Flujo de análisis en clasificación supervisada

Métodos supervisados

Laura Rodríguez Navas

Septiembre 2020

Contents

Comenzamos cargando los paquetes necesarios.

```
library(tidyverse)
library(stringi)
library(tm)
library(irlba)
library(RColorBrewer)
library(wordcloud)
library(gridExtra)
library(caret)
library(doParallel)
library(syuzhet)
library(ggcorrplot)
```

Análisis Exploratorio de los Datos

Para la realización del ejercicio propuesto se ha elegido la competición en Kaggle: **Real or Not? NLP with Disaster Tweets**. El dataset de la competición se puede encontrar en el siguiente enlace: https://www.kaggle.com/c/nlp-getting-started/data. Este dataset, con 10.876 instancias, contiene 4 variables explicativas: **id**, **keyword**, **location** y **text**, y dos valores en la variable clase **target** (0 y 1). La variable clase es binaria, así que, vamos a aprender un modelo de clasificación binaria. El objetivo de este modelo será predecir si dado un tweet, este tweet trata sobre un desastre real o no. Si un tweet trata sobre un desastre real, se predice un 1. Si no, se predice un 0.

La métrica de evaluación esperada por la competición es F1. Y se calcula de la siguiente manera:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

donde:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

La partición inicial train-test, no se tiene que realizar, ya que las instancias de train y test ya vienen definidas en el dataset de la competición (ficheros **train.csv** y **test.csv**).

A continuación, cargaremos el conjunto de datos de train y test, nombrando los valores perdidos como **NA** para que los podamos tratar más adelante, y mostraremos sus dimensiones.

```
train <- read.csv("train.csv", na.strings=c("", "NA"))
test <- read.csv("test.csv", na.strings=c("", "NA"))
dim(train)
```

```
## [1] 7613    5
```

```
dim(test)
```

```
## [1] 3263    4
```

El conjunto de datos de train contiene 7613 instancias y el conjunto de datos de test contiene 3263 instancias. Cada instancia de estos conjuntos contiene la siguiente información:

- **id**: un identificador único para cada tweet.
- **keyword**: una palabra clave del tweet.
- **location**: la ubicación desde la que se envió el tweet.
- **text**: el texto del tweet.
- **target**: solo en el conjunto de datos de train porqué es la variable clase a predecir. Indica si un tweet es sobre un desastre real (1) o no (0).

```
str(train, width = 85, strict.width = "cut")
```

```
## 'data.frame':    7613 obs. of  5 variables:
##  $ id      : int  1 4 5 6 7 8 10 13 14 15 ...
##  $ keyword : chr  NA NA NA NA ...
##  $ location: chr  NA NA NA NA ...
##  $ text    : chr  "Our Deeds are the Reason of this #earthquake May ALLAH Forgive "..
##  $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

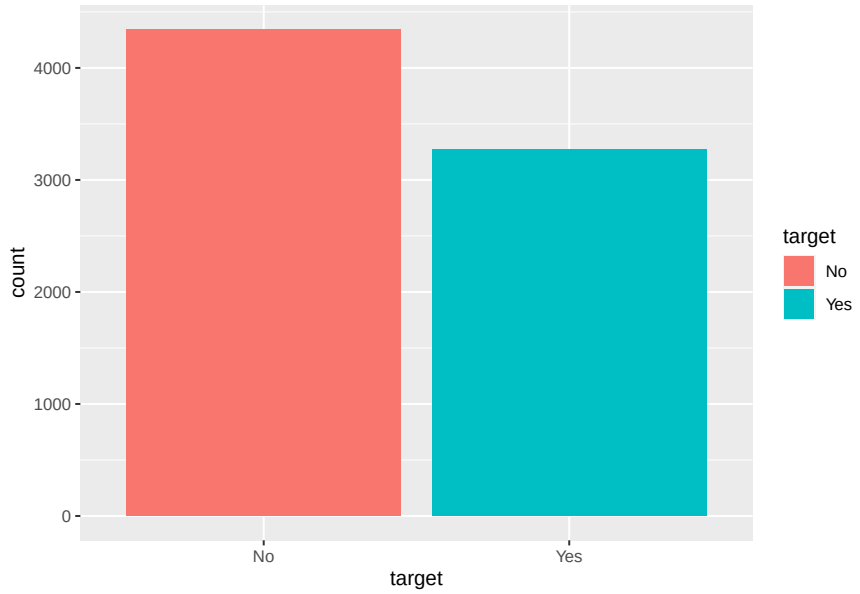
```
str(test, width = 85, strict.width = "cut")
```

```
## 'data.frame':    3263 obs. of  4 variables:
##  $ id      : int  0 2 3 9 11 12 21 22 27 29 ...
##  $ keyword : chr  NA NA NA NA ...
##  $ location: chr  NA NA NA NA ...
##  $ text    : chr  "Just happened a terrible car crash" "Heard about #earthquake is"..
```

Variable *target*

Categorizamos la variable a predecir, ya que inicialmente es de tipo entero, y observamos su distribución.

```
train$target <- as.factor(ifelse(train$target == 0, "No", "Yes"))
ggplot(train, aes(x=target)) + geom_bar(aes(fill=target))
```



La distribución no está muy sesgada y vemos que hay menos tweets que se refieren a desastres reales. La distribución de la variable a predecir está relativamente equilibrada, donde el 43% de las observaciones son desastrosas y el 57% no.

```
sum(train$target == "Yes") / dim(train)[1] * 100
```

```
## [1] 42.96598
```

```
sum(train$target == "No") / dim(train)[1] * 100
```

```
## [1] 57.03402
```

Tampoco presenta un problema notable de *desbalanceo de clase* porqué contamos con muchas muestras del caso minoritario.

Variable *keyword*

La variable **keyword** representa una palabra representativa de cada tweet, se muestran las primeras 10.

```
train %>% select(keyword) %>% unique() %>% head(10)
```

```
##                  keyword
## 1                    <NA>
## 32                ablaze
## 68              accident
## 103           aftershock
## 137 airplane%20accident
## 172            ambulance
## 210          annihilated
## 244         annihilation
## 273           apocalypse
## 305           armageddon
```

Ahora veremos si la asociación de cada **keyword** con un sentimiento indica una relación con la variable a predecir. Para ello realizaremos un análisis de sentimientos de cada palabra clave.

El análisis de sentimientos es una técnica de Machine Learning*, basada en el procesado del lenguaje natural, que pretende obtener información subjetiva de una serie de textos. Su aplicación es este caso, consiste en resolver si un tweet es real o no real en relación a un desastre.*

Para ello usaremos los paquetes **syuzhet**, **ggcorrplot** y **doParallel**.

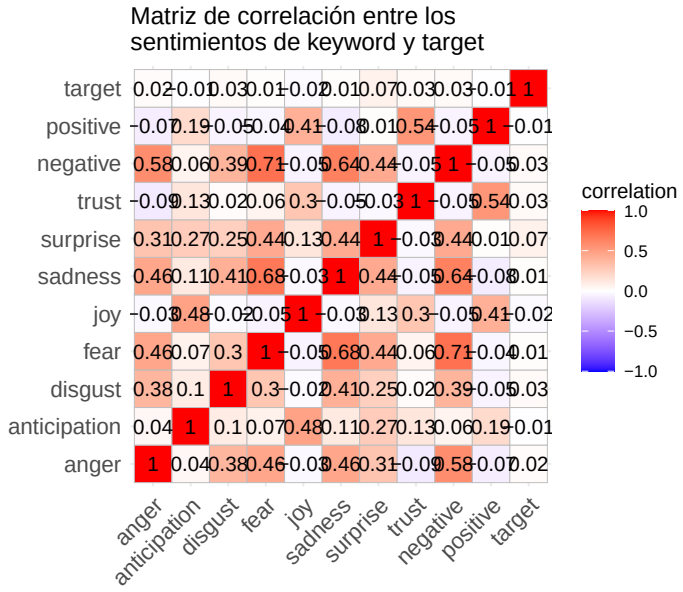
- **syuzhet** cuenta con la función **get_nrc_sentiment** que calcula la presencia de los diferentes sentimientos dado un conjunto de textos. Los parámetros de esta función son:
  - **char_v**. Un vector de caracteres que en este caso contiene todas las palabras clave.
  - **language**. Define el lenguaje.
  - **cl**. Para análisis paralelo. Es opcional, pero en este caso lo usaremos porqué hay muchas palabras clave.
- **ggcorrplot** muestra una visualización gráfica de una matriz de correlación usando *ggplot2*.
- **doParallel** proporciona una computación paralela. Los parámetros de esta función son:
  - **makePSOCKcluster**. Crea un clúster de sockets paralelos.
  - **registerDoParallel**. Registra el número de *cores* que usará el clúster creado.
  - **stopCluster**. Detiene la computación paralela.

Análisis de correlaciones entre **keyword** y **target**:

```
cl <- makePSOCKcluster(4, setup_strategy="sequential")
registerDoParallel(cl)

emocion.df <- get_nrc_sentiment(char_v = gsub("_", " ", train$keyword),
                                language = "english", cl=cl)
emocion.df <- emocion.df %>% data.frame(target = train$target)
emocion.df$target <- as.numeric(emocion.df$target)

cor(emocion.df) %>%
  ggcorrplot(lab = TRUE,
             title = "Matriz de correlación entre los \nsentimientos de keyword y target",
             legend.title = "correlation")
```




```
stopCluster(cl)
```

Al observar la matriz de correlaciones, se observa una correlación nula con cada uno de los sentimientos. Esto hace que esta variable explicativa no sea buena para hacer una predicción.

Variable *location*

La variable **location** representa la ubicación desde donde se generaron los tweets, se muestran las primeras 10.

```
train %>% select(location) %>% unique() %>% head(10)
```

```
##                          location
## 1                            <NA>
## 32                     Birmingham
## 33 Est. September 2012 - Bristol
## 34                          AFRICA
## 35               Philadelphia, PA
## 36                     London, UK
## 37                       Pretoria
## 38                  World Wide!!
## 40                Paranaque City
## 41                Live On Webcam
```

```
count(train %>% select(location) %>% unique())
```

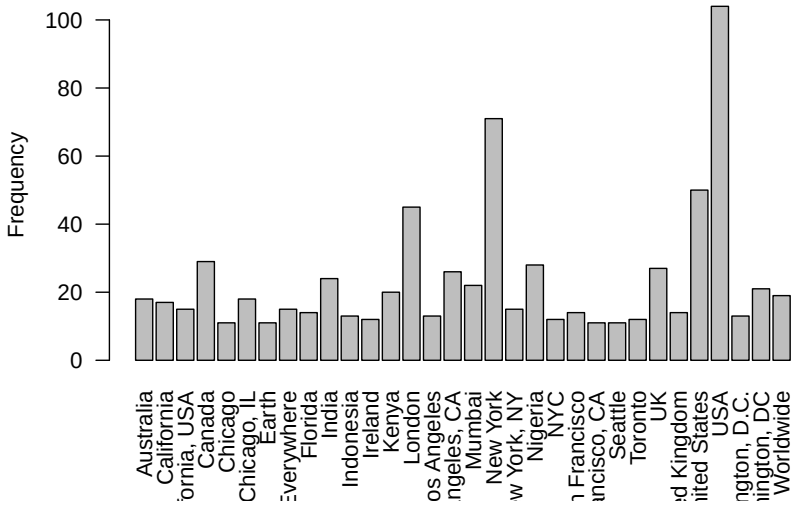
```
##      n
## 1 3342
```

En total hay 3342 ubicaciones. A continuación, mostramos las ubicaciones más frecuentes:

```
location.freq <- table(unlist(train %>% select(location)))
location.freq[which(location.freq > 10)]
```

```
##
##          Australia         California   California, USA             Canada
##                 18                 17                 15                 29
##            Chicago       Chicago, IL            Earth        Everywhere
##                 11                 18                 11                 15
##            Florida              India          Indonesia            Ireland
##                 14                 24                 13                 12
##              Kenya             London        Los Angeles   Los Angeles, CA
##                 20                 45                 13                 26
##             Mumbai           New York       New York, NY            Nigeria
##                 22                 71                 15                 28
##                NYC      San Francisco San Francisco, CA            Seattle
##                 12                 14                 11                 11
##            Toronto                 UK     United Kingdom      United States
##                 12                 27                 14                 50
##                USA  Washington, D.C.     Washington, DC          Worldwide
##                104                 13                 21                 19
```

```
barplot(location.freq[which(location.freq>10)], las = 2,
        ylab = "Frequency")
```



Del total de ubicaciones (3342), la mayoría de ellas cuenta con menos de 10 observaciones. Esto hace que esta variable explicativa tampoco sea buena para hacer una predicción.

Variable *text*

Hemos considerado que las variables explicativas **keyword** y **location** no son buenas para hacer una predicción, así que nos centraremos en la variable **text**.

Llegados a este punto unimos los conjuntos de train y test (*7613 + 3263 observaciones*) para poder extraer los sentimientos más adelante.

```
complete_df <- bind_rows(train, test)
dim(complete_df)
```

```
## [1] 10876     5
```

Echamos un vistazo más de cerca a las variables del nuevo conjunto de datos **complete_df**.

```
summary(complete_df)
```

```
##        id          keyword            location            text
##  Min.   :    0   Length:10876       Length:10876       Length:10876
##  1st Qu.: 2719   Class :character   Class :character   Class :character
##  Median : 5438   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5438
##  3rd Qu.: 8156
##  Max.   :10875
##   target
##  No  :4342
##  Yes :3271
##  NA's:3263
##
##
##
```

La variable **id** es solo un identificador único y la eliminaremos.

```
complete_df$id <- NULL
```

Observamos si existen valores perdidos.

```
colSums(sapply(complete_df, is.na))
```

```
##  keyword location     text   target
##       87     3638        0     3263
```

Las variables explicativas **keyword** y **location** contienen valores perdidos. Sobretodo hay una gran cantidad de tweets, para los cuales falta su ubicación. No existen valores perdidos para la variable explicativa **text**, tampoco para la variable a predecir **target**. Los 3263 valores perdidos de la variable a predecir provienen del conjunto de datos de test. Nos ocuparemos de los valores perdidos más adelante.

Parece que la variable explicativa **text** es una buena elección para una buena predicción y basaremos los siguientes pasos en ella.

Procesamiento de texto

Como en todo procesamiento de lenguaje natural, realizaremos el procesamiento de un conjunto de textos. En este caso realizaremos un procesamiento de los textos de los tweets y los prepararemos para el modelado.

Comencemos por crear un corpus de los mensajes de texto de los tweets. Para ello usaremos la función **Corpus** del paquete **tm**, que creará nuestro corpus a partir de un vector de textos. La función **VectorSource** interpretará cada mensaje de texto de los tweets como un elemento del vector de textos.

Un corpus lingüístico se define como "un conjunto de textos de un mismo origen" y que tiene por función recopilar un conjunto de textos. El uso de un corpus lingüístico nos permitirá obtener información de las palabras utilizadas con más o menor frecuencia.

```
myCorpus <- Corpus(VectorSource(complete_df$text))
```

Durante el procesamiento de texto seguiremos la transformación de un mensaje de tweet específico para ver como se modifica a medida que avanzamos en el procesamiento de texto. Este mensaje es:

```
paste0(myCorpus[[400]])
```

```
## [1] "Jewish leaders prayed at the hospital where a Palestinian family is being treated after arson h
```

Dividimos el procesamiento de texto en 7 pasos.

1. Eliminar enlaces.

```
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
paste0(myCorpus[[400]])
```

```
## [1] "Jewish leaders prayed at the hospital where a Palestinian family is being treated after arson
```

Hemos eliminado: *http://t.co/Wf8iTK2KVx*.

La función **gsub** busca y reemplaza desde la primera hasta todas las coincidencias de un patrón (que normalmente representa una *regular expression*). La función **tm_map** es la encargada de aplicar las diferentes transformaciones de los textos al corpus creado.

2. Convertir a minúsculas.

```
myCorpus <- tm_map(myCorpus, content_transformer(stri_trans_tolower))
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed at the hospital where a palestinian family is being treated after arson
```

3. Eliminar los nombres de usuario.

```
removeUsername <- function(x) gsub("@[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeUsername))
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed at the hospital where a palestinian family is being treated after arson
```

Hemos eliminado: *@huffpostrelig*.

4. Eliminar todo excepto el idioma y el espacio en inglés.

```
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed at the hospital where a palestinian family is being treated after arson  v
```

No se observan cambios en en ejemplo.

5. Eliminar palabras irrelevantes (eliminación de redundancias).

```
myStopWords <- c((stopwords('english')),
           c("really", "tweets", "saw", "just", "feel", "may", "us", "rt", "every", "one",
             "amp", "like", "will", "got", "new", "can", "still", "back", "top", "much",
             "near", "im", "see", "via", "get", "now", "come", "oil", "let", "god", "want",
             "pm", "last", "hope", "since", "everyone", "food", "content", "always", "th",
             "full", "found", "dont", "look", "cant", "mh", "lol", "set", "old", "service",
             "city", "home", "live", "night", "news", "say", "video", "people", "ill",
             "way",  "please", "years", "take", "homes", "read", "man", "next", "cross",
             "boy", "bad", "ass"))

myCorpus <- tm_map(myCorpus, removeWords, myStopWords)
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed   hospital   palestinian family   treated  arson   "
```

Hemos eliminado: at the where a is being after via.

Las palabras irrelevantes que hemos eliminado se denominan *stop words o palabras vacías*. Cada idioma tiene sus propias palabras vacías. Como los textos están en inglés hemos eliminado los *stop words* que pertenecen al inglés usando la función **stopwords**, además hemos añadido aleatóriamente alguna de las palabras vacías más usadas en los mensajes de texto de los tweets (ver https://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/).

Las stop words o palabras vacías son todas aquellas palabras que carecen de un significado por si solas. Suelen ser artículos, preposiciones, conjunciones, pronombres, etc.

6. Eliminar palabras de una sola letra.

```
removeSingle <- function(x) gsub(" . ", " ", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeSingle))
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed hospital palestinian family treated  arson "
```

No se observan cambios en en ejemplo.

7. Eliminar espacios en blanco adicionales.

```
myCorpus <- tm_map(myCorpus, stripWhitespace)
paste0(myCorpus[[400]])
```

```
## [1] "jewish leaders prayed hospital palestinian family treated arson "
```

Terminamos con el procesamiento de texto. A continuación, crearemos dos *Term Document Matrix* (matriz que describe la frecuencia de las palabras que se producen en una colección de textos) para un análisis de sentimientos más detallado. Usaremos la función **TermDocumentMatrix** y dividiremos el corpus en dos, según el número de elementos de los conjuntos de datos train y test. Recordamos que el conjunto de datos de train contiene 7613 observaciones, y el conjunto de datos de test contiene 3263 observaciones. El parámetro **control** evalúa cada texto de la matriz, concretamente se evaluaran todas las palabras de cada texto (no aplicamos ningún filtro).

```
train_tdm <- TermDocumentMatrix(myCorpus[1:7613],
                                control= list(wordLengths= c(1, Inf)))
test_tdm <- TermDocumentMatrix(myCorpus[7614:10876],
                               control= list(wordLengths= c(1, Inf)))
train_tdm
```

```
## <<TermDocumentMatrix (terms: 14825, documents: 7613)>>
## Non-/sparse entries: 58707/112804018
## Sparsity           : 100%
## Maximal term length: 49
## Weighting          : term frequency (tf)
```

```
test_tdm
```

```
## <<TermDocumentMatrix (terms: 8966, documents: 3263)>>
## Non-/sparse entries: 25434/29230624
## Sparsity           : 100%
## Maximal term length: 35
## Weighting          : term frequency (tf)
```