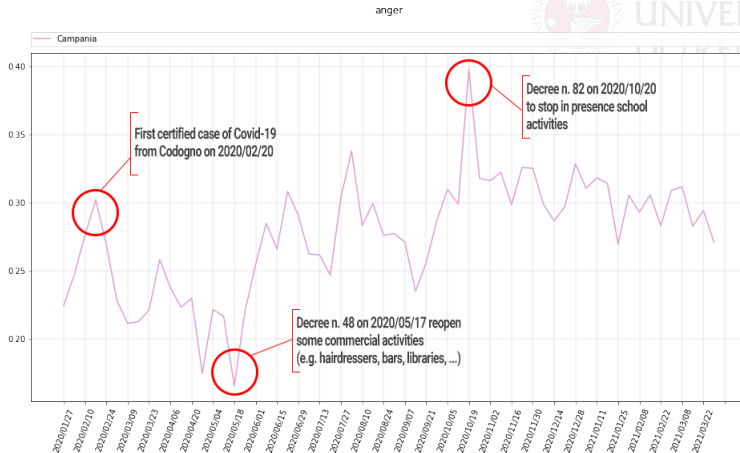


Figure 15: Italian tweets from Campania expressing anger per week

# Tweets from Campania expressing joy per week

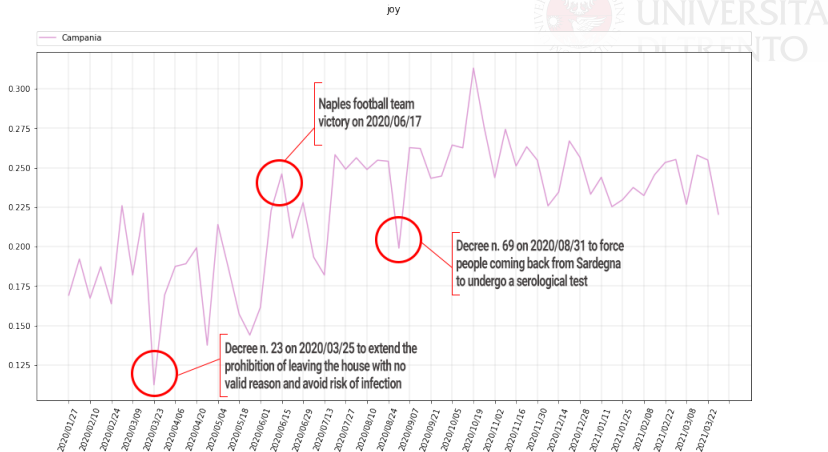


Figure 16: Italian tweets from Campania expressing joy per week

## Conclusions

---

During the project we were able to understand users' emotion in different ways:

- ▶ **categories analysis** showed that women in Italian tweets seems to express more joy through the whole period, and this make sense if we consider the fact that men expressed more negative emotions (e.g. anger, sadness)
- ▶ **locations analysis** was very useful to link certain emotional peaks to real world events

I was only able to scratch the surface of this research field and this impressive amount of data from Twitter, but I hope that my contribution could be a good starting point for further studies.



## Tools used

---



- Python, as main programming language to write the code for the project
- Pandas, to perform small operation on the datasets
- Matplotlib and Plotly, for data visualization
- Twarc, to retrieve (hydrate) tweets from Twitter using TweetIDs
- m3inference, a deep learning system for demographic inference (gender, age, and person/organization)
- geopy and Nominatim, to geocode the locations of the users
- NRC Word-Emotion Association Lexicon (aka EmoLex), to perform sentiment analysis on the tweets of the users

## Bibliography

---

## References

---



UNIVERSITÀ  
DI TRENTO



Aiello, Luca Maria et al. *How Epidemic Psychology Works on Social Media: Evolution of responses to the COVID-19 pandemic*. 2020. arXiv: 2007.13169 [cs.CY].



Chen, Emily, Kristina Lerman, and Emilio Ferrara. "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set". In: *JMIR Public Health and Surveillance* 6.2 (2020), e19273.



*NRC Emotion Lexicon*. URL:

<https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>. (accessed: 2021/06/03).



OpenStreetMap contributors. *Planet dump retrieved from <https://planet.osm.org>*. <https://www.openstreetmap.org>. 2017.



Wang, Zijian et al. "Demographic inference and representative population estimates from multilingual social media data". In: *The World Wide Web Conference*. ACM. 2019, pp. 2056–2067.