# Asignatura Text Mining en Social Media.
# Master Big Data 2018

**Jose Ignacio Marqués Ortega**
jomaror4@etsii.upv.es

## Abstract

Como tarea evaluable de la asignatura de Text Mining en Social Media del programa del Máster en Big Data Analytics de la UPV, se propone realizar un ejercicio de *Author-Profiling* utilizando tweeter como fuente de datos.

El ejercicio propuesto consiste en abordar dos características concretas de un autor, su sexo y variedad de idioma Espaol que utiliza. La hipótesis de partida se centra en que las personas se expresan de manera distinta en función de sus propias caracter´sticas, como pueden ser sexo y variedad del idioma.

Con este objetivo, se dispone de un corpus obtenido en Twitter, con una colección de textos de miles de autores y cientos de tuits por autor, de una gran variedad de temas, etiquetados por sexo (2 clases) y por variedad de lenguaje (7 clases).

## 1 Introducción

Se propone realizar un ejercicio de ***Author-Profiling***, donde se pretende clasificar al autor de un tweet por sexo (Hombre, Mujer) y variedad de idioma español (castellano, argentino, mexicano, colombiano, chileno, venezolano, peruano).

Para ello, se dispone de un corpus de entrenamiento de trecientos autores con cien tweets por autor, de los cuales se dividen en doscientos para entrenamiento y cien para validar el modelo generado. Adicionalmente se generan bolsas de palabras a partir de la frecuencia de aparicin de las mismas en el corpus, para entrenar modelos de aprendizaje automático se emplearán diversos algoritmos de clasificación supervisada.

## 2 Dataset

Se proporciona el *dataset* PAN-AP17 construido de la siguiente forma:

1. Se recopilan tweets enmarcados en un área geográfica determinada.
2. Se seleccionan los autores únicos que han escrito tweets.
3. Se recuperan los *time-lines* de los autores preseleccionados.
4. Se vuelve a seleccionar los autores que hallan escrito más de 100 tweets, en el idioma correspondiente.
5. Se revisan personalmente los tweets para determinar el sexo del autor.
6. Por último, se seleccionan los 100 tweets por autor, almacenados de la siguiente forma:
    - Dos ficheros txt, uno el train y otro el test con el id, sexo y variedad separados por tabuladores.
    - Fichero json por autor, con cada uno de los tweets escritos por el mismo en formato xml.

El corpus de entrenamiento lo componen 8200 archivo .xml mientras que el de test 1400. Cada xml pertenece a un autor con sus tweets. Se dispone de 2800 autores de los cuales:

- Clasficación por sexo: 1400 para cada clase (2 clases).
- Clasficación por variedad: 200 por cada clase (7 clases).

Las probabilidades apriori de clasificar bien un tweet son:

- Para el caso del Sexo 0,5
- Para el caso de la variedad 0,2

## 3 Propuesta de resolución del equipo

En primer lugar, se procedió a abordar los problemas de clasificación por separado, es decir, tratar el sexo y variedad de forma independiente. Posteriormente se plantea el realizar clasificador cruzado para ambos problemas con un total de 14 clases, ya que se generaría una clase por la combinación de los dos problemas.

Para el caso del sexo, se partió de la hipótesis "*Las mujeres tienden a escribir más que los hombres* ". Por ello, se obtuvieron variables de longitudes y distancias a los máximos y mínimos sobre las longitudes de los tweets, separados por hombre y mujer.

Las variables obtenidas para enriquecer el modelo fueron:

- Distancia del tweet a la longitud media de tweets por sexo.
- Distancia del tweet a la longitud máxima de tweets por sexo.
- Distancia del tweet a la longitud mínima de tweets por sexo.

Estas variables son de tipo numérica y no tiene que ver con el contenido, únicamente con la longitud de tweet. Por otra parte, para mejorar la clasificación, se planteó adicionalmente utilizar a la bolsa de palabras facilitada en el código del baseline, donde variariamos el número de palabras de la bolsa. Además de generar otra bolsa de palabras frecuentes para el caso del sexo.

Para el caso de la variedad de dialecto espaol se optó por utilizar unas bolas de palabras frecuentes por dialecto y palabras únicas, además de las variables antes calculadas y se valoró la utilización de bolsas de n-gramas.

En cuanto a los algoritmos de clasificación se probaron:

- SVM
- Regresión Logística.
- k-NN
- Ranger (Implementación ágil de *Random forest*)

Para la ejcución de los modelos se utilizó la librería caret de R, mediante el id de programación RStudio.

## 4 Resultados experimentales

La medida de evaluación utilizada ha sido accuracy, siendo el porcentaje de tweets correctamente clasificados.

Estos son los resultados obtenidos al ejecutar los algoritmos nicamente variando el numero de elementos de la bolas de palabras facilitada en el baseline.

| N_Words | Model | Gender | Variety | Joint |
|---|---|---|---|---|
| 50 | SVM | 0,666 | 0,376 | 0,252 |
| 100 | SVM | 0,678 | 0,489 | 0,326 |
| 500 | SVM | 0,709 | 0,758 | 0,539 |
| 1000 | SVM | 0,659 | 0,782 | 0,515 |
| 100 | Ranger | 0,661 | 0,498 | 0,331 |
| 1000 | Ranger | 0,716 | 0,885 | 0,629 |
| 100 | RegLog | 0,673 | | |
| 500 | RegLog | 0,706 | | |

Como se puede observar en la tabla anterior la combinación ganadora es, el modelo Ranger con una bolsa de 1000 palabras.

Por ello, a esta combinación se le añadirán las variables de distancias de tweets antes calculadas, con el fin de ajustar el modelo.

El resultado de esta combinación es el siguinte:

- Accuracy Clasificación gender: **0,723**
- Accuracy Clasificación variety: **0,885**

De forma visual tenemos:

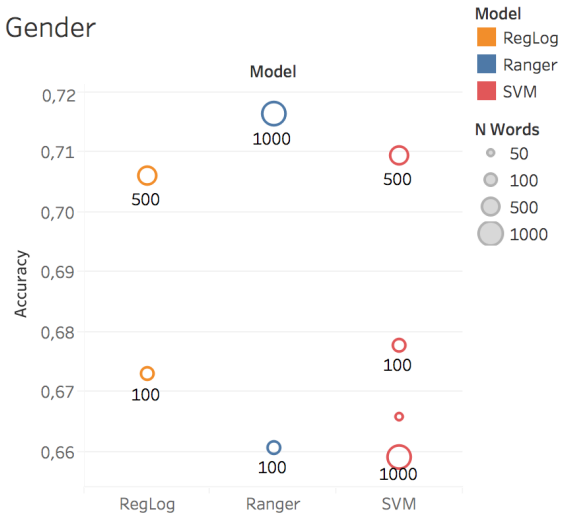


Figure 1: visualización sexo.

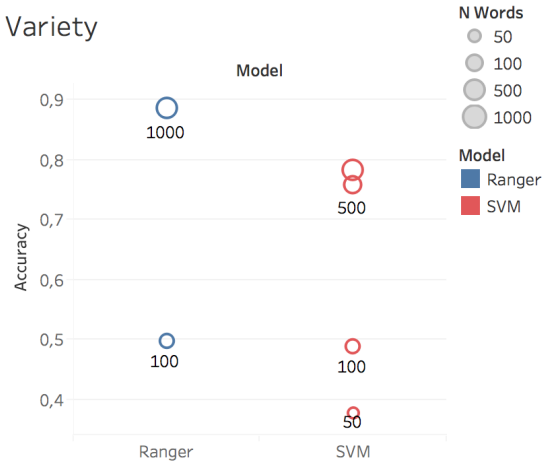


Figure 2: visualización variedad.

A la luz de estos resultado se puede decir que la obtencin de variables derivadas para enriquecer el entrenamiento produce un mejor grado de ajuste en el modelo.

En cuanto al tiempo de ejecución se observa que cuantas más palabras en el vocabulario se incrementa considerablemente el tiempo de entrenamiento y procesado del modelo.

## 5 Conclusiones y trabajo futuro

Análisis de los resultados obtenidos:

1. El mejor clasificador se obtiene al entrenar el modelo Ranger de la librería Caret, con un a bolsa de 1000 palabras y adicionando variables derivadas. (ver tabla).

2. En general, el clasificador conjunto presenta peores resultados que los clasificadores separados.

3. Las mejoras más sustanciales se observan al incrementar el número de palabras en la bolsa, en casi todos los modelos se ha mejorado el ajuste únicamente aumentado el vocabulario.

Propuestas de trabajos futuros:

- Realizar diferentes bolas de palabras para la clases.
- Utilización de N-gramas.
- Probar otro tipo de modelos como el ExtraTree Classifer, Naïve Bayes o redes neuronales.
- Añadir utilización de emoticonos.

## References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.