Università di Trento

Department of Information Engineering and Computer Science

Bachelor's Degree in Computer Science

Final dissertation

The emotional impact of the Covid-19

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

Supervisor
Alberto Montresor

Co-Supervisor
Cristian Consonni
David Laniado

Student
Simone Alghisi

Academic year 2020/2021

Contents

Abstract

This dissertation describes in detail the activity performed during my two-month traineeship at the Big Data Department of Eurecat - Centro Tecnológico de Catalunya, which was supervised by Cristian Consonni and David Laniado.

The purpose of the project was to analyze the emotions emerging from Twitter messages during the pandemic, in order to understand how people felt over the whole period. Based on the result obtained from this research, it may be possible to determine which countermeasures better handled the situation while offering the best possible trade-off between people's satisfaction and the reduction of the spread of the disease.

In general, my contribution to the research mostly regarded:

- retrieving and organizing the data
- processing the tweets to understand users' emotions
- plotting graphics in order to visualize more clearly the results
- normalizing the results obtained to compare different emotions or categories
- inferring demographic information about the users
- geocoding the location of the user

The dataset used for the project is the echen102/COVID-19-TweetIDs, a collection of over 1 billion tweet IDs available on GitHub. The selected tweets are either related to specific accounts, or sampled real-time from the Twitter API because they matched a defined set of keywords.

In order to start the analysis, I was asked to retrieve the tweets from January 2020 to March 2021 using Twarc. In fact, to comply with Twitter's term of service, the dataset contains only the ID of the original tweet; however, is possible to get the associated information using the Twitter's API and a Twitter Developer Account.

After collecting the data, we decided to group the tweets first based on their language, to perform a targeted analysis on a restricted set (Catalan, English, Italian and Spanish); secondly per week, for better data visualization and to average the results.

In order to understand which emotions were expressed in a single tweet, we decided to use the NRC Word-Emotion Association Lexicon (aka EmoLex). Emolex is 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).

To reduce the possible bias of particularly active users, we decided to follow one of the approaches discussed by Aiello et al., 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)

For the first sentiment analysis, tweets belonging to a given language were analyzed over the whole period. In particular, we decided to normalize the obtained results using the z-score and to manually retrieve some peaks to study the most used words for that particular language.

To understand how differently men and women perceived the pandemic, we decided to use m3inference, a deep learning system for demographic inference (gender, age, and person/organization)

implemented on PyTorch. Only those users that the system inferred with a confidence grater or equal to 0.95 were considered valid and used for the next sentiment analysis.

Finally, we used Twitter location field to analyze users from the same geographical area. To overcome the absence of constraints to specify a location, we retrieved the position of the users using address geocoding, the process of taking a text-based description of a location and returning its geographic coordinates. In particular, we used Nominatim to access the data made available by OpenStreetMap(OSM).

The analysis of the data revealed some first interesting results:

- first of all, there are cases when the course of the emotions seems to be the same. In particular, when one emotion increases, so do the others. This could happen because on some days users are simply more emotional and tend to use more words. Because of that, the probability of conveying more emotions increases. Furthermore, more emotions could be associated to a single word, so this also explains why some emotions are always more frequently expressed than the others.

- secondly, if we consider the English tweets it seems that women tend to express joy more frequently in their tweets. The same goes for sadness and fear, but in this case the difference is less marked. Finally, anger seems to be slightly more expressed by men, but the lines tend to overlap for the majority of the considered period of time.

- finally, we were able to understand the proportion of Italian users for each state. However, even if it is indeed possible to map manually certain emotional peaks to particular events that occurred around the same time, we still lack a way to perform this method in a more valid and reliable way.

During the course of the project I had the possibility to personally contribute to the improvement of m3inference on GitHub, by opening a pull request to solve some issues while downloading images from Twitter.

In the end, I was only able to scratch the surface of this research field, because the amount of data to analyze was really impressive. In any case, I hope that my contribution could be a good starting point for further studies and I would really like to continue researching about this topic in the future.

1 Introduction

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.

Context and motivations During this year, everyone's daily life changed significantly and we had to adapt to restrictive measures in order to stop the disease: whether we liked it or not. This research proposed by Eurecat - Centro Tecnológico de Catalunya, really caught my eye: the possibility to study how people perceived all of this situation, and better understand which measures were more welcomed than others, was really fascinating and, above all, may be useful in the case of some other unfortunate event.

1.1 Project description

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. Moreover, users have been divided based on their gender, to study the different emotional response of males and females, and also based on their location, to analyze users' emotions considering a particular place.

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.

1.2 Twitter

Twitter is an American microblogging and social networking service created by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in March 2006 and launched in July of that year [10].

Registered users can perform operations similar to those available on other social networks, e.g. create (tweet), like, or share (retweet) a post; however, these are public, and even unregistered users can read them.

In particular, users post and interact with particular messages known as "tweets", which have a limited number of characters. Tweets were originally restricted to 140 characters, but the limit was doubled to 280 for non-CJK languages.

As of Q1 2019, Twitter had more than 330 million monthly active users, and is considered a some-to-many microblogging service because the vast majority of tweets are written by a small minority of users.

Why did we use Twitter data? The main reason behind this critical choice, is the fact that retrieving data from Twitter is particularly easy. That is because, in the majority of the case, posts are public and everyone can see them (i.e. there are less privacy related issues). Furthermore, it is possible to have access to a considerable amount of data, which is fundamental for this kind of research.

Obviously, the data from other platforms (e.g. Facebook, Instagram, Reddit, other minor blogs, ...) could have been interesting. However, it is either too difficult to get the data (due to particular limitations) or to get enough data. For this reason, Twitter was the best possible option.

Furthermore, given its relevant in social and political debate, Twitter can be considered like a kind of online "public sphere" [2].

On the other hand, Twitter maximum number of characters per tweets limits the possibilities of the users to express their feelings: this could have a negative impact on the performance of sentiment analysis. However, with a sufficient amount of data, is possible to reduce this to a bare minimum.

1.3 Sentiment analysis

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information [9]. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

The objective and challenges of sentiment analysis can be shown through some simple examples:

- I do not dislike carrots. (Negation handling)
- There are times when I regret not being a cat (Adverbial modifies the sentiment)
- It's all day that I was waiting to clean my room! (Possibly sarcastic)
- I think that the best part of the movie is when the villain dies. (Negative term used in a positive sense in certain domains).
- ...

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level - whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as enjoyment, anger, disgust, sadness, fear, and surprise.

Emotion detection Also called emotion recognition, is the process of identifying human emotions [7].

To solve this particular task, different approaches have been developed. In particular, I had the possibility to use knowledge-based techniques (sometimes referred to as lexicon-based techniques), where domain knowledge and the semantic and syntactic characteristics of the language are used to detect emotions. In practice, it is possible to use dictionaries to map terms to the emotions.

One of the advantages of this approach is the accessibility of such knowledge-based resources. On the other hand, is limited and, for example, cannot properly handle complex linguistic rules.

2 Data collection

The dataset used for the project is the echen102/COVID-19-TweetIDs GitHub repository [3]. The repository contains an ongoing collection of tweet IDs starting on the January 28th, 2020.

The IDs in the dataset are either tweets posted by a specific account, or sampled real-time from the Twitter API because the text contained specific keywords, such as Coronavirus, Epidemic, covid-19, Social Distancing, panic buy, lockdown, ...

From the followed accounts list, it is worth mentioning the following ones:

- CDCgov, CDC's official Twitter source for daily credible health & safety updates from Centers for Disease Control & Prevention
- WHO, the United Nations health agency
- HHSGov, News and information from the U.S. Department of Health & Human Services (HHS)
- NIAIDNews, National Institute of Allergy and Infectious Diseases (NIAID)
- drtedros, Twitter's profile of Tedros Adhanom, the Director-General of the World Health Organization

The fact that these accounts are presents in the dataset is a crucial point: given their relevance, we are sure that some tweets are not biased and the statements are valid. However, at the same time, it means that there may be a lot of retweets from various people. This is not necessarily a bad thing, it depends on which type of analysis we are interested in.

Dataset structure Data is organized in the following way inside of the dataset:

- at the top layer, the IDs are sorted into YEAR-MONTH folders
- then, in each folder, tweet-IDs are grouped into files with a prefix "coronavirus-tweet-id-" followed by YEAR-MONTH-DAY-HOUR

2.1 Analyzed period

COVID-19 was declared a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March 2020. However, depending on the considered country, restrictive measures were applied on different dates and for different time periods.

For this reason, we decided to analyze tweets from January 2020 to March 2021. In this way, we were able to: a) consider more countries; b) better understand when people expressed more negative (or positive) emotions w.r.t the whole period.

Considering this time frame, I have processed and worked with:

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 2.1: Dataset general statistics

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 2.2: Top 10 languages with the most tweets

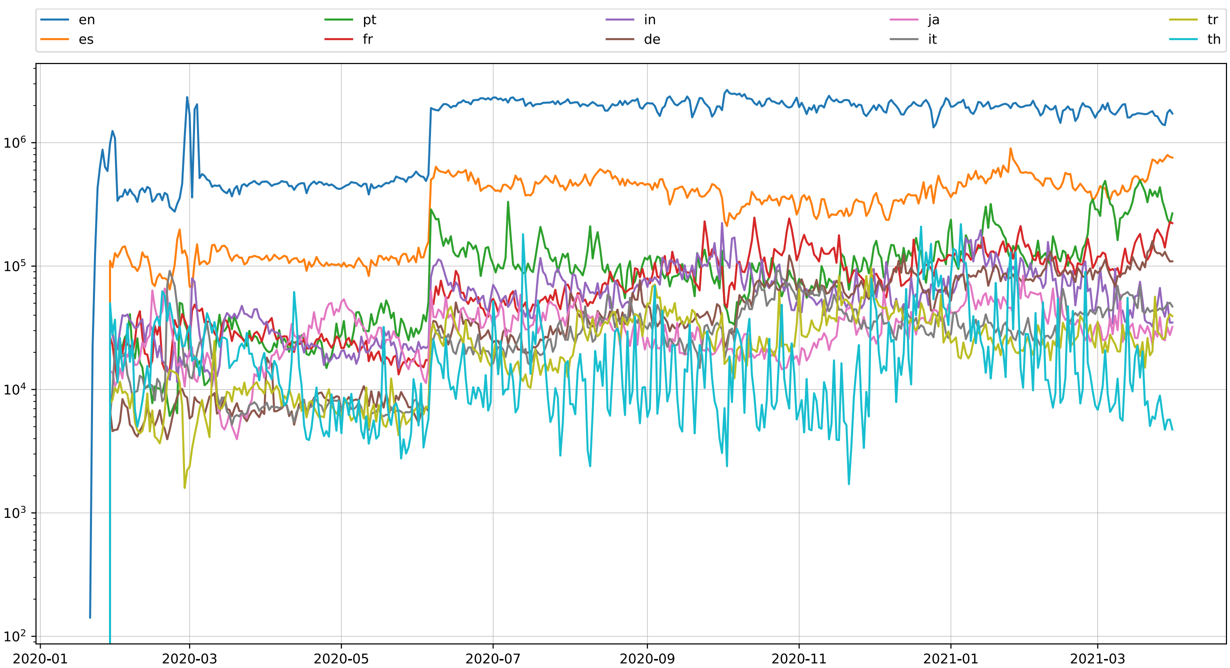

Figure 2.1: Number of tweets in logarithmic scale over time for the top 10 languages

2.2 How to retrieve the data

To comply with Twitter's terms of service,[1] tweets cannot be released publicly: the repository is in fact a collection of tweet-IDs.

The original tweets can be retrieved, or *hydrated*, using the Python library Twarc with a Twitter Developer Account. In fact, to be able to use the Twitter API, it is mandatory to apply to Twitter Developer.[2] When the application has been accepted, the developer will be entitled to access the API using tokens.

Given the id of a tweet, Twarc uses the tokens of the associated developer account to contact the API, and returns the corresponding information as a json object. However, Twarc also tries to maximize the number of possible IDs per request and, at the same time, makes sure to be compliant with the API usage limits.

In the case of this dataset, the data came along with a script to hydrate the tweets automatically to facilitate the procedure.

2.3 tweets

The structure of the json object for the associated tweet depends on the version of the API used: for this project, we have used Standard v1.1 to hydrate the tweets.[3]

In general, a tweet is characterized by a number of different fields. For the scope of the project, we decided to keep only the most relevant fields, such as:

- `id`, the integer representation of the unique identifier for this tweet
- `created_at`, UTC time when this tweet was created
- `full_text`, the actual UTF-8 text of the status update (not truncated)
- `user`, the user who posted this tweet
- `retweeted_status`, the presence of this attribute distinguishes retweets from typical tweets
- `favorite_count`, indicates approximately how many times this tweet has been liked by Twitter users
- `retweet_count`, number of times this tweet has been retweeted
- `lang`, indicates a BCP 47 language identifier corresponding to the machine-detected language of the tweet text, or und if no language could be detected

Further information about the user that posted the tweet are available in the `user` field, such as:

- `id`, the integer representation of the unique identifier for this user
- `name`, the name of the user, as they have defined it
- `screen_name`, the screen name, handle, or alias that this user identifies themselves with
- `location`, the user-defined location for this account's profile. Not necessarily a location, nor machine-parseable
- `description`, the user-defined UTF-8 string describing their account
- `default_profile_image`, when true, indicates that the user has not uploaded their own profile image and a default image is used instead
- `profile_image_url_https`, a HTTPS-based URL pointing to the user's profile image

[1] https://developer.twitter.com/en/developer-terms/agreement-and-policy
[2] https://developer.twitter.com/en/apply-for-access
[3] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

- `followers_count`, the number of followers this account currently has

- `statuses_count`, the number of tweets (including retweets) issued by the user

Given that, each json object associated with a tweet can be represented as follows:


```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 2.1: Final json object for a tweet

3 Methods

This chapter aims to describe all the different methodologies used to analyze the data collected during the previous phase, and explained in details in Chapter 2.

The scope of the project is to understand and measure the emotions of the users through sentiment analysis. In particular, we would like to identify, given a certain set of tweets scattered across our considered period of time, when the users conveyed more feelings and which was the emotion expressed the most (e.g. from the Italian tweets it is possible to notice a peak of anger on 21st of February 2020).

Given the heterogeneity of the tweets, we would also like to conduct a more in-dept analysis, by considering the users, over the whole time frame, based on their gender and also on their location.

3.1 Lexicons

Until now, we have discussed that we would like to identify and quantify the emotions expressed in the tweets, but we still lack a way to achieve this result. During the project I have used *Lexicons* for this particular task.

The idea behind lexicons is quite simple: we take a particular word in our dictionary, and we assign zero or more emotions or sentiments to it.

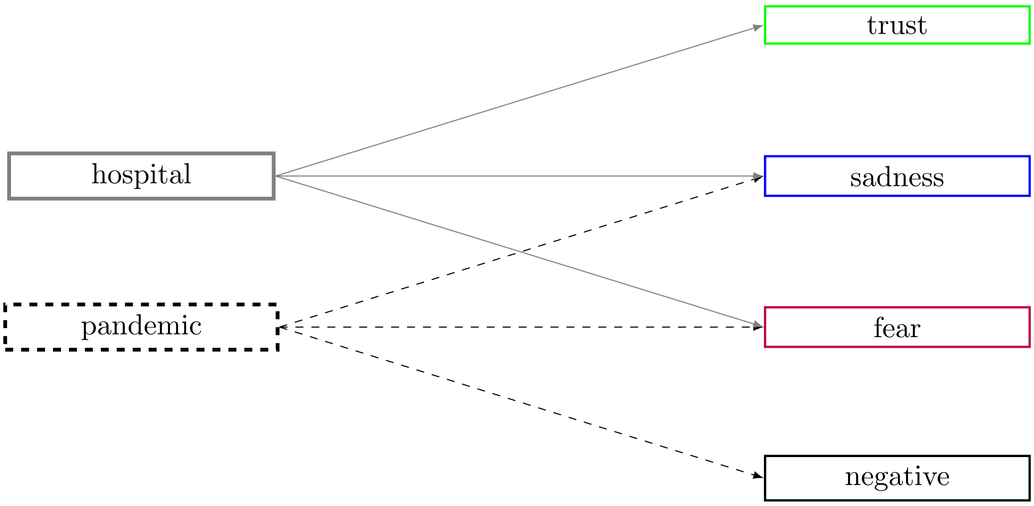

Figure 3.1: Word-emotions/sentiments association

However, if we look at Figure 3.1, we can notice that, while the word pandemic is associated with only negative emotions/sentiments, the word hospital is associated with fear and sadness but, at the same time, with trust. While this is technically correct, because our lexicon does not know in which context the word is used (i.e. it must consider all the possible meanings), it also introduces a bias.

To simplify even more, even a negation can totally change the meaning of a particular phrase:

I am fine / I am not fine

On the other hand, there are different schools of thought relatively at sentiment analysis: someone tries to understand the meaning of the phrase, while other ones focus on the emotions conveyed by a certain set of words (without caring about the meaning).

Let us now consider the following example:

This house is not an obscure prison

In this case, we are probably going to associate a negative feeling to the phrase because, even if there is a negation (not), the words "obscure" and "prison" convey fear or sadness.

In any case, the results obtained with lexicons should be checked and contextualized to get a clear understanding of the situation.

EmoLex For the research, I have used the NRC Word-Emotion Association Lexicon (aka EmoLex) for emotion detection [4]. EmoLex is 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).

The peculiarity of this lexicon is the fact that, even if it has been designed for the analysis of English words, it has been translated in over one hundred languages using Google Translates. Given that, the number of different languages identified in the dataset is more than 60, this particular dictionary could perfectly fit our problem.

3.2 Data organization

Before performing emotion detection on our tweets, we decided to define a new data organization strategy. In particular,

- at the top layer, we sorted the tweets into LANGUAGE folders using the `lang` field of the json object
- then, into YEAR-MONTH folders

- finally, we grouped the tweets into files with a prefix “coronavirus-tweet-id-” followed by YEAR-MONTH-DAY

The idea behind this partition, aside from considering tweets of the same language, was to get rid of the per hour aggregation: for the purpose of the project, this kind of granularity was simply too much. Given that, we have preferred to aggregate in a single file the tweets posted on the same day.

Figure 3.2 shows a visual representation of the new data organization: now it is possible to access directly to tweets of the same language posted on a particular date.

```
/
├── en
│   ├── 2020-01
│   │   ├── coronavirus-tweet-2020-01-21.json.gz
│   │   ├── coronavirus-tweet-2020-01-22.json.gz
│   │   ├── ...
│   │   ├── coronavirus-tweet-2020-01-30.json.gz
│   │   └── coronavirus-tweet-2020-01-31.json.gz
│   ├── ...
│   └── 2021-03
│       ├── coronavirus-tweet-2021-03-01.json.gz
│       ├── ...
│       └── coronavirus-tweet-2021-03-31.json.gz
├── it
└── ...
```

Figure 3.2: First tweets organization

However, this first data organization strategy led to very noisy results: this happened because, depending on the language considered, the number of tweets can drastically change. To solve this problem, we thought about grouping together tweets of the same week: in this way, we were able to obtain better averages over the results, get stabler data for languages with fewer tweets, and obtain a clearer visualization.

```
/
├── en
│   ├── coronavirus-tweet-2020-01-20.json.gz
│   ├── coronavirus-tweet-2020-01-27.json.gz
│   ├── ...
│   ├── coronavirus-tweet-2021-03-22.json.gz
│   └── coronavirus-tweet-2021-03-29.json.gz
├── it
└── ...
```

Figure 3.3: Weekly tweets organization

In this case, the syntax of the files showed in Figure 3.3 has a slightly different meaning. In particular, tweets are grouped into files with a prefix “coronavirus-tweet-id-” followed by YEAR-MONTH-FIRST_WEEK_DAY.

3.3 Data filtering

Section 2.1 shows that the number of retweets in our dataset is much greater than the number of tweets. This could be a problem because, if we perform emotion detection with this data, we may end up with biased results.

Let us suppose to have a particularly positive tweet, for example:

Figure 3.4: An example of a particularly positive tweet.

If John Doe is the only happy person on the 14th of June 2021, then the emotional impact produced from this tweet will not be so significant. Maybe, during that day a lot of people complained on the social network that a streaming platform was not working properly. In this particular case, we will probably find a peak of anger.

However, if John Doe is a celebrity, then the tweet will be surely retweeted thousands of times, and the emotional impact will be much grater. In practice, this means that the previous anger, associated to the malfunctioning of the streaming platform, could pass unnoticed.

To clarify, the fact that a person could retweet a post of a user is not a bad thing: maybe they liked what the original author had written. However, in the case of emotion detection, every words count: if a person agrees with someone's opinion, they will not necessarily use the same exact set of word. In practice, they could still agree with the author of the tweet, but at the same time not convey the same emotion.

For the reasons explained above, we decided to avoid this bias by removing from our dataset all the retweets. As explained in Section 2.3, this particular operation can be performed by checking the presence of the `retweet_status` field.

3.4 Emotion detection metrics

We know from Section 1.2 that Twitter's constraint on the number of characters per tweet limits the users. For this reason, emotions are also limited because the users cannot express themselves to the fullest. As a consequence, the emotion detection process could fail because the small set of words used does not convey any emotion.

Fortunately, the effect produced from this problem can be minimized with a considerable amount of data. Obviously, with many tweets, the possibility of having meaningful data increases.

However, we have not still defined how users' emotions should be considered: should we count the number of occurrences (i.e. words) for each emotion inside of a tweet? Should we consider independently each tweet posted by a user?

Regarding this particular topic, we have decided to follow one of the approaches discussed by by Aiello et al. [1]

The idea that they have proposed is to consider

- 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)

We can then formalize the metrics used for the project as follows:

Definition 3.4.1. A tweet t contains an emotion e, if at least one of the tweet's words or stems w belongs to e,

$$t \in e \Leftrightarrow \exists w \in t \text{ s.t. } w \in e$$

First of all, given that there are only 280 characters inside of a tweet, multiple emotions occurrences are rare. Secondly, they do not necessarily reflect the intensity of a tweet.

Definition 3.4.2. 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.

Considering the number of unique users, instead of the number of total tweets, is a way to reduce the bias introduced by particularly active users (or bots), who could otherwise mask other important events.

3.5 Emotion detection over time - by language

For the first emotion detection, we have decided to analyze, over the whole period, the course of the users' emotions for a particular language. Specifically, our studies regarded four languages: Catalan, English, Italian and Spanish.

Table 3.1 shows the total number of unique tweets and users over the considered period.

language	tweets	tweets %	users
Catalan	1 377 225	0.4%	379 193
English	195 645 826	60.4%	21 979 558
Italian	5 256 748	1.6%	558 224
Spanish	35 533 886	10.9%	5 220 714

Table 3.1: Number of tweets and users for Catalan, English, Italian and Spanish

It is important to underline the following: given that, as explained in Section 3.2, we are considering tweets per week, $f_e(t)$ will correspond to the proportion of users that expressed emotion e during the analyzed week. For this particular reason, we will keep track of the weekly users and, once the week is over, we will reset them and we will consider a new set of users. In practice, we will aggregate the emotions that a user conveyed only if they tweeted multiple times in the same week.

3.5.1 Data normalization

The issue with the results obtained with this emotion detection, is the fact that the emotions course are not easily comparable. For example, we can notice which emotion was the one expressed more in a certain week considering $f_e(t)$. However, it is more difficult to say, for example, whether users expressed in a certain week the most joy over the whole period.

To solve this problem, we have decided to apply a data normalization, in particular the *z-score*.

Definition 3.5.1. 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 $\sigma_{[0,T]}(f_e) = \sqrt{\frac{1}{\mid T \mid} \sum_{t=0}^{T} \left(f_e(t) - \mu_{[0,T]}(f_e)\right)^2}$

The z-score is very useful in this particular case because, given the fact that $f_e(t)$ is normalized w.r.t. its mean value over the whole period, it is easier to determine which weeks have the highest (or lowest) emotional impact. Moreover, this operation is applied to each considered emotion, so we are also able to compare the different courses and determine which emotion has the highest relevance in a certain week.

3.6 Emotion detection over time - by gender

The second analysis of the data regarded the gender of the users. In particular, we were interested in finding, and understanding, the differences between the emotions expressed by men and women during the considered period of time.

However, Twitter does not allow users to specify their gender explicitly. This means that there is no way to recover this information from a specific field. Besides, searching through the user's description is not a valid option because they may have not given that kind of information.

Fortunately, even if this information is not explicitly available, is possible to infer it.

3.6.1 Users inference

To solve this particular issue, we used m3inference, a deep learning system for demographic inference implemented on PyTorch [6].

Given the user's `name`, `screen_name`, `description`, `default_profile_image`, and `profile_image_url_https` fields, m3inference will predict, with a certain probability ([0, 1]), their gender, their age, and if the user's profile belongs to a person or organization.

In particular, m3inference will detect the user's language and then examine their description. Moreover, it will also check if the user has a profile picture with the `default_profile_image` field. If that is the case, it will first download it using the `profile_image_url_https`, and then it will resize it. Otherwise, it will use the Twitter's default profile image.

Given the fact that m3inference uses both text and images, operates in 32 major languages, and computes multiple output predictions, is classified respectively as a Multimodal, Multilingual, and Multi-attribute system (M3).

The result of the prediction performed by m3inference is showed in the following example:

```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 3.1: Json object returned by m3inference

Given the fact that the inference process takes some time, it was decided to take into account only a subset of users for each considered language. In particular, we analyzed the users with at least n tweets over all the dataset.

Furthermore, for the emotion detection process, we assigned the users that respected Definition 3.6.1 to their corresponding category (male, female or org) .

Definition 3.6.1. 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

Finally, we obtained the following statistics from the previous considerations:

language	n **tweets**	**inferred users**	**valid users**	**males %**	**females %**	**orgs %**
Catalan	2	98 132	73 835	67.30	22.95	9.75
English	10	3 099 883	2 335 112	63.88	32.16	3.96
Italian	2	217 340	167 093	67.21	27.96	4.83
Spanish	5	2 555 941	2 029 765	63.88	33.23	2.89

Table 3.2: General users statistics for Catalan, English, Italian and Spanish tweets (gender)

3.6.2 Data normalization

The results obtained from the inferred users analysis allowed us to check the emotions of men and women separately. However, 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 have applied Definition 3.6.2 to our data:

Definition 3.6.2. 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 $[0, T]$,

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

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

To clarify, if $v_{e,c}(t) = 0.20$, with $e =$ joy and $c =$ female, then it means that females expressed 20% more joy at time t w.r.t. the mean joy value over the whole period.

3.7 Emotion detection over time - by age

Given the fact that the information produced by m3inference (see Section 3.6.1 for further details) not only regards the gender of a user, we also thought about considering the emotions expressed by a particular age bracket.

Definition 3.7.1. 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.

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

m3inference outputs a prediction confidence for 4 different age brackets: `<=18`, `19-29`, `30-39`, and `>=40`. However, we noticed that it could be more effective to consider a binary separation. In particular, we reduced the possible age brackets to `<40` and `>=40`.

To consider only the users that comply with Definition 3.7.1, we applied the following methodology:

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

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

3.8 Emotion detection over time - by location

Finally, we decided to study the course of the emotions over the whole time frame, considering the users from the same location.

Fortunately, Twitter allows users specify their location in the `location` field. However, it does not impose any constraint or predefined format. It follows that: a) not all the users specified their location; b) some locations could be fake or misspelled; c) the same location could be written in a different syntax.

3.8.1 Geocoding

To solve the issue introduced previously and link users to a specific place, we need to perform *address geocoding*. Address geocoding is the process of taking a text-based description of a location and returning its geographic coordinates. In particular, different geocoding web services are available, such as Bing, GoogleV3, TomTom, ...

For this project, we decided to use OpenStreetMap (OSM) and the data made available by this particular service [5]. OpenStreetMap is a collaborative project to create a free editable map of the world [8]. Created by Steve Coast in the UK in 2004, it was inspired by the success of Wikipedia and the predominance of proprietary map data in the UK and elsewhere.

However, OSM only provides the data: to perform address geocoding, Nominatim or similar tools are required. In practice, Nominatim is used to search OSM data by name and address. Furthermore, it allows developers to use the public API for various tasks, as long as they comply with the usage policy.[1]

To reduce the possible workload, we decided to perform address geocoding only for the most frequent locations. For this reason, we grouped together users that specified the same location, hoping to maximize the results. In particular, we removed additional spaces (e.g. white spaces, tabs, ...) from the string, and we made it lowercase.

```
  Milano, Lombardia → milano,lombardia
  Milano, lombardia → milano,lombardia
Milano,     Lombardia → milano,lombardia
```

Finally, we used Geopy, a Python client for several geocoding web services, to contact Nominatim public API. The result of the address geocoding operation using Geopy with Nominatim is showed below:


```
{
  "place_id": 317098601,
  "licence": "Data \u00a9 OpenStreetMap contributors, ODbL 1.0. https://osm.
     org/copyright",
  "boundingbox": [
    "45.3867381",
    "45.5358482",
    "9.0408867",
    "9.2781103"
  ],
```


[1] https://operations.osmfoundation.org/policies/nominatim/

```
  "lat": "45.4668",
  "lon": "9.1905",
  "display_name": "Milano, Lombardia, Italia",
  "address": {
    "city": "Milano",
    "county": "Milano",
    "state": "Lombardia",
    "country": "Italia",
    "country_code": "it"
  }
}
```

Listing 3.2: Json object returned by Geopy given “Milano, Lombardia” as input

Link the users to a certain location Given the fact that the geographical coordinates are too specific, we used instead the information contained in the `address` field. In particular, we obtained our final set of users using the following method:

- we optionally specify a country (or a state) to focus on a restricted area
- then, we choose a certain location type (e.g. country, state, county, city, ...)
- finally, we consider all the users that have a geocoded location coherent with the place specified in 1. and with enough information in the `address` field to satisfy 2.

In practice, if we take Listing 3.2 as example, and we decide to map all the users from Italy based on their county, all the ones with “Milano, Lombardia” as their original location will then be assigned to “Milano”. Instead, all the users that have a country different from “Italy”, or they do not have a county, will be discarded.

4 Results and discussion

This chapter is meant to show and discuss the results obtained from the application of the methods introduced in Chapter 3.

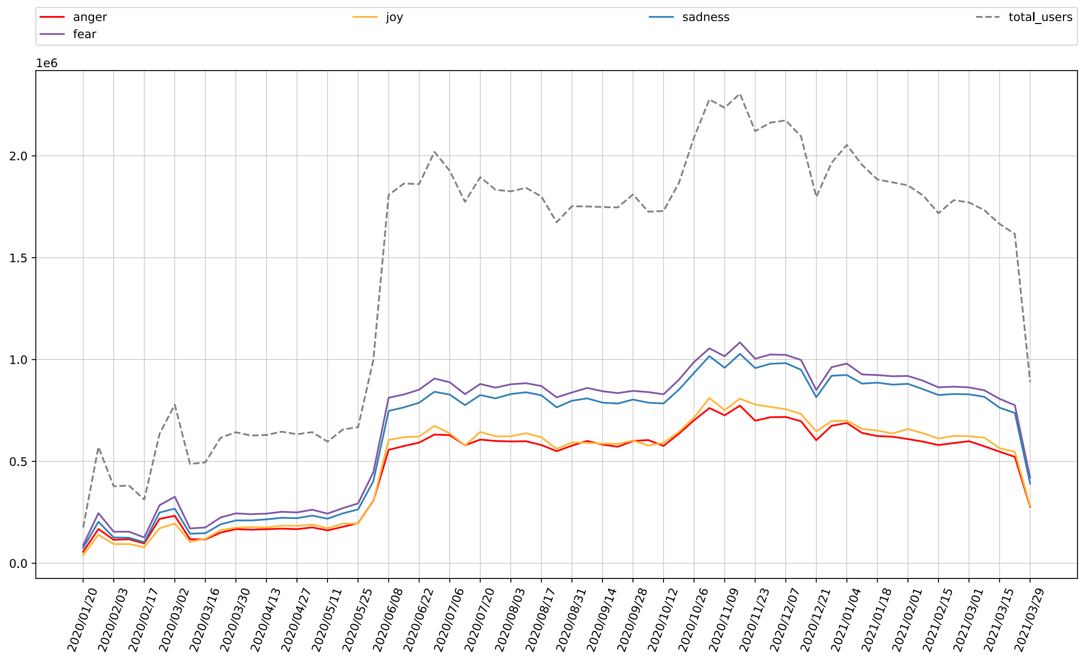

Figure 4.1: Number of weekly users per emotion in the English tweets

Figure 4.1 displays the number of users that expressed a particular emotion in a given week, considering the English tweets. Instead, the gray dashed line indicates the number of users that posted at least one tweet during a week. The fact that the data collection migrated to AWS around June 2020, explains why the number of tweets almost doubled from 2020/06/01 to 2020/06/08.

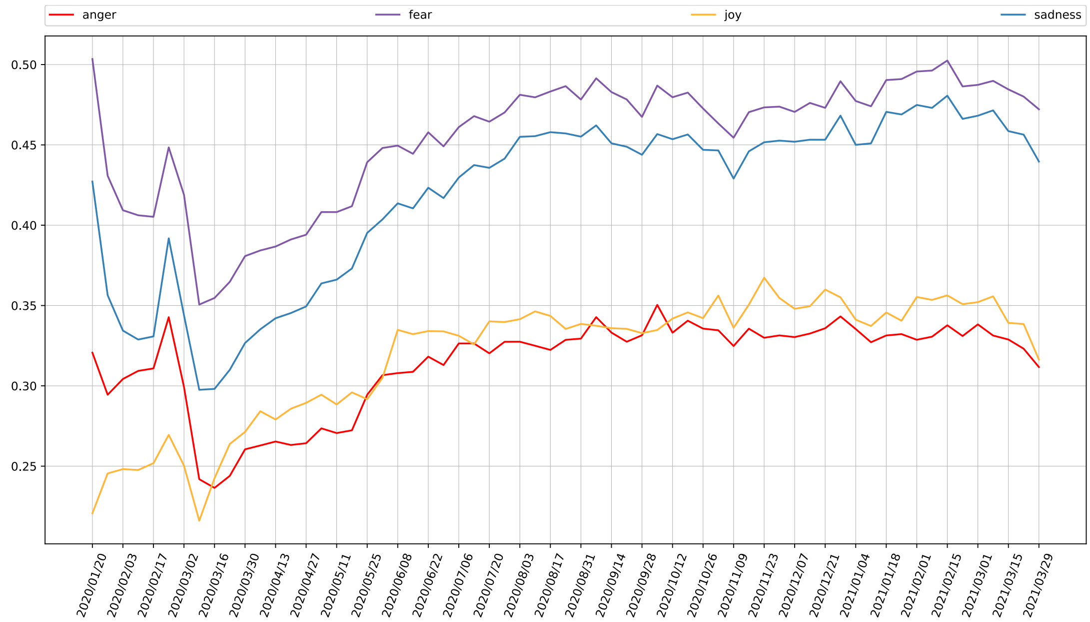

Figure 4.2: Proportion of weekly users per emotion in the English tweets

Figure 4.2 instead shows the proportion of users that expressed a particular emotion in a given week, considering the English tweets. Here we can notice some interesting results: first of all, there are some cases where all the emotions seem to have the same course. We would have expected that, if negative emotions increased, positive emotions would decrease (and vice versa). Instead, during the week starting on 2020/02/24, the proportion of users that expressed anger, joy, fear, and sadness raised. This probably happened because users during that time were particularly emotional and used several more words within a tweet. However, that is not always the case: for example, during 2020/07/13 the joy decreased, while fear, and sadness increased.

Secondly, it is also possible to observe how time frames of greater joy alternate with periods of greater anger. However, that is not the case for fear and sadness. This probably happens because there are a lot of words that convey both these emotions at the same time. In practice, when one increases, so does the other.

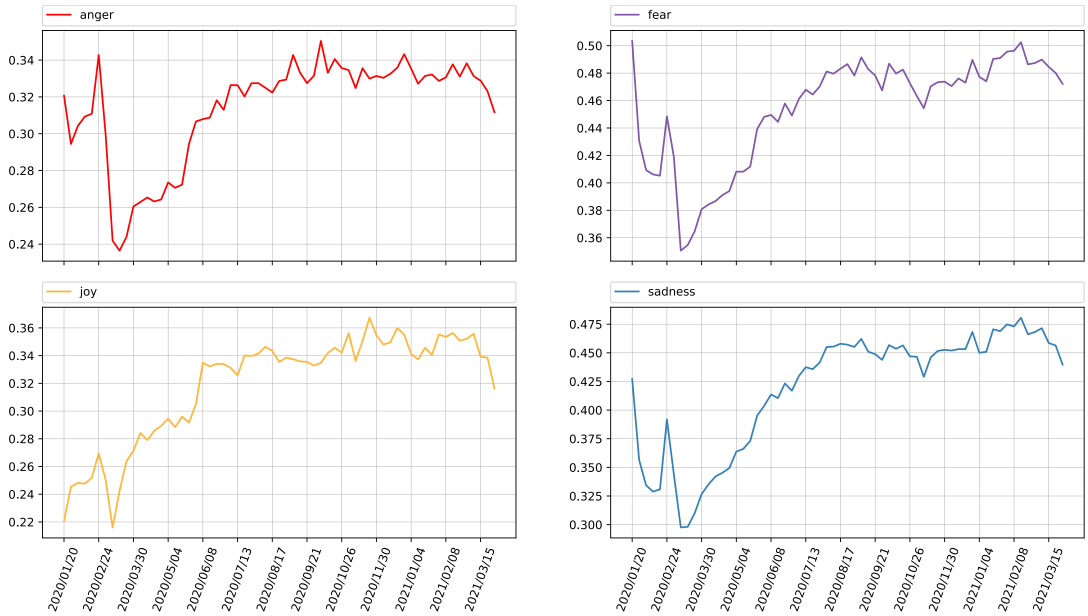

Figure 4.3: Proportion of weekly users expressing a particular emotion in the English tweets

Before approaching the data normalization, we have also tried to divide the different emotions courses into subplots. If we take a look at Figure 4.3, both global (and local) maxima and minima are visible to the naked eye. However, the comparison between different emotions becomes really difficult.

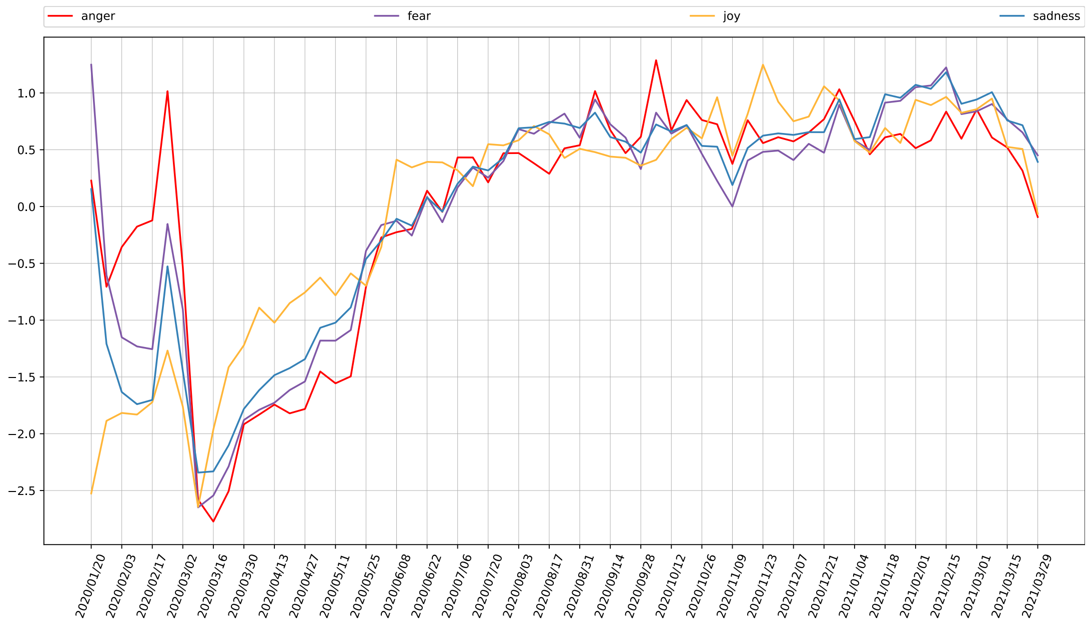

Figure 4.4: Z-score of weekly users per emotion in the English tweets

The z-score instead proved to be a very useful data normalization. As we can see from Figure 4.4, not only peaks are noticeable even if we are considering the emotions altogether, but we can also understand which time frames are characterized by a greater (or lower) emotion value w.r.t. the mean value of that particular emotion. For example, it is possible to see that, starting on 2020/03/16, the proportion of users that expressed joy w.r.t. the mean value of joy over the whole period, surpassed all the other emotions w.r.t. their mean value over the whole period.

From the comparison of the course of the emotions we have selected some peaks (i.e. week) manually to understand which words were the most used.

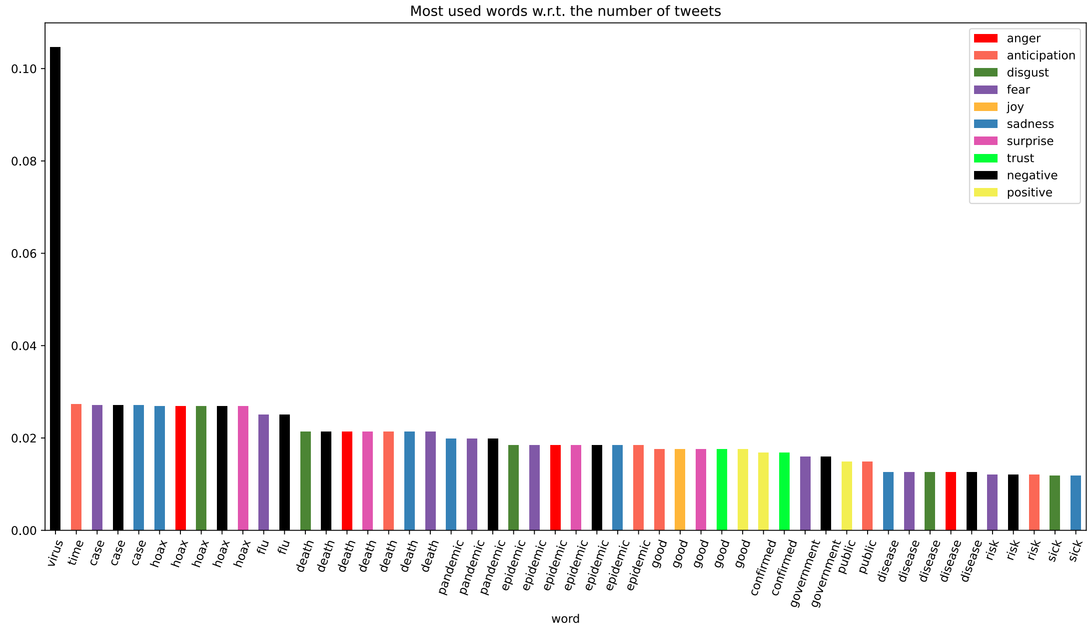

Figure 4.5: Proportion of most used words on 2020/02/24 that express emotion/sentiment in the English tweets

Figure 4.5 shows which are the 50 most used words in the English tweets during the week starting on 2020/02/24.

Here, we can see that some words, such as “virus”, are only related to a sentiment and not to any other emotion. On the other hand, “epidemic” is associated to almost the totality of the emotions.

It is also very interesting that the word “virus” was present in 10% of the tweets. Moreover, we can also see that this particular graphics validates the previous hypothesis: indeed, there are lot of words that convey both sadness and fear. In particular, we can notice that it is the case for “case”, “death”, “pandemic”, “epidemic”, “die”, and “disease”.

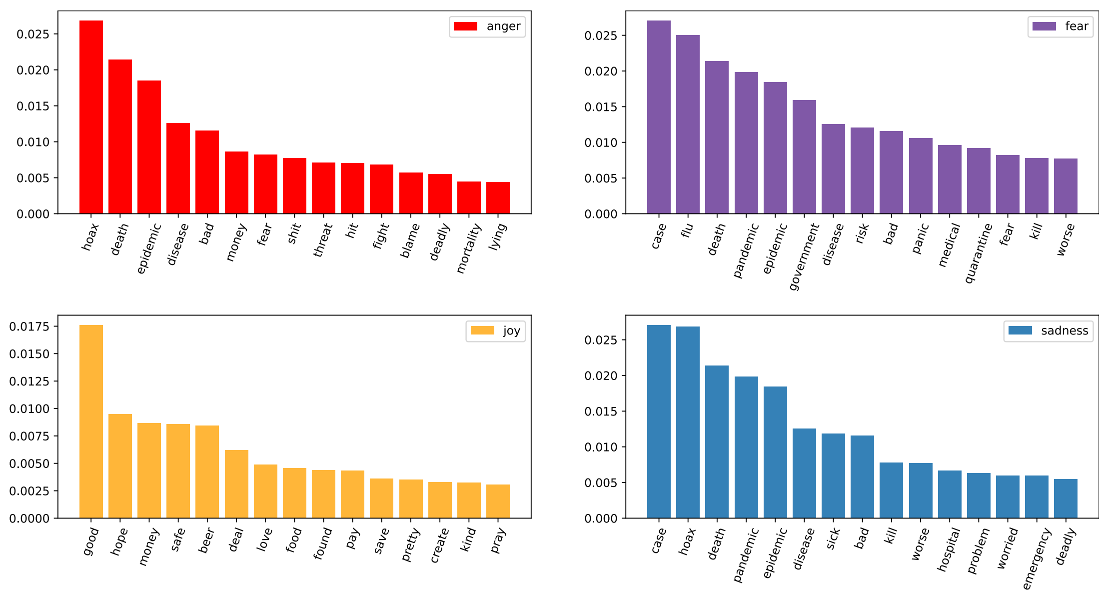

Figure 4.6: Proportion of 15 most used words on 2020/02/24 per emotion in the English tweets

Instead, Figure 4.6 shows the 15 most used words for anger, fear, joy, and sadness. if we look at the subplots of fear and sadness, we can clearly see that the tweets contain more words expressing fear than sadness. This explains once again the course of the emotions in Figure 4.2.

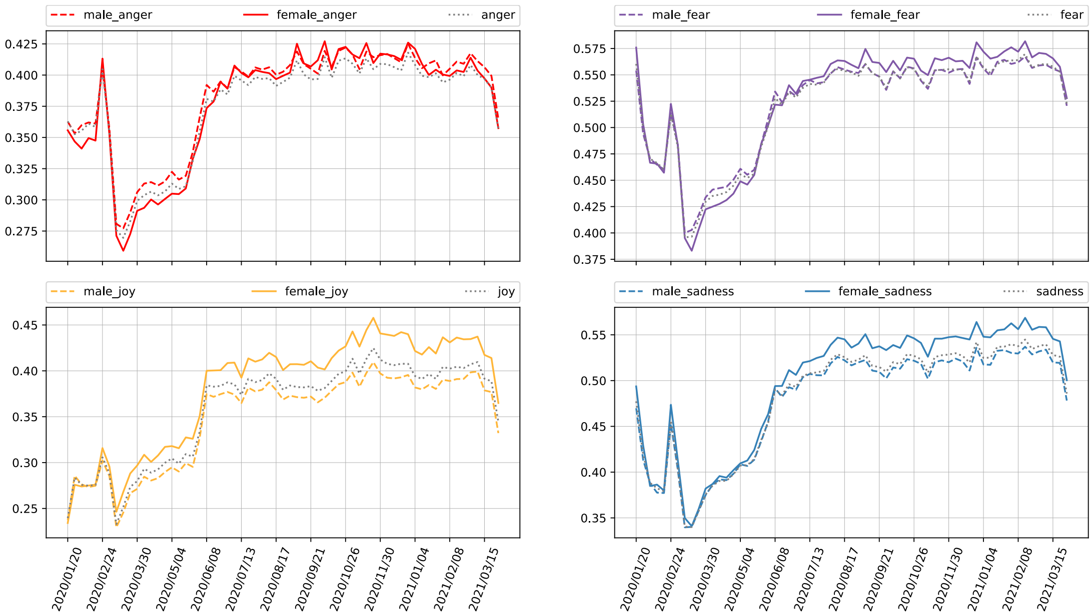

Figure 4.7: Proportion of weekly men/women expressing a particular emotion in the English tweets

Figure 4.7 introduces the results of the analysis of the emotions expressed by females and males in the English tweets during the whole period. In this case, the dotted gray line represents the proportion of users that, independently from the category they belong to, expressed a certain emotion on a given week. However, this line differs from the results obtained in Figure 4.3, because we are now considering only the users with a certain number of tweets (see Section 3.6.1 for further details).

We can also see that women tend to write a lot more tweets that convey joy in most cases. This seems to be also the case for sadness and fear. However, if we consider for example the fear, the gap is very small. Instead, men seem to express slightly more anger over the considered period, but the two lines mostly overlap.

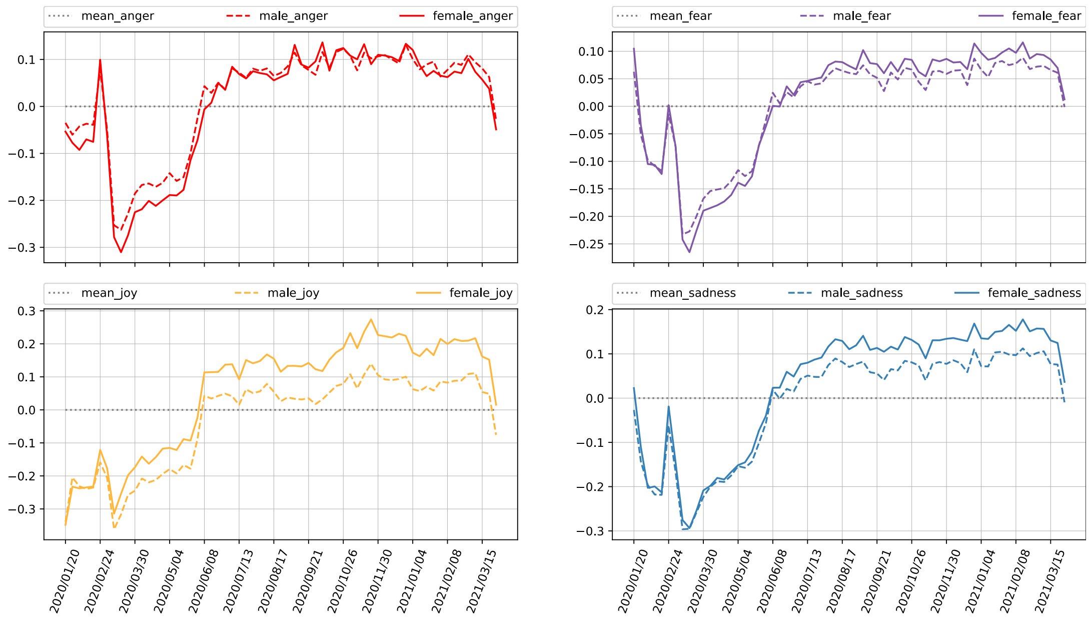

Figure 4.8: Men/Women expressing the weekly proportion of a particular emotion w.r.t. the average value among all users

Figure 4.8 is a variation of Figure 4.7, where 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. Instead, the proportion reported by the other lines represents how much a certain gender expressed a particular emotion in a given week w.r.t. the mean value among the users. For example, both males and females expressed 10% more anger during the week starting on 2020/02/24.

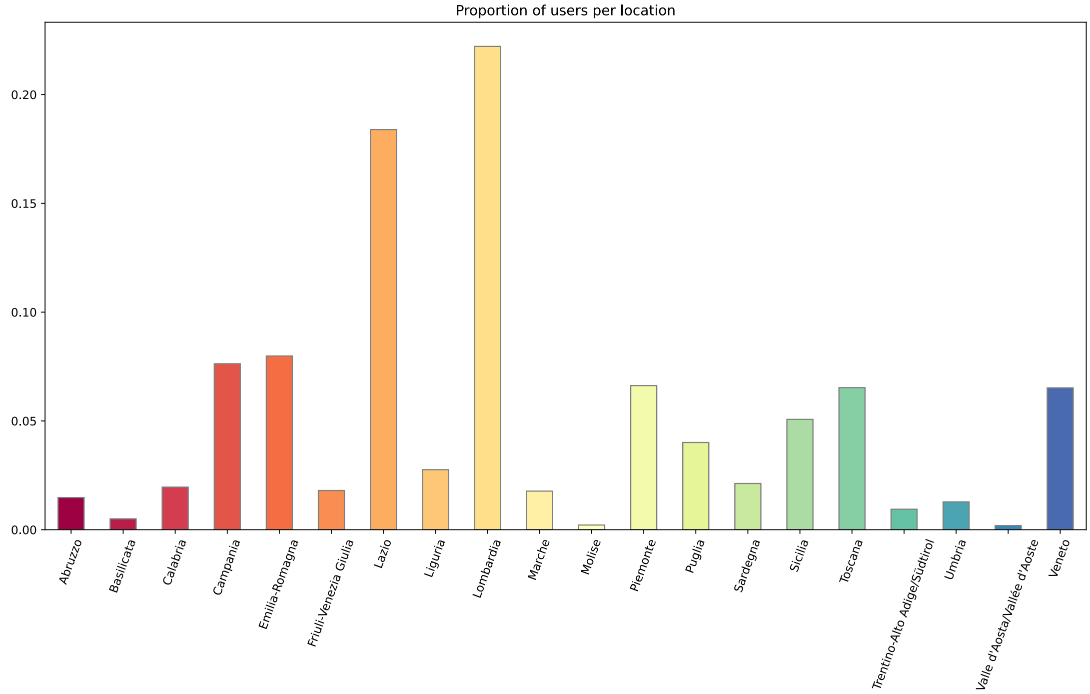

Figure 4.9: Proportion of users per state in Italy

Finally, we can now move to the analysis of the users' emotions from a specific location. Figure 4.9 shows the proportion of users per state in Italy. From the results that we have obtained, we decided to consider only Lombardia, Lazio, Emilia Romagna, and Campania for further studies. In this way, we are able to better visualize the data and we have more possibilities to get stabler results.

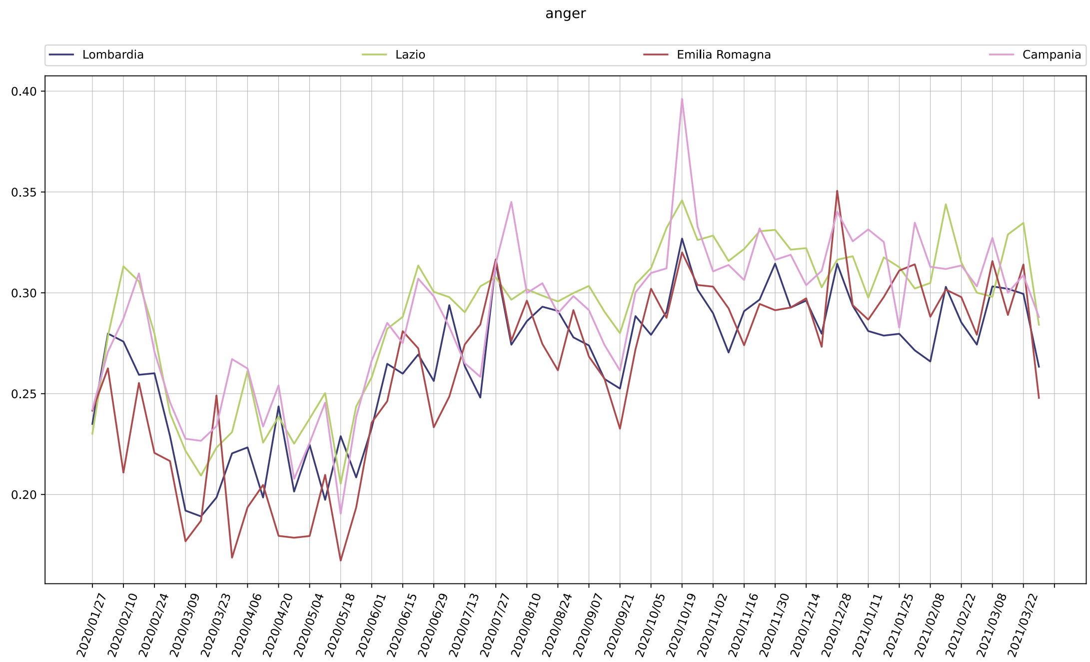

Figure 4.10: Proportion of weekly users that expressed anger in Lombardia, Lazio, Emilia Romagna and Campania

From Figure 4.10 we can actually understand which events had an impact on every state. For example, it is possible to notice a peak of anger on the week starting on the 2020/10/19. Here it is a bit more difficult to understand what actually happened without more complex techniques. However, from the analysis of the news, it is possible to link that particular time frame with the first of a series of different decrees, issued to stop the spreading of the disease once more after the first lockdown.

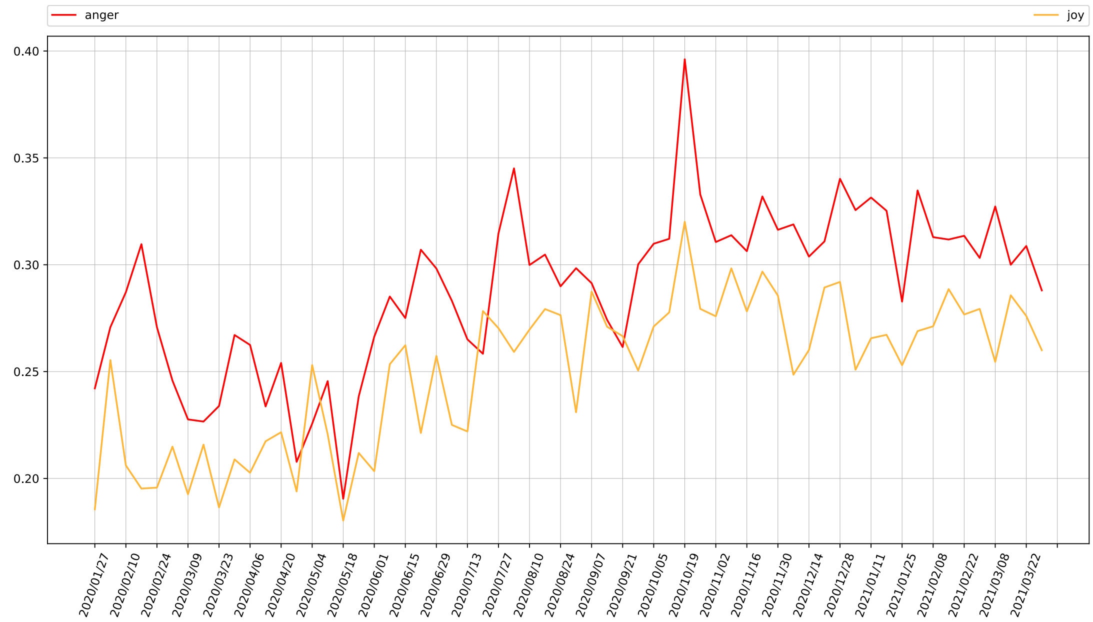

Figure 4.11: Proportion of weekly users that expressed anger and joy in Campania

Of course, it is also possible to consider a single state, for example Campania, as showed in Figure 4.11. Here we can notice even more how, by looking at the week starting on 2020/02/17, the news of the first coronavirus case in Codogno, had an impact that people's emotions. Unfortunately, given that these results were obtained with a lexicon that has not been originally designed for the Italian language, they must be taken with extreme care. Moreover, trying to manually associate events cannot really be considered a best practice to end up with valid and reliable results.

5 Conclusions

Bibliography

[1] Luca Maria Aiello, Daniele Quercia, Ke Zhou, Marios Constantinides, Sanja Šćepanović, and Sagar Joglekar. How epidemic psychology works on social media: Evolution of responses to the covid-19 pandemic, 2020.

[2] Julian Ausserhofer and Axel Maireder. National politics on twitter. *Information, Communication & Society*, 16(3):291–314, 2013.

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[4] Saif Mohammad. Nrc emotion lexicon. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Online; accessed 13-June-2021].

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

[6] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067. ACM, 2019.

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

[8] Wikipedia contributors. Openstreetmap — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=OpenStreetMap&oldid=1028995235, 2021. [Online; accessed 17-June-2021].

[9] Wikipedia contributors. Sentiment analysis — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=1024880646, 2021. [Online; accessed 12-June-2021].

[10] Wikipedia contributors. Twitter — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Twitter&oldid=1027840990, 2021. [Online; accessed 12-June-2021].