Trabajo asignatura Text Mining en Social Media. Master Big Data Analytics

Jose Joaquín Rodríguez García
yhoatsu@gmail.com

Abstract

Siendo un grupo de seis miembros lo primero que hicimos fue debatir en grupo con los posibles enfoques al trabajo. A raíz de esa reunión nos dividimos el trabajo y por parejas o de forma individual nos repartimos las tareas. Al final reunimos los resultados y los reflejamos en la presentación final. Aún teniendo ya un script para trabajar con R también optamos usar Python para aquellas tareas que nos fue posible hacerlo en R.

1. Introducción

El problema a abordar es clasificar miles de Twitts teniendo en cuenta únicamente el texto presente presente en cada Twitt. La clasificación es por un lado determinar el sexo de la persona que ha escrito el Twitt y por otro determinar la variedad lingüística que ha sido usada para escribir el Twitt. Nosotros hemos abordado ambos problemas obteniendo muy buenos resultados.

2. Dataset

En primer lugar lo que tenemos son dos carpetas con archivos *xml* que denotan a autores que han escrito en Twitter al menos cien mensajes, de estos mensajes se han seleccionado cien, y esto es justamente lo que hay en cada archivo, el texto de cada uno de esos Twitts. Las carpetas tienen como nombre *training* y *test*, conteniendo cada una 2800 y 1400 autores respectivamente. Esta será la información que usaremos en el trabajo, la primera carpeta *training* contiene los datos que usaremos para generar el algoritmo que prediga las categorías de clasificación y en la carpeta *test* encontramos los datos que usaremos para comprobar la precisión de las predicciones.

Dado que usaremos el dataset de *training* vamos a hacer un estudio exploratorio del dataset. Como lo que tenemos son twitts, es decir texto plano y *emoticonos*, una visualización rápida de esta información es por ejemplo llevar a cabo un conteo de las palabras más usadas para cada categoría de clasificación posible. Para ello lo que se ha llevado a cabo han sido dos gráficos distintos tanto para genero como para variedad, donde por un lado se representan las cien palabras más frecuentes, de modo que el tamaño de las palabras de la imagen refleja el número de apariciones, a mayor tamaño mayor número de apariciones. Por otro lado se ha representado un diagrama de caja donde se puede apreciar la distribución del número de aparición de cada palabra. En la figura 1 se muestra el gráfico de palabras y en la figura 2 el diagrama de caja para la clasificación por genero. De forma análoga para la variedad lingüistica en la figura 3 se muestra el gráfico de palabras para cada variedad y en la figura 4 se puede observar el diagrama de caja para la variedad lingüística.

Figura 1: A la izquierda se muestran las cien palabras más frecuentes en hombre y a la derecha las cien más frecuentes en mujeres.

3. Propuesta del alumno

En conjunto hemos hecho varias propuestas:

- **N-gramas**: esta propuesta se hizo tanto para determinar el género como la variedad lingüística. Lo que hicimos fue modificar el

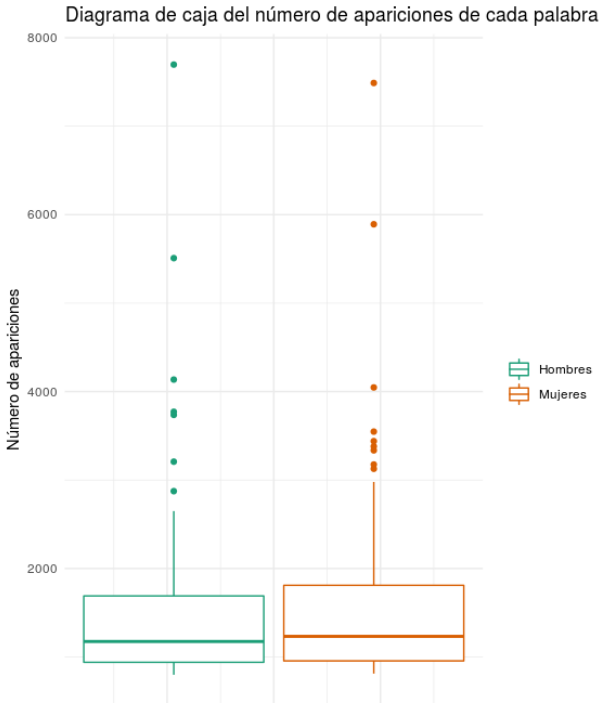

Figura 2: Diagrama de cajas por genero para las cien palabras más frecuentes.

script dado en R para generar una bolsa con los n-gramas más frecuentes. Debido a la cantidad de información en R no conseguimos llegar a un resultado ya que no terminaba de ejecutarse. La hipótesis fue que habiamos visto en una asignatura del máster como el aplicar 2-gramas y 3-gramas mejoraba el resultado usando solamente bolsa de palabras.

- **Terminaciones de cada palabra**: esta propuesta se hizo para el problema de determinación del genero. Lo que hicimos fue implementar un script en R para quedarse con los sufijos más discriminantes, como las terminaciones -o -a, pero al igual que con los n-gramas no conseguimos llegar a un resultado en R. La hipótesis es que como en el idioma español existen sufijos distinto para reflejar el genero masculino o femenino, bajo el supuesto de que un autor al referirse a sí mismo usará las palabras de acuerdo a su sexo, entonces el usar la frecuencia de aparición de los sufijos seleccionados podría discriminar el sexo del autor del twitt.

- **Longitud del Twitt**: esta propuesta se hizo para el problema de determinación del genero. La idea es determinar el genero únicamente usando la variable que determina el número de letras, símbolos de puntuación, emoticonos y espacios en blanco. La hipótesis está basada en estudios psicológicos que muestran que en términos generales las mujeres hablan más que los hombres, trasladando esta hipótesis a nuestro caso, supusimos que las mujeres escribirían twitts más largos.

- **Bolsa de palabras con las n palabras más frecuentes**: esta propuesta se hizo tanto para determinar el género como la variedad lingüistica. Usamos la baseline usando el algoritmo de clasificación *RandomForest* y variando el número n de palabras. La hipótesis fue que dado que la bolsa de palabras ya aportaba buenos resultados usando el algoritmo *Support Vector Machines* (SVM) y con la experiencia de los casos vistos en el máster, de que el algoritmo *RandomForest* proporciona mejores resultados, decidimos cambiar el algoritmo.

- **Bolsa de palabras determinadas previamente en base a temáticas**: esta propuesta se hizo para el problema de determinación del genero. Se buscaron palabras de temas como deportes, creyendo que esta bolsa discriminaría a los twitts escritos por hombres, y otros como moda, bajo la hipótesis de que discriminaría mucho mejor a los twitts escritos por mujeres. Al ser una elección subjetiva de temas para hombre y mujeres, al igual que es complejo elegir unas palabras que representen estos temas, la aproximación fue muy superficial. La hipótesis fue que los hombres y mujeres hablan de temas diferentes concretos.

- **Term frequency – Inverse document frequency (Tf-idf)**: esta propuesta se hizo para el problema de determinación del genero así como para la variedad lingüistica. El método es similar al de bolsa de palabras, salvo por la diferencia de que en este se tiene en cuenta también la frecuencia de ocurrencia del término en la colección de documentos, así que lo que tenemos es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Este método lo implementamos y aplicamos en Python, ya que nos resultó más fácil al haber trabajado antes con este método en esa plataforma. La

hipótesis estuvo basaba en que hay muchas palabras cuyo uso es común por ejemplo tanto en hombre como mujeres, entonces se trata de usar solo aquellas palabras frecuentes pero que sean lo más discriminantes posible.

4. Resultados experimentales

En esta sección no comentaremos todos los resultados obtenidos para no cargar en exceso el artículo con números. En caso de querer conocerlos se dejarán en la presentación disponible en el repositorio. Así, cabe destacar que el mejor resultado ha sido obtenido aplicando la propuesta *Term frequency – Inverse document frequency (Tf-idf)*, cuya precisión obtenida para la clasificación del genero ha sido de un 93 % y un 74 % para la clasificación de la variedad lingüística.

5. Conclusiones y trabajo futuro

Los resultados nos muestran que la técnica *Tf-idf* es muy efectiva para la discriminación del genero del autor de un twitt, ya que se ha mejorado consideramente el mejor resultado de la *baseline*, el cual es del 75 %, al igual que es efectiva para la determinación de la variedad lingüística, ya que el resultado de la *baseline* es 0,7275 %.

Como trabajo futuro y dado el buen resultado obtenido con *Tf-idf*, sería implementar esta técnica para n-gramas y/o con la propuesta *Bolsa de palabras determinadas previamente en base a temáticas*, *Terminaciones de cada palabra*, *Longitud del Twitt*. Ya que estamos convencidos de que la combinación una o varias técnicas con *Tf-idf* proporcionará unos resultados aún mejores.

References

Luís Felip Hurtado Oliver. 2018. *Material de la asignatura del Máster en Big Data Analytics: Procesamiento del Lenguaje Natural*

Paolo Rosso., Francisco Manuel Rangel Pardo 2018. *Material de la asignatura del Máster en Big Data Analytics: Text Mining en Social Media*

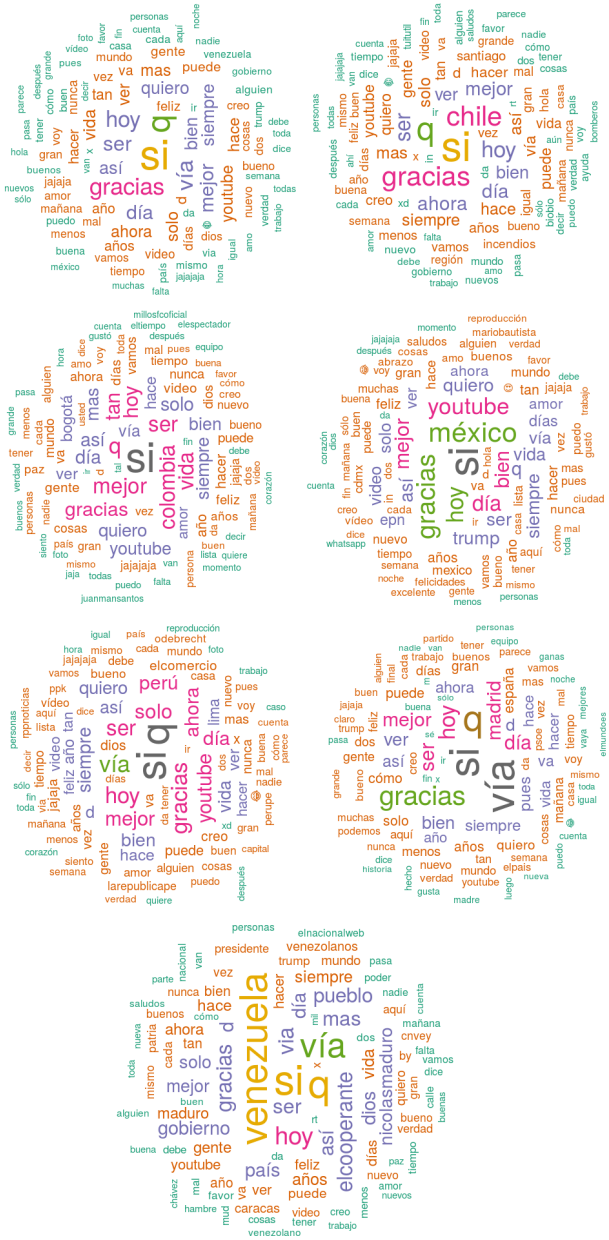

Figura 3: Siguiendo el orden de izquierda a derecha y de arriba a abajo las variedades lingüística son: Argentina, Chile, Colombia, México, Perú, España, Venezuela.

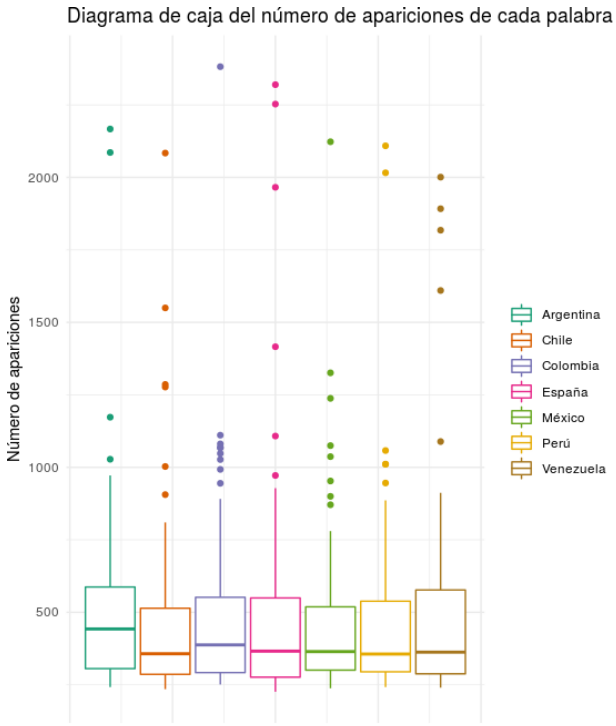

Figura 4: Diagrama de cajas por variedad lingüística para las cien palabras más frecuentes.