Davì\_Report

Valentina Davì

## Capstone project – Valentina Davì

### Introduction

I conducted this project together with my colleague Federica Giua. We decided that we wanted to try and collect data from Twitter, and we chose a topic related to the current situation in our country, Italy. We chose to focus on tweets relating to the phenomenon of Coronavirus.

What we first hypothesized was that, due to the big increasing of media attention to this virus, we could expect to observe an important amount of Italian tweets that linked the virus to the phenomenon of fake news, which is a diffused one in the society we live in. We had no idea of how big a dataset we could expect to observe, so we decided to obtain Italian tweets that contained both the keywords “coronavirus” and “fake news”, written in several ways.

### Getting data

In order to do that, we read the documentation of the rtweet package, and we tried retrieving the tweets using the function search\_tweets. To our surprise, we did not seem to be able to recollect the amount of tweets that we were asking for (about 20 000). The download of the data started, then kept getting to 5/6% of the process, and then it interrupted without any warning or error, giving us about 500 tweets. We repeated the attempt several times, changing bits of the code, but we kept having the same results over and over again.

After some days of research, we found some explanation in the fact that the package doesn’t return any error or message if it interrupts the research because it cannot find the exact amount of tweets asked. So we understood that, while we were asking for 20.000 tweets, R was giving us only 500 because they were the only ones existing at said time.

We decided to keep what we had, about 500 tweets containing both the keywords “coronavirus” and “fake news” written in different ways (“fakenews”, “fake news”, “#fakenews”), and we decided to construct a second dataset, with the Italian tweets that contained only the keyword “coronavirus”.

It should be noted that the R package rtweets doesn’t require a Twitter developer account to be run, and that Twitter limits the amount of data that one can obtain without a Premium API to those created in the last 6 – 9 days. We tried to get a little around this limit by repeating the download a second time after some days. We collaborated in writing the code to download the data, and then we decided to divide the job: I took on the dataset for the coronavirus only, while Federica was working on the dataset with both the keywords.

We downloaded the tweets a first time on February 10, then, we repeated the download after a week, retrieving only tweets that were more recent than the ones we had already downloaded, using the max\_id argument of the search\_tweets function. We then merged the two datasets each we had, creating only one dataset for the word coronavirus, and only one dataset with both the keywords. The dataset containing only “coronavirus” as a keyword had 87189 observations, while the one with both the keywords had 658.

### Frequency plot

After merging the datasets, each one of us did a plot showing the time series of the tweets. In my case, the tweets containing coronavirus reach a peak on the first day of our analysis, February 2nd, with more than 800 tweets in the middle of the day. We can see that during nighttime the observations reach every day a minimum, which is quite expected. We can also observe that the overall trend is decreasing from the 2nd to the 8th of February, then it grows again until the 11th, and has a fluctuant trending in the following days. It is also interesting to notice that on the 14th we see a peak in the evening hours. This should be due to the fact that two things were happening on the 14th: first, Italy got the news that Niccolò, the young Italian who was in quarantine in Wuhan, was finally able to leave China to return to Italy; and second, the news of the first case of coronavirus in Egypt arrived. These two events, and probably the first in particular, were quite emphasized on the news, which, we could hypothesize, influenced the entity of the conversation about them on Twitter.

### Text analysis

With the plotting of the frequencies, we had more information about when did people talk about the virus, we then wanted to understand more about what they were saying. Therefore, we tried to do some text analysis.

### Most used words

In order to do that, I had to do some preprocessing, such as bringing everything to lower case and removing punctuation.

Once I did that, I saved the cleaned dataset, and used it to visualize the 10 most used words in these tweets. Its unsurprising that coronavirus, our keyword, is the most used one, followed by Cina and cinesi, which have been linked to the virus from the beginning. Many Italian newspapers had “the Chinese virus” in their title, and so on. This has drawn a lot of attention, as linking the virus to a specific country was risky, because it could create a stigma, as it happened with AIDS in the past.

In the most used words we also find “Spallanzani” and “isolato” which refers to the Italian hospital where the virus was isolated to be studied. This was also a big topic on the news, since it leveraged the feeling of being proud of the “Italian excellences”. Aside from the words referring to China in different ways, we find that the 10th most used words is “morti”, which stresses the fact that the conversation is often about the fear that this virus brings along, and of the feeling of uncertainty that revolves around it.

### Wordcloud

After that, we chose to visualize a wordcloud of the 50 words that are most used when talking about the phenomenon. Some of them are expected, they refer to China (cina, cinese, cinesi, wuhan), some others refer to the spreading of the disease and the cure that scientist are looking for it (contagio, casi, test, epidemia, ricerca…). Other words need to be explained, for instance, we see that “Niccolò” is in these words. Niccolò is the name of the Italian student that couldn’t leave Wuhan for some time, and his story was quite followed in the country. We also find “nave”, “Africa”, “giappone”, that refer to the spreading of the virus outside China. It is interesting to notice that we have “razzismo” among the words, probably referring to the stigma we talked about before, as some days after the news about the virus broke out, several Chinese people in Italy revealed that they were almost forced to close their restaurants, and that people looked at them with suspicion, seeing them as the cause of the spreading of the virus. Another interesting element in the wordcloud is “salvini”, as he doesn’t belong to the sphere of what is discussed, but he is somehow very present in the discussion. An independent analysis could be conducted to understand what actually is discussed when Salvini is cited.

### Sentiment analysis and plot

Going further with the project, we tried to conduct a sentiment analysis, to understand something more about the discussion. We did it by parsing the data and creating a dictionary to be able to read the text. This passage was not easy, and it required us to share our doubts and problems with our classmates, who were doing something similar. The package quanteda was used in this phase.

Once we managed to create the dictionary and the DFM, we plotted the results to see what we could find about the sentiment of these tweets. The related plot tells us that the majority of the sentiment expressed in the tweets is neutral, which could be due to the fact that information about the disease is being conveyed. It is a bit surprising to find out that the negative sentiment is the one that is less expressed. Further investigation should be done to better understand why, and if this depends on what is actually in the tweets, or if the dictionaries used could be improved.

### Emotions analysis and plot

In the last part, we tried to conduct an emotion analysis, to see which emotions were most expressed in the tweets. To do that, we constructed five categories of emotions: outraged, worried, sad, entertained and pleased. We created the DFM for that and then we plotted the grouped emotions in the last plot.

From that we can see that the emotions expressed are quite varied. Being sad is the emotion that is portrayed less often, while being entertained is the one that results the most. Once again, this could be due both to the fact that people joke a lot about different situations, and therefore we could understand the “entertained” results, but it could also be due to the robustness of the analysis itself. Further research could clarify that doubt, but it would also be very time consuming, and requiring of some more advanced skills.

### Conclusions and further research

This project has been very interesting, I’m quite proud of the fact that in the beginning it seemed to be something impossible, but in the end, it has brought us to what I think is a good result. One of the things that was very difficult was to understand what was not working in the beginning, because we thought that the thin amount of data we were able to collect was due to some error in our code or some fault in our connection. Resolving that problem was time consuming, we spent several days to figure that out. We could think to share our experience with people, as we saw that this happens quite often, but the explanation is very scarce. We could also think to signal to the developers of rtweet that a warning message could be useful (but we need to understand how to do that).

Of course, our sentiment and emotion analysis is not the best that can be done, but it is our first try and I’m satisfied that we were able to figure something out. We surely can think of going more in deep with the sentiment analysis once we gather more knowledge about it.

One last comment: in the last couple of days, Italy saw a lot of things happening regarding the Covid-19 (which is now the official name of the virus), as the contagion started spreading in northern Italy, having as a result the suspension of several activities, among which there is also the suspension of university and school lessons all over Lombardy. This results in a huge amount of news being spread every moment of the day, and I expect it to reflect on the number of tweets about it. It would be very interesting to continue scraping the tweets in the following days and see what happens next. As now we have the code, we could think to actually do that and see where the discussion brings us.