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

Bachelor Degree Thesis Presentation, (TeX)

Simone Alghisi

June 6, 2021

Università degli Studi di Trento

Contents

Project description

Introduction to the problem

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

Dataset

The tweets to analyse were retrieved from echen102/COVID-19-TweetIDs GitHub repository.[1] The content of the repository along with their data collection strategy is described as follows:

The repository contains an ongoing collection of tweets IDs associated with the novel coronavirus COVID-19 (SARS-CoV-2), which commenced on January 28, 2020. [...] *We leveraged Twitter's streaming API to follow* ***specified accounts*** *and also collect in* ***real-time tweets that mention specific keywords****.*

[1] 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.

Valid Tweets

The analyzed period was **from January 2020 to March 2021**, for a total of

1 055 843 481 tweets

To avoid bias, we have considered each tweet only once, discarding **retweets**. This reduced the dataset to

323 504 667 tweets

Dateset stats

Number of files	**10 402**
Number of identified languages	**65**
Dataset compressed size	**865GB**
Dataset estimated uncompressed size	**6.252TB**

Table 1: General statistics regarding the dataset

Tweet Structure

Thanks to Twarc, it was possible to retrieve a tweet given its id. After some reasoning, it was decided that only a subset of the fields of the json object needed to be actually saved.

Final JSON object


```json
{
  "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\/
          lEODkLTh_normal.jpg"
  }
}
```

Partitioning

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

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

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

Sentiment analysis over time - by language

Lexicon

For the sentimental analysis the NRC Word-Emotion Association Lexicon (aka EmoLex)[2] was used, which can be defined as *" . . . 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 negative
abundant	positive

Word-Emotion Associations

abacus	trust
abandon	fear sadness
abandoned	anger fear sadness

We have chosen this lexicon because it has been translated in over one hundred languages and could easily fit our problem with **65 identified languages**.

[2] *NRC Emotion Lexicon*. URL: `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`. (accessed: 2021/06/03).

Languages considered

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 2: Number of tweets for a specific language

Methodology

However, sentiment analysis could not be performed using the tweets as they are: in fact, our statistics could end up biased because of **particularly active** (or emotional) **users**.

Following one of the approaches discussed by Aiello et al.,[3] we decided to consider

- **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.

[3] 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.

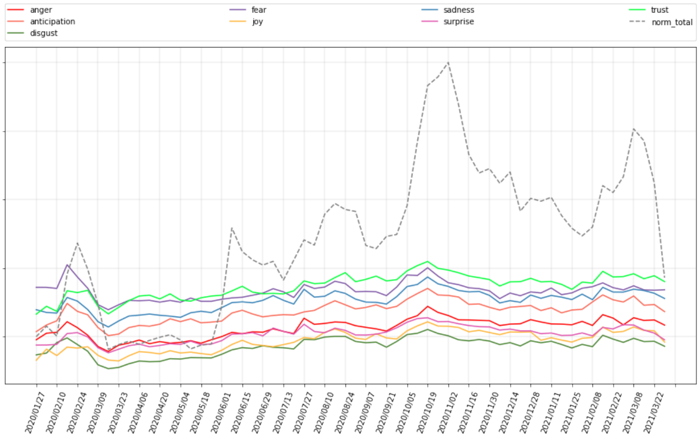

Figure 1: Emotions expressed in Italian tweets per week.

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

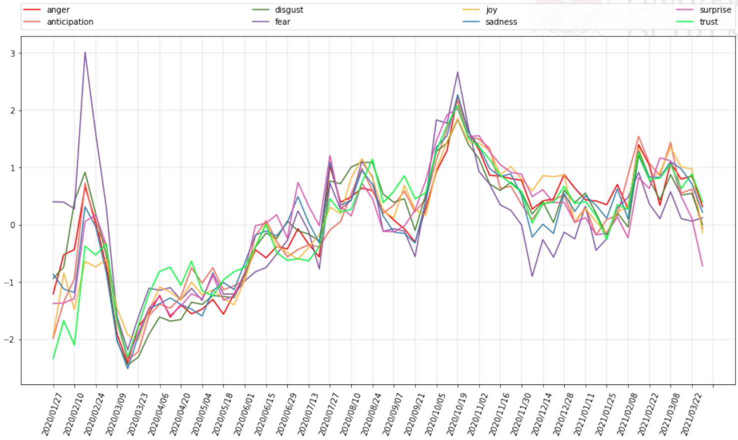

Figure 3: Emotions expressed in Italian tweets per week, normalized w.r.t. the displayed period.

Peaks

If we take a closer look at fig. 3, it is possible to notice the presence of different **peaks**, e.g. the week labeled as 2020/10/19.

We decided to sample some of them manually and conduct a more in depth study to

- visualize and better understand the most used words during that particular week
- validate the lexicon-based method and detect possible bias

Most used Italian words 2020/02/17 - 23

The result of a peak analysis is showed below: notice that the same word in the lexicon has been associated with different emotions. In fact, it is very difficult to understand the sentiment that the user wanted to convey without knowing the context.

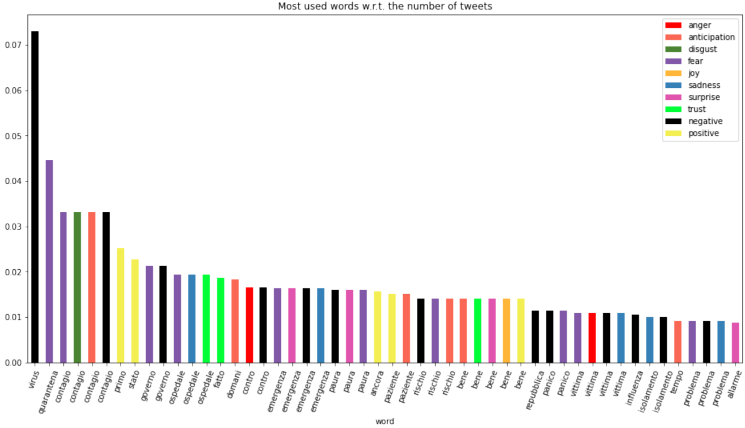

Figure 4: Most used Italian words from 2020/02/17 to 2020/02/23

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

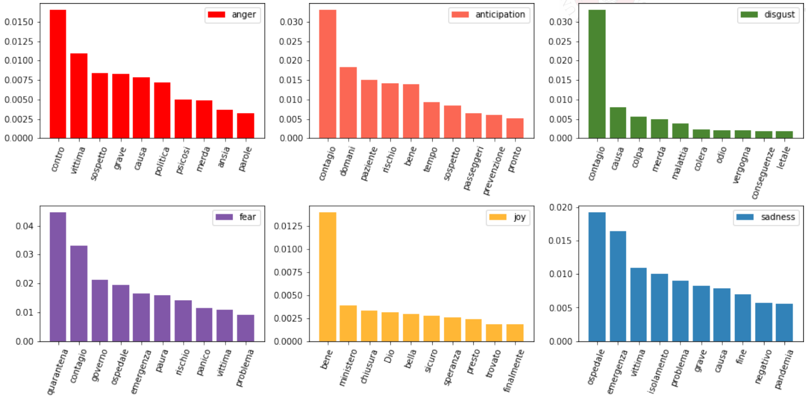

Figure 5: Most used Italian words from 2020/02/17 to 2020/02/23 per emotion #1

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

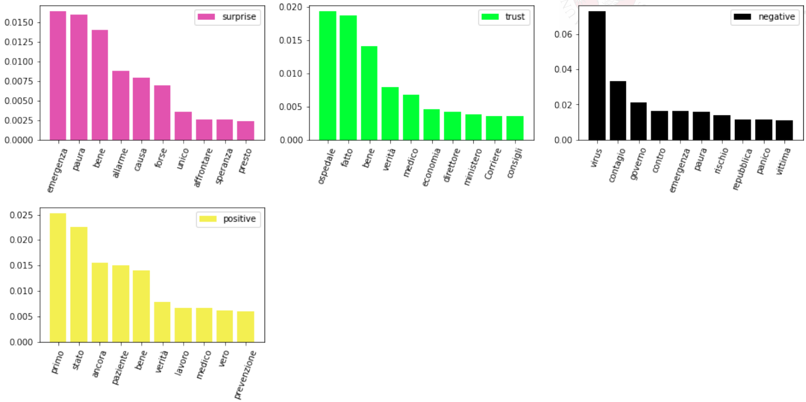

Figure 6: Most used Italian words from 2020/02/17 to 2020/02/23 per emotion #2

Sentiment analysis over time - by gender

Users selection

We decided to **analyze the emotions of the users w.r.t. the gender** to take the research a step further, so we extracted **unique users with more than one tweet** for each analysed language.

Users inference

To infer data about the users, we used m3inference.[4] m3inference is 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

- 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}
}
```

[4] 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[5] to contribute to the project.

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

Merged **computermacgyver** 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 Pull Request was accepted almost immediately, while the version of m3inference on the packet manager was updated when the same issue was notified by another user[6].

[5] Pull request: fix urllib errors while trying to fetch a profile image from twitter #20

[6] issue: Error fetching images will fail the infer method #21

Prediction confidence

We decided to consider valid all the users that respected the following definition:

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 to select valid users:

- check if the user account belongs to an organization by comparing the results obtained from m3inference, otherwise
- check if the user is male (or female) by comparing the results obtained from m3inference, otherwise
- if none of the previous constraints were satisfied, we do not consider this user

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 $[0, T]$,

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

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

Italian tweets per week with user categories

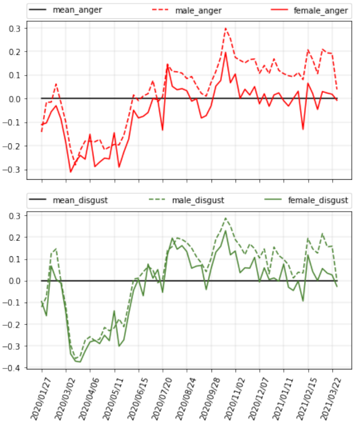

Figure 9: Value of the emotions per category w.r.t the average value among all users #1

Italian tweets per week with user categories

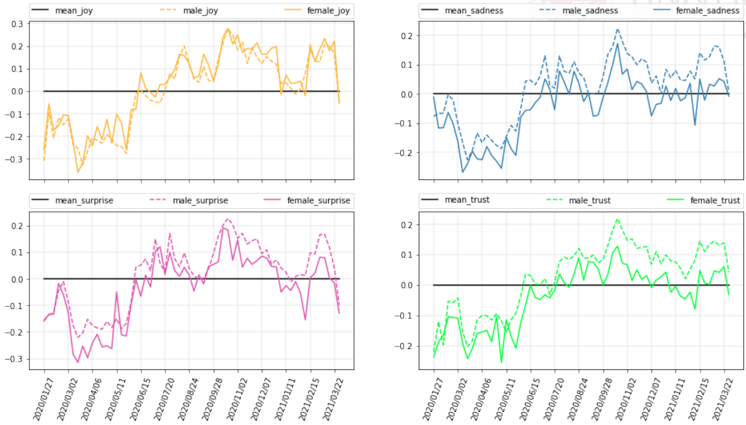

Figure 10: Value of the emotions per category w.r.t average value among all users #2

Sentiment analysis over time - by region

Users locations

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)**,[7] a collaborative project to create a free editable map of the world.

[7] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org. 2017.

Request results

In particular, given a location we used

- **geopy**[8] to contact the Nominatim public API
- **Nominatim**[9] to get the coordinates and the address

Result obtained given "Milano" as location


```json
{
    "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"
    }
}
```


[8] Python client for several geocoding web services
[9] tool to search through OSM data by name and address

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

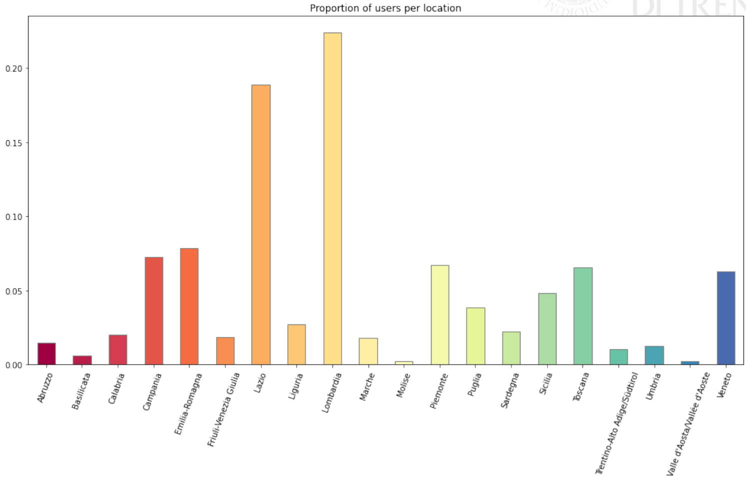

Figure 12: Italian users distribution in Italy per state

From the analysis of fig. 12, we decided to consider **Lombardia, Lazio, Emilia Romagna** and **Campania**, because they could give us stabler results.

The following fig. 13 and fig. 14 show us the weekly course of anger and joy for the four considered states, instead fig. 15 and fig. 16 **focus only on Campania to analyze some peaks**.

Italian tweets expressing anger per week and state

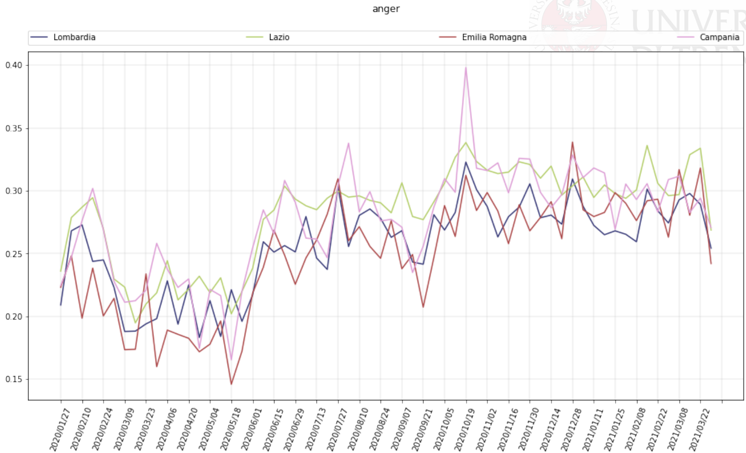

Figure 13: Italian tweets expressing anger per week from Lombardia, Lazio, Emilia Romagna and Campania

Italian tweets expressing joy per week and state

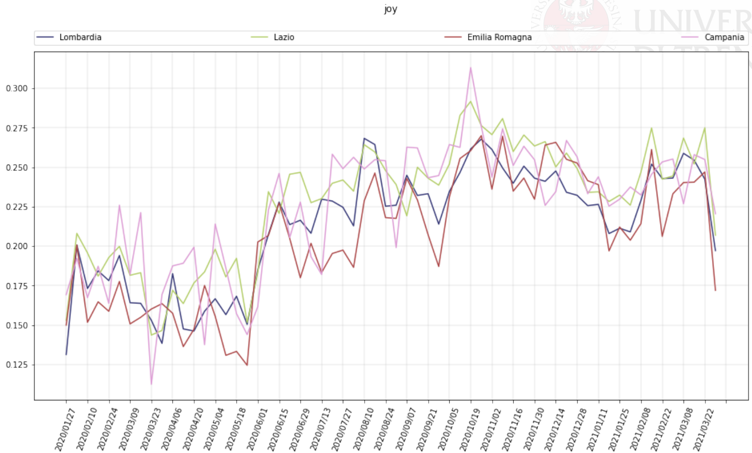

Figure 14: Italian tweets expressing joy per week from Lombardia, Lazio, Emilia Romagna and Campania

Tweets from Campania expressing anger per week

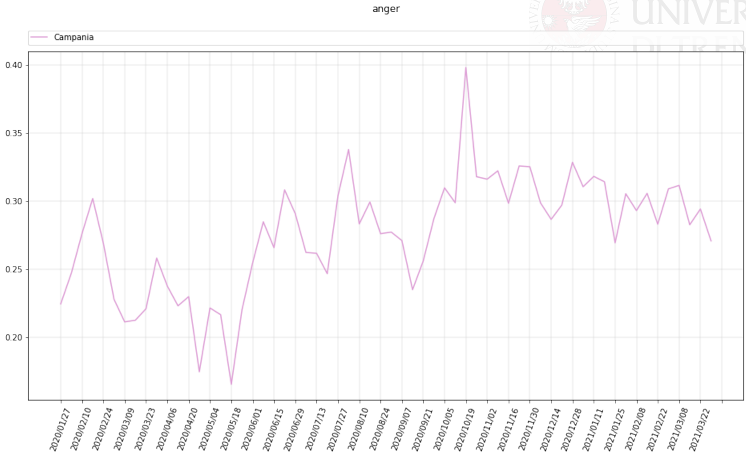

Figure 15: Italian tweets from Campania expressing anger per week

It is possible to associate the events on week

- 2020/02/17 with Decree n. 1 on 2020/02/24 to alert people about the pandemic and reduce contacts as much as possible
- 2020/05/18 with Decree n. 48 on 2020/05/17 to reopen some commercial activities (e.g. hairdressers, bars, libraries, ...)
- 2020/10/19 with Decree n. 82 on 2020/10/20 to stop in presence school activities

Tweets from Campania expressing joy per week

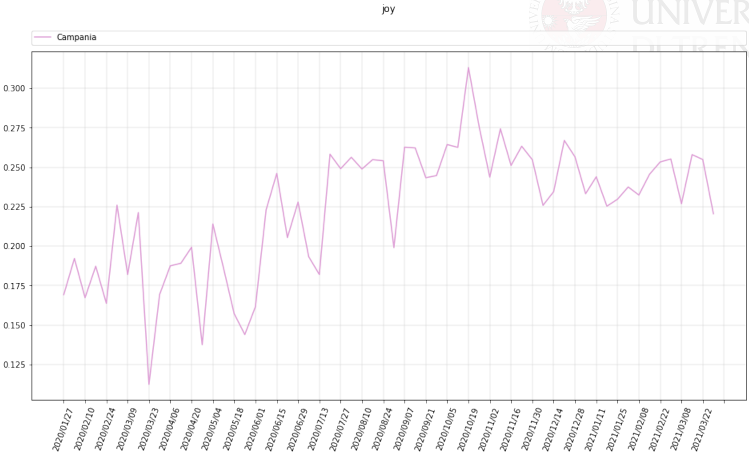

Figure 16: Italian tweets from Campania expressing joy per week

Joy analysis of Campania

It is possible to associate the events on week

- 2020/03/23 with Decree n. 23 on 2020/03/25 to extend the prohibition of leaving the house with no valid reason and avoid risk of infection
- 2020/06/15 with Naples football team victory on 2020/06/17
- 2020/08/31 with Decree n. 69 on 2020/08/31 to force people coming back from Sardegna to undergo a serological test

Conclusions

Final comments

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

Tools and Libraries 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

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.