Sentiment Analysis of metropolitan Covid-19 related tweets

Anastasia PENKOVA
Master in Computer Science
1st semester
University of Passau
penkov01@ads.uni-passau.de

Bochra SMIDA
Master in Computer Science
1st semester
University of Passau
smida02@ads.uni-passau.de

Amal ABDERRAHMANI
Master in Computer Science
1st semester
University of Passau
abderr01@ads.uni-passau.de

Abstract

In process... After phase 4 it will be added

1 Introduction

Nowadays social media is a crucial part of people's lives. People build communities and share their thoughts, opinions, ideas and news about everything in social media such as Twitter, Facebook, Instagram, etc. One of the most relevant social networks to collect data is Twitter.

From the length of messages limited to 140 characters to the simplicity of the interface, everything in Twitter has been designed to benefit directly from the exponential development of mobile telephony. In addition, the choice to implement open source tools and to make the programming interface (API) of the service accessible to third parties was decisive for its success and, at the same time, favored the emergence of an ecosystem made up of dozens of interoperable services that make it possible to consult Twitter without going through a web browser: geolocation (FourSquare), link shorteners (Bit.ly), photo and video hosting (Twitpic), search engines (Topsy), etc. Twitter is therefore primarily neither a closed platform like Facebook, nor a portal, but a news feed updated in real time and viewable on a wide variety of media as well as a raw information dissemination medium.

Since the pandemic of COVID-19 began and the world wide organisation suggested the self quarantine, people are exposing their social related issues through social media . Twitter itself welcomes developers and data scientists to study conversations and disputes about COVID-19 in real-time. Especially for this situation, developers of Twitter adapted Twitter API to this topic. The COVID-19 Twitter endpoint does not have data volume and throughput limitations and also it is completely reliable compared to other endpoints. In that way, Twitter became the best solution to gather these contents and use them for analysis purposes.

2 Sentiment analysis

Many works and research papers took advantage of the twitter content to study the effect of the pandemic on different aspects using text mining and sentiment analysis. Actually, Text mining is the process that uses natural language processing (NLP) to transform unstructured data into ordered and understandable information. Its advances in hardware and software technologies lead to the availability of different types of data. This field encourages also to create large numbers of data which are easy to store and process.

A part of text mining is sentiment analysis also known as opinion mining. It aims to determine and classify a variety of opinions, sentiments and attitudes in subjective data, mainly found in texts. In fact, understanding sentiments from online social networking can help to understand the dynamics of the network related to location, time or events.

3 Related work

Since the pandemic began and countries started to apply necessary measures to stop spreading the virus, many researches were published on the topic of the influence of COVID-19 on social media and especially on Twitter.

In the first studies "Is working from home the new norm?", analysis is based on a large geotagged COVID-19 Twitter dataset over the USA [**1**]. The study presented an overview of tweet distribution by periods of day, found geographical patterns, evaluated work engagement comparing tweets during weeks and classified emotions using sentiment analysis and emoji classification. Sentiment analysis was applied on specific cases, for example, when the 1000th death was reported.

They also studied work engagement patterns after the announced lockdown.

In the second paper [**2**] Twitter Sentiment Analysis During COVID-19 Outbreak in Nepal, Bishwo Prakash identified emotions of tweets which was published during the coronavirus pandemic by users who live in Nepal.

Anna Kruspe et al. (2020) [**3**] in their research focused on automatic methods for cross-language sentiment analysis of European Twitter messages during COVID-19 pandemic. They studied about 5 million COVID-19 related tweets in different languages. They grouped tweets by countries and applied their method on each country separately and averaged them over time. Tweets determined to have one of 3 sentiments: positive, neutral or negative.

In the fourth research "In the Eyes of the Beholder: Analyzing Social Media Use of Neutral and Controversial Terms for COVID-19" [**4**] Long Chen et al. analyzed 2 groups of COVID-19 related tweets with:

- controversial terms, such as "Chinese virus". These tweets were more about Chinese impact of this pandemic rather than just analyzing the situation

- non-controversial, where tweets are more often about how to stay safe and statistics of spreading of virus.

However, in both groups of tweets the sentiment analysis declared negative sentiment, in the second group there was a slight more positive mood of tweets.

4 Problem domain

We find it therefore relevant and of utmost importance to follow in these researchers' footsteps and answer the following questions:

- How does polarity in sentiments correlate with the number of cases of Corona Virus in different metropolises?
- What effects do news and other subtopics have on people's moods?

We approach that by applying text mining and sentiment analysis techniques to collected twitter messages, grouping them by big cities and correlating the results with the evolution of the number of COVID-19 cases. In a second phase, we correlate those sentiments to specific events happening since the beginning of the pandemic. This will allow us to add knowledge to our understanding of the psychological reaction of people with the spread of the pandemic.

acl2015 times url latexsym hyperref

biblatex sample.bib

graphicx [super]nth pgfplots float

5 Description of data

We chose 2 datasets: one from Kaggle https://www.kaggle.com/gpreda/covid19-tweets and another one was provided by our tutor Sahib Julka where Covid-19 related tweets were collected using Twitter API. The principal of collection was to take those tweets that consist hashtag #covid19. The second dataset includes tweets and their metadata, which were published in Germany.

We merged these 2 datasets and chose those columns, that were included in both datasets.

The number of tweets by the moment of analyzing was 252 961 tweets with 7 columns: username, user location, number of followers, friends and favorites, date of publishing tweet, text of tweet itself.

Counting how many unique records we have, we got overall 116 571 unique users, 33275 unique locations:

Dataset was reduced due to the lack of some lines of text, which is the main requirement for further revealing sentiments in the texts. We got 213 474 entries.

Then we found top 10 active users in the dataset:

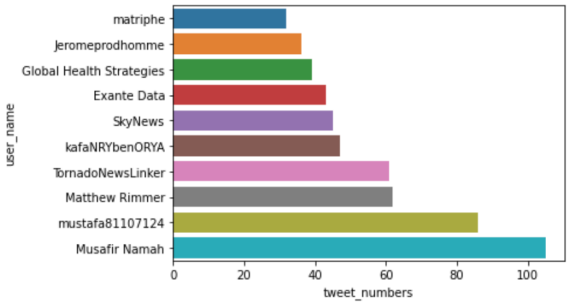

Figure 1: Top 10 active users

Next step was to choose 4 cities in order to apply our model in further step. We chose 4 big cities: Berlin, London, Los Angeles and New York.

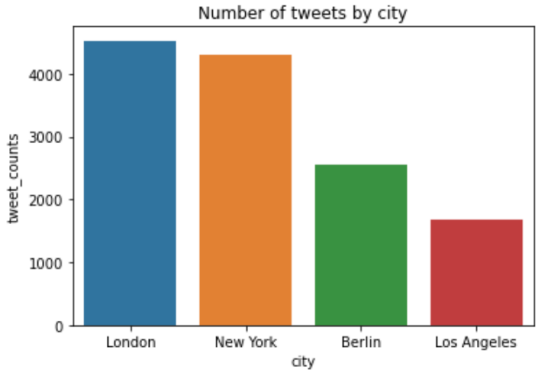

Figure 2: Number of tweets by 4 cities: Berlin, London, Los Angeles and New York

In the plot below we can see the activity of publishing tweets day by day from 24th of January to 9th of August. Since we have merged 2 datasets, there is a very sharp growth on July 27. Because the first dataset was focused to collect tweets from 27th of July until 9th of August. And the second one collected tweet from January 24, 2020 until April 21, 2020

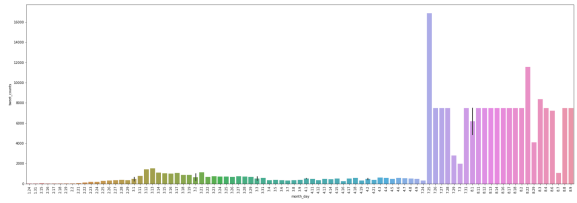

Figure 3: Number of tweets by month and day of the month

Below we can see how the activity of publishing tweets was in each chosen city. As most of tweets in Berlin was collected from the second dataset, that was focused on German tweets, it shows that the biggest number of tweets was published in the first half of March.

Figure 4: Number of tweets by city, month and day of the month

Then we analyzed what time the most activity is. The most active period of day is early morning and period from 15:00 until 18:00:

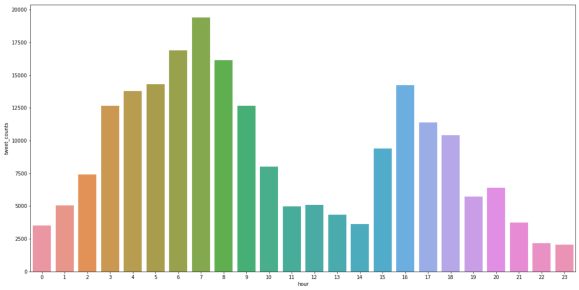

Figure 5: Number of tweets by hour

Here the plot shows activity of publishing by days of week. The most active day is Saturday:

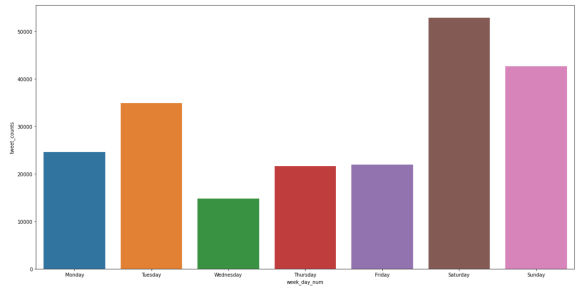

Figure 6: Number of tweets grouped by day of the week

In the plot below we got results grouping dataset by day of the week and city. The result as in the previous plot shows us that the most active day of publishing is Saturday:

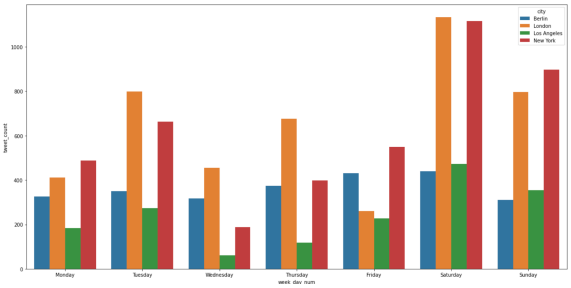

Figure 7: Number of tweets grouped by city and day of the week

6 Pre-processing Phase

The pre-processing phase consists mainly of cleaning the tweets and vectorizing them.

- **Cleaning the tweets:**
  The tweet texts were also lower-cased and cleaned using a function that removed hashtags, links, numbers and information irrelevant to us.

```
In [10]: train_data.head()
```

Out[10]:

	polarity	text
0	0	- awww, that's a bummer. you shoulda got da...
1	0	is upset that he can't update his facebook by ...
2	0	i dived many times for the ball. managed to s...
3	0	my whole body feels itchy and like its on fire
4	0	no, it's not behaving at all. i'm mad. why am...

Figure 8: Cleaned tweet texts

- **Vectorizing the tweets:**
  We chose TF-IDF (“Term Frequency and Inverse Document Frequency”) to create the vectors we will be using later. It is a popular method that selects the most interesting words in a document. It was our choice because thanks to the frequency score it calculates, it lowers the value of words such as “this” and “in” since they are present in almost every tweet. The implementation is done thanks to the open source python library sickit-learn.

7 Rule-Based Sentiment Analysis

As a first approach, we will be trying to identify emotions using a rule-based method, in other words using manually crafted rules and lexicons.

7.0.1 The NRC lexicon:

The NRC Emotion Lexicon is a list of 14182 English words and their associations (1 means ”is associated” and 0 means ”is not associated”) with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.”

	word	positive	negative	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
0	aback	0	0	0	0	0	0	0	0	0	0
1	abacus	0	0	0	0	0	0	0	0	0	1
2	abandon	0	1	0	0	0	1	0	1	0	0
3	abandoned	0	1	1	0	0	1	0	1	0	0
4	abandonment	0	1	1	0	0	1	0	1	1	0

Table 1: An exemple of entries in the NRC lexicon

7.1 Methodology used:

In this phase, The tweet texts are represented by a dataframe containing the following columns: the text,the location, the date and other informations about the tweets. The text is cleaned in the pre-processing phase. Using the sickit-learn count Vectorizer and a vector of vocabulary from the NRC Lexicon, we can identify the sentiments present in each text while parsing it and detecting the words present in it then summing the occurrence of the words associated to the same sentiment.

	user_location	date	text	is_retweet	Subjectivity	Polarity	anger	anticipation	disgust	fear	joy	sadness	surprise	positive	negative
30	United States	2020-07-25 12:26:21	fema acknowledges puertorico lacks rebuilt hom...	False	0.000000	0.000000	0.0	0.2	0.2	0.2	0.4	0.4	0.0	0.4	0.4
75	India	2020-07-25 12:25:05	covid to shrink power sector growth, take disc...	False	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
76	India	2020-07-25 12:25:05	saturdayvibes: the current situation calls for...	False	0.322222	0.111111	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
120	India	2020-07-25 12:23:44	bihar witnesses biggest single-day spike of 2...	False	0.454545	0.136364	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.5
151	United States	2020-07-25 12:22:55	how about everyone can spread covid19. another...	False	0.633333	-0.250000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Sentiments in tweet texts

The best way to have the final sentiment is to weigh each sentiment in a text. We have chosen to define the weight in our work as the frequency of appearance of this sentiment in the texts of the dataset.

The final labeled sentiment to a tweet is thus the one that has the maximal weight and is present in the parsed text.

freq_anger	freq_anticipation	freq_disgust	freq_fear	freq_joy	freq_sadness	freq_surprise
0.176556	0.304371	0.118278	0.337748	0.18755	0.26	0.162781

Table 3: Sentiments frequency in tweet texts

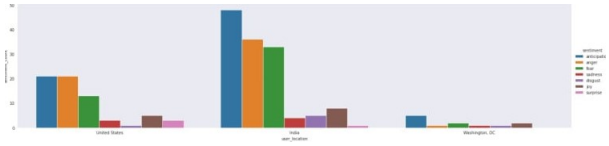

Figure 9: Sentiment counts per city

8 Machine learning approach to classify emotions

Machine Learning classification methods consist in representing each tweet as a set of variables, then building a model from examples of tweets for which we already know the label, in our case the label would be the emotions. The template is then used to classify a new untagged tweet. To do so, we need to gather data for training, vectorize it and create a machine learning model (in our case and for this phase we have chosen the SVM model) to train and test.

8.1 Classifying tweets into negative, positive and neutral tweets:

- **Training dataset:**
  We were able to find an already trained dataset called Sentiment140 [documentation: For Academics]commonly used for this type of classification that gives an insight of the sentiment of a brand, product, or topic on Twitter. This labeled dataset was obtained using the Maximum Entropy classifier. The data is a CSV without emoticons organised as follows:
  - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
  - the id of the tweet
  - the date of the tweet
  - the query
  - the user that tweeted
  - the text of the tweet

  First, we updated the polarity so that 0 is negative and 1 is positive (instead of 4) for easier labeling and understanding. We then removed unneeded columns (Id of the tweets, query, user, date) leaving only the tweet text and the polarity associated. The tweet texts were then preprocessed like mentioned in the previous part.

- **Training the classifier**:
  We have chosen the Support Vector Machine model to classify our tweets. It is a supervised machine learning algorithm that performs classification by finding the hyperplane that differentiate the classes we plotted in n-dimensional space. We use for its implementation scikit-learn, an open source machine learning library. The output of the algorithm is the label (0 for negative or 1 for positive) for each tweet.

8.2 Classifying tweets into emotion categories

Since there were no training datasets that have as labels the different emotions we need, we tried to create our own dataset using the rule-based method to determine sentiment in a text that we have discussed earlier. Our training dataset will contain texts which are labeled in one of the emotion categories. The SVM algorithm is used in this case also.

9 Deep learning approach to classify emotions

As a third approach, we have decided to try to get a better understanding of the text using BERT (Bidirectional Encoder Representations from Transformers), a pre-trained deep learning natural language framework developed by GOOGLE. BERT has been a great advancement in NLP since it allows the language model to discern word context in a text by learning information from both the left and right side of a token's context during the training using the word embedding.

BERT learns in an unsupervised way. The entry, in our case the training dataset (Sentiment140), is self-sufficient and there is no need to label it. BERT will try to find similarities between the different tweets and class them together. The learning is done in two main phases:

- **Pre-training:**
  the model is trained on unlabelled data over different pre-training tasks. The aim of this step is to understand the language of the input.

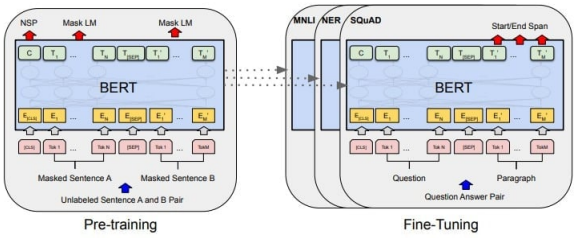

Figure 10: Modelisation of BERT Pre-training and Fine-Tuning

- **Fine-tuning:**
  this is the stage where we continue to update the weights trained from the pre-training phase, on the dataset in use for our problem. That means that in this phase the model will learn a specific task to answer the question we want.

10 Experiments and results

In this part we will be presenting the experiments our deep learning and machine learning models allowed us to do. The curves we will be presenting are the result of the prediction of the pretrained BERT model that we fine-tuned in accordance to our task. The next part will entail a comparison between the two models, but for now, we will be focusing on the observations our predictions tell us. Our analysis will be time-based and location-based. First, we will be visualising and commenting on the evolution of covid-related sentiment all over the world and in some countries, namely London, Stockholm and India. Then we'll proceed to country, or rather city based observations of emotion density. We should mention that only English tweets are analyzed, due to the fact that our models are trained on English emotion-labeled datasets.

10.1 Stockholm

The Stockholm tweets dataset dates from January 2020 to May 2020. It contains around 22000 tweets. For reasons of clarity, we chose to show sentiment evolution separately for each emotion. Figures 11, 12 and 13 show the evolution of anger, joy and sadness sentiments all over that time period. It seems that joy is relatively higher during the first two pandemic months but know a decrease starting in March and the following months. On the other hand, we can notice that the evolution of sadness reaches its highest levels in March. Anger, though existing throughout the tweets time period, also knows a peak in March. This lead us to try to understand the reason behind these fluctuations. In fact, it seems that Sweden, and Stockholm in particular, experienced a big rise in the number of death cases as shown in figure 14 and the spread of the pandemic had probably reached its maximum on March 5Th.

10.2 India

The Indian dataset contains around 30000 tweets dating from August 2020 to December 2020. We chose to plot the fear and sadness evolution and the two show similarly high peaks in the middle of the September (figures 15 and 16). That date coincides with the beginning of the great rise in corona virus cases. As shown in figure 17, we notice that the curve pertaining the number of Covid 19 cases starts growing considerably starting September, and continue doing so until January. That explains as well why the fear curve has another peak in November. When we visualize the joy curve, we notice that the joy levels are relatively low, except in September with a small peak. It seems that around that date, two popular yearly events take place in India: the International Day of Democracy and the Engineers Day

10.3 London

Using a dataset containing around 30000 tweets dating from January 2020 to May 2020, we were able to plot emotion curves that show as usual some fluctuations around particular dates. If we take a look at the fear evolution in London (figure 18) throughout these months, we can notice a slight relative peak in fear in the middle of March. The same observation is made about the sadness evolution in London (figure 19), though the sadness in London Covid-related tweets is almost always high. The beginning of the fear and sadness growth goes hand in hand with the rise of the daily new confirmed Covid19 cases as shown in figure 20. The optimism levels remain overall constant during this period.

10.4 Location-based Analysis

In this part, we try to figure some location patterns to the emotions' evolution in the month of July and August. For this, we used a dataset collected from different locations and containing around 180000 tweets. As shown in figure 21, it seems that in

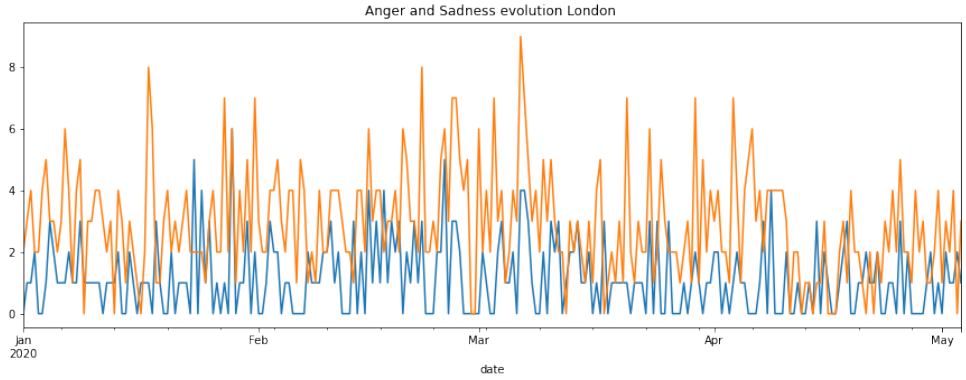

Figure 11: Anger and Sadness Evolution in Stockholm

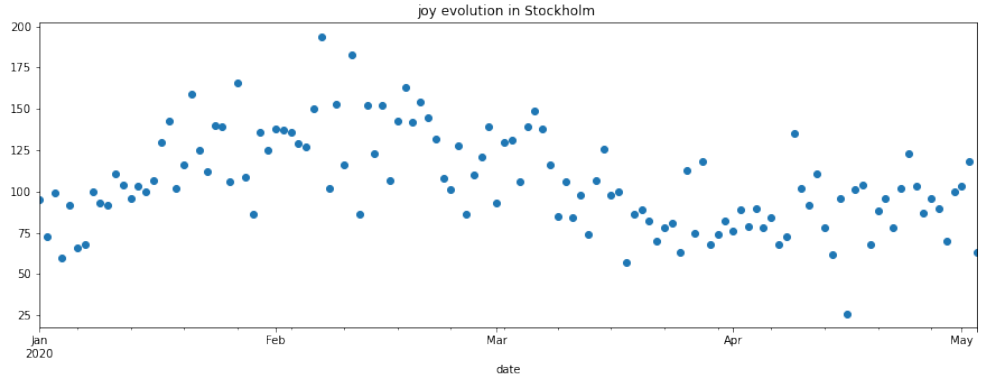

Figure 12: Joy evolution in Stockholm

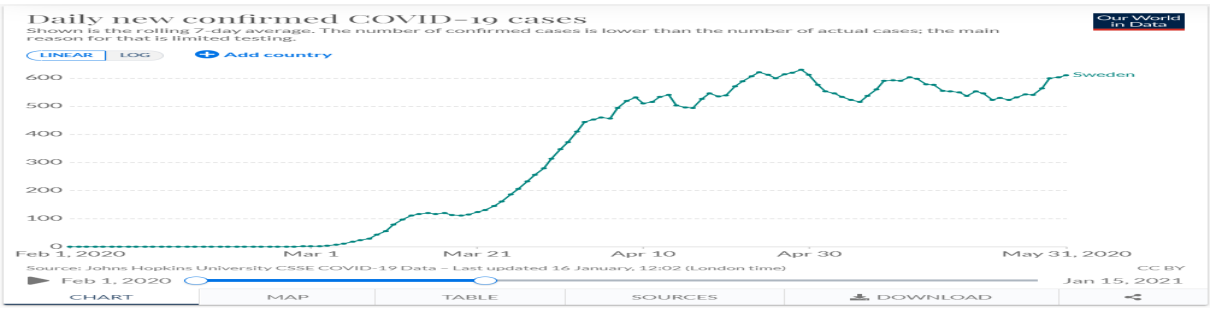

Figure 13: New daily cases evolution in Sweden

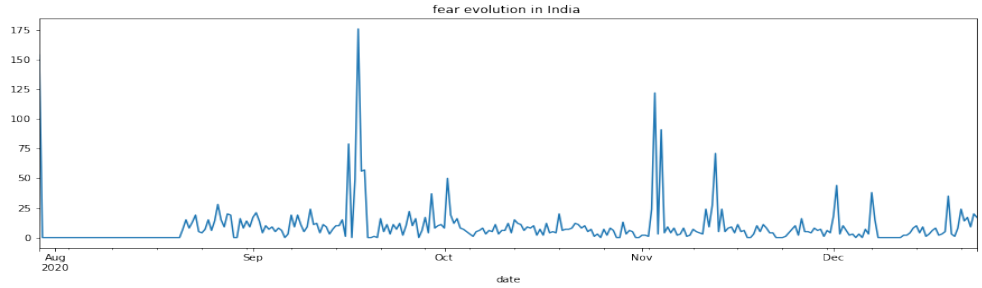

Figure 14: Fear Evolution in India

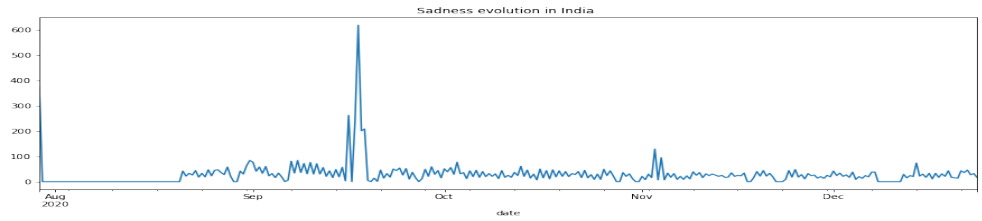

Figure 15: Sadness Evolution in India

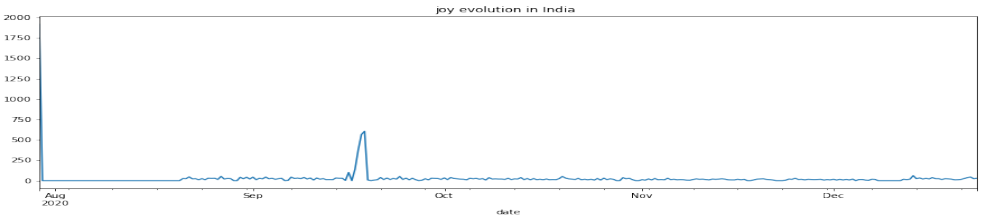

Figure 16: Joy evolution in India

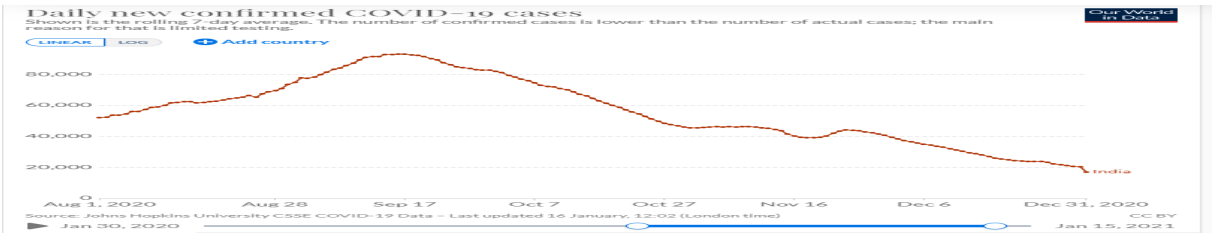

Figure 17: New daily cases evolution in India

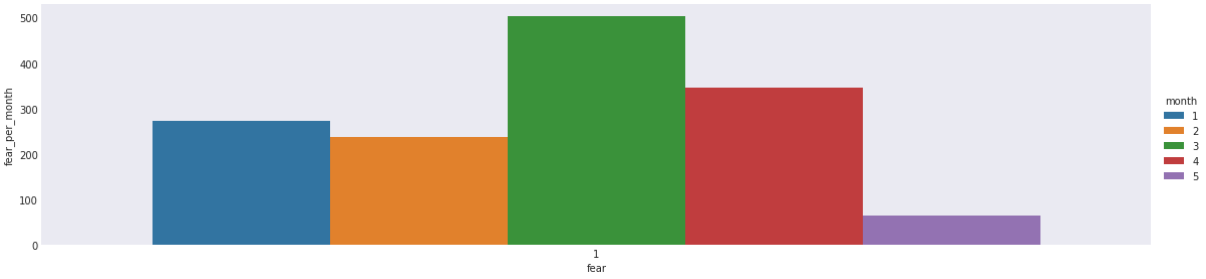

Figure 18: Fear Evolution in London

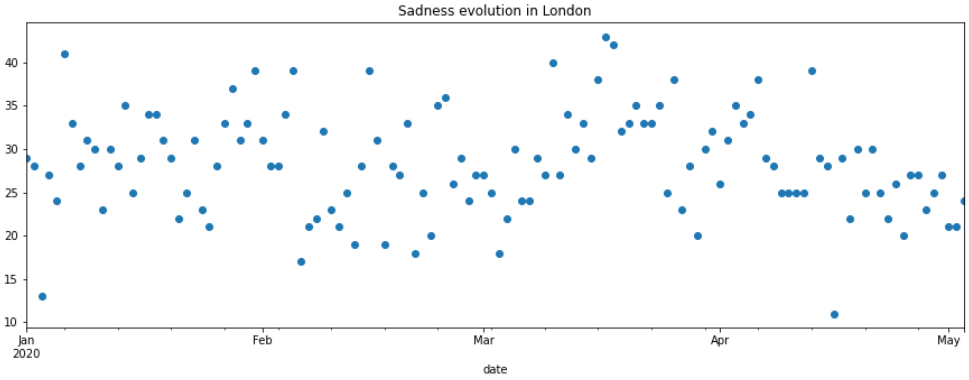

Figure 19: Sadness Evolution in London

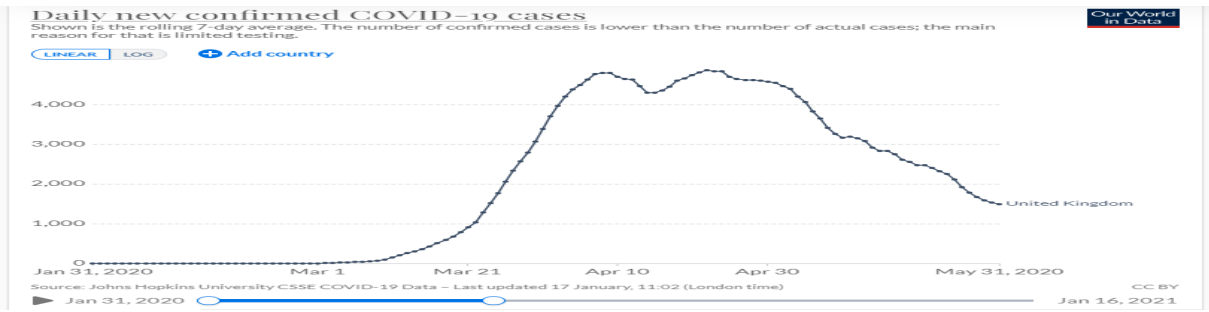

Figure 20: Daily new Covid 19 cases evolution in London

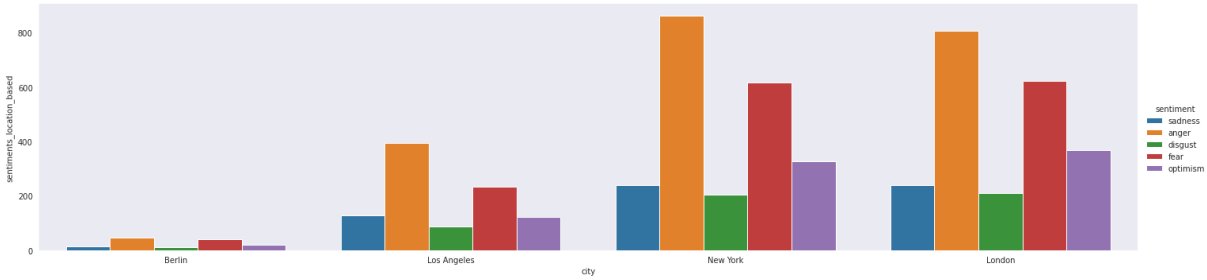

Figure 21: Emotions in different cities

most cities, the ambient sentiments are fear and anger, which is perfectly acceptable in the time of a pandemy. Berlin is the city with the least level of fears and anger. In fact, looking at the news in Germany during those two months, it seems that the population was accepting and understanding. The economy quickly recovered from the Corona crash and the people were able to adapt to the measures of online schooling and home office. On the other hand, New York city shows significant levels of fear and anger. According to Our Town, a blog pertaining NYPD news, at that time, the city faced a Covid spike among young adults. That period also witnessed anti-police protests, which explains the highest level of anger. Los Angeles knew a disturbing surge in coronavirus cases in mid-July, however the county started to see those numbers drop in the beginning of August. Less social problems were reported in LA according to abc7 News

References

[1] Y.Feng, W.Zhou 2020. *Is working from home the new norm*, Prentice-Hall, Englewood Cliffs, NJ. arxiv.org:2006.08581.

[2] B. Pokharel. 2020. *Twitter Sentiment analysis during COVID-19 Outbreak in Nepal*, Nepal Open University, Nepal.

[3] A.Kruspe, M. Haberle, I. Kuhn, X. Zhu. 2020. *Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic* Germany. arxiv.org:2008.12172.

[4] L.Chen, H.Lyu, T.Yang, Y.Wang, J.Luo. 1981. *In the Eyes of the Beholder: Analyzing Social Media Use of Neutral and Controversial Terms for COVID-19* arXiv:2004.06307.

[5] Y.Feng, W.Zhou 2020. *Is working from home the new norm*, Prentice-Hall, Englewood Cliffs, NJ. arxiv.org:2006.08581.