Trabajo Text Mining en Social Media. Master Big Data

Eusebio Soriano Benegas
eusesoriano@gmail.com

Abstract

En este documento se expondrá brevemente el concepto de Author Profiling y se presentará una solución a la detección de género del autor de un determinado tuit. A partir de dos datasets y un documento de texto que relaciona tuits con autores aplicaremos diversas técnicas de minado de texto con el fin de obtener el mayor número de aciertos posibles. Posteriormente se analizarán los resultados y se expondrán soluciones alternativas para un trabajo futuro.

1 Introducción

Author Porfiling se trata de una técnica que consiste en el análisis de texto con el fin de obtener la mayor información relevante posible de su autor. De esta forma, mediante estos datos los cuales suelen basarse en el estilo y el contenido gramatical generalmente, es posible caracterizar al autor de un texto en mayor o menor medida.

El objetivo de este trabajo consistirá en identificar con la mayor precisión y menor tiempo posible el género del autor de un tuit determinado. Cabe destacar que uno de los criterios a tener en cuenta en la evaluación es el tiempo de ejecución del proceso. A continuación, se analizará en primer lugar el dataset, tras comprender su estructura, se planteará una serie de técnicas a emplear las cuales se aplicarán a los datos con el fin de obtener el mayor porcentaje de aciertos.

2 Dataset

Básicamente, el dataset está formado por dos corpus. Uno de test y otro de training. Por una parte el corpus de train está formado por 2800 ficheros xml los cuales estan formados a su vez de 100 tuits. Por lo que se refiere al test, este estará formado por 1400 ficheros xml. Cada fichero xml tiene como nombre el id del tuit. Ademas, tenemos un archivo de texto el cual relaciona el identificador del tuit con el género de su autor y la variedad lingüística del español empleada. En nuestro caso únicamente nos centraremos en la información proporcionada en cuanto al género del autor.

3 Propuesta del alumno

En este apartado expondremos las técnicas de análisis y preparación de los datos propuestos. Una de las ténicas a las que podemos recurrir como estudio inicial sería el uso de bag of words. Básicamente se trata de contar la frecuencia de aparición de un grupo determinado de palabras. Para ello, primero cargamos el corpus de tuits que queramos procesar, de esta forma podremos aplicarle las transformaciones que se consideremos necesarias.
Después sobre el corpus ya procesado generamos un vocabulario de las palabras mas frecuentes que aparecen en un grupo de tuits.

Seguidamente, generaremos una bolsa de palabras del mismo tamaño que el vocabulario. El proceso que hemos seguido ha consistido en ir probando diferentes configuraciones de las funciones del código empleado hasta encontrar un equilibrio en los parámetros de modo que se ajuste eficientemente a nuestro estudio.

Posteriormente, eliminamos los elementos del texto que carecían de importancia en primera instancia como los signos de puntuación, tildes, stop words como determinantes y artículos. Además se eliminaron ciertas palabras extremadamente genéricas las cuales no supondrían una mejora en la decisión final y se transformarán todas las mayúsculas a minúsculas para no hacer distinción entre palabras iguales. Como modelo se ha empleado Support Vector Machines con un Cross

Validation de 4 capas.

4 Resultados experimentales

Como se ha expuesto anteriormente, en primer lugar se empleó el baseline con un vocabulario de 100 palabras y un Bag of Words de 10 palabras. Sin embargo, en algunos entornos de desarrollo del equipo los documentos no eran leídos de forma correcta por lo que se fuerza una conversión a UTF-8.

Analizando el código se decidió optimizarlo para que la fase de generación de vocabulario sea un poco más rápida. Por otra parte, en la implementación actual las transformaciones al texto se realizan en el momento de leerlo. Esto lo hacemos para no volver a recorrer el corpus completo por cada transformación que queramos hacer.

En las funciones de nuestro código que requieren el calculo de frecuencias, se han calculado las frecuencias relativas, ya que reflejara un mejor comportamiento de la muestra (siguen una densidad explicita).

En este primer caso no se aplicó ninguna transformación. Como modelo de clasificación se han empleado Support Vector Machines. Los resultados obtenidos son los siguientes:

```
Accuracy : 65.29
Kappa : 0.3057
```

Seguidamente, y como segunda prueba, se implementó una función para obtener los términos mas frecuentes usados por hombres o por mujeres, de esta forma podríamos comparar que términos usan mas unos u otros.

Haciendo este muestreo decidimos borrar del corpus los términos que mas se repitiesen tanto en hombres como en mujeres. Estas palabras las añadimos a nuestra lista de palabras a eliminar. Tras eliminar los acentos y las palabras más comunes de ambos géneros, manteniendo una Bag of Words de 10 palabras conseguimos unos resultados de:

```
Accuracy : 67,36
Kappa : 0.3471
```

En nuestro ultimo y mejor resultado seguimos borramos también las palabras muestreadas mas comunes en hombres y mujeres como en el apartado anterior, Se decidió no eliminar los acentos ni convertir las palabras a minúsculas de esta forma se podrá mantener el ”tipo” de escritura de distintos autores (hombres o mujeres).Se eliminaron todos los signos de puntuación, números, espacios en blanco y stop words del español.Los resultados del test fueron:

```
Accuracy 70,00
Kappa 0,40
```

5 Conclusiones y trabajo futuro

Como hemos podido observar, los resultados obtenidos pueden ser mejorados en gran medida, si bien el análisis de texto para Author Profiling puede llegar a ser bastante complejo, podríamos aplicar otras técnicas efectivas y obtener una mayor ión. Entre estas técnicas tendríamos el estudio de los emoticonos utilizados como información extra al texto del tuit.

Al fin y al cabo el objetivo del trabajo es extraer toda la información de un tuit determinado, de esta forma es posible analizar los nicknames de las personas etiquetadas en ese tuit de forma que pueda ayudar a caracterizar mejor al autor.

Otra técnica similar sería el análisis de los enlaces a fotos, páginas web. No obstante, será interesante profundizar en el análisis de las palabras o grupos de palabras utilizadas, además tener en cuenta los temas o topics que son tratados en el tuit.

Además, podríamos generar un contador con el numero de repeticiones de palabras plenamente de mujeres y otro contador de palabras frecuentes de hombres. Es decir a la bolsa de palabras añadiríamos dos parametros más, uno indicaría la abundancia de palabras que tiene de mujeres y el otro de hombres, esto podría ayudar al modelo.

Tras una buena preparación de las variables y un clasificador de palabras según a temática o el ámbito semántico podríamos obtener una mejora sustancial en cuanto al análisis de género del autor.