Estudio del género en Twitter. Asignatura Text Mining en Social Media. Master Big Data

David Pérez Lázaro
davix1603@gmail.com

Abstract

Esta tarea ha sido centrada en el estudio y modelado del género de los autores en Twitter. Haciendo uso de una extensa colección de tweets, se han extraído un conjunto de características diferenciadoras del género: emojis, preposiciones, palabras usadas, media de palabras por tweet... Con Python como lenguaje, y haciendo uso de diferentes técnicas, algunas de ellas avanzadas (Machine Learning), el objetivo ha sido predecir en base a un tweet el género de la persona que lo escribe, con el mayor acierto posible.
Uno de los principales retos encontrados ha sido la mala escritura presente en general en las redes sociales, y más concretamente, en Twitter. Las faltas gramaticales, las repeticiones de caracteres, faltas de signos de puntuacin, etc, han ocasionado numerosos problemas a la hora de realizar el procesamiento del texto y la extracción de las características buscadas.
Las predicciones se han realizado con modelos de Machine Learning, realizando un ajuste de los parámetros de control en busca de un mejor resultado. Entre los modelos usados para evaluar encontramos los siguientes: SVM (Support Vector Machine), Árboles de Decisión, Random Forest y Redes Neuronales.

1 Introducción

El problema de Author Profiling se basa en la existencia de diferencias, tanto en el lenguaje como en el estilo, en los textos escritos debidas, entre otras cosas, en la edad, género o variedad del lenguaje. Es decir, existen diferencias entre lo que puede escribir un niño de 10 años, al igual que entre un hombre y una mujer o una persona de España y otra de Venezuela. Haciendo uso de este profundo conocimiento del lenguaje, se puede predecir, con un grado de acierto bastante elevado, las características básicas de la persona detrás de un texto.

Como ya se ha comentado, en esta tarea nos hemos centrado en trabajar con el problema del género, es decir, predecir si la persona que ha escrito un tweet es un hombre o una mujer. La lingüistica nos proporciona algunas características diferenciadoras entre ambos géneros. Los hombres usan más determinantes y adjetivos, mientras que las mujeres hacen lo mismo con los pronombres, el uso de la negación o el uso de los verbos en presente.

Un aspecto que no se puede perder de vista tampoco es la plataforma que se está analizando, ya que dependiendo de la misma, una misma persona escribe de manera diferente. ¿ O acaso escribes de la misma forma un artículo para un periódico, un email a tu superior, un post en tu blog personal, un tweet o un mensaje en el chat con tus amigos ?

2 Dataset

Para esta tarea se utiliza un subconjunto reducido de todo el corpus utilizado en PAN-AP'17. En concreto, este subconjunto reduce el texto a sólo un idioma, el español, además del número total de muestras. Como es habitual, los datos están separados en dos, training y test.

Para los datos de training, contamos con 2800 autores diferentes, con un total de 100 tweets para cada uno de ellos que no son retweets. Excepto para un caso puntual, correspondiente al archivo 7d8903efb5d6a919ba15c49b24f446c8.xml, que cuenta únicamente con 99 tweets. Un archivo training/truth.txt nos permitirá el entrenamiento del modelo.

Por otro lado, en la parte de test tenemos un conjunto más reducido, con la mitad de autores, y por consiguiente, de tweets. 1400 autores en total, no repetidos con los datos de training. El número de

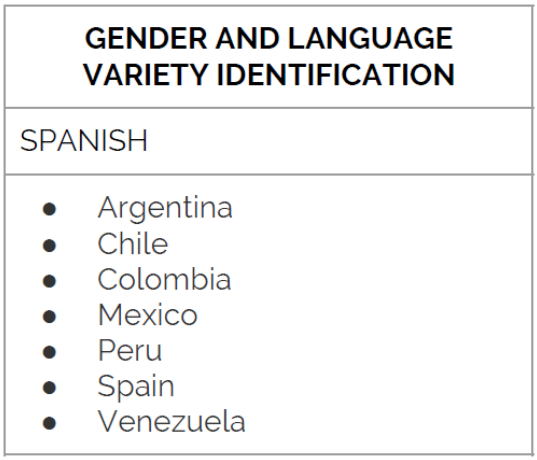

Figure 1: Distribución del dataset

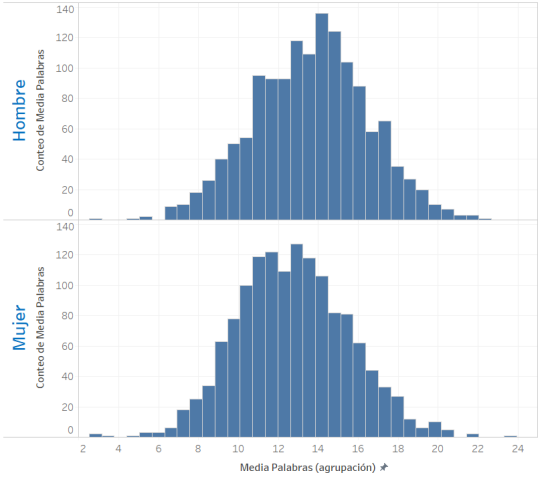

Figure 2: Media de palabras por tweet.

autores de cada género está perfectamente equilibrado, así como el reparto de los autores entre las diferentes variedades del lenguaje, a pesar de que no vayamos a utilizar esta parte para una primera aproximación a la tarea presente. Esta distribución del dataset queda perfectamente visualizada en la Figura 1.

A continuación, se analizarán algunas métricas del corpus para tener una idea de la importancia de las diferentes características que se ha pensado que pueden ser las más interesantes. Se ha tenido en cuenta la plataforma donante.

3 Propuesta del alumno

Para abordar este problema, el de la predicción de género, se han seguido dos enfoques o caminos distintos, que finalmente han sido unidos para el modelo final. Se trata de la extracción de características importantes para el contexto, y la utilización de bolsas de palabras.

3.1 Características

Lo primero que nos vino a la cabeza cuando estudiamos las características que podrían ser más útiles fue la pregunta ¿ De verdad escriben más las mujeres que los hombres ? Pues según dicen los datos, a partir de un gráfico de distribución en la Figura 2, es justo al revés, aunque no por mucha diferencia. A partir de los datos de training, obtenemos que la media de palabras por tweet es casi un punto superior en los hombres que en las mujeres. 13'63 frente a 12'75.

Analizando el contenido de Twitter, también llama la atención el elevado número de emojis utilizado. Hemos creído relavante seleccionarlo

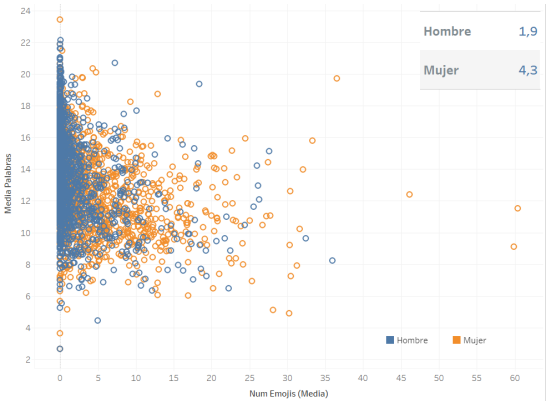

Figure 3: Uso de emojis.

como una característica para incluir al modelo, ya que como se observa en la Figura 3, hay diferencias entre los dos géneros en el uso de los emojis, usando las mujeres más del doble que los hombres.

Por otro lado, al principio se comentaba que las preposiciones también diferían en su uso entre mujeres y hombres. Según la Figura 4, esto es cierto. Los hombres usan más número de preposiciones de media en los tweets.

Por último, hemos querido ver si había diferencias en el uso de adjetivos masculinos y femeninos por parte de cada uno de los dos géneros. ¿ Usan las mujeres más adjetivos femeninos que los hombres, y viceversa ? Según se observa en la Figura 5, hay leves diferencias entre los géneros, pero por regla general se utilizan más adjetivos masculinos que femeninos, por ambos géneros. Dentro de cada uno de ellos, se usan ligeramente más los propios que los ajenos.

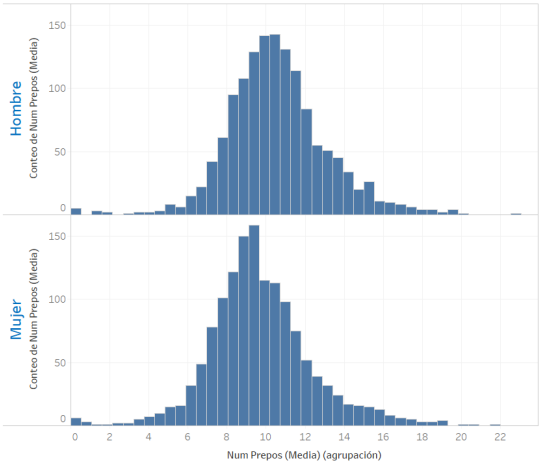

Figure 4: Uso de preposiciones.

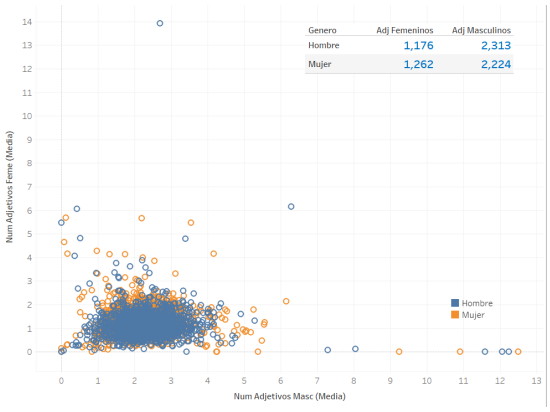

Figure 5: Uso de adjetivos por género.

3.2 Bolsa de palabras

Además de las características anteriormente vistas, se ha completado y enriquezido el modelo con bolsas de palabras para cada uno de los dos géneros. Para esta tarea, se ha optado por la creación de bolsas de palabras por conteo simple de las mismas (TF), dejando para un futuro otra aproximación, potencialmente mejor, con conteo inverso (TF-IDF).

Como se verá en la siguiente sección de resultados, se han realizado pruebas del modelo con diferentes tamaños de las bolsas de palabras, para ver diferencias en las precisiones obtenidas por los modelos.

4 Resultados experimentales

En esta sección se revisan los resultados de los cuatro modelos aplicados a este problema, así como diferentes métricas de tiempo. Los modelos analizados son Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) y Neural Net

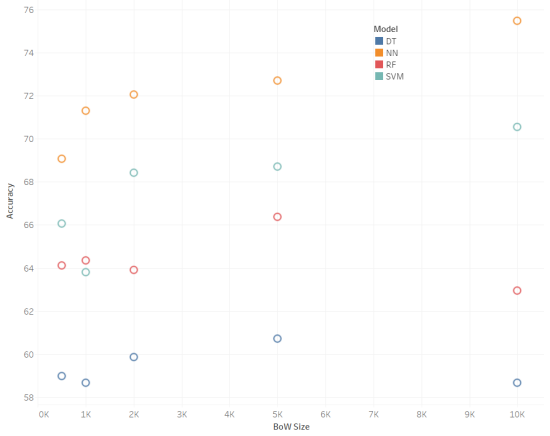

Figure 6: Accuracy vs BoW Size.

(NN). Se han realizado 5 ejecuciones diferentes de todos los modelos, variando el tamaño de la bolsa de palabras utilizada. Previamente se había realizado un ajuste rápido de los parámetros de los modelos, para quedarse con los más óptimos.

Los tamaños de la bolsa de palabras utilizada, para cada uno de los dos géneros son: 500, 1000, 2000, 5000 y 10000. De cara a evaluar el rendimiento de los modelos, nos fijaremos en los siguientes parámetros:

- Tamaño de la bolsa de palabras.
- Tiempo de procesamiento de los datos. Extracción de características.
- Tiempo de entrenamiento y aplicació de los modelos.
- Precisión alcanzada.

Según los datos obtenidos, mostrados en la Figura 6, el tamaño de la bolsa de palabras utilizada afecta a la precisión de los modelos. Cuanto más grande, mejor precisión dan los modelos de manera generalizada, hasta un punto de inflexión situado sobre los 5000, a partir del cual ya no todos los modelos mejoran. Es posible que los que mejoran experimenten un ligero overfitting.

Sin embargo, no todo son ventajas, ya que ese incremento de tamaño tiene un coste en forma de tiempo de procesamiento de los datos, extracción de características empleadas y entrenamiento y aplicación de los modelos. Como se puede observar en la Figura 7 y en la Figura 8, tanto los tiempos de ejecución de los modelos como los de extracción de las características utilizadas se incrementan conforme se aumenta el tamaño de la bolsa de palabras.

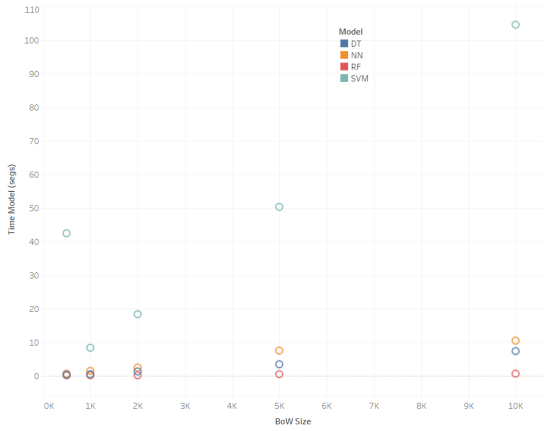

Figure 7: Accuracy vs Model Execution Time.

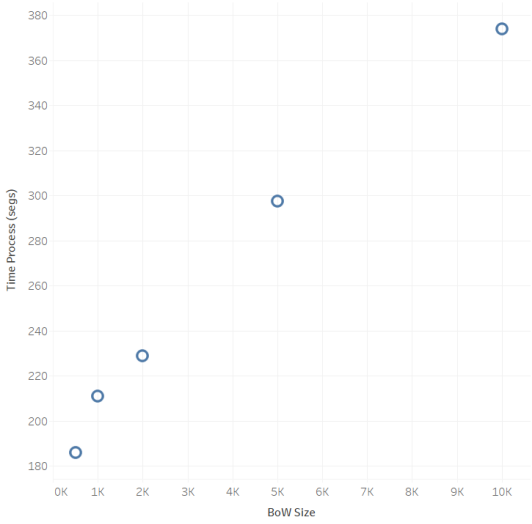

Figure 8: Accuracy vs Feature Extraction Time.

Habrá que determinar en qué punto se puede parar, porque no merezca la pena la ligera mejoría de los modelos, sacrificando los tiempos. A excepción de la SVM, los tiempos de entrenamiento y aplicación del resto de modelos son muy bajos, por lo que no debería suponer un problema esta parte. Hay que tener más encuenta los tiempos de extracción de las características deseadas, ya que es mucho mayor. El procesamiento de 280.000 tweets, utilizando una bolsa de palabras para cada género de 5000 palabras, tiene un coste de 297 segundos, prácticamente 5 minutos. Mientras que esa cantidad de tweets se genere cada más tiempo, no hay problema.

Las mejores precisiones las han obtenidos modelos de redes neuronales, con bolsas de palabras de 5000 y 10000 palabras, para unas precisiones del **72,7%** y del **75,48%**, respectivamente.

5 Conclusiones y trabajo futuro

A lo largo de esta tarea se ha abordado el problema de la predicción del género en Twitter. No obstante, por limitaciones de tiempo, no se ha podido abordar el problema en su totalidad. Entre las tareas que se quedan pendientes para una futura iteración se encuentran:

- Aplicar los modelos entrenados a los datos de testing.
- Mejorar el procesamiento de los tweets, eliminando signos de puntuación, caracteres repetidos... Se han escapado algunos.
- Utilizar un análisis sintáctico de los tweets para detectar los adjetivos y su género, en lugar de una lista limitada y una separación en función a la terminación de los mismos.
- Probar más modelos, buscando una mejor aproximación, más rápida.
- Implementar un TF-IDF en lugar de un simple TF en la creación de las bolsas de palabras más repetidas en cada género. Creemos que puede marcar una gran diferencia.
- Abordar el problema de la identificación de la variedad del lenguaje. Al margen de los mismas características ya empleadas, creemos que el uso de los dominios de las webs acortadas en los enlaces puede ser realmente interesante, aunque potencialmente costoso en tiempo.

Finalmente, estamos muy satisfechos con los resultados obtenidos, superiores al **70%**. Y creemos que aún hay margen de mejora.

References

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