Asignatura Text Mining II. Master Big Data Analytics

Patricia Roca Saura
patrirocasaura@gmail.com

Abstract

Para la resolución del problema primero se hizo una análisis del dataset de training para comprobar que respecto al número de usuarios por país estaba balanceado, es decir, teníamos el mismo número de usuarios para cada país, mientras que para la división por sexo vimos que para todos los países teníamos mas hombres que mujeres. De todo la información disponible para cada usuario de test únicamente se utilizo el texto de los tweets pensando que en ampliaciones futuras se podría utilizar la otra información disponible para mejorar los resultados. Debido a la distinta naturaleza del problema de detectar el país del autor respecto al problema de detectar el sexo del autor se decidió tratar los dos problemas por separado. Para detectar el país se utilizaron las palabras utilizadas, mientras que para distinguir el sexo se intento utilizar el tipo de palabras utilizadas, ya que tal y como se vio en clase el uso de preposiciones, sustantivos y pronombres se puede utilizar para distinguir el sexo del autor.

1. Introducción

El problema planteado consiste en entrenar un modelo para detectar el país y el sexo de un autor a partir de la información que se puede extraer de él en twitter.

Para entrenar el modelo se dispone de información de autores de 7 países de habla hispana.

2. Dataset

En el dataset facilitado, disponemos de autores de 7 países, de cada país tenemos 389 autores con lo siguiente distribución entre hombres y mujeres y desconocidos.

El dataset esta bien balanceado para el problema de distinguir el país pero muy mal balanceado para el problema de distinguir el sexo ya que muchos de los autores tienen sexo desconocido y de los que conocemos el sexo la mayoría son hombres.

3. Propuesta del alumno

Para acotar problema, utilizaremos solo el texto de los tweets y no tendremos en cuenta ninguna información más de la disponible. Separemos el problema, entrenaremos un modelo para detectar el país y otro para detectar el sexo.

Para distinguir el país lo primero que haremos es buscar una bolsa de palabras para generar un vector de características de cada autor para entrenar el modelo. Decidimos centrarnos en las palabras que se utilizan ya que hay muchas palabras que se utilizan en unos países y nunca se utilizan en otro. Además los temas de actualidad, que son los que se suelen tratar en twitter son distintos en unos países que en otros por lo que algunas palabras claves de estos temas nos permitirán distinguir entre unos y otros. Nuestro objetivo es buscar palabras que se usen bastante y que aparezcan en unos países y no en los otros. Primero eliminamos las menciones de usuarios, quitamos las repeticiones de mas de 3 letras y pasamos todo a minúsculas. Además eliminamos las llamadas palabras vacías utilizando el modulo stopwords de las nltk, ya que no creemos que sean relevantes para distinguir entre países y se usan mucho por lo que si cogemos las palabras más usadas nos iban a aparecer en nuestra bolsa de palabras. Después buscamos de las 500 palabras más usadas de cada país cuales no se usan en los otros países, con esto obtenemos una lista de 461. A estas 461 palabras le añadimos las 500 palabras más usadas en general, sin distinguir entre países y con esto construimos nuestra bolsa de 961 palabras. Para obtener el vector de características de cada autor utilizamos la representación binaria, cada componente es un 0

o un 1 dependiendo si el autor utiliza o no esa palabra en sus tweets. Luego entrenamos un SVM con el conjunto de entrenamiento y comprobamos los resultados para los autores de test. Para estos obtenemos también el vector de características con la misma bolsa de palabras y obtenemos el país pasando este vector de características por el modelo entrenado.

Para distinguir el sexo vamos a utilizar, además de la palabras que usa un sexo y no el otro, el tipo de palabras que usamos ya que hay varios estudios que indican que la distinción de hombres y mujeres es más fácil detectarla por la forma de escribir y no por las palabras utilizadas. Para cada autor calculamos un vector de características que contiene el numero de determinantes, el numero de adjetivos, el número de veces que aparece la palabra de, el número de pronombres, el numero de veces que aparece la palabra para, el número de veces que aparece la palabra con, el numero de verbos y el numero de veces que aparece la palabra no en los tweets de ese autor, todo ello normalizado por el número total de palabras de los tweets de ese autor y además una bits que indican si aparece o no una palabra de la bolsa de palabras calculada buscando que palabras usa un sexo y no otro.

Para determinar el tipo de palabra, si es un pronombre, determinante, verbo, etc. se utiliza un corpus de palabras etiquetas del nltk y para el caso de distinguir el sexo no eliminamos ninguna palabra.

Con los vectores de características obtenidos se entrena un SVM.

4. Resultados experimentales

El modelo entrenado para distinguir el país obtiene un acierto del 90 %.

En el caso de detectar el sexo se encontraron muchas dificultades para determinar el tipo de una palabra por lo que en trabajos futuros se podría intentar mejorar los resultados mejorando el método de etiquetado de las palabras.

El modelo entrenado obtiene un porcentaje de acierto del 50 % para la detección del sexo.

Como curiosidad a continuación se muestran algunas de las las palabras que aparecen en los tweets de un país y no en los otros.

Figura 1: España

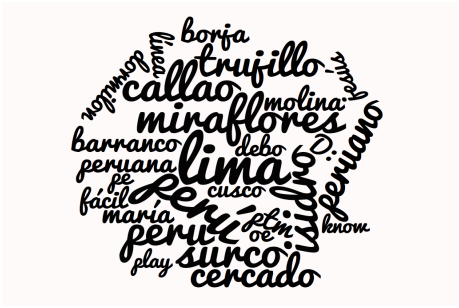

Figura 2: Perú

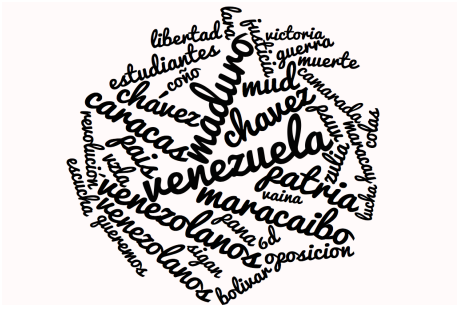

Figura 3: Venezuela

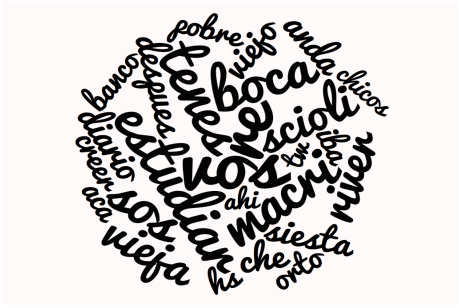

Figura 4: Argentina

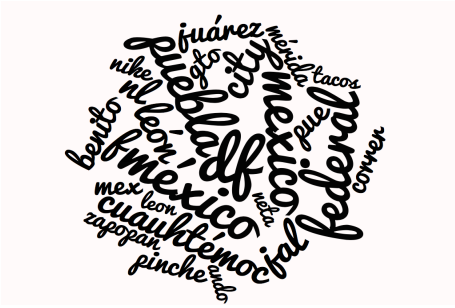

Figura 5: México

Figura 6: Colombia

Figura 7: Chile

5. Conclusiones y trabajo futuro

Los resultados son buenos pero habría que trabajar más en depurar el código para hacerlo de forma más eficiente ya que el código realizado es un código de test para probar si las ideas planteadas daban buenos resultados.

También se podría profundizar más en el etiquetado de los tipos de palabras para discriminar entre hombres y mujeres utilizando librerías como freeling.

References

Rosso, Paolo 2016. *Author profiling in social media* UPV

Pla, Ferran 2016. *Analisis sentimiento* UPV