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	Co-Supervisor	Student
Alberto Montresor	Cristian Consonni	Simone Alghisi
	David Laniado	

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 counter measures better handled the situation while offering the best possible trade off between people's satisfaction and reducing 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
- 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
- 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 certain 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 place. 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).

Questa ultima sezione ha bisogno di essere rivista successivamente, una volta deciso se usare i risultati di LWIC o di Emolex

The analysis of the English datatset revealed some first interesting results: it seems that females are more inclined to express joy and sadness; males instead, more anger.

During the course of the project I had the possibility to personally contribute to m3inference improvement 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 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 [4].

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 and only reason behind this critical choice, is the fact that it was the only possibility. Obviously, the data from other platforms (e.g. Facebook) could have been interesting. However, it is either too difficult to get the data (due to particular limitations) or get enough data. For this reason, Twitter was the only option.

Nonetheless, Twitter remains a very interesting social network where, in the majority of the case, posts are public and everyone can see them (i.e. there are less privacy related issues), and it is particularly easy to have access to a considerable amount of data.

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 [3]. 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.

2 Data collection

The dataset used for the project is the echen102/COVID-19-TweetIDs GitHub repository [1]. The repository contains an ongoing collection of tweet IDs starting on the 28th of January 2020.

The IDs in the dataset are either tweets

- from one of the following accounts

  - PneumoniaWuhan
  - CoronaVirusInfo
  - V2019N
  - CDCemergency
  - CDCgov
  - WHO
  - HHSGov
  - NIAIDNews
  - drtedros

- sampled real-time, because they mention specific keywords, such as
  - Coronavirus
  - Epidemic
  - covid-19
  - Social Distancing
  - panic buy
  - lockdown
  - ...

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

- consider more countries
- 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

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 described by the following fields:

- `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
- `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

[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`

- `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
- `profile_image_url_https`, a HTTPS-based URL pointing to the user's profile image

For the scope of the project, we decided to keep a simpler representation of the original json object, with only the most relevant fields:

```
{
  "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 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.

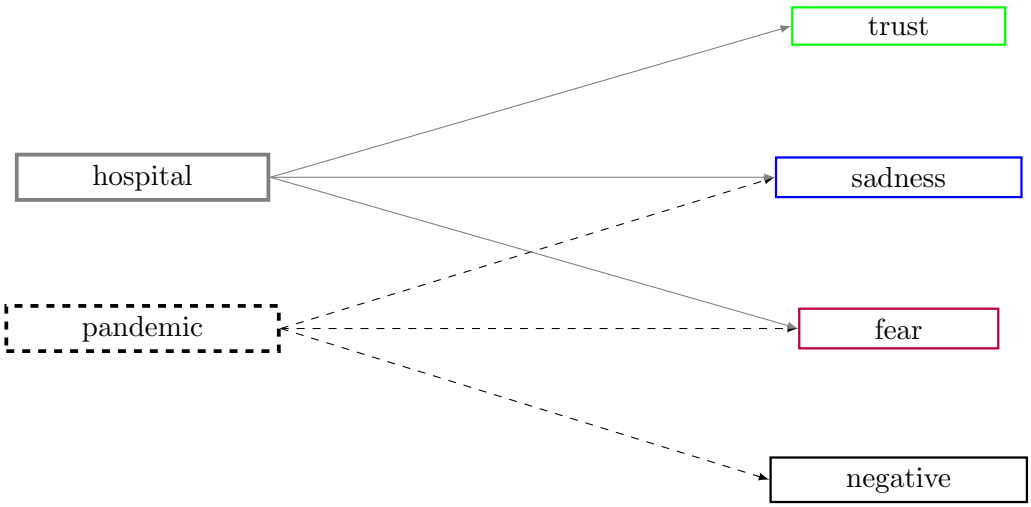

Figure 3.1: Word-emotions/sentiments association

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 simple negation can totally change the meaning of a particular quote:

I am fine / I am not fine

For this reason, results obtained 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 sentiment analysis [2]. 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 sentiment analysis 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 is 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.

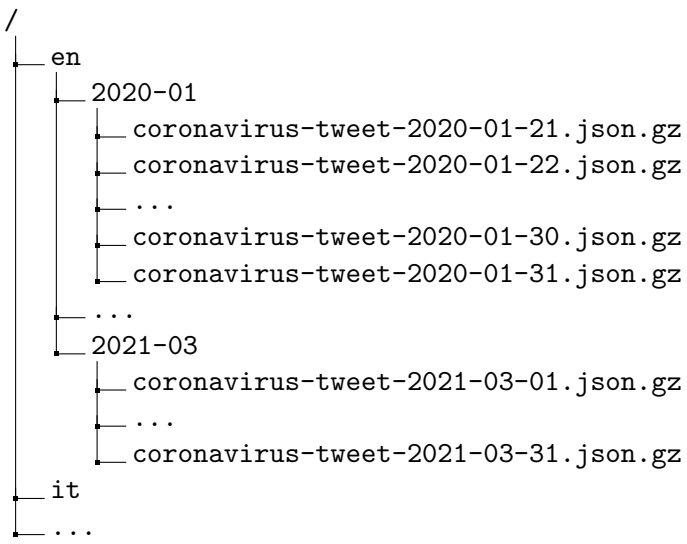

Figure 3.2: First tweets organization

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 average the results more, get stabler data for languages with fewer tweets, and obtain a clearer visualization.

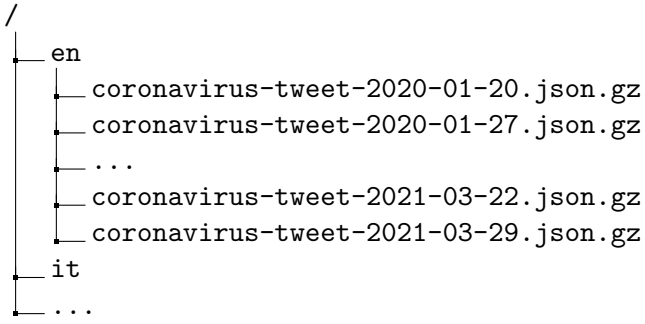

Figure 3.3: Weekly tweets organization

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.

Bibliography

[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. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

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

[3] 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].

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