UNIVERSITÀ DI TRENTO

Department of Information Engineering and Computer Science

Bachelor's Degree in Computer Science

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

Special thanks

...thanks to...

Contents

Abstract

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. For this reason, I immediately thought that the research proposed by Eurecat could be very interesting: the possibility to study how people perceived all of this situation, and better understand which measures were more welcomed than others, seemed really fascinating and, above all, may be useful in the case of some other unfortunate event. Moreover, I never had the chance to explore the field of Data Science, to use tools such as Lexicon to perform an analysis of the emotions based on the text, or worked with such an impressive amount of data to obtain valid results.

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.

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) available on

Python. 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).

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.

Questo ultimo trafiletto ha bisogno di essere rivisto 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.

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.

1.1 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 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.

1.2 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 into into two different groups w.r.t. their gender, to study the different emotional response of males and females, and also based on their location.

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.

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 tweets IDs, starting on the 28th of January 2020, from specified accounts and also real-time tweets that mention specific keywords.

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%
Undefined	und	20 560 392	8 590 707	29 151 099	2.76%
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%

Table 2.2: Top 10 languages with the most tweets

2.1 Tweets

To comply with Twitter's Term of Service, tweets cannot be released publicly: the repository is in fact a collection of tweets IDs. The original tweets can be retrieved, or hydrated, using the Python library Twarc with a Twitter Developer Account. Given an id, Twarc simply uses the token of the associated developer account to contact the API, and returns the corresponding tweet as a json object.

The original structure of the tweets was changed, in order to consider only the 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"
  }
}
```

2.2 Analyzed period and languages

We have decided to consider the period from January 2020 to March 2021

3   Conclusioni

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

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.

Attachment A   Titolo primo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1   Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1.1   Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

Attachment B   Titolo secondo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1   Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1.1   Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.