ANÁLISIS

1. Descripción del dataset

A través del hashtag **#Feminismo**, que es una forma de relacionar directamente los mensajes de los usuarios en Twitter, se podrán observar las opiniones y sentimientos que este tema provoca en la sociedad.

Las operaciones a realizar a través de **RStudio**, son las siguientes:

1. Acceder a la API pública de Twitter.
2. Descargar el conjunto de datos en un dataset.
3. Limpieza de los datos y crear el dataset para el análisis.
4. Análisis de los datos:
   4.1. Análisis de Sentimientos.
   4.2. Análisis estadísticos de los resultados
5. Representación de los resultados:
6. Modelización y Predicción.
   6.1 Machine Learning: 'RTexTools'

Para obtener unos datos suficientemente representativos, se ha optado por descargar 2000 tweets, con 16 variables por tweet.

Los datos serán exportados a un archivo: "*1-tweetsDF-ORIGINAL.csv*".

	text	favorited	favoriteCount	replyToSN	created	truncated
1	cuando dicen "feminismo era el de antes ahora sólo quiere...	FALSE	0	NA	2018-01-06 01:28:49	TRUE
2	RT @NachoGentili1: Tengo los huevos por el piso de que ca...	FALSE	0	NA	2018-01-06 01:28:38	FALSE
3	RT @ale_mdd: Lo creáis o no el feminismo también lucha c...	FALSE	0	NA	2018-01-06 01:28:38	FALSE
4	RT @_magnetismo: la gente dice q el feminismo te llena de ...	FALSE	0	NA	2018-01-06 01:28:33	FALSE
5	RT @ale_mdd: Lo creáis o no el feminismo también lucha c...	FALSE	0	NA	2018-01-06 01:28:30	FALSE
6	@MartinSoria99 Tincho , te quiero loco. Entendiste todo. Di...	FALSE	0	MartinSoria99	2018-01-06 01:28:30	FALSE
7	RT @ale_mdd: Lo creáis o no el feminismo también lucha c...	FALSE	0	NA	2018-01-06 01:28:17	FALSE
8	RT @yoongihxbits: no sé q me da más pena la gente q cree...	FALSE	0	NA	2018-01-06 01:28:08	FALSE
9	RT @ale_mdd: Lo creáis o no el feminismo también lucha c...	FALSE	0	NA	2018-01-06 01:28:07	FALSE
10	RT @ale_mdd: Lo creáis o no el feminismo también lucha c...	FALSE	0	NA	2018-01-06 01:28:05	FALSE
11	RT @NachoGentili1: Tengo los huevos por el piso de que ca...	FALSE	0	NA	2018-01-06 01:27:43	FALSE

ing 1 to 13 of 2,000 entries

2. Limpieza de los datos.

Es en el campo "text", se podrán observar las opiniones y sentimientos que este tema provoca. Se pueden descartar el resto de variables del tweet al no ser de interés para esta práctica.

Posteriormente con los datos del campo "text", realizaremos las siguientes operaciones:

A. Limpieza del texto, mediante la aplicación de una función diseñada *ad hoc*.

B. Realizar un análisis de sentimientos, mediante la aplicación de la correspondiente función.

En un tweet existen diferentes caracteres que no tienen significación sentimental alguna. En R existen varias funciones, si bien en este caso se ha utilizado *gsub ()*, la cual reemplaza la aparición de una subcadena con otra subcadena dentro de un vector.

Como resultado de la limpieza de datos, el campo "text" estará en condiciones de serle aplicado el análisis de sentimientos.

3. Análisis de los datos.

Puesto que el objetivo es analizar las palabras contenidas en el texto limpio de los tweets, será necesario disponer de un **diccionario de palabras ponderadas según sentimientos** relacionados con la palabra "Feminismo", para comparar cada palabra de los tweets con las del diccionario y si coinciden asignarle el valor correspondiente. El archivo del diccionario de sentimientos es: "*LISTA-Palabras-Sentimientos.txt*".

Una vez que se tiene el texto de los tweets limpio y el diccionario de sentimientos confeccionado, se crea una FUNCIÓN para clasificar palabras del texto de los tweet, en 4 categorías: muy Negativo, negativo, positivo y muy Positivo.

Seguidamente se creará una matriz "*scores_final*", la cual incluirá el campo "text" y los cuatro campos correspondientes a la clasificación de las palabras. Se calcula la puntuación de cada tweet y se añaden dos nuevas columnas, la del valor alcanzado por cada tweet y el del sentimiento asignado. El *data frame* obtenido y exportado es: "*2-tweet.Resultado.csv*".

	texto	muyNeg	Neg	Pos	muyPos	Valor	Sentimiento
1	cuando dicen feminismo era el de antes ahora sólo quieren...	0	0	0	0	0.0000000	Neutro
2	Tengo los huevos por el piso de que cada tweets uno sea d...	0	0	0	0	0.0000000	Neutro
3	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo
4	la gente dice q el feminismo te llena de odio y re q te llena ...	0	1	2	0	0.3571429	Positivo
5	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo
6	Tincho te quiero loco Entendiste todo Dio un discurso sobr...	0	2	1	0	-0.3571429	Negativo
7	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo
8	no sé q me da más pena la gente q cree q el feto siente a l...	0	1	0	0	-0.5000000	Negativo
9	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo
10	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo
11	Tengo los huevos por el piso de que cada tweets uno sea d...	0	0	0	0	0.0000000	Neutro
12	Lo creáis o no el feminismo también lucha contra esto Si lo...	0	1	0	0	-0.5000000	Negativo

Showing 1 to 13 of 2,000 entries

Finalmente, se crea una tabla de sentimientos 5x2, correspondiendo a la primera columna sentimientos (cuatro categorías de sentimientos más el neutro -ausencia de sentimientos-) y a la segunda columnas la frecuencia observada para cada sentimiento. La tabla de sentimientos es exportada a un archivo csv: "*3-Tabla-Conteo-Sentimientos.csv*".

	Var1	Freq
1	Muy Negativo	84
2	Muy Positivo	17
3	Negativo	944
4	Neutro	588
5	Positivo	367

ng 1 to 5 of 5 entries

3.1. Análisis estadísticos de los resultados.

Seguidamente se realizará la prueba de *machine learning,* utilizando la librería RtexTools, para crear una tabla de predicciones y calcular la precisión de la recuperación de los datos y otra segunda prueba, de comparación de dos muestras (antes-después), donde se estudiará el ajuste de datos reales entre ambas.

Como se puede observar en la tabla que figura más abajo, los resultados muestran que la totalidad de las predicciones son negativas, calculando una precisión de la recuperación de los datos del 46,63%.

```
269 # Modelo de Predicciones
270 tweets_ponderados <- tweetResultado[,c(1,7)]
271
272 matrix <- create_matrix(tweets_ponderados[,1],minWordLength=4,language="spanish",
273                         maxWordLength=15, removeStopwords = F, removeNumbers = T,
274                         removePunctuation=F, toLower=T, stemWords = F)
275
276 mat <- as.matrix(matrix)
277 classifier <- naiveBayes(mat[1:(num.tweets*0.8)], as.factor(tweets_ponderados[1:(num.tweets*0.8),2]))
278
279 # Función genérica para predicciones a partir de los resultados de ajuste del primer argumento.
280 predicted <- predict(classifier, mat[(num.tweets*0.8):num.tweets]); predicted
281
282 # Crea la tabla de predicción
283 table(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
284
285 # Calcula la precisión de recuperación de los datos clasificados
286 recall_accuracy(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
287
265:20  (Untitled)                                                    R Script
```

```
Console ~/
[386] Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo
[397] Negativo Negativo Negativo Negativo Negativo
Levels: Muy Negativo Muy Positivo Negativo Neutro Positivo
>
> # Crea la tabla de predicción
> table(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
              predicted
               Muy Negativo Muy Positivo Negativo Neutro Positivo
  Muy Negativo            0            0       13      0        0
  Muy Positivo            0            0        7      0        0
  Negativo                0            0      187      0        0
  Neutro                  0            0      130      0        0
  Positivo                0            0       64      0        0
>
> # Calcula la precisión de recuperación de los datos clasificados
> recall_accuracy(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
[1] 0.4663342
```

3.2. Comparación de los grupos de datos.

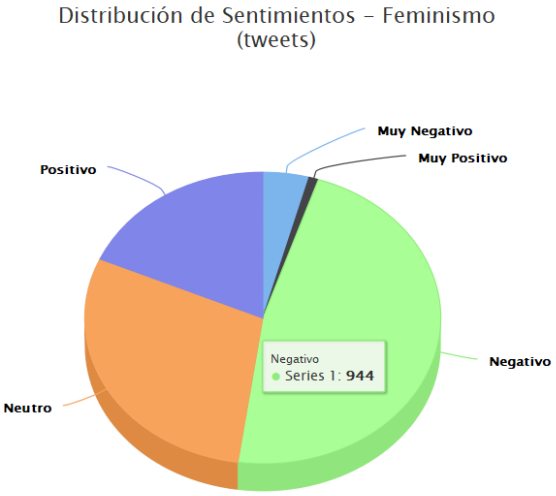
También compararemos las dos muestras (antes-después), donde se observará el ajuste de datos reales entre ambas.

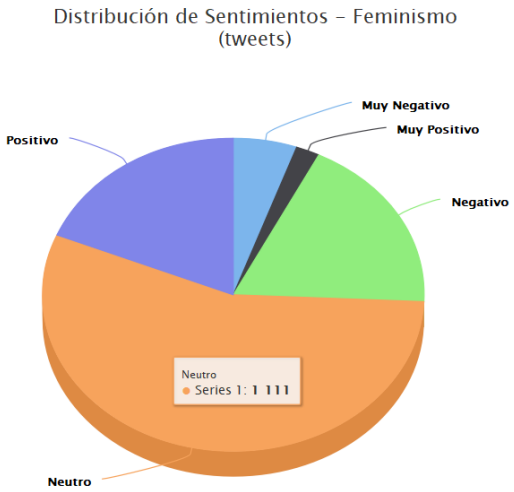
	Var1	Freq	Freq2
1	Muy Negativo	84	108
2	Muy Positivo	17	40
3	Negativo	944	366
4	Neutro	588	1111
5	Positivo	367	375

A diferencia de la prueba de *machine learning*, ahora los datos son muestras reales y por lo tanto, nos indicarán con mayor fiabilidad, no solo como funciona el ajuste del modelo establecido para el análisis de sentimientos sino también la variabilidad de los sentimientos recogidos en los tweets en tan solo 48 horas de diferencia.

4. Comparativa de análisis de sentimientos antes-después.

	Var1	Freq	Freq2
1	Muy Negativo	84	108
2	Muy Positivo	17	40
3	Negativo	944	366
4	Neutro	588	1111
5	Positivo	367	375





5. ¿Cuáles son las conclusiones?

Tras el hashtag #Feminismo, los usuarios en Twitter en sus mensajes manifiestan opiniones y sentimientos mayoritariamente negativos o neutros, no siendo despreciable tampoco la cantidad de mensajes muy negativos.

El diccionario de sentimientos es muy sensible, y mejoraría incorporando palabras más representativas del feminismo. Se ha podido comprobar que las opiniones que se vierten contra el feminismo son negativas o muy negativas en un número considerable.