# Whatsapp sentiment analysis project

Gabriel Scozzarro

19/12/2020

## 1. Introduction

This project aim to perform a sentiment analysis on the content of a Whatsapp chat based on words and emoji used. The time frame of the analysis is Jan 2020 - Dic 2020

## 2. Toolbox

Importing the libraries that the project will use.

```
library(rwhatsapp)
library(lubridate)
library(tidyverse)
library(tidytext)
library(kableExtra)
library(knitr)
library(ggimage)
library(RColorBrewer)
```

## 3. Preparation and reading data

After importing the contet of the chat as dataframe, some operation were performed: 1. Delete first row with the Whatsapp privicy disclaimer 2. Change one chat user name to simplify usage in code 3. Create a new feature to assign to each messages the correct season

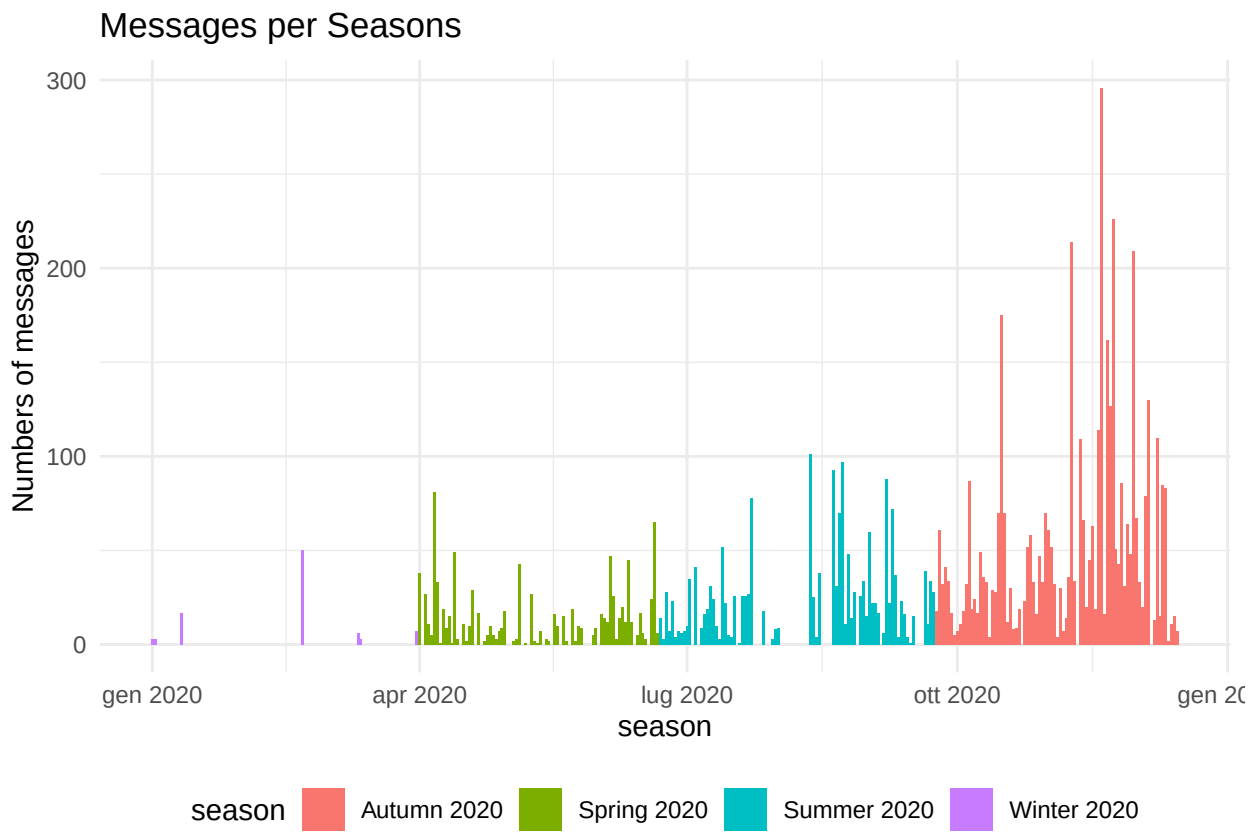
```
mychat<- rwa_read('chat_A_G.txt')
```

This is the preview of the content inside our dataframe after the processing:

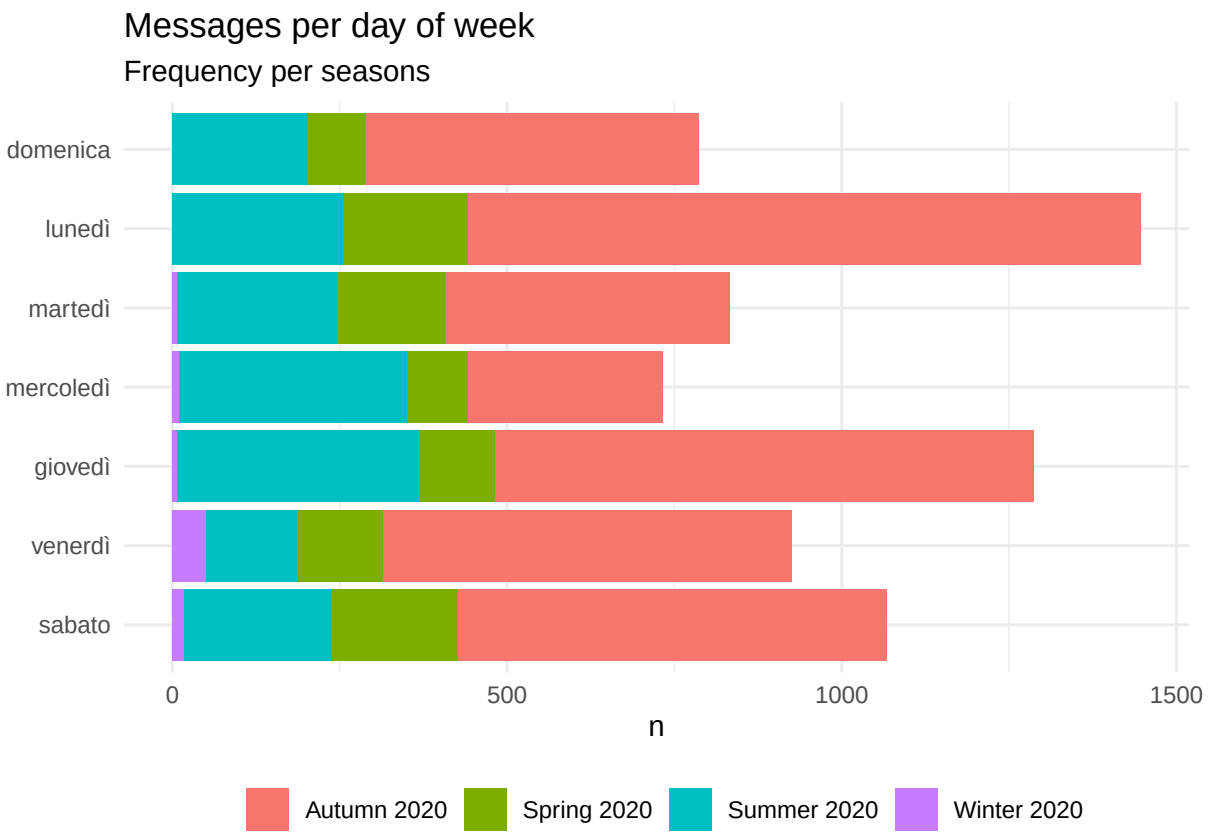
| time | author | text | source | id | emoji | emoji_name | day | season |
|---|---|---|---|---|---|---|---|---|
| 2020-01-01 00:05:55 | Gabriel | Auguroni bastasooooo<U+0001F618> | chat_A_G.txt | 15 | <U+0001F618> | face blowing a kiss | 2020-01-01 | Winter 2020 |
| 2020-01-01 00:06:00 | Gabriel | Famni il nipote | chat_A_G.txt | 16 | NULL | NULL | 2020-01-01 | Winter 2020 |
| 2020-01-01 02:09:55 | Andrea Marciano | <U+2665><U+FE0F><U+2665><U+FE0F><U+2665 | chat_A_G.txt | 17 | <U+2665><U+FE0F>, <U+2665><U+FE0F>, <U+2665><U+FE0F> | heart suit, heart suit, heart suit | 2020-01-01 | Winter 2020 |
| 2020-01-02 15:17:43 | Gabriel | <U+200E>audio omesso | chat_A_G.txt | 18 | NULL | NULL | 2020-01-02 | Winter 2020 |
| 2020-01-02 15:24:46 | Andrea Marciano | Puoi sentire perrone lui ne ha fatti mille di questi | chat_A_G.txt | 19 | NULL | NULL | 2020-01-02 | Winter 2020 |
| 2020-01-02 16:07:11 | Gabriel | Okioki | chat_A_G.txt | 20 | NULL | NULL | 2020-01-02 | Winter 2020 |
| 2020-01-11 21:14:46 | Andrea Marciano | Amo | chat_A_G.txt | 21 | NULL | NULL | 2020-01-11 | Winter 2020 |
| 2020-01-11 21:14:53 | Gabriel | Pisello | chat_A_G.txt | 22 | NULL | NULL | 2020-01-11 | Winter 2020 |
| 2020-01-11 21:15:13 | Gabriel | Ma nn stai al ristopizza con la sciarpa | chat_A_G.txt | 23 | NULL | NULL | 2020-01-11 | Winter 2020 |
| 2020-01-11 21:15:21 | Gabriel | Mentre mangi | chat_A_G.txt | 24 | NULL | NULL | 2020-01-11 | Winter 2020 |

## 4. EDA

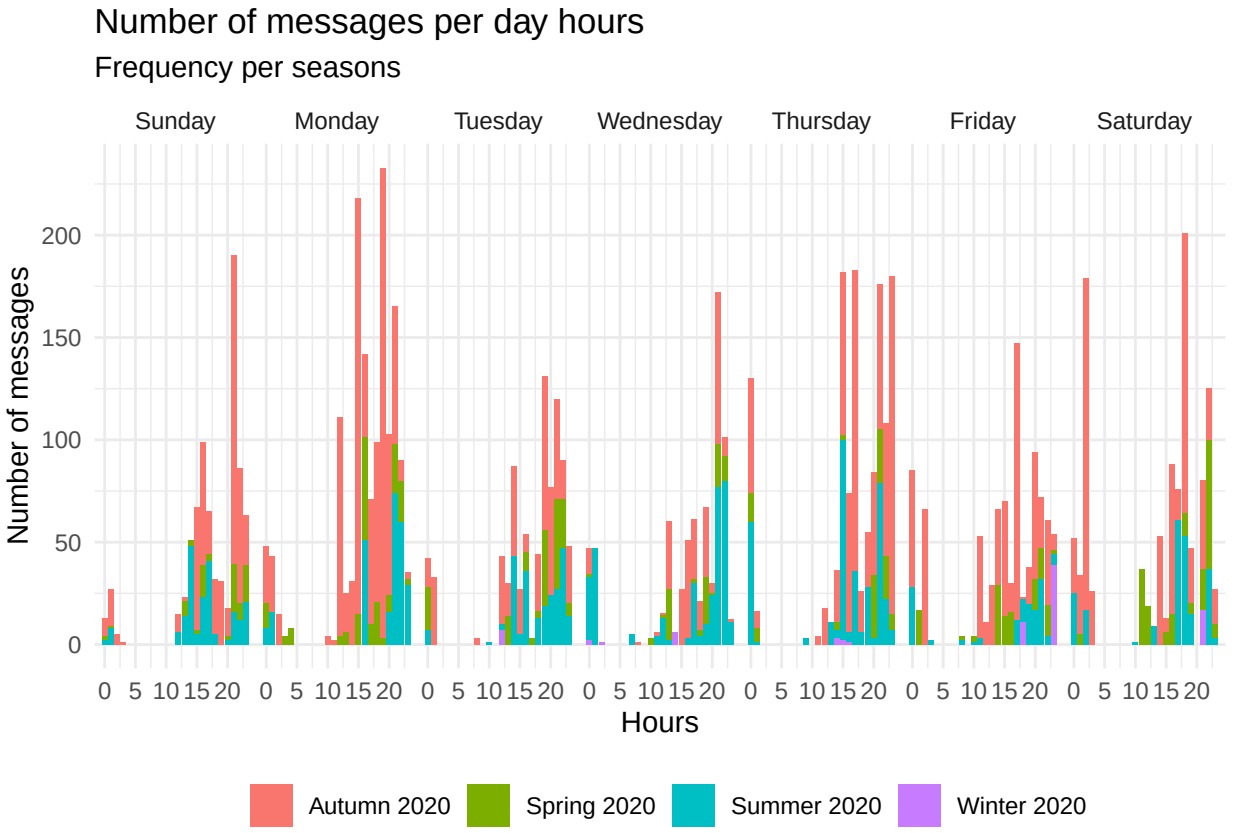
I start the exploration and the data analysis looking for the daily messages frequency divided by season



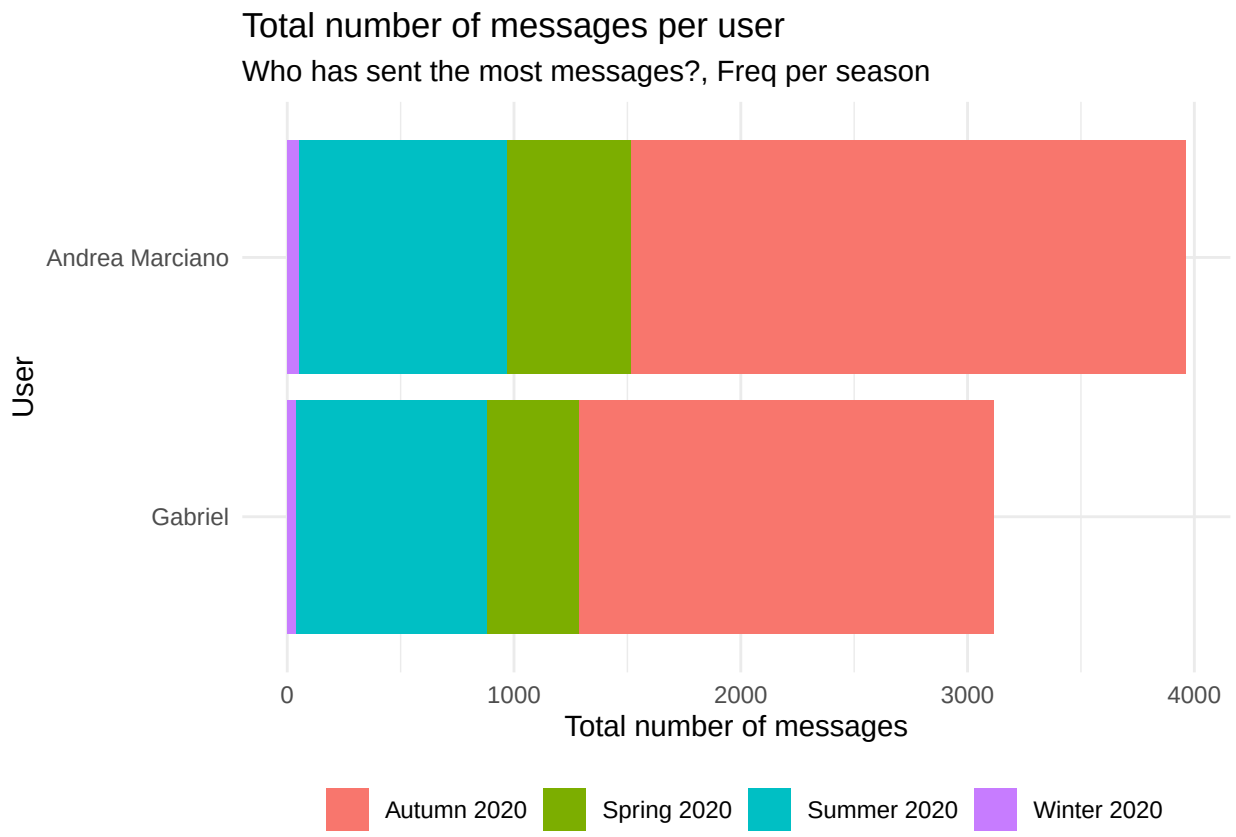

We can observed a rise in number of messages through the season with some spikes but with essentially an exponential trend, starting from April 2020 which was in the middle of the first lockdown. Then I explore the frequency of messages per day of week and also the frequency of messages per hour of the day.



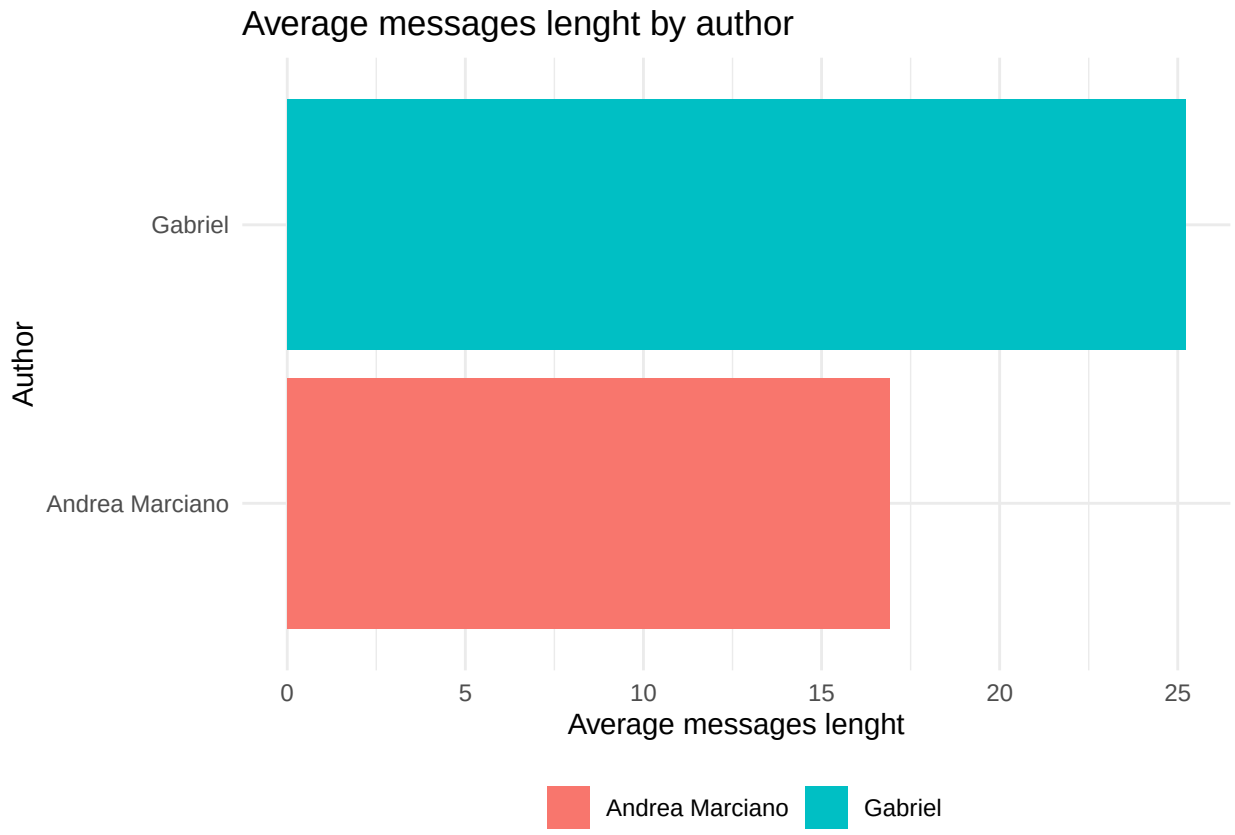

We can appreciate the consistency during the summer of messages, instead during the autumn 2020 we have clearly more messages per day but with 2 noticeable spikes in monday and thursday.



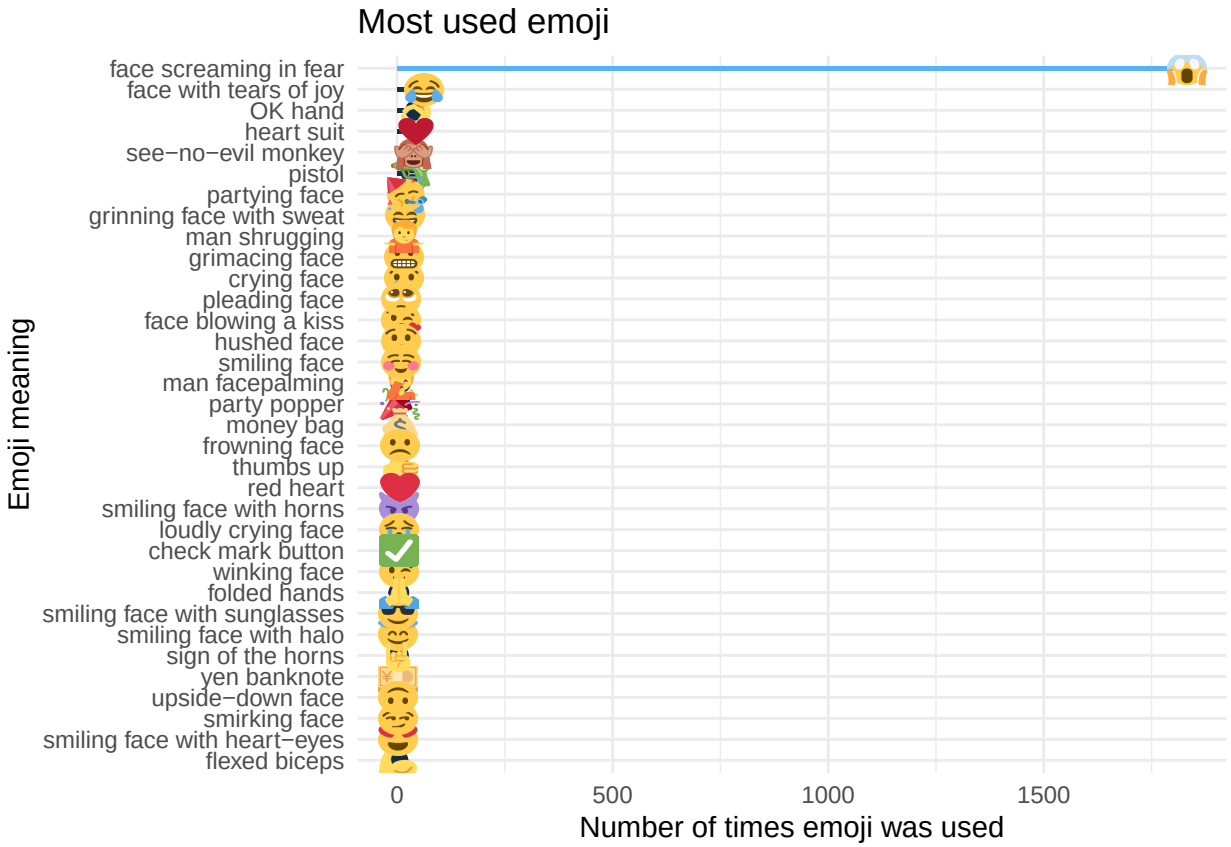

Now I can move on exploring who sent more messages during the all time frame in analysis and also divided by season



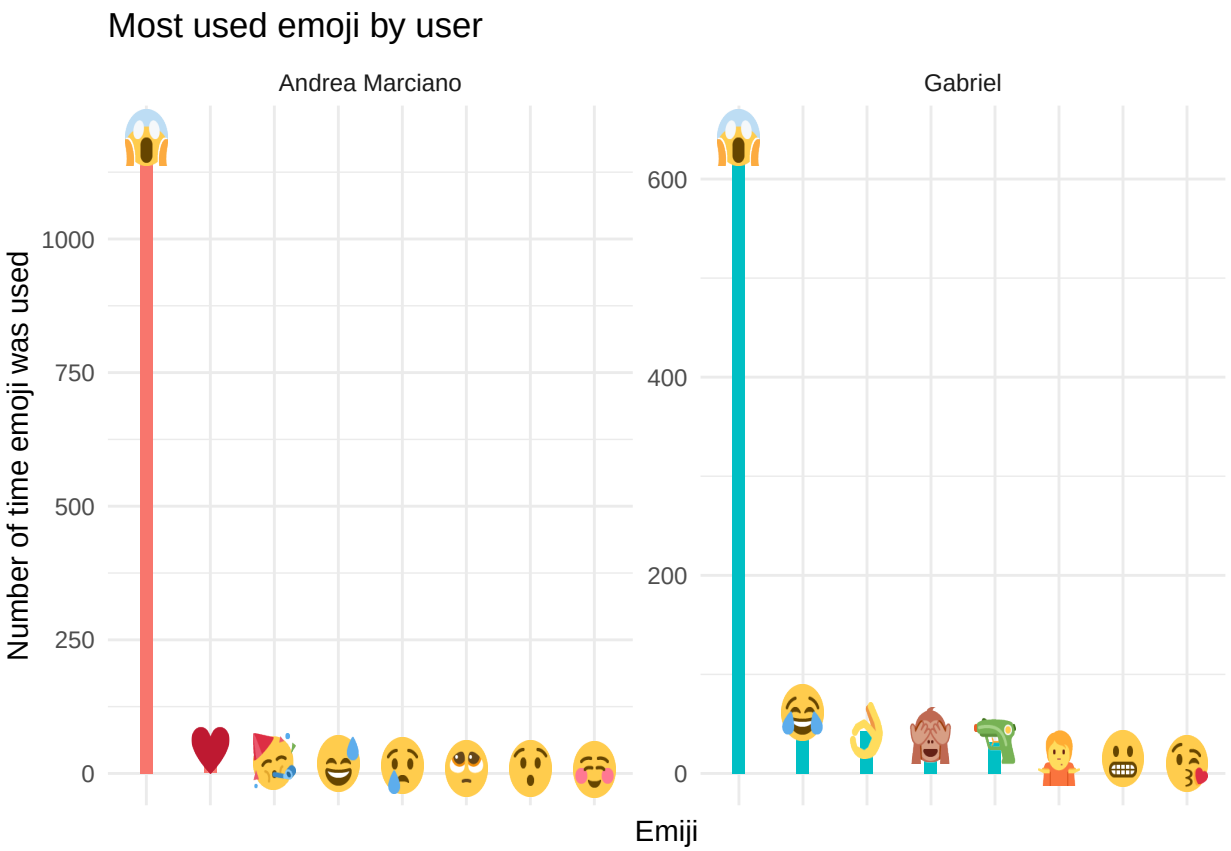

One of the user sent sensibly more messages but this can be due to the fragmentation of the content in several more small messages. We can investigate further to see how it is.



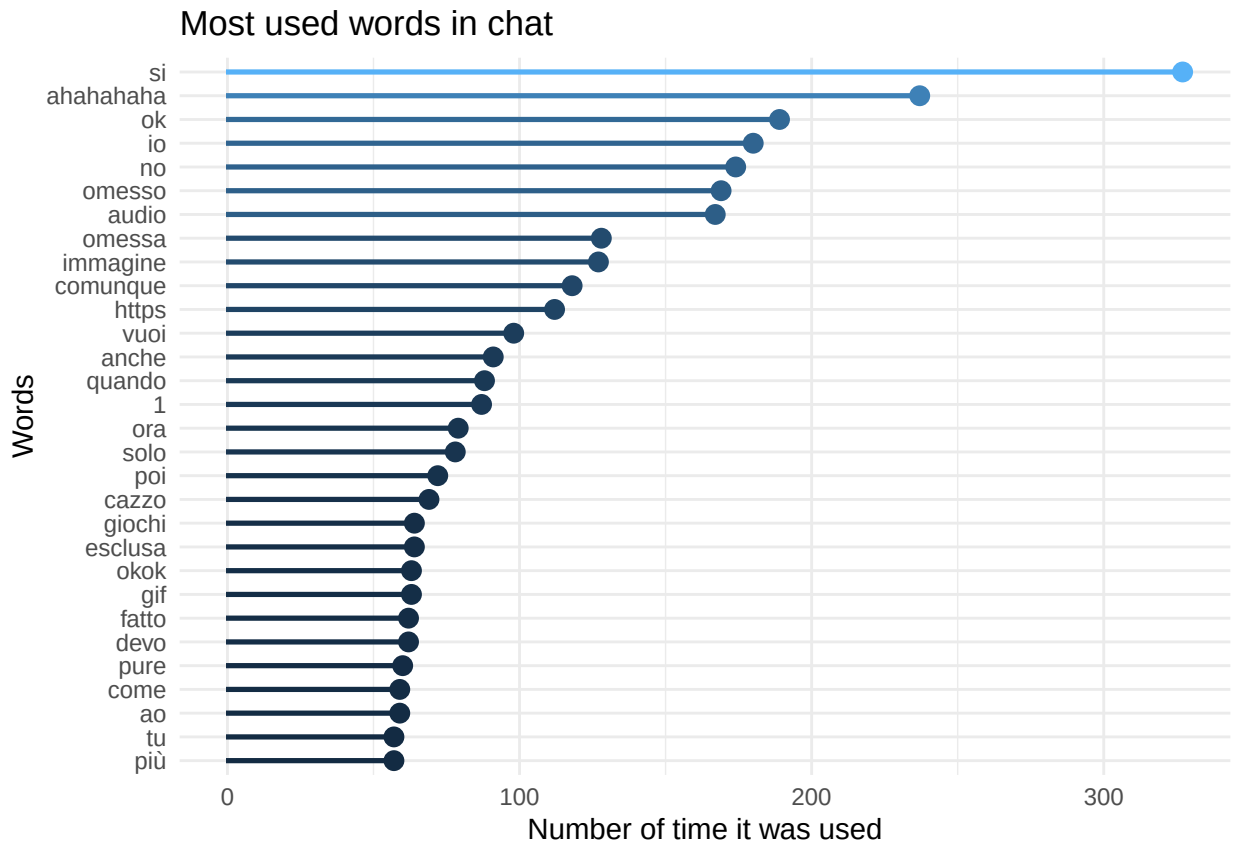

As we suspected the average length of messages of the other user is longer sacrificing the quantity of messages. This can improve also the clarity of the content. Moving forward to the content inside the chat we can start with the non text related content which are emojis.



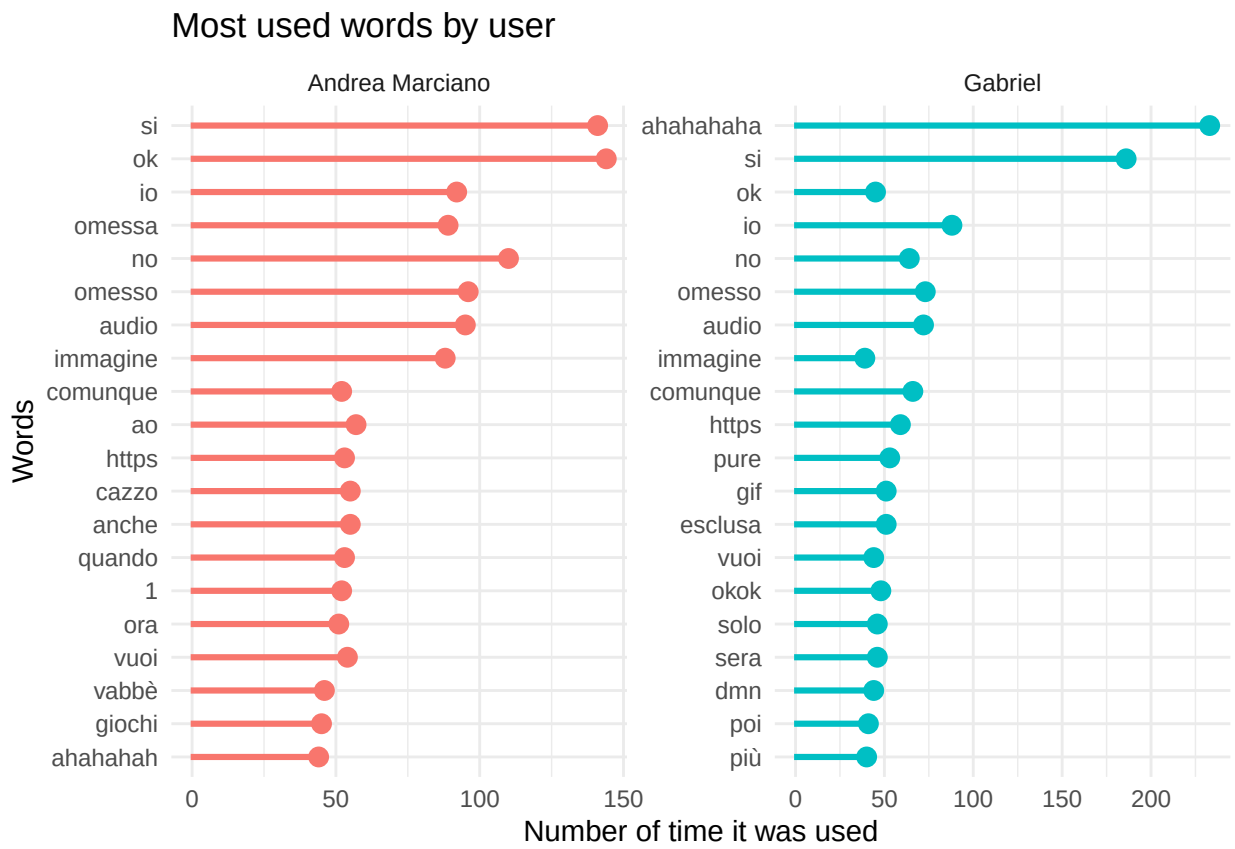

Those are the 30 most used emojis overall inside the chat. In the next plot we can also appreciate the most used ones by user.




Now is time for text related content. I define a list of words which will not be taken in consideration because are pronouns or articles and ecc.




Overall surprisingly the most used word taking in account both user is “yes”. I further divided the most used words by the user.

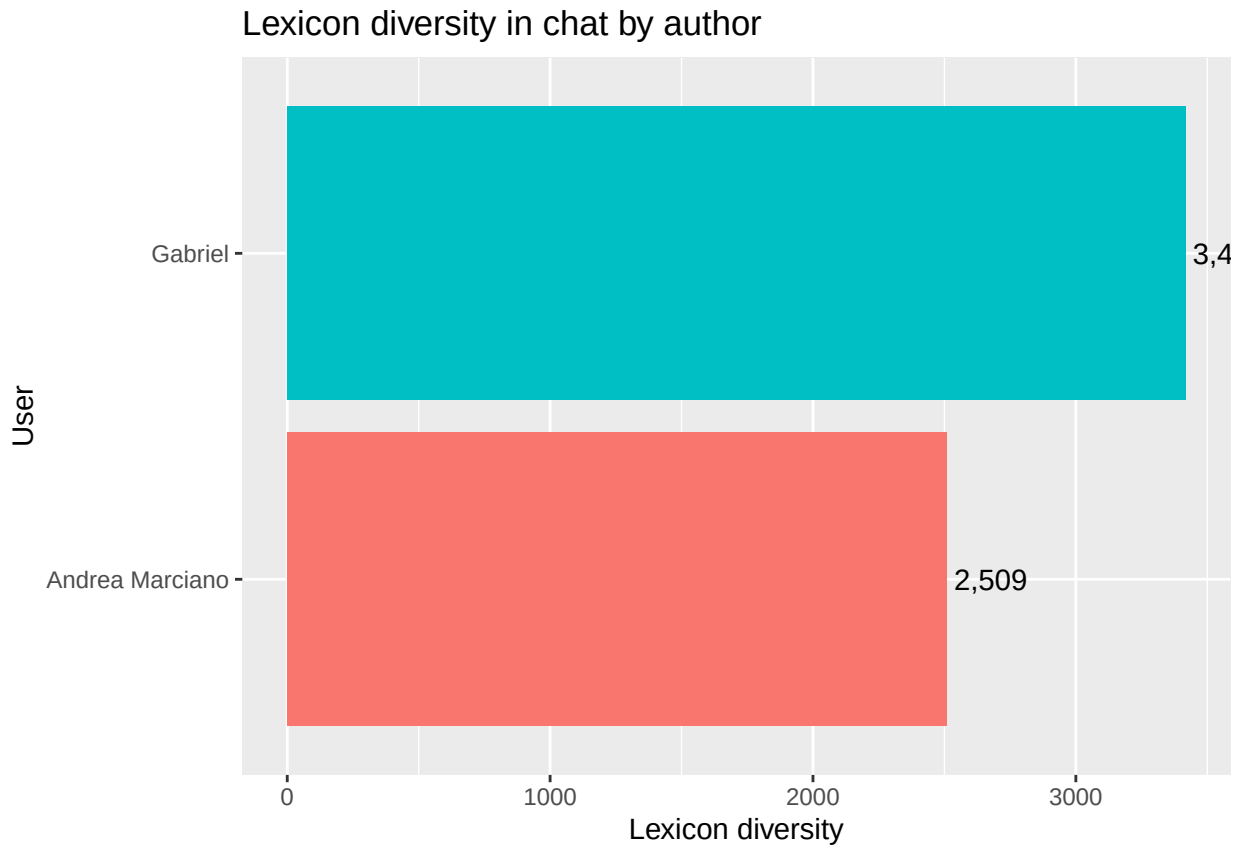



Surprisingly for one of the user the 2 most user words are synonyms, meaning "yes".

## 5. Lexicon analysis

Let's see the user lexicon diversity to understand if the counting of words is due to vocabulary poorness.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

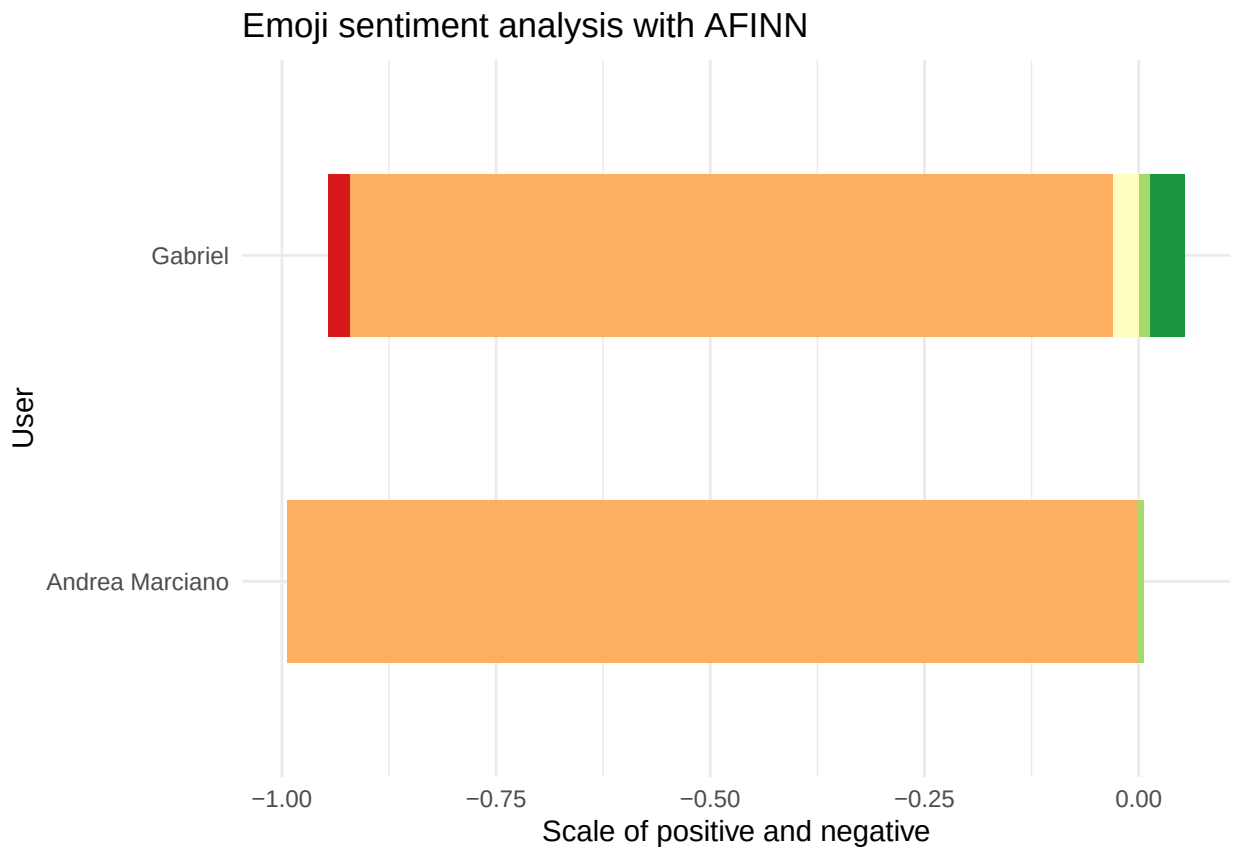



As we see one user has a more rich vocabulary to choose on and so a few chances to repeat the same word.

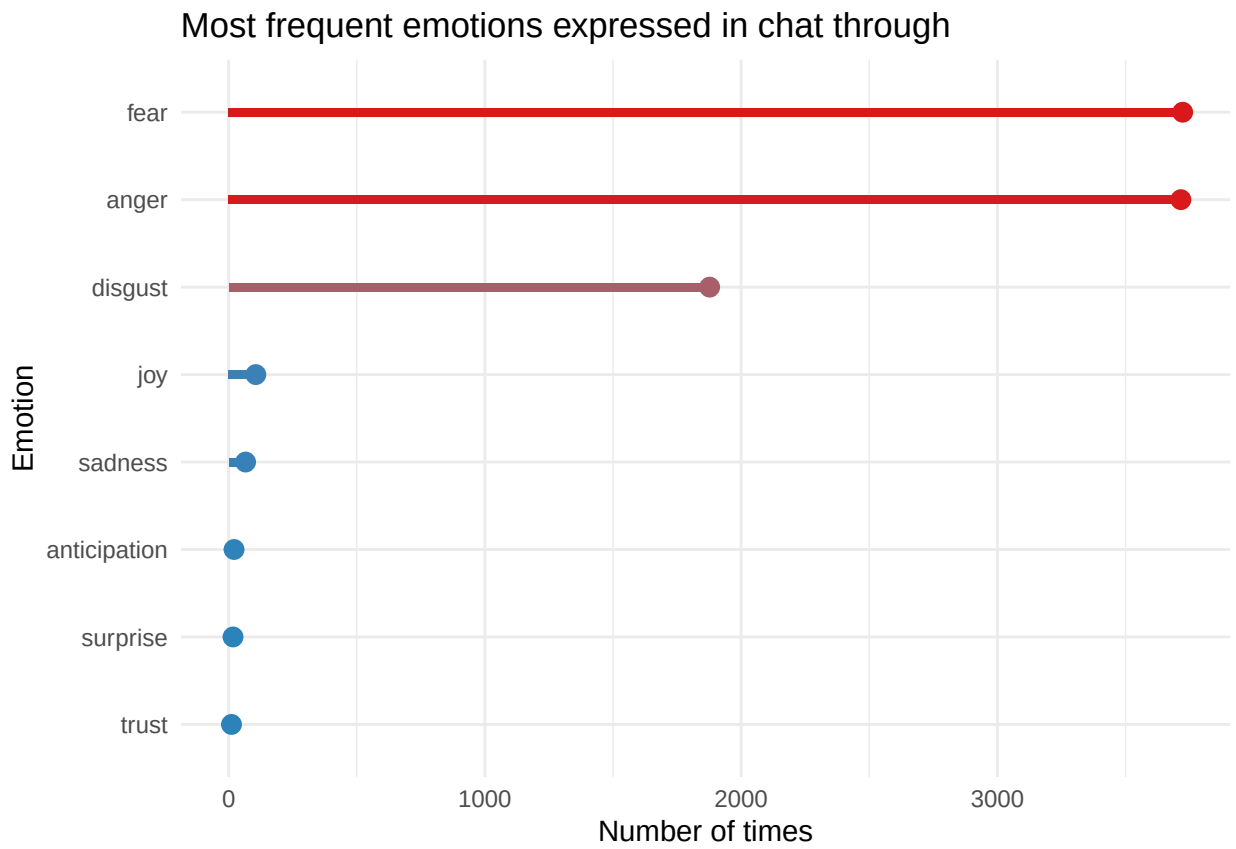
## 6. Sentiment analysis

Now is time to investigate the sentiment analysis. I started with the sentiment express in the non textual content which are emojis. For this analysis I used the AFINN method.

| emoji...1 | emoji_words...2 | value...3 | emoji...4 | emoji_words...5 | value...6 | emoji...7 | emoji_words...8 | value...9 | emoji...10 | emoji_words...11 | value...12 | emoji...13 | emoji_words...14 | value...15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <U+0001F618> | kiss | 2 | <U+0001F631> | fear | -2 | <U+0001F648> | evil | -3 | <U+0001F62D> | crying | -2 | <U+0001F607> | smiling | 2 |
| <U+0001F613> | downcast | -2 | <U+2639> | frowning | -1 | <U+0001F60D> | smiling | 2 | <U+0001F60A> | smiling | 2 | <U+0001F625> | sad | -2 |
| <U+0001F602> | tears | -2 | <U+0001F622> | crying | -2 | <U+0001F608> | smiling | 2 | <U+0001F60A> | smiling | 2 | <U+0001F625> | relieved | 2 |
| <U+0001F602> | joy | 3 | <U+263A> | smiling | 2 | <U+0001F601> | smiling | 2 | <U+0001F929> | struck | -1 | <U+0001F51D> | top | 2 |
| <U+0001F631> | screaming | -2 | <U+0001F648> | no | -1 | <U+0001F60E> | smiling | 2 | <U+0001F629> | weary | -2 | <U+0001F645> | no | -1 |




You can see previously a preview table with the meaning of the emoji and the corresponding polarity value, then a plot with the overall sentiment analysis in the chat. In the next plot you will see the most frequent emotion express in the chat.




You can see as the emoji used by one of the user are practically monotone expressing a mild negative sentiment. The other user used mostly mild negative emojis too but with more shades.