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.