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

**Javier Montagud Gómez**

javiganval@gmail.com

## 1 Introducción

A continuación voy a tratar de analizar un conjunto de tweets escritos por diferentes autores con el objetivo de conseguir predecir ante un tweet nuevo la siguiente información:

- El sexo del autor, si es hombre o mujer.
- La variedad española del autor.

Para ello, partimos de un total de 2800 de autores que han escrito 100 tweets cada uno. Aplicando Machine Learning voy a tratar de resolver los problemas anteriores.

Al tratarse de 2 problemas diferentes, los estudiaré de forma separada.

## 2 Dataset

Se dispone de un vocabulario con las 1000 palabras más frecuentes en todos los tweets de los autores. Se dispone además de una bolsa de palabras que contienen la frecuencia de cada una de estas palabras en los tweets de cada autor.

A raíz de aquí voy a tratar de realizar mejoras de predicción en función del problema.

## 3 Propuesta del alumno

He tratado tanto el género como la variedad como 2 problemas diferentes.

Para el problema de la variedad he creado el fichero **Gender.R**.

Sobre la propuesta inicial he realizado las siguientes modificaciones:

En la fase de generación de la bolsa de palabras, he añadido las siguientes columnas, que contienen las frecuencias de las siguientes palabras en los tweets de cada autor:

- Adjetivos: palabras localizadas en el fichero adjetivos.txt. En principio las mujeres utilizan más los adjetivos que los hombres.
- Pronombres: palabras localizadas en el fichero pronombres.txt. En principio las mujeres utilizan más los pronombres que los hombres

- Palabras cariñosas: palabras localizadas en el fichero palabrasb.txt. En principio las mujeres utilizan más los adjetivos que los hombres
- Los horóscopos.
- Las palabras más frecuentes de las mujeres.
- Las palabras más frecuentes de los hombres.
- Las palabras de la revista Superpop

Estas frecuencias las he dividido por el número de palabras de los 100 tweets de cada autor.

No es lo mismo que la palabra “casa” aparezca 5 veces en un total de 100 palabras escritas por un autor que en un total de 200 palabras de otro autor. No tendría el mismo peso. Es por ello que lo dividimos por el total de palabras de cada autor

Adicionalmente he añadido una columna más con el número de palabras de todos los tweets de cada autor.

Para el problema de la variedad he creado el fichero **Variety.R**.

Sobre la propuesta inicial he realizado las siguientes modificaciones:

En la fase inicial, donde se preprocesan los datos, he quitado los acentos para conseguir que no se tuviesen en cuenta a la hora de entrenar el modelo.

En la fase posterior, donde se genera la bolsa de palabras, he añadido 8 nuevas columnas.

7 de las columnas hacen referencia a cada una de las 7 variedades posibles. Los valores de cada una de estas 7 columnas son:

- Frecuencia de las 300 palabras más repetidas en cada variedad dividido por el numero de palabras de los 100 tweets de cada autor

La última columna hace referencia al número de palabras de todos los tweets de cada autor.

## 4 Resultados experimentales

Para el problema del género se realizan diferentes pruebas con diferentes métodos que tras las pruebas se desestiman.

El método más favorable es el “Penalized Multinomial Regression”.

Sobre dicho método realizamos diferentes pruebas con diferentes combinaciones.

Los resultados obtenidos de dichas pruebas son los siguientes:

| MÉTODO | DESCRIPCIÓN | ACCURACY | KAPPA |
|---|---|---|---|
| Multinomial | Sin tener en cuenta los archivos mujeres.txt y hombres.txt | 0.7321 | 0.4643 |
| Multinomial | Con acentos | 0.7336 | 0.4671 |
| Multinomial | Añadiendo Superpop2.txt | 0.7329 | 0.4657 |
| Multinomial | Añadiendo el logaritmo del tamaño | 0.7336 | 0.4671 |
| Multinomial | Dividiendo el tamaño por la variedad | 0.73 | 0.46 |
| **Multinomial** | **Añadiendo superpop.txt** | **0.7379** | **0.4757** |
| Multinomial | Poniendo deportes.txt | 0.7364 | 0.4729 |

El mejor resultado obtenido se puede observar en rojo en la tabla anterior.

Para el problema de la variedad se toman como datos los siguientes valores:

- **Método**: método utilizado para entrenar el modelo
- **Train Control**: ajustes para el método a utilizar
- **Palabras vocabulario**: número de palabras más frecuentes
- **Palabras variedad**: número de palabras más frecuentes en cada una de las 7 variedades posibles
- **Accuracy y Kappa: resultados**

Los resultados obtenidos son los siguientes:

| MÉTODO | TRAIN CONTROL | PALABRAS VOCABULARIO | PALABRAS VARIEDAD | ACCURACY | KAPPA |
|---|---|---|---|---|---|
| SVMLinear | NONE | 1000 | 300 | 0.7829 | 0.7467 |
| SVMLinear | NONE | 500 | 300 | 0.7364 | 0.6925 |
| SVMLinear | NONE | 0 | 300 | 0.7029 | 0.6533 |
| **RandomForest** | **NONE** | **1000** | **300** | **0.8636** | **0.8408** |
| RandomForest | NONE | 1000 | 500 | 0.8536 | 0.8292 |
| Multinomial | CV | 1000 | 500 | 0.7893 | 0.7542 |

El mejor resultado obtenido es el indicado en rojo en la tabla anterior

## 5 Conclusiones y trabajo futuro

Como conclusiones destaco las siguientes:

- Para solucionar el problema del género es más influyente buscar posibles patrones que den respuesta al ¿cómo?, por ejemplo ¿cómo se expresa la mujer?, mientras que para el problema de la variedad es más influyente dar respuesta a preguntas relacionadas con el ¿qué?. Por ejemplo, ¿qué palabras son las más frecuentes en Colombia?

- El hecho de que hayan menos opciones en una clase a predecir no implica que por ello se obtengan mejores resultados y de forma más fácil. Para este caso se

obtienen mejores resultados y más fácilmente cuando se trata de solucionar el problema de la variedad, teniendo ésta muchas más posibles opciones que el problema del género, que sólo tiene dos posibles. Hombre o mujer.

- A nivel de resultados cabe destacar que el método de RandomForest ha sido el mejor de todos los utilizados con diferencia.

Como trabajo futuro, lo desglosaría en los 2 problemas a resolver:

Para el problema del género:

- Detección de los emoticonos. Creemos que las mujeres los utilizan más y por tanto podría obtenerse mejoras.
- Etiquetar cada palabra con algún etiquetador. De esta forma sabríamos que es un adjetivo, un pronombre y sobre todo que es un verbo y si es presente o pasado. Los verbos en presente son más utilizados por las mujeres y esta sería una forma de detectarlos.

Para el caso de la variedad:

- Etiquetar cada palabra con algún etiquetador.