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

Bachelor Degree Thesis Presentation, (TeX)

Simone Alghisi

Supervisor: Alberto Montresor

Co-Supervisors: Cristian Consonni, David Laniado

June 26, 2021

Università degli Studi di Trento

Contents

Introduction

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

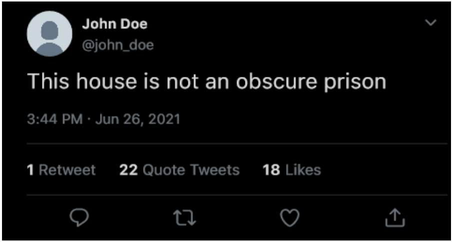


Figure 1: An example of a tweet.

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.

Also called emotion recognition, is the **process of identifying human emotions**.[1] To solve this task, it is possible to use lexicon-based techniques, where each word is assigned to a set of zero or more emotions/sentiments.

Figure 2: Emotion detection for a particular sentence.

[1] Wikipedia contributors. *Emotion recognition — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Emotion_recognition&oldid=1023798177. [Online; accessed 14-June-2021]. 2021.

Data collection

Dataset description and stats

The **echen102/COVID-19-TweetIDs** GitHub repository contains an ongoing collection of tweet IDs starting on January 28th, 2020.[2]

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	865 GB
Dataset estimated uncompressed size	6.252 TB

Table 1: Dataset general statistics

[2] 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.

Tweet Structure

We decided to analyze the tweets **from January 2020 to March 2021**, keeping only the most relevant information.


```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"
  }
}
```


Listing 1: Final json object for a tweet

Methods

We used

- **EmoLex**, 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),[3] to detect the emotions
- **LIWC**, a widely used computerized text analysis program that outputs the percentage of words in a given text that falls into one or more **linguistic, psychological, and topical categories**,[4] to validate the results

[3] Saif Mohammad. *NRC Emotion Lexicon*. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Online; accessed 13-June-2021].

[4] Wikipedia contributors. *James W. Pennebaker — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=James_W._Pennebaker&oldid=1023542720. [Online; accessed 24-June-2021]. 2021.

To correctly **detect the users' emotions**, we decided to follow one of the approaches discussed by Aiello et al.[5]

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

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.

[5] 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].

Emotion detection over time

We decided to analyze for the whole time frame the emotions of different set of users. In particular we grouped them

- by **language** (e.g. Catalan, Italian, English, Spanish, . . .)
- by **gender** (male or female)
- by **age** (≥ 40 or < 40)
- by **location** (e.g. per state, country, county, . . .)

The issue is that the **emotions course are not easily comparable**. For this reason, it was decided to normalize the data using the *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 $\sigma_{[0,T]}(f_e) = \sqrt{\frac{1}{\mid T \mid} \sum_{t=0}^{T} \left(f_e(t) - \mu_{[0,T]}(f_e)\right)^2}$

To infer data about the users, we used m3inference, a **deep learning system for demographic inference** (gender, age, and person/organization) available on Python.[6]


```json
{
  "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
  }
}
```


Listing 2: Json object returned by m3inference

[6] 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.

To link users to a specific place, we need to perform **address geocoding**, the process of taking a text-based description of a location and returning its geographic coordinates.

For this project, we decided to use **OpenStreetMap (OSM)** and the data made available by this particular service.[7]


```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"
  }
}
```


Listing 3: Json object returned by Geopy given “Milano, Lombardia” as input

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

Results and discussion

TODO

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.

Thanks for the attention.

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.

Mohammad, Saif. *NRC Emotion Lexicon*. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Online; accessed 13-June-2021].

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.

Wikipedia contributors. *Emotion recognition — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Emotion_recognition&oldid=1023798177. [Online; accessed 14-June-2021]. 2021.

– .*James W. Pennebaker — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=James_W._Pennebaker&oldid=1023542720. [Online; accessed 24-June-2021]. 2021.

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

Languages with the most tweets

language	ISO	unique tweets	retweets	total	percentage
English	en	195 645 826	473 950 322	669 596 148	63.41%
Spanish	es	35 533 886	111 464 189	146 998 075	13.92%
Portuguese	pt	15 459 760	29 912 427	45 372 187	4.30%
French	fr	9 547 251	23 635 273	33 182 524	3.14%
Indonesian	in	9 029 012	16 479 537	25 508 549	2.41%
German	de	8 091 516	11 447 554	19 539 070	1.85%
Japanese	ja	3 228 542	10 220 609	13 449 151	1.27%
Italian	it	5 256 748	7 173 234	12 429 982	1.18%
Turkish	tr	3 347 597	6 698 252	10 045 849	0.95%
Thai	th	350 268	9 028 730	9 378 998	0.89%

Table 2: Top 10 languages with the most tweets

Tweets over time

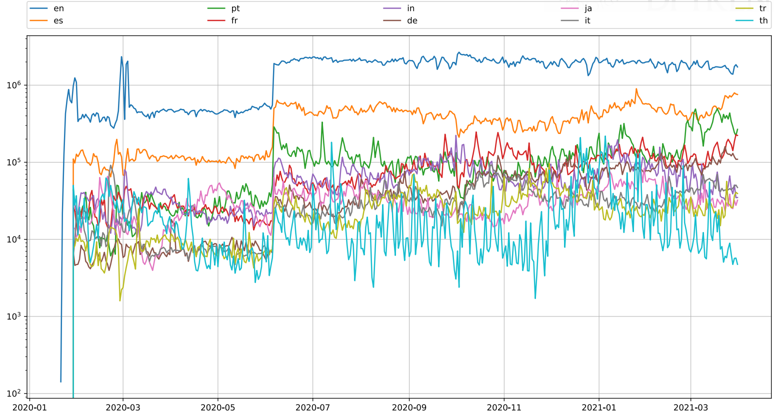


Figure 3: Number of tweets in logarithmic scale over time for the top 10 languages

Data organization

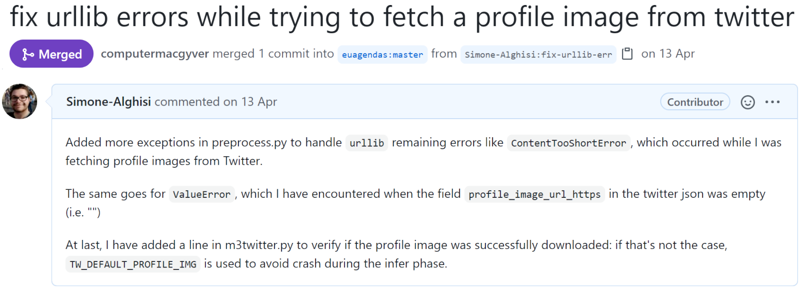
To access directly to a specific category of tweets (e.g. the tweets written in English on the 3rd of June 2020), we grouped them

- 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 a clearer visualization and to average the results.

Personal contribution to m3inference

During the project I encountered several errors while using m3inference, for this reason I decided to open a Pull Request on GitHub[8] to contribute to the project.



The version of m3inference on the packet manager was updated when the same issue was notified by another user[9].

---

[8] Pull request: fix urllib errors while trying to fetch a profile image from twitter #20
[9] issue: Error fetching images will fail the infer method #21

Valid users - gender

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.

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

In particular, the following methodology was applied:

- first we 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

Valid users - age

Definition

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

$$u \in a \iff pc(u, a) \geq 0.95$$

To consider only the users that comply with Theorem 4, we applied the following methodology:

- first we check if $pc(u, \text{>=40}) \geq 0.95$
- then, if $1 - pc(u, \text{>=40}) \geq 0.95$ (i.e. if they have less than forty years)
- finally, if none of the previous constraints were satisfied, we do not consider this user

Valid users - statistics

language	inferred users	valid users	males %	females %	orgs %
Catalan	98 132	73 835	67.30	22.95	9.75
English	3 099 883	2 335 112	63.88	32.16	3.96
Italian	217 340	167 093	67.21	27.96	4.83
Spanish	2 555 941	2 029 765	63.88	33.23	2.89

Table 3: General users statistics for Catalan, English, Italian and Spanish tweets (gender)

language	inferred users	valid users	<40 %	≥ 40 %
Catalan	98 132	42 383	57.83	42.17
English	3 099 883	1 717 733	68.84	33.16
Italian	217 340	122 271	64.54	35.46
Spanish	2 555 941	1 542 935	84.82	15.18

Table 4: General users statistics for Catalan, English, Italian and Spanish tweets (age)

Data normalization (categories)

We also wanted to study 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.

For this reason, we applied Theorem 5 to our data:

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 $[0, T]$,

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

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

Tweets from Campania expressing anger per week

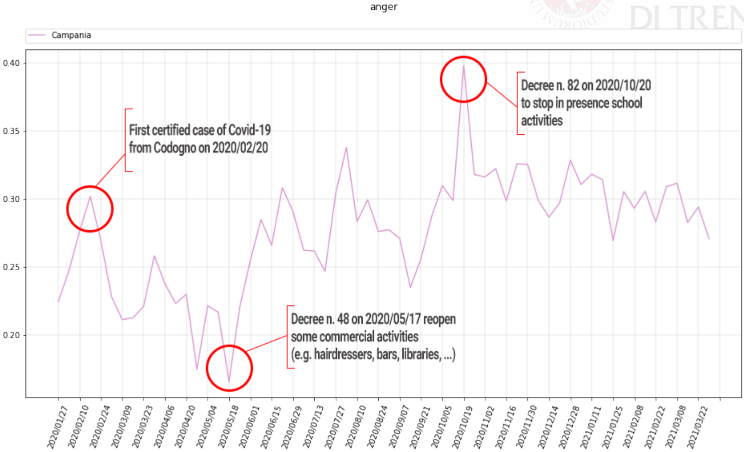


Figure 4: Italian tweets from Campania expressing anger per week

Tweets from Campania expressing joy per week

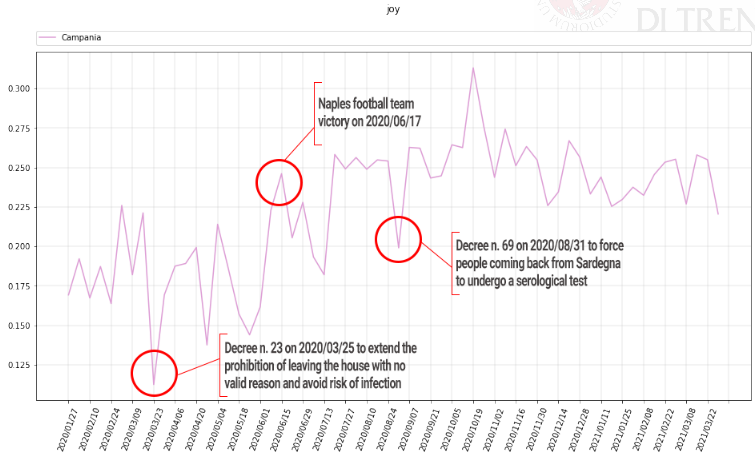


Figure 5: Italian tweets from Campania expressing joy per week