



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

July 9, 2021

Università degli Studi di Trento

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



Figure 1: President Biden's tweet - June 24, 2021.

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.

Emotion detection, 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. In particular, we have used two state-of-the-art libraries: **LIWC** and **EmoLex**.

Here's the **deal**: The Delta variant is more **contagious**, it's deadlier, and it's spreading quickly around the world - leaving **young**, unvaccinated people more vulnerable than ever.

JOY TRUST DISGUST
SURPRISE FEAR
JOY
SURPRISE

Figure 2: Emotion detection for a particular sentence.

Only the highlighted words, which convey at least one emotion, are considered in the process.

¹Wikipedia 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.



I have contributed to the project by

- ▶ retrieving over 6TB of data from Twitter
- ▶ analyzing users' emotions using state-of-the-art libraries (EmoLex, LIWC)
- ▶ studying which emotions were expressed by a large group of users in 4 languages (ca, en, es, it)
- ▶ studying how the emotions expressed by users can vary based on their demographic characteristic, including inferred age and gender
- ▶ studying how the emotions expressed by users can vary based on their geographical provenance (at national and regional level)

Furthermore, I had the chance to contribute to m3inference, an open source, state-of-the-art library used to analyze tweets.



Chen, Lerman, and Ferrara have built an ongoing collection of tweet IDs starting on January 28th, 2020^{2,3}. In particular, we analyzed the tweets **from January 2020 to March 2021**.

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

²echen102/COVID-19-TweetIDs: <https://github.com/echen102/COVID-19-TweetIDs>

³Emily 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.

To correctly **detect the users' emotions**, we decided to follow one of the approaches discussed by Aiello et al.⁴

- categorizing emotions in a binary way (e.g. whether at given time the user expressed joy or not)
- counting 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 .

To compare the emotions course all together we normalized the data using **z-score**.

⁴Luca Maria Aiello et al. *How Epidemic Psychology Works on Social Media: Evolution of responses to the COVID-19 pandemic*. 2020. arXiv: [2007.13169](https://arxiv.org/abs/2007.13169) [cs.CY].



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

To infer data about the users, we used **m3inference**, a deep learning system for demographic inference.⁵

Instead, geocoding was performed using the data made available by **OpenStreetMap (OSM)**.⁶

⁵Zijian 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.

⁶OpenStreetMap contributors. *Planet dump retrieved from <https://planet.osm.org>*.
<https://www.openstreetmap.org>. 2017.

Emotions in the English tweets

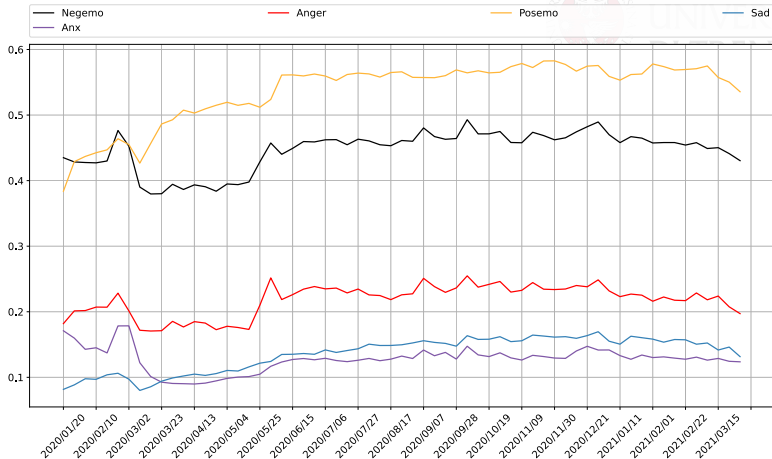


Figure 3: Proportion of weekly users per emotion in the English tweets

Normalized emotions in the English tweets

The z-score is the number of standard deviations by which a datapoint is above (or below) the mean value.

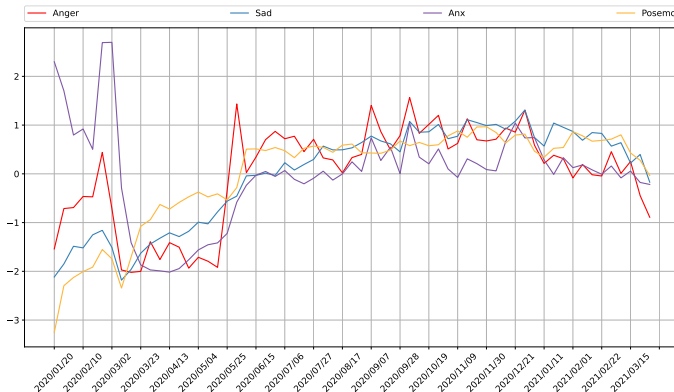


Figure 4: Z-score of weekly users per emotion in the English tweets

Emotions expressed by men/women in the English tweets

The gray dotted line indicates the average value of a particular emotion expressed by all the users (i.e. independently from the category) over the whole period.

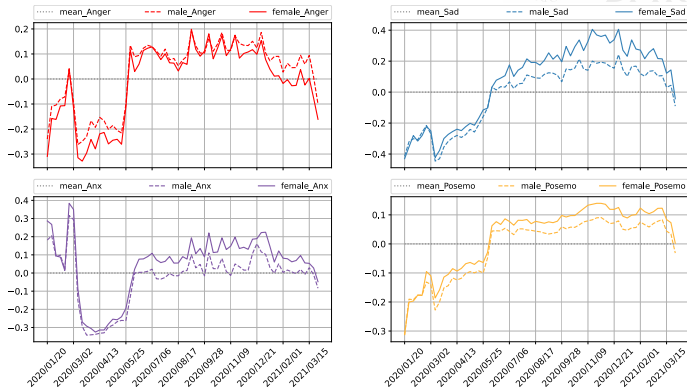


Figure 5: Men/Women expressing the weekly proportion of a particular emotion w.r.t. the average value among all users

Emotions expressed per age in the English tweets

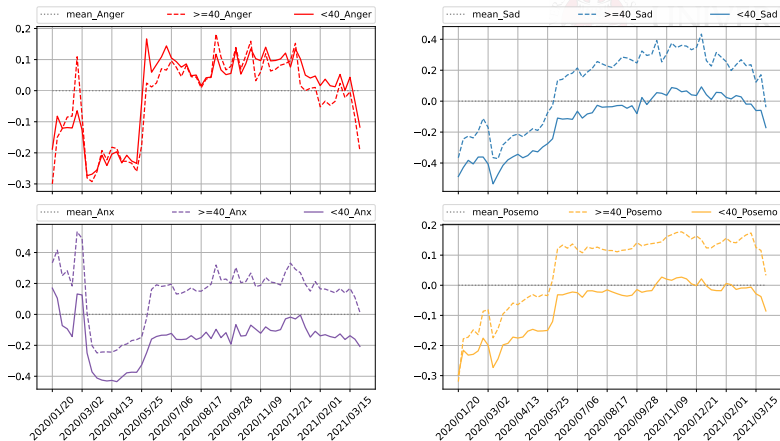


Figure 6: Users per age expressing the weekly proportion of a particular emotion w.r.t. the average value among all users

Emotions expressed in Lombardia from the Italian tweets with events

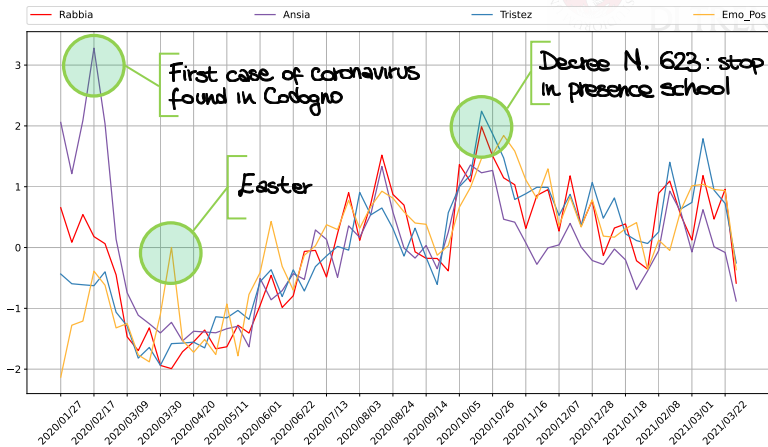


Figure 7: Z-score of weekly users per emotion in Lombardia with possible associated events










The analysis of the data revealed some first interesting results:

- ▶ sometimes the course of the emotions seems to be the same. This happens because on some days users are simply more emotional and tend to use more words.
- ▶ women tend to write more tweets conveying sadness. The difference is less marked for anxiety and positive emotions, while anger seems to be slightly more expressed by men.
- ▶ users below forty years old seem to be slightly more aggressive. Instead, those at least forty years old express more sadness, anxiety, and positive emotions with a significant gap
- ▶ finally, we were able to notice how, after the discovery of the first coronavirus case, people's anxiety dropped. This probably happened because news and various decrees helped to calm people down.

Aside from the results obtained, it must be underlined that there is always some space left for improvement that could bring even more value to this research. It could be possible to

- ▶ take into account the context of the words in a sentence using NLP algorithms to avoid biases and misclassification
- ▶ take into account the whole sentence instead of single words
- ▶ link the emotional peaks to real events (e. g. lockdowns, public health restrictions, etc.)

Thanks for the attention.

- 
- Aiello, Luca Maria et al. *How Epidemic Psychology Works on Social Media: Evolution of responses to the COVID-19 pandemic*. 2020. arXiv: [2007.13169](https://arxiv.org/abs/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.
- 
- Pennebaker, J. W. et al. *Linguistic Inquiry and Word Count: LIWC2015*.
<https://www.liwc.net>. [Online; accessed 28-June-2021].
- 
- 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.



- 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



UNIVERSITÀ
DI TRENTO

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



UNIVERSITÀ
DI TRENTO

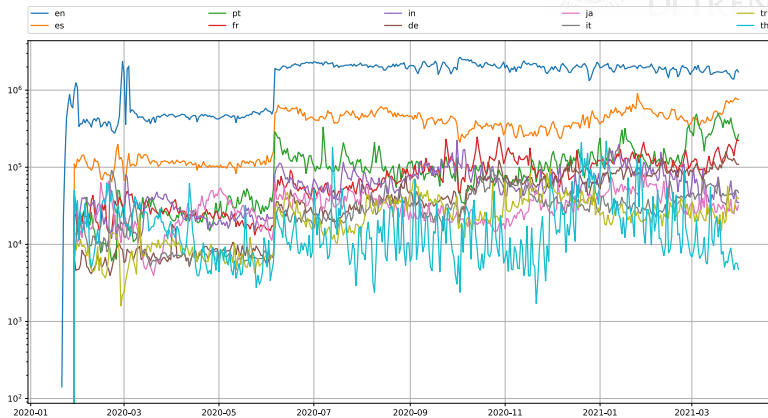


Figure 8: Number of tweets in logarithmic scale over time for the top 10 languages



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.



```
{
  "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"
  }
}
```

Listing 1: Final json object for a tweet

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),⁷ 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**,⁸ to validate the results

⁷Saif Mohammad. *NRC Emotion Lexicon*.

<https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>. [Online; accessed 13-June-2021].

⁸J. W. Pennebaker et al. *Linguistic Inquiry and Word Count: LIWC2015*. <https://www.liwc.net>. [Online; accessed 28-June-2021].



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



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

```
{  
  "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

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.

```
{  
  "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



During the project I encountered several errors while using m3inference, for this reason I decided to open a Pull Request on GitHub⁹ 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 version of m3inference on the packet manager was updated when the same issue was notified by another user¹⁰.

⁹Pull request: [fix urllib errors while trying to fetch a profile image from twitter #20](#)

¹⁰issue: [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.

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

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



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 $pc(u, \geq 40) \geq 0.95$
- then, if $1 - pc(u, \geq 40) \geq 0.95$ (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)

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 $[0, T]$,

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

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

Comparison between EmoLex and LIWC

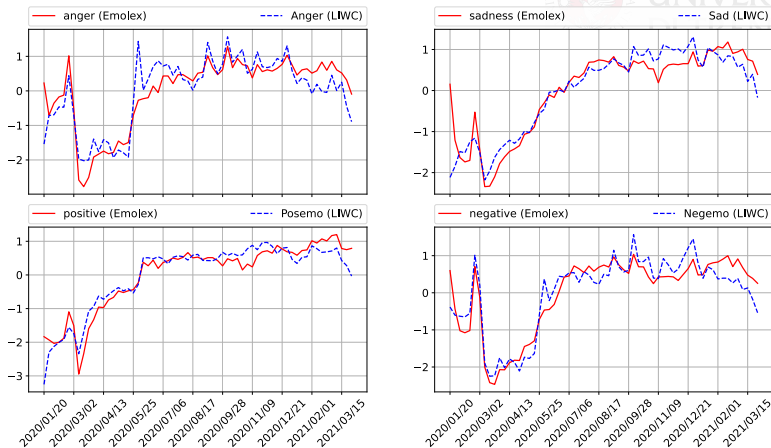


Figure 9: Comparison between the z-score of the emotions and the z-score of the LIWC categories

Emotion in the English tweets (subplots)

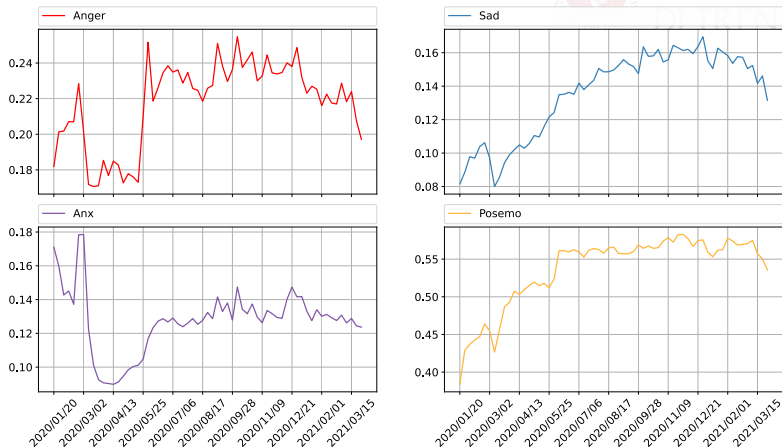


Figure 10: Proportion of weekly users expressing a particular emotion in the English tweets

Most used words on 2020/02/24 in the English tweets (LIWC)

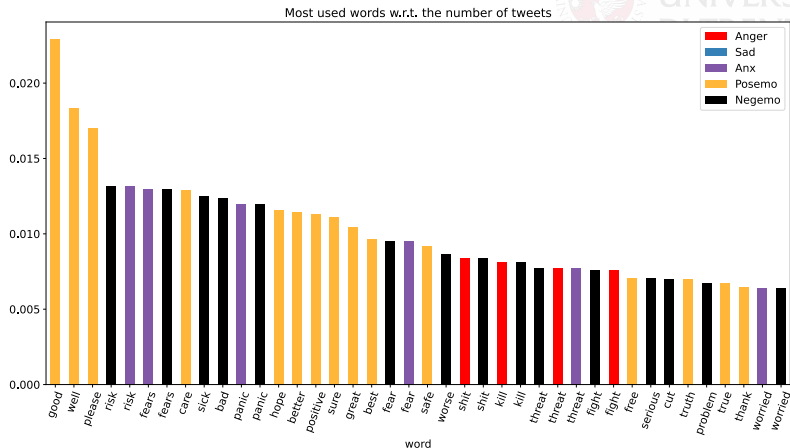


Figure 11: Proportion of most used words on 2020/02/24 that express an emotion in the English tweets

Most used words on 2020/02/24 in the English tweets (Emolex)

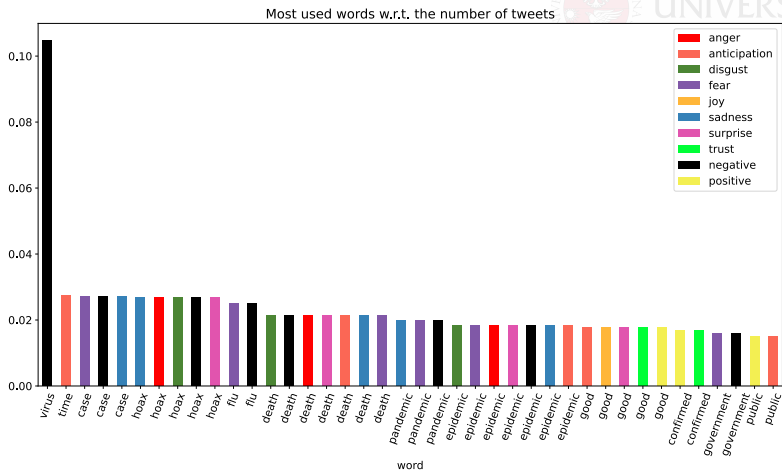


Figure 12: Proportion of most used words on 2020/02/24 that express emotion/sentiment in the English tweets

Most used words per emotions on 2020/02/24 in the English tweets

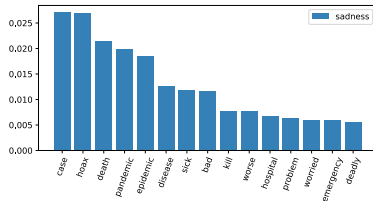
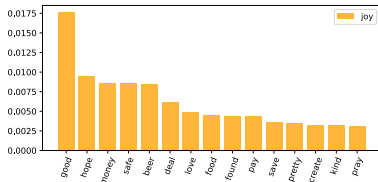
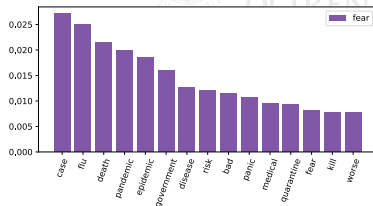
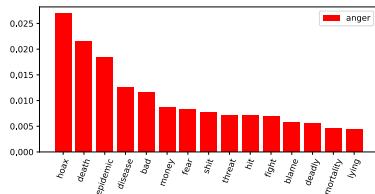


Figure 13: Proportion of 15 most used words on 2020/02/24 per emotion in the English tweets

Users per state in the Italian tweets

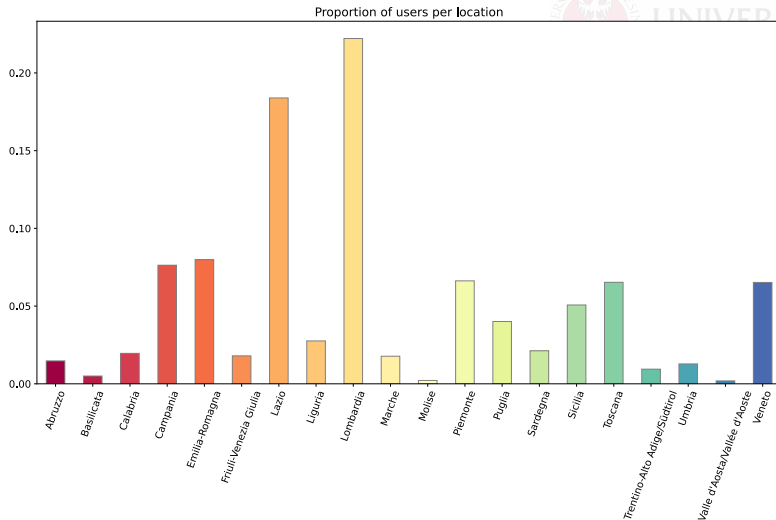


Figure 14: Proportion of users per state in Italy

Emotions expressed per region in the Italian tweets

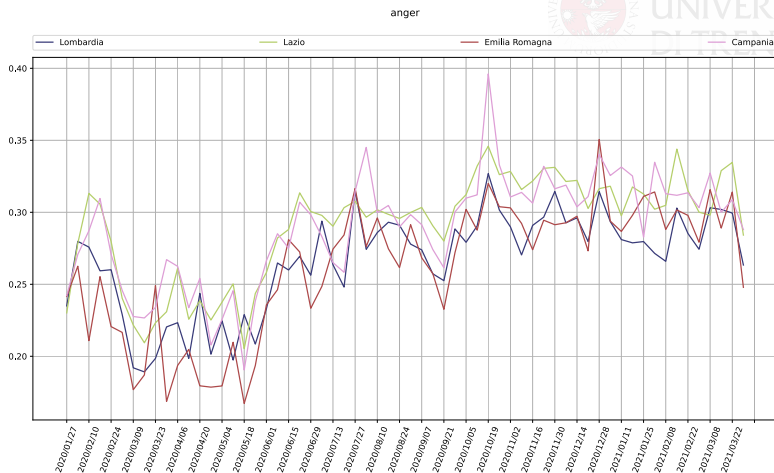


Figure 15: Proportion of weekly users that expressed anger in Lombardia, Lazio, Emilia Romagna and Campania

Emotions expressed in Lombardia from the Italian tweets

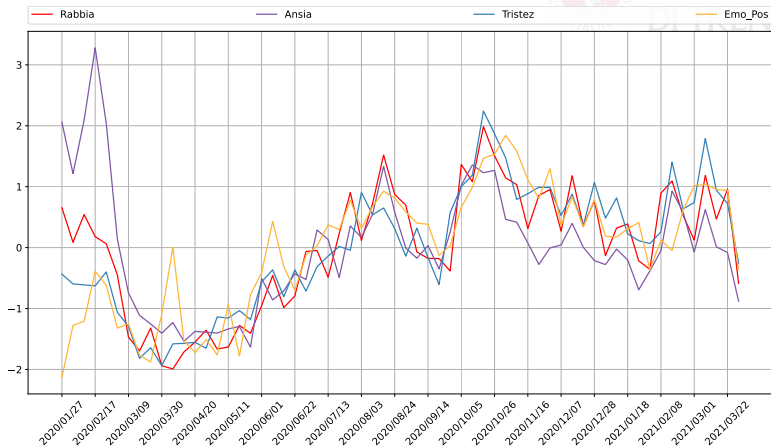


Figure 16: Z-score of weekly users per emotion in Lombardia