# Esquema de paper. Asignatura Text Mining en Social Media. Master Big Data

**Araceli Teruel Domenech**
araceliteruel@gmail.com

## Abstract

La tarea realizada y que se detallará en este papaer consiste en hacer un análisis de text mining para identificar a partir de unos tweets dados, el sexo del autor del tweet, así como la variedad del lenguaje.

## 1 Introducción

El objetivo del estudio es identificar el perfil del autor de un texto, en nuestro caso de Tweets.

Hoy en día las redes sociales están presentes de una manera muy importante en nuestra sociedad. Como sabemos, los usuarios de las redes sociales, son un simple "nick" que puede desenvolverse de manera totalmente anónima por las redes, sin que sepamos ni su sexo, ni su edad, ni tan siquiera su nacionalidad. Problemas como la pederastia online, los asaltos sexuales a menores y no menores, el "bulling" virtual etc, son problemas tan graves que se hace necesario construir un sistema de detección de los perfiles de los usuarios que navegan por internet. Es por ello que conseguir a través del análisis del texto del Tweet, saber datos como el sexo, la variedad lingüística o hasta la personalidad del autor es un reto que se nos hace posible de afrontar gracias a los métodos de Machine Learning.

Así, clasificaremos los tweets utilizando técnicas de Machine Learning. La clasificación se realizará en función de las características extraídas del tweet que se usarán después para entrenar al clasificador. En este trabajo, dado que nuestro principal objetivo es abordar el problema del "Author Profiling", nos enfocaremos en la predicción del género y la variedad lingüística de los autores de los tweets.

Usaremos distintos métodos, y analizaremos los tweets dados a partir de un dataset de

## 2 Dataset

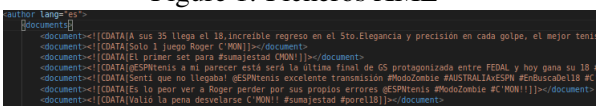
Los datos que se han usado para enseñar al modelo son los proporcionados por PAN-AP 2017 y consisten en un conjunto de tweets en español. El dataset está formado por 2800 líneas etiquetados por género (male y female) y por variedad lingüística (Argentina, Chile, Colombia, Mexico, Perú, Spain, Venezuela). Una primera visualización de los datos puede ser la que podemos observar en la Tabla 1

| **Sexo variedad** | **female** | **male** |
|---|---|---|
| Argentina | 200 | 200 |
| Chile | 200 | 200 |
| Colombia | 200 | 200 |
| Mexico | 200 | 200 |
| Peru | 200 | 200 |
| Spain | 200 | 200 |
| Venezuela | 200 | 200 |

Table 1: Resumen datos

Los datos los tenemos en formato xml. Cada usuario está guardado en un fichero distinto. Dentro de cada fichero nos encontramos con los tweets del usuario estructurado de manera que cada tweet empieza con la etiqueta document. Un ejemplo, lo podemos ver en la Figura 1

Figure 1: Ficheros XML




Para trabajar con los datos en R, hemos usado el paquete "XML" de R. Para obtener la clasificación de los autores del dataset train, los resultados están guardados en un txt, con el nombre del fichero del autor, la clasificación por género, y la clasificación

por variedad, tal y como podemos ver en la Figura 2

Figure 2: Clasificación

```
74bcc9b0882c8440716ff370494aea09:::female:::colombia
4639c055f34ca1f944d0137a5aeb7914:::female:::colombia
92ffa98bade702b86417b118e8aca319:::female:::colombia
```

## 3 Propuesta del alumno

La primera decisión que tomamos fue realizar el estudio usando R, dado que teníamos implemetadas las funciones necesarias para leer los ficheros XML que contienen los tweets.

Se tomaba como baseline los resultados que se obtenían recogiendo la bolsa de palabras más usadas por los autores de los tweets, y después entrenando un modelo de Supper Vector Machine lineal tanto para la predicción por género como por variedad lingüística.

Para la predicción por género, pensamos en añadir a la bolsa de palabras que se generaba como palabras más usadas con stopwords, palabras relacionadas con el deporte. Las palabras que se añadieron fueron

(messi, ronaldo, cristiano, madrid, barça, real, futbol, arbitro, gol, atletico, numancia, roja, expulsion, basket, canasta, partido, eurocopa, mundial, pique, champions, entrenador)

Con esta bolsa de palabras y usando el mismo modelo que con el baseline obtuvimos una mejora de resultados notable.

Otra característica que nos planteamos fue añadir a la bolsa de palabras inicial, un conjunto de palabras que como encontramos en [**1**] y usando el modelo de entrenamiento de Random Forest, conseguimos mejorar el baseline por sexo.

Una vez llegados a este punto, cada vez se nos hacía más complicado continuar con R, por lo que decidimos pasarlo todo a python y seguir buscando características adecuadas. Una vez teníamos la lectura de los XML en python, utilizamos el método TFIDF para obtener las palabras más usadas por cada género y jugamos con ellos. Así buscamos las palabras más usadas por los hombres que las mujeres no usa y viceversa, y así creamos una bolsa de palabras con lo más usado por cada uno de ellos.

Un ejemplo puede ser las palabras más usadas por los hombres, que las mujeres no usan y viceversa, como podemos ver en las Figuras 3 y

A pesar de que este modelo no ha sido el que mejor resultado nos ha dado, es una buena característica para añadir al resto de características.

Figure 3: Palabras más usadas por los hombres que las mujeres no utilizan

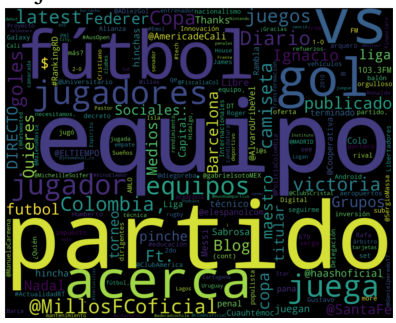



Figure 4: Palabras más usadas por las mujeres que los hombres no utilizan


## 4 Resultados experimentales

El modelo que mejor nos ha funcionado y mejores resultados nos ha dado ha sido haciendo un TFIDF para obtener las palabras más usadas por hombres y mujeres, y así hacer un modelo Random Forest con 500 árboles. Los resultados obtenidos para la clasificación por sexo han sido los mostrados en la Figura 5

Figure 5: Clasificación por género usando Random Forest sobre una bolsa de palabras obtenida haciendo un TFIDF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| female | 0.74 | 0.72 | 0.73 | 700 |
| male | 0.73 | 0.75 | 0.74 | 700 |
| avg / total | 0.74 | 0.73 | 0.73 | 1400 |

En cuanto a la variedad lingüística, si aplicamos el mismo modelo usado para clasificar los tweets por genero, obtenemos los resultados que se pueden observar en la Figura

Como podemos observar, funciona mucho mejor para la variedad lingüística que para el género.

Figure 6: Clasificación de la variedad lingüística usando Random Forest sobre una bolsa de palabras obtenida haciendo un TFIDF

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| argentina | 0.93 | 0.96 | 0.94 | 200 |
| chile | 0.97 | 0.96 | 0.97 | 200 |
| colombia | 0.92 | 0.94 | 0.93 | 200 |
| mexico | 0.89 | 0.93 | 0.91 | 200 |
| peru | 0.98 | 0.88 | 0.93 | 200 |
| spain | 0.89 | 0.95 | 0.92 | 200 |
| venezuela | 0.97 | 0.93 | 0.95 | 200 |
| avg / total | 0.94 | 0.94 | 0.94 | 1400 |

## 5 Conclusiones y trabajo futuro

El trabajo futuro que se plantea es intentar juntar las características que mejores resultados nos han dado, y entrenar un modelo que tenga en cuenta a todas ellas. Estas características son

- TFIDF con las palabras más usadas por género
- Bolsa de palabras añadidas [**1**]
- Bolsa de palabras de deporte
- Longitud de tweets
- Signos de puntuación

Así mismo se queda pendiente probar algunas de las características como pueden ser las negaciones usadas, los pronombres usados, la terminación de las palabras...

Breve presentación de las conclusiones sobre el trabajo realizado e ideas de futuro para mejorar los resultados.

## References

Las palabras más utilizadas por hombres y mujeres en internet, Sarah Romero https://www.muyinteresante.es/tecnologia/articulo/las-palabras-mas-utilizadas-por-hombres-y-mujeres-351464876138