Proyecto semestral: Hito 1

Nicolás Herrera, Yesenia Marulanda, Franco Migliorelli, Samuel Sánchez, Sebastián Urbina

Octubre 2020

Contents

Motivacion:

Twitter es una de las redes sociales más utilizadas para comentar, compartir o debatir temas de actualidad y tendencias. El primer trimestre de 2020 tuvo un aumento de un 23% en comentarios diarios con respecto al mismo periodo de tiempo del año 2019 [1], siendo una de sus principales causas la pandemia del COVID-19.

Es por este contexto que analizar datos de twitter es interesante, además se puede obtener información en tiempo real de sucesos que están ocurriendo, se pueden ocupar métodos de la minería de datos sobre esta información y es fácil manipular grandes volúmenes de información. Por estas razones, nace la motivación por explorar tweets relacionados con coronavirus en un intervalo de tiempo acotado, desde la perspectiva de los sentimientos y relacionandolos con el contexto pais desde donde se emiten; buscando establecer si este ùltimo influencia la percepcion de las personas acerca de la pandemia.

Dicho lo anterior, los puntos claves a analizar en el desarrollo del proyecto son:

- Identificar como se relaciona el sentimiento identificado con el contexto pais (segun mayor o menor presencia del sentimiento).
- Categorizar paises segun mayor o menor presencia de sentimiento y relacionarlos con algun indice de felicidad publicado en el ultimo año.
- Identificar palabras que son clave a la hora de categorizar el sentimiento.
- Establecer algoritmos para predecir sentimientos de forma sistematizada.
- Entrenar modelos de clasificaciòn en base a tweets usando un Dataset de entrenamiento y un dataset de evaluacion.

Descripción de base de datos

La base de datos, extraída desde la plataforma *Kaggle*[2], esta compuesta en principio por 44955 tweets relacionados con el tema COVID 19 y que fueron publicados del 2 de marzo al 14 de abril de 2020. Ademas de encontrarse el texto publicado se encuentran en la base de datos los atributos de fecha exacta de publicacion,

ubicacion desde la cual se realizla publicacion, un identificador para el usuario y la asignacion de sentimiento para cada tweet. La asignacion de sentimiento a cada Tweet fue realizada de forma manual por el propietario de la base de datos. Esta variable de sentimiento sería un punto de comparación para un posible modelo de clasifiación de los sentimientos de los tweets.

Exploración de datos

El objetivo de esta sección es describir la base de datos seleccionadas, mostrando estadísticas de resumen, gráficos relevantes para la descripción y un breve análisis sobre estos.

El primer paso consiste en cargar todas las librerías que se utilizarán en este trabajo, las cuales permiten realizar gráficos y trabajar con los datos de una forma más simple y eficiente.

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidytext)
library(stopwords)
library(wordcloud)
library(wordcloud2)
```

A continuación se procece a cargar la base de datos para poder realizar el análisis respectivo.

```
train  <- read.csv("data/Corona_NLP_train.csv", encoding="Latin-1")
test <- read.csv("data/Corona_NLP_test.csv", encoding="Latin-1")
df <- rbind(train,test)
```

Tal y como se mencionó anteriormente, esta base de datos contiene 6 columnas y 44955 filas de datos. Para poder tener una referencia de cómo son estos campos, a continuación se puede observar una vista previa de las primeras filas del set de datos.

```
head(df)
```

```
##   UserName ScreenName                   Location    TweetAt
## 1     3799      48751                     London 16-03-2020
## 2     3800      48752                         UK 16-03-2020
## 3     3801      48753                  Vagabonds 16-03-2020
## 4     3802      48754                            16-03-2020
## 5     3803      48755                            16-03-2020
## 6     3804      48756 ÃœT: 36.319708,-82.363649 16-03-2020
##
## 1
## 2                                                                        advice Talk to your neighbours
## 3
## 4      My food stock is not the only one which is empty...\n\nPLEASE, don't panic, THERE WILL BE ENOU
## 5 Me, ready to go at supermarket during the #COVID19 outbreak.\n\nNot because I'm paranoid, but becau
## 6                                                  As news of the regionÂ's first confirmed CO
##            Sentiment
## 1            Neutral
## 2           Positive
## 3           Positive
## 4           Positive
## 5 Extremely Negative
## 6           Positive
```

Para poder trabajar con los datos es necesario covnertir algunos formatos de las variables, en particular en este caso, se procede a transformar las variables "TweetAt" a un formato de fecha, dado que corresponde al

momento en donde se realizó el tweet, y la variable "OriginalTweet" que corresponde al contenido del tweet realizado.

```
df$TweetAt <- as.Date(df$TweetAt, format="%d-%m-%y")
df$OriginalTweet <- as.character(df$OriginalTweet)
```

Los formatos de cada variable del set de datos son mostrados a continuación:

```
str(df)
```

```
## 'data.frame':    44955 obs. of  6 variables:
##  $ UserName     : int  3799 3800 3801 3802 3803 3804 3805 3806 3807 3808 ...
##  $ ScreenName   : int  48751 48752 48753 48754 48755 48756 48757 48758 48759 48760 ...
##  $ Location     : Factor w/ 13128 levels "","- ? ? universe ? + ? ",..: 6149 11045 11317 1 1 1017 51
##  $ TweetAt      : Date, format: "2020-03-16" "2020-03-16" ...
##  $ OriginalTweet: chr  "@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6
##  $ Sentiment    : Factor w/ 5 levels "Extremely Negative",..: 4 5 5 5 1 5 5 4 5 3 ...
```

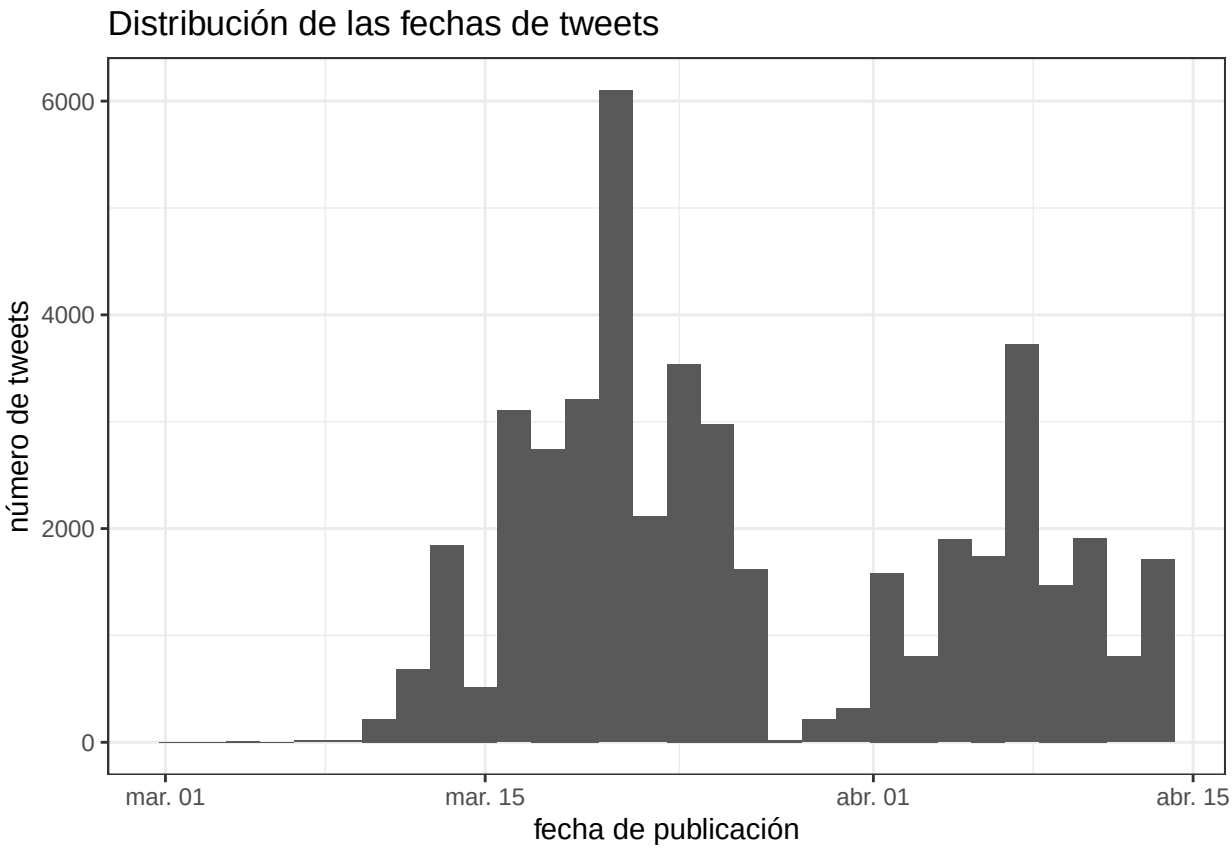
Además, es importante mencionar con qué periodos se estará trabajando, por lo que el resultado de la siguiente línea de código arroja la fecha mínima y máxima del set de datos, siendo el 2 de marzo de 2020 y el 14 de abril de 2020 respectivamente.

```
summary(df$TweetAt)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2020-03-02" "2020-03-19" "2020-03-23" "2020-03-26" "2020-04-06" "2020-04-14"
```

Para ver cómo están distribuidos los tweets en el tiempo, a continuación se realiza y muestra un histograma con las distribuciones de las fechas de los tweets realizados y registrados en esta base de datos.

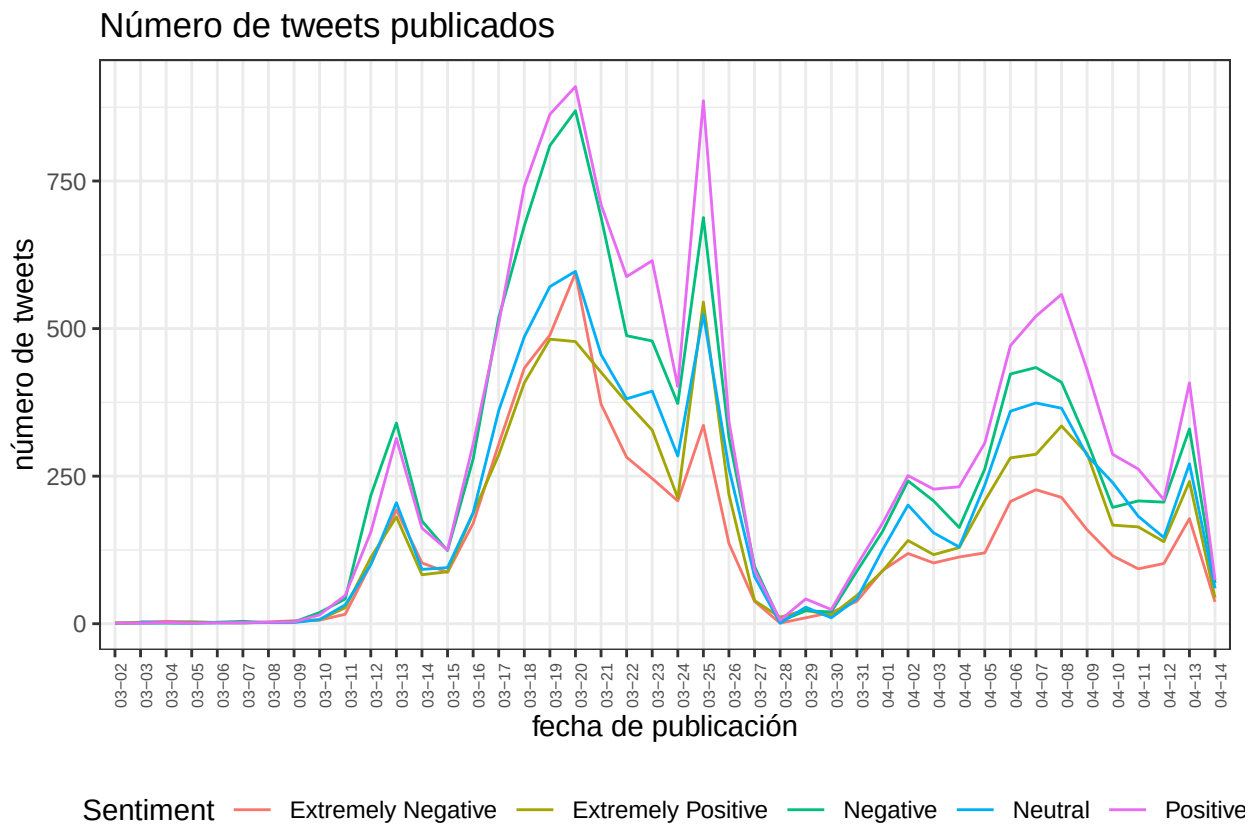
```
ggplot(data=df, aes(x=TweetAt)) + geom_histogram(position="identity", bins=30) +
  labs(title = "Distribución de las fechas de tweets", x = "fecha de publicación",
       y = "número de tweets") + theme_bw()
```



Un comportamiento particular de los datos es que a finales de marzo la cantidad de tweets registrados disminiye considerablemente llegando a ser nulo, mientras que la mayor cantidad de tweets se encuentra concentrada durante la tercera semana de marzo y la segunda semana de abril.

Un campo importante que posee esta base de datos es la columna "Sentiment", la cual representa el sentimiento asociado al tweet registrado. A continuación se presenta la cantidad de tweets segun tipo de sentimiento ("Extremely Negative", "Extremely Postive", "Negative", "Neutral", "Positive") durante el periodo registrado en el set de datos.

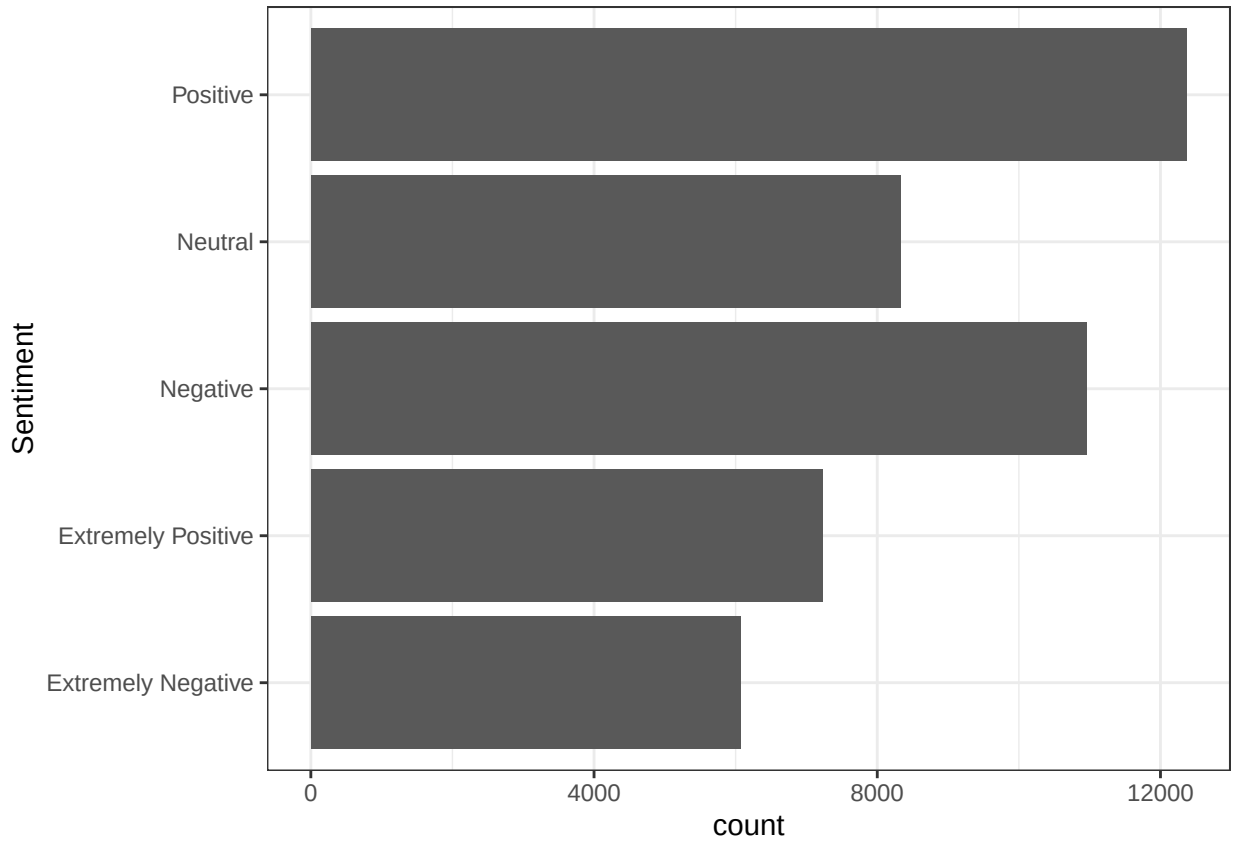
```
tweets_mes_dia <- df %>% mutate(mes_dia = format(TweetAt, "%m-%d"))
tweets_mes_dia %>% group_by(Sentiment, mes_dia) %>% summarise(n = n()) %>%
  ggplot(aes(x = mes_dia, y = n, color = Sentiment)) +
  geom_line(aes(group = Sentiment)) +
  labs(title = "Número de tweets publicados", x = "fecha de publicación",
       y = "número de tweets") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, size = 6),
        legend.position = "bottom")
```



Se puede observar en este último gráfico que en general aquellos sentimientos que prevalecen son los de "Positive" y "Negative". Por otro lado, todos los tipos de sentimientos registrados tienen un comportamiento similar a lo largo del tiempo. Mientras que el sentimiento menos registrados en el tiempo corresponde al de "Extremely Negative", lo cual es interesante dado el contexto de la base de datos, en donde se esperaría que hubiese una mayor cantidad de tweets asociado a un sentimiento negativo o extramademente negativo.

En el siguiente gráfico se puede observar la cantidad de tweets por cada tipo de sentimiento durante todo el periodo de observación.

```
# library(tidyverse)
#tweets_sentiment <- df %>% group_by(Sentiment) %>% summarise(n = n())
df %>% ggplot(aes(x = Sentiment)) + geom_bar(stat="count") + coord_flip() + theme_bw()
```



Se observa que tal y como se mencionó anteriormente, los sentimientos que destacan son "Positive" y "Negative", alcanzando un total aproximado de 13000 y 11000 tweets respectivamente. El sentimiento con una menor cantidad de registros es "Extremely Negative", con un aproximado de 6000 tweets.

Si se comparación la proporción de estos tweets en comparación al total del periodo de observación, se tiene que el porcentaje asociado a cada sentimiento son los siguientes:

```
df %>% group_by(Sentiment) %>% summarise(Proporcion = n()/nrow(df)) %>% arrange(-Proporcion)
```

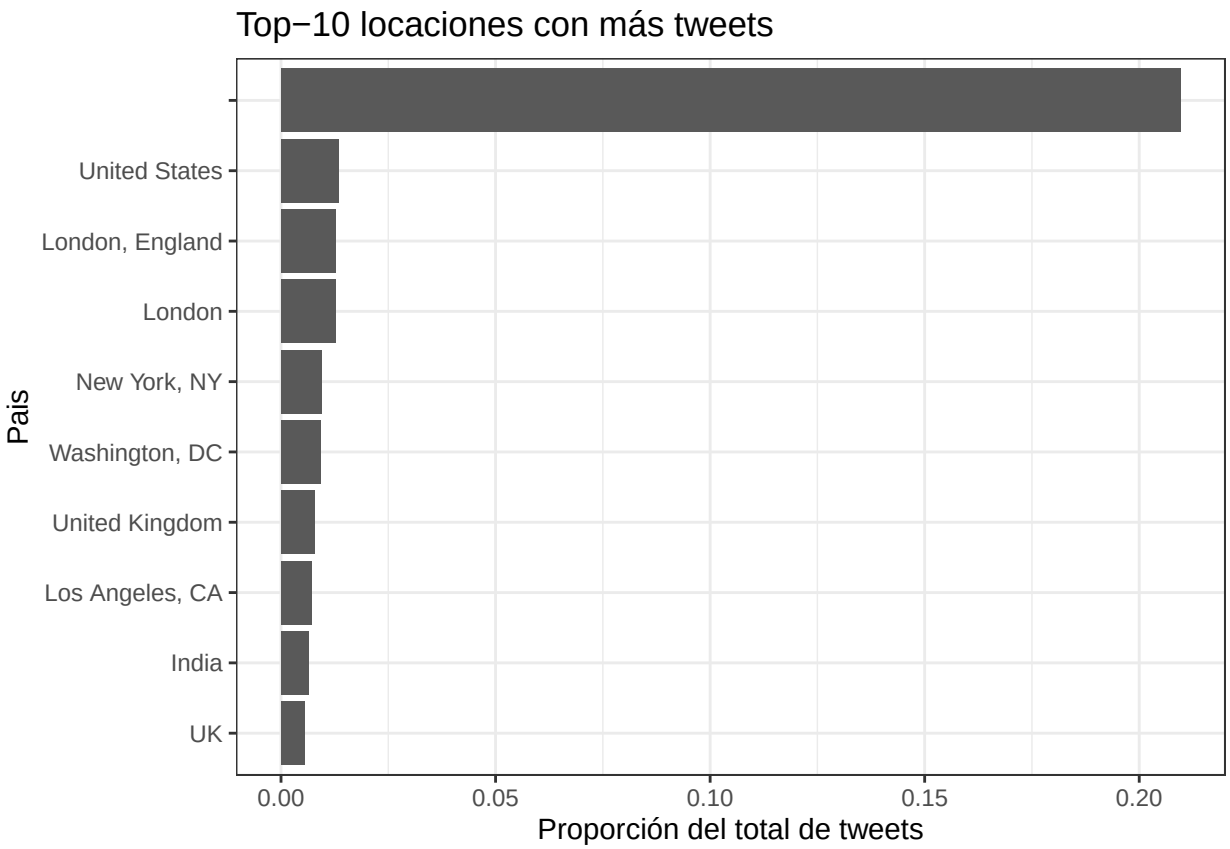
```
## # A tibble: 5 x 2
##   Sentiment          Proporcion
##   <fct>                   <dbl>
## 1 Positive                0.275
## 2 Negative                0.244
## 3 Neutral                 0.185
## 4 Extremely Positive      0.161
## 5 Extremely Negative      0.135
```

Es decir, el sentimiento "Positive" representa al 27,5% de los tweets del set de datos, mientras que el 13,5% de los tweets están categorizados bajo el sentimiento "Extremely Negative".Es importante notar que los sentimientos extremos tanto positivo como negativo se presentan en menor proporcion.

Otro punto importante a caracterizar es la locación en donde se emiten los tweets. En la siguiente tabla, se puede observar la cantidad total de tweets registrados para el top 10 de localidades.

```
top_10 <- df %>% group_by(Location) %>% summarise(N=n()/nrow(df)) %>% arrange(-N)
top_10 <- top_10[1:10,]
top_10$N <- round(top_10$N,4)
ggplot(top_10,aes(x=reorder(Location,N), y =N )) +
```

```
  geom_bar(stat="identity") +
  coord_flip() +
  labs(x="Pais", y = "Proporción del total de tweets", title="Top-10 locaciones con más tweets") +
  theme_bw()
```



En este último gráfico se puede observar que un 20% de los tweets no registran alguna locación, mientras que del porcentaje restante, los lugares más comunes son de países como Estados Unidos, Reino Unido e India.

Para poder analizar los contenidos de los tweets es necesario realizar una limpieza y normalización de estos. A continuación se crea una función que permite corregir algunos patrones de los textos tales como números, puntuación y espacios en blanco.

```
limpiar_texto <- function(texto){
    # Se convierte todo el texto a minúsculas
    nuevo_texto <- tolower(texto)
    # Eliminación de páginas web (palabras que empiezan por "http." seguidas
    # de cualquier cosa que no sea un espacio)
    nuevo_texto <- str_replace_all(nuevo_texto,"http\\S*", "")
    # Eliminación de signos de puntuación
    nuevo_texto <- str_replace_all(nuevo_texto,"[[:punct:]]", " ")
    # Eliminación de números
    nuevo_texto <- str_replace_all(nuevo_texto,"[[:digit:]]", " ")
    # Eliminación de espacios en blanco múltiples
    nuevo_texto <- str_replace_all(nuevo_texto,"[\\s]+", " ")
    # Tokenización por palabras individuales
    nuevo_texto <- str_split(nuevo_texto, " ")[[1]]
    # Eliminación de tokens con una longitud < 2
```

```
    nuevo_texto <- keep(.x = nuevo_texto, .p = function(x){str_length(x) > 1})
    return(nuevo_texto)
}
```

Para entender como procede esta última función, en la siguiente línea se muestra un ejemplo junto con los resultados de la aplicación de esta, en donde se puede observar que se extrajo cada palabra del objeto *text*.

```
text = "Hola mi nombre es https://www.google.cl. Como. no sé xd6666 ASDA"
limpiar_texto(text)
```

```
## [1] "hola"   "mi"     "nombre" "es"     "como"   "no"     "sé"     "xd"
## [9] "asda"
```

En el siguiente paso, se aplica la función *limpiar_texto* al contenido de los tweets de la base de datos, en donde cada resultado de cada tweet es almacenado en un vector de palabras, por lo que cada tweet tendría asociado uno de estos vectores.

```
tweets <- df %>% mutate(texto_vector = map(.x = OriginalTweet, .f = limpiar_texto))
```

```
tweets %>% select(texto_vector) %>% head()
```

```
##
## 1
## 2              advice, talk, to, your, neighbours, family, to, exchange, phone, numbers, create, con
## 3
## 4 my, food, stock, is, not, the, only, one, which, is, empty, please, don, panic, there, will, be, e
## 5  me, ready, to, go, at, supermarket, during, the, covid, outbreak, not, because, paranoid, but, be
## 6                                      as, news, of, the, regionâ, first, confirmed, covid, case, ca
```

```
#Cada valor de la columna texto_vector es un vector con cada palabra del texto
tweets$texto_vector[1]
```

```
## [[1]]
## [1] "menyrbie" "phil"     "gahan"    "chrisitv" "and"      "and"
```

En donde cada valor de la columna texto_vector es un vector con cada palabra del texto

```
#unnest() nos permite realizar una expansión de los vectores de palabras que creamos, esto aumenta la d
tweets_expand <- tweets %>% select(-OriginalTweet) %>% unnest()
```

```
## Warning: `cols` is now required.
## Please use `cols = c(texto_vector)`
```

```
tweets_expand <- tweets_expand %>% rename(word = texto_vector)
head(tweets_expand)
```

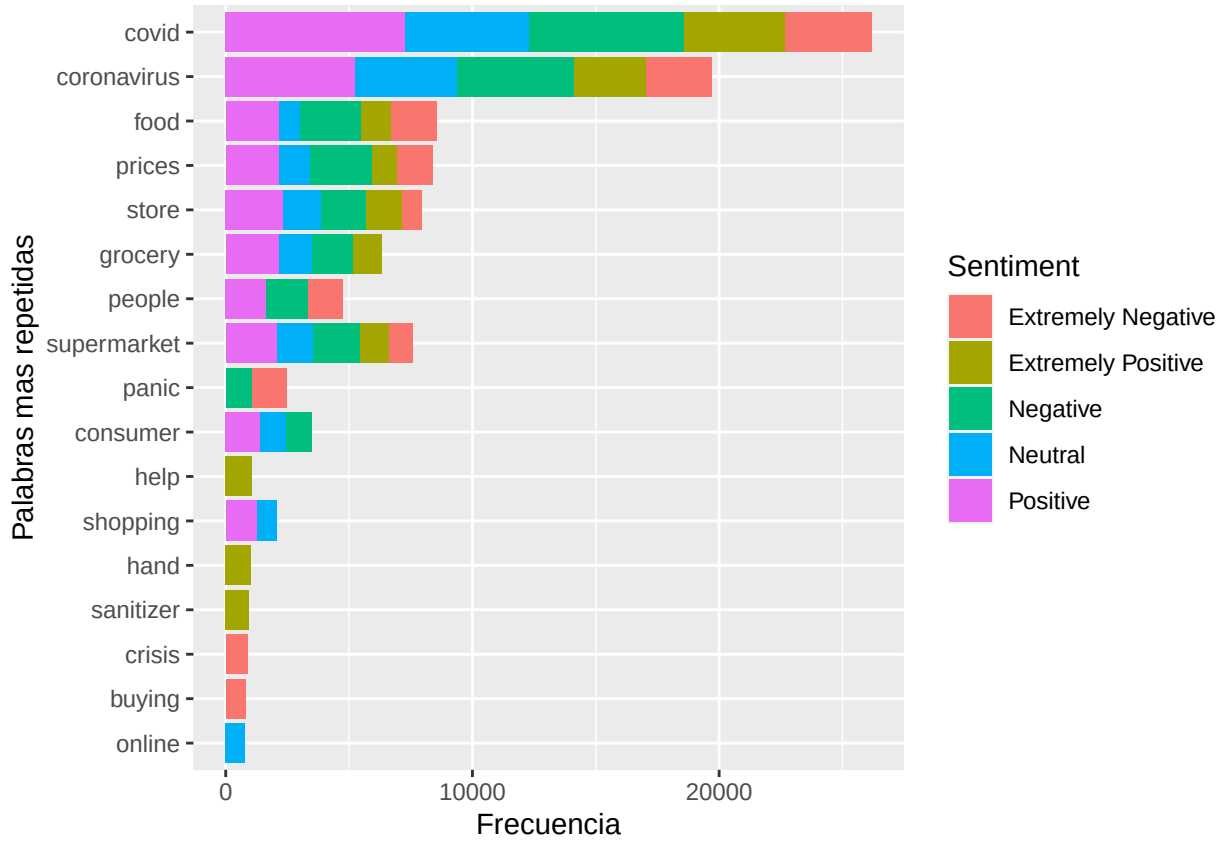
```
## # A tibble: 6 x 6
##   UserName ScreenName Location TweetAt    Sentiment word
##      <int>      <int> <fct>    <date>     <fct>     <chr>
## 1     3799      48751 London   2020-03-16 Neutral   menyrbie
## 2     3799      48751 London   2020-03-16 Neutral   phil
## 3     3799      48751 London   2020-03-16 Neutral   gahan
## 4     3799      48751 London   2020-03-16 Neutral   chrisitv
## 5     3799      48751 London   2020-03-16 Neutral   and
## 6     3799      48751 London   2020-03-16 Neutral   and
```

Se utilizan *stopwords* para filtrar algunas palabras propias del ingles (lenguaje dominio de los comentarios) como artículos, pronombres, preposiciones, adverbios e incluso algunos verbos. Estas palabras no tienen un significado por si solas, sino que modifican o acompañan a otras

```r
# "word" %in% vector -> true or false
lista_stopwords <- stopwords("english")
lista_stopwords <- c(lista_stopwords, "amp","can")
```

Luego se representa en un grafico de barras, las 10 palabras mas repetidas en los comentarios segun el sentimiento asignado al Tweet en el que se encuentran.

```r
tweets_expand <- tweets_expand %>% filter(!(word %in% lista_stopwords))

tweets_expand %>% group_by(Sentiment, word) %>%
  count(word) %>%
  group_by(Sentiment) %>%
  top_n(10,n) %>%
  arrange(Sentiment, desc(n)) %>%
  ggplot(aes(x=reorder(word,n),y=n,fill=Sentiment)) +
  geom_col() +
  labs(y = "Frecuencia", x = "Palabras mas repetidas") +
  coord_flip()
```



Se puede observar que, como era de esperarse, las palabras mas comentadas en todas las categorias de sentimiento son las referentes directamente a la pandemia: “covid” y “coronavirus”. Se destaca de igual forma la alta popularidad de las palabras “food” y “prices”, posiblemente debido al parcial desabastecimiento de productos y la subida de precios producto de la cuarentena.

Un analisis similar se presenta a continuacion, pero esta vez en word clouds:

```r
#Listas de colores utilizadas en las nubes de palabras
pal_neg <- c("#FC9272", "#FB6A4A", "#EF3B2C", "#CB181D", "#A50F15")
pal_neu <- c("#A1D99B", "#74C476", "#41AB5D", "#238B45", "#006D2C")
pal_pos <- c("#9ECAE1", "#6BAED6", "#4292C6", "#2171B5", "#08519C")

top <- 400
```

```r
#Palabras mas repetidas en comentarios negativos y extremadamente negativos
neg_tweets <- tweets_expand %>%
  filter(Sentiment == "Negative" | Sentiment == "Extremely Negative") %>%
  count(word)
neg_tweets_top <- neg_tweets%>%
  arrange(desc(n)) %>%
  top_n(top,n)


set.seed(1234)
wordcloud(words = neg_tweets_top$word, scale=c(5,0.7), freq = neg_tweets_top$n, min.freq = 1,max.words=
```


```r
#Palabras mas repetidas en comentarios neutrales
neu_tweets <- tweets_expand %>%
  filter(Sentiment == "Neutral") %>%
  count(word)
neu_tweets_top <- neu_tweets%>%
  arrange(desc(n)) %>%
  top_n(top,n)
```

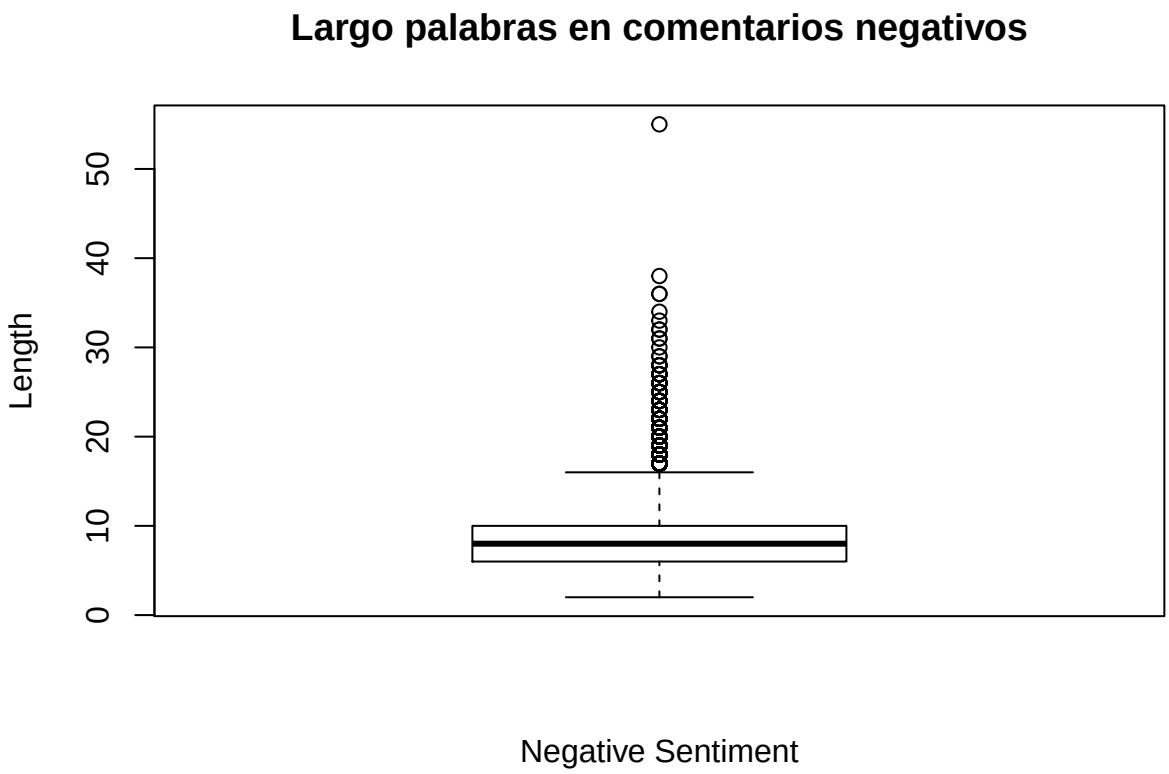
```
set.seed(1234)
wordcloud(words = neu_tweets_top$word, scale=c(5,0.7), freq = neu_tweets_top$n, min.freq = 1,max.words=
```


```
#Palabra mas repetidas en comentarios positivos y extremadamente positivos
pos_tweets <- tweets_expand %>%
  filter(Sentiment == "Positive" | Sentiment == "Extremely Positive") %>%
  count(word)
pos_tweets_top <- pos_tweets%>%
  arrange(desc(n)) %>%
  top_n(top,n)


set.seed(1234)
wordcloud(words = pos_tweets_top$word, scale=c(5,0.7), freq = pos_tweets_top$n, min.freq = 1,max.words=
```

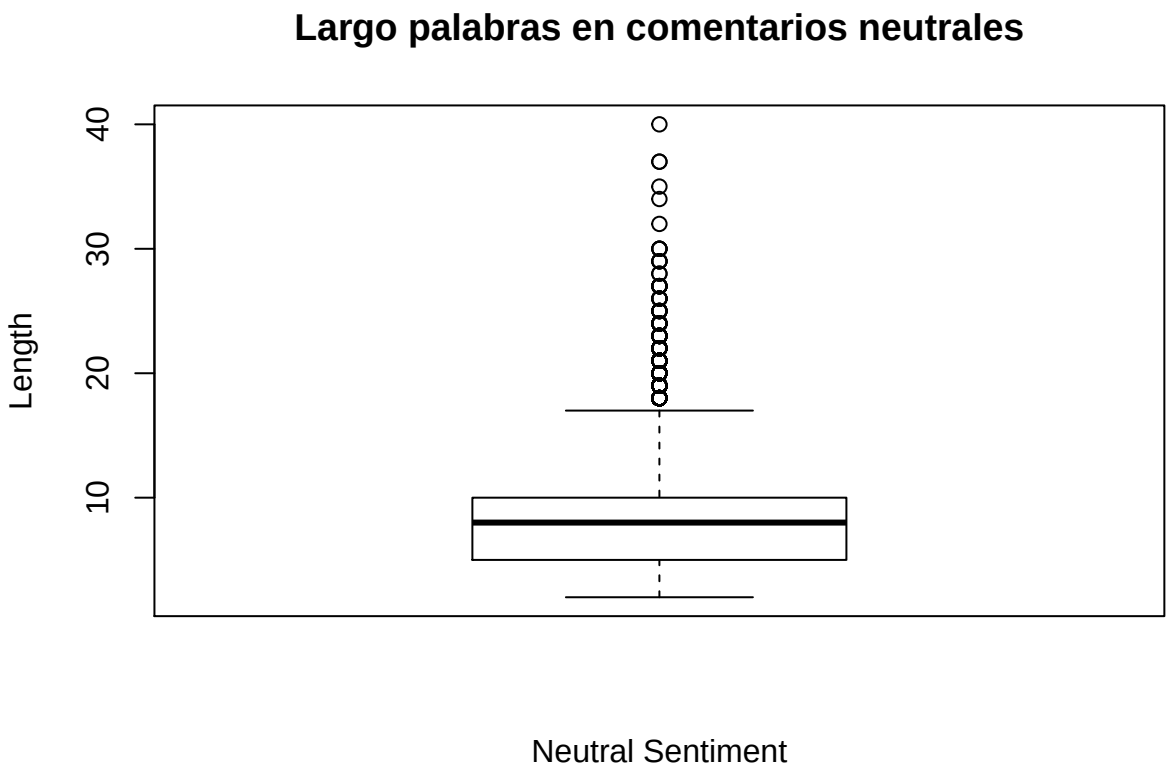
Se puede destacar que el tipo de palabras utilizadas en los 3 contextos de sentimiento (negativo, neutral y positivo) no varia en gran manera, repitiendose tipicamente las mismas dentro de las mas populares: “covid”, “coronavirus”, “supermarket”, “food”, “prices”, “store” y “grocery”. Dejando de lado la palabra “covid” las palabras mas repetidas se relacionan con el abastecimiento de productos basicos, lo que permite inferir de forma preliminar que este tema fue relevante para los usuarios durante la pandemia (entre marzo y abril).

En base a los datos obtenidos categorizados por sentimiento, puede analizarse numero promedio de sus palabras, y asi ver si existe alguna relacion entre esas variables. Para esto se hace uso de los boxplot, con el fin de comparar promedios y distribuciones, y detectar algunos outliers segun el sentimiento.

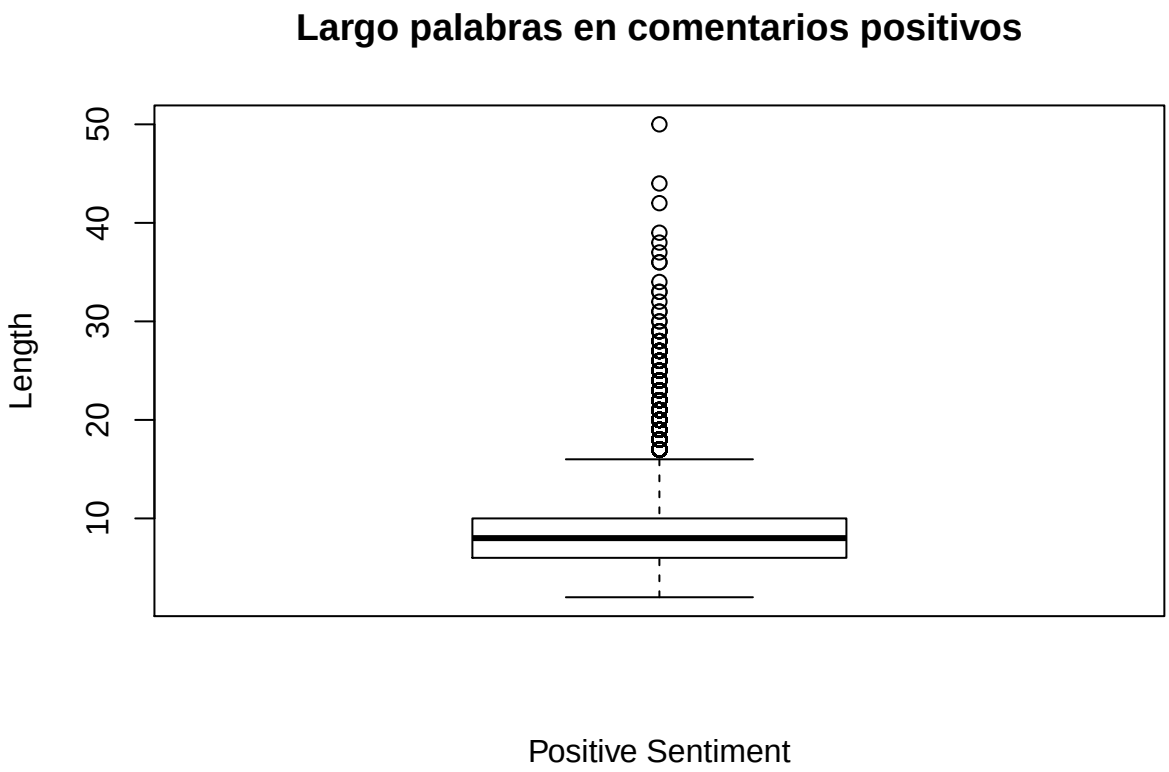
```
#Promedio en el largo de palabras por sentimiento. (Intentar hacerlos todos en un unico grafico)
library(stringr)

boxplot(str_length(neg_tweets$word),xlab = "Negative Sentiment", ylab = "Length", main="Largo palabras
```




```
boxplot(str_length(neu_tweets$word),xlab = "Neutral Sentiment", ylab = "Length", main="Largo palabras en
```




```
boxplot(str_length(pos_tweets$word),xlab = "Positive Sentiment", ylab = "Length", main="Largo palabras
```




```
cat("Negative:",mean(str_length(neg_tweets$word)),"\n")
```

```
## Negative: 8.105155
```

```
cat("Neutral:",mean(str_length(neu_tweets$word)),"\n")
```

```
## Neutral: 8.175911
```

```
cat("Positive:",mean(str_length(pos_tweets$word)),"\n")
```

```
## Positive: 8.322604
```

Se puede observar que el largo de las palabras no parece ser determinante para el valor de sentimiento del comentario en general, pues en promedio todas miden muy parecido: aproximadamente 8 caracteres.

Propuesta de hipótesis

A partir del análisis exploratorio se proponen las siguientes hipótesis y preguntas que se podrían abordar con este set de datos:

1. ¿El sentimiento general sobre el COVID-19 varía por la localidad registrada?
2. Dada las características del COVID-19 y sus consecuencias, más del 50% de los tweets están asociados a un sentimiento negativo o extremadamente negativo.
3. ¿Existe alguna relacion entre el largo promedio de las palabras utilizadas en comentarios, y el sentimiento asociado al tweet de donde provino?
4. Hipótesis/pregunta: ¿Los sentimientos mayormente expresados en los Tweets tienen relacion con el contexto social del lugar desde donde son publicados?

5. ¿Se puede asociar un sentimiento a una palabra dependiendo de las otras palabras mencionadas en un tweet?
6. ¿Existe un alza o descenso de comentarios positivos al avanzar de los dìas?¿Si es asì, a què se debe?

Referencias

[1] Twitter suma 166 millones de usuarios durante el Coronavirus. (2020, 3 junio). REBOLD, Data-Driven Marketing & Communication. https://letsrebold.com/es/blog/twitter-suma-166-millones-de-usuarios-frente-al-coronavirus/#:%7E:text=Asimismo%2C%20Twitter%20comunic%C3%B3%20en%20la,el%20primer%20trimestre%20de%202019.

[2] Coronavirus tweets NLP - Text Classification. (2020, 8 septiembre). Kaggle. https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_test.csv

Contribuciones del equipo

1. Nicolás Herrera:
2. Yesenia Marulanda:
3. Franco Migliorelli:
4. Samuel Sánchez:
5. Sebastián Urbina: