Text Mining en Social Media. Master Big Data

Salvador Villalba Marin
salvilma@gmail.com

Abstract

El presente artículo explica la solución propuesta a un ejercicio planteado en la asignatura Text Mining en Social Media de máster Big Data Analytics de la UPV. El objetivo del ejercicio es determinar el género y el país de origen de un autor a partir de un conjunto de sus tuits.

Este ejercicio de sacar conclusiones sobre los rasgos de una persona a partir de la expresión escrita se conoce como author profiling. Y se aborda aplicando técnicas de machine learning, en nuestro caso para inferir el género y el país de origen. Las predicciones de estas dos variables se han realizado con modelos difentes, es decir, no se ha hecho un modelo que prediga los dos resultados al mismo tiempo. Ya que en cada caso se han utilizado unas estrategias diferentes. Y en ambos casos la información utilizada para entrenar los modelos ha sido el texto escrito en los tuits.

En el presente documento se explican las diferentes estrategias utilizadas para inferir el género y el país de origen, así como la precisión de acierto que se ha obtenido. También se incluyen otros posibles planteamientos que no han sido aplicados pero sí se han considerado interesantes para explorar en un posible futuro trabajo de continuación del problema.

1. Introducción

El problema a resolver consiste en determinar el género y el país de origen del autor de un tuit basándose en el texto escrito. Para ello se nos proporcionan datos de 2.800 autores con 100 tuits cada uno, y de los que se conoce su género y su país de procedencia.

Con esta información se nos pide construir un modelo que dado un tuit haga una predicción del género y país del autor con un acierto superior a los siguientes:

- 66,43 % para el género
- 77,021 % para el país de origen

Como se explica en los siguientes apartados, se han utilizado técnicas de machine learning de aprendizaje supervisado. Esto significa que partimos de un conjunto de datos que entre otra información también contienen las variables que se quieren predecir. En nuestro caso disponemos del texto del tuit junto con el género y país del autor.

Adicionalmente se proporciona un conjunto de datos de test donde también son conocidos el género y el país del autor, y sobre los que se aplicará el modelo predictivo y se podrá determinar el acierto que proporciona el modelo.

Como parte del problema se pone a disposición del alumno código en R, con el objetivo de completarlo y hacer uso de funciones de tratamiento y procesamiento de los tuits que se dan implementadas.

2. Dataset

Los datos que se utilizarán para el ejercicio son un subconjunto de PAN-AP'17, que es un dataset que contiene gran cantidad de tuits de autores de diferentes países, y no solo en lengua española. Al descargar los datos se crean dos directorios:

- Training: con los tuits para entrenar el modelo
- Test: con los tuits para hacer la predicción aplicando el modelo y obtener el acierto a comparar.

Los tuits se obtienen en ficheros XML, teniendo un fichero por autor y siendo el nombre del fichero el identificador del mismo. De modo que cada

XML contiene los tuits de ese autor cuyo identificador es el propio nombre del fichero. En ambos directorios se dispone de un fichero de texto llamado *truth.txt* donde se indica para cada identificador de autor su género y país de origen.

La primera parte del trabajo con los datos consiste en prepararlos para su explotación, para ello el texto de los tuits almacenado en los ficheros XML es tratada en R en un proceso que estaría enmarcado en lo que se conoce como *data wrangling*, y que en nuestro caso consiste en eliminar tildes, pasar el texto a minúsculas, quitar signos de puntuación, eliminar números y espacios en blanco, y la posibilidad de eliminar las palabras del stopwords del español (conjunto de palabras que en principio no aportan información).

Respecto a la distribución de la información sobre las variables a inferir en el conjunto de training podemos decir que están homogéneamente distribuidos, los datos son los siguientes:

- Género

Hay 50 % hombres, 50 % mujeres. En concreto de los 2.800 autores, 1.400 son hombres y 1.400 mujeres.

- País de origen

Hay 7 países de habla española: Argentina, Chile, Colombia, Venezuela, México, Perú y España. Con 400 autores por cada uno de los países.

En el conjunto de test, las proporciones son las mismas. Hay 1400 autores, siendo 700 hombres y 700 mujeres. Y respecto al país de origen, hay 200 autores de cada país.

Además en ambos conjuntos, para cada país la mitad de los autores son hombres y la otra mitad mujeres.

3. Propuesta del alumno

Tal y como se ha comentado, para cada variable a predecir, sexo y país de origen, se hacen modelos diferentes en vez de hacer un modelo conjunto que dé una predicción con un resultado con los dos valores.

Para resolver el problema nos basamos en la implementación del código R que se proporciona, y que consiste en generar una bolsa de palabras partiendo de las utilizadas en los tuits. Además, en esta generación se aplican las técnicas de data wrangling indicadas en el apartado anterior. Esta bolsa de palabras contendrá 1000 elementos, que serán las palabras que aparezcan con mayor frecuencia en los tuits.

A continuación se describen las estrategias utilizadas para crear el modelo de predicción del género del autor (en el apartado resultados experimentales se detallan los resultados de acierto):

Estrategia 1

Generar una bolsa de palabras más frecuentes para mujeres y entrenar el modelo.

Estrategia 2

Como la estrategia anterior pero sólo para autores hombres

Estrategia 3

Utilizar las dos bolsas de palabras generadas anteriormente, y a partir de ellas generar dos nuevos vocabularios:

- Uno con palabras exclusivas del vocabulario de mujeres (palabras que no aparecen en el diccionario de hombres)

- Y otro con palabras exclusivas del vocabulario de hombres (no aparecen en el diccionario de mujeres)

Después se genera un nuevo vocabulario con la suma de estos dos vocabularios exclusivos. Y utilizamos este nuevo vocabulario para entrenar el modelo

Estrategia 4

Se genera una bolsa con la intersección de las bolsas generadas en las hipótesis 1 y 2, y se añade a la bolsa de la hipótesis 3.

Estrategia 5

Se toma la bolsa generada en la hipótesis 3 y se añaden una serie de palabras elegidas que en el equipo pensamos que son determinantes para identificar el género del autor

A continuación se describen las estrategias utilizadas para crear el modelo de predicción del país de origen del autor (en el apartado resultados experimentales se detallan los resultados de acierto):

Estrategia 1

Se genera un vocabulario propio de cada país, filtrando por el país de origen del autor. Así obtenemos 7 diccionarios con las palabras más frecuentes de cada país. Para cada uno de estos diccionarios se quitan las palabras que también aparecen los diccionarios de los otros 6 países, es decir, para cada país queda un diccionario en el que solo aparecen palabras exclusivas de los tuits de ese país y no aparecen en tuits de otros países. Después se crea un nuevo diccionario con la unión de estos 7 diccionarios y es el que se utiliza para el modelo

Estrategia 2

Se utiliza el mismo diccionario generado en la estrategia anterior pero en vez de utilizar como modelo de clasificación Support Vector Machine (el utilizado en todas las estrategias anteriores) se utiliza Random Forest con 50 árboles.

Estrategia 3

Generamos un vocabulario propio de cada país, filtrando por el país de origen del autor. Así obtenemos 7 diccionarios con las palabras más frecuentes de cada país. De cada diccionario nos quedamos con las 100 palabras más frecuentes y generamos un nuevo diccionario que tendrá 700 palabras.

4. Resultados experimentales

A continuación se muestran los resultados de *accuracy* para cada una de las estrategias descritas.

Predicción del género del autor:

Estrategia	Accuracy
1	0,67
2	0,68
3	0,6957
4	0,6907
5	**0,6979**

Predicción del país de origen del autor:

Estrategia	Accuracy
1	0,8393
2	**0,8507**
3	0,7271

5. Conclusiones y trabajo futuro

Los mejores resultados se han obtenido cuando se han utilizado palabras que eran exclusivas de cada uno de los grupos para entrenar los modelos, ya sea el género o el país de origen. Es decir, términos que solo aparecen cuando la variable a predecir tiene un valor concreto, por ejemplo, palabras que solo están en los tuits de los hombres. Al parecer esto favorece la discriminación en los algoritmos de clasificación y se obtiene mejor resultado.

También hemos observado que bolsas de palabras más pequeñas han dado mejor resultado que aquellas que tienen más número de términos, pese a lo que podría pensarse a priori. Sospechamos que en nuestro caso esto es debido a que cuando hemos utilizado menos palabras ha sido porque estás eran más discriminantes, tal y como se ha indicado anteriormente, es decir, es más importante una buena selección de términos para entrenar el modelo que utilizar bolsas de palabras más grandes sin un filtrado o estrategia previa.

Así, la continuación del trabajo iría encaminado a esta selección de palabras que tuvieran un poder discriminatorio más grande, por ejemplo, la palabra “boludo” es propia de los argentinos y rara vez aparecerá en autores de otra nacionalidad, o la palabra “ahorita” para los mexicanos. Y añadir estos términos a las bolsas de palabras que nos han dado mejores resultados, que han sido aquellas donde hemos generado diccionarios exclusivos de cada variedad, por ejemplo, las palabras utilizadas solo por mujeres.

Además, deberíamos probar con diferentes algoritmos de clasificación, no solo con SVM y Random Forest, es posibles que con redes neuronales, Naive Bayes, etc. Se obtuvieran mejores resultados.

References

https://stackoverflow.com/questions/tagged/r
http://www.statmethods.net