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

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 the Big Data Department of Eurecat - Centro Tecnológico de Catalunya, and supervised by *Cristian Consonni* and *David Laniado*, could be very interesting: the possibility to study, during my traineeship, 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.

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) 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).

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.

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 - 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.

1.2 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 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.

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.