Text Mining en Social Media. Master Big Data

Natalia Cervelló Fenollar
nacerfe@fiv.upv.es

Abstract

Clasificación de tuits en base a dos criterios: Sexo y Variedad. Entendiendo como variedad el lenguaje en el que está escrito el tuit, dentro de las variantes del idioma español. La tarea se ha implementado en el lenguaje R.

1. Introducción

Se va a abordar el problema de Author Profiling. Este caso se basa en analizar los mensajes que los usuarios de Twitter van escribiendo en la plataforma de la red social, para sacar conclusiones dependiendo de los mensajes que hay en su perfil.

Los usuarios están tratados de forma anónima, pero identificados por un número. Con los datos del autor y lo tuits que se han recopilado de su perfil, podemos sacar información que nos puede servir para clasificar a ese usuario.

En nuestro caso vamos a usar dos factores para clasificar los tuits, y de esa forma identificar al autor:

- Sexo: Masculino o Femenino.

- Variedad: Entendiendo esta como la variedad lingüística en la que está escrito el tweet, para este problema solo distinguiremos entre variedades de dialectos dentro del español.

Para poder hacer una buena clasificación he identificar para cada factor, cual pertenece al autor, debemos hacer un estudio de palabras y expresiones que se usan dependiendo de cada factor.

2. Dataset

En el problema que nos compete se va a utilizar un subconjunto del dataset llamado PAN-AP'17, que contiene tweets en varios idiomas del mundo y los dialectos asociados a cada lenguaje.

El dataset se ha obtenido de Twitter, contiene miles de autores y para cada uno se han extraido cierta cantidad de tweets, del orden de cientos, por lo tanto estos tweets contienen una gran cantidad de temas e información.

Cuando se descarga el subconjunto del dataset, dentro encontramos una carpeta 'training'y otra llamada 'test', dentro de cada una se encuentran unos ficheros en formato XML (cada fichero es un autor), y dentro de los XML encontramos el autor y los tweets asociados a ellos.

En este subconjunto tenemos los siguientes datos:

- 300 autores, por género y variedad de lenguaje español.
  - 200 en training.
  - 100 en test.
- Cada autor tiene un conjunto de 100 tweets.
- En total tenemos 30.000 tweets.

3. Propuesta del alumno

Antes de centrarnos en un tipo de clase, sexo o variedad, hemos hecho una limpieza de datos sin tener en cuenta si estabamos viendo una clase u otra, es decir, la solución que se ha para construir el train ha sido la misma para las dos clasificaciones.

Se ha utilizado la opción basada en bolsa de palabras, usando las frecuencias de uso de palabras en los tuits.

En primer lugar y utilizando las funciones que se han proporcionado para recoger los tuits, limpiarlos y sacar el vocabolario de palabras, se ha añadido una nueva función, con la cual se han eliminado los acentos de todas las palabras.

Se ha usado la función de R:

chartr("áéíóú","aeiou",corpus.preprocessed)

Procedemos a hacer una elección de palabras que no nos aportan nada a nuestro modelo de train,

ya que no aportan información que ayude a la clasificación.

Lanzamos el generador de vocabulario, basado en las frecuencias de las palabras que aparecen en los tuits, en un principio se lanza con una bolsa de palabras de 1000, al salir el gráfico no podemos distinguir bien, cuales son las palabras, así que reducimos la bolsa de palabras a 100 para seleccionar a partir de esas palabras, cuales no nos sirven para el modelo.

Decidimos eliminar las siguientes palabras de nuestra bolsa:

swlist=c("youtube", "video", "q", "si", "x", "jajaja", "trump", "d", "mas")

Y volvemos a generar el vocabulario con 1000 palabras, eliminando esa lista de stopwords que se han seleccionado.

De esta lista podemos destacar:

- Palabras que son iguales en todas las variedades y sexos:
  - youtube
  - video
  - trump
  - mas (palabra con más frecuencia)
- Abreviaciones de escritura:
  - q
  - x
  - d
- Onomatopeyas comunes en todos los dialectos:
  - jajaja

Más tarde, para intentar mejorar la clasificación, se decide aumentar la bolsa de palabras de 1000 a 2000 palabras.

Con todos estos cambios se decide probar con distintos modelos de clasificación para nuestro conjunto de training, estos son los modelos por orden que se han utilizado:

- SVM (Support Vector Machine)
- Naive Bayes con Cross-Validation
- Random Forest

Probando estos métodos se han sacado unos modelos (uno para sexo y otro para variedad, en cada uno de ellos), con los que se ha hecho una predicción, para sacar un valor de áccuracy', que nos muestra el porcentaje de acierto en la clasificación.

A continuación se van a mostrar los resultados que se han ido obteniendo al aplicar todos los puntos descritos en este apartado.

4. Resultados experimentales

Mientras se iban haciendo cambios en los datos, y se iba investigando para limpiar el dataset, se han hecho distintas pruebas para ver como mejoraba el accuracy de la clasificación.

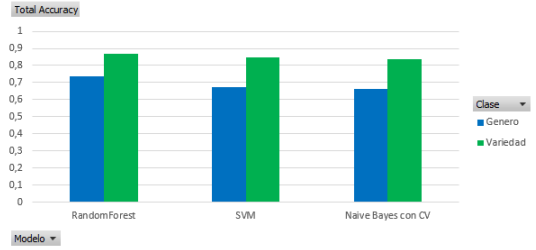
Al hacer todas las modificaciones que se han descrito en el apartado anterior, y probando los distintos modelos para la predicción, se han obtenido los siguientes valores de accuracy:

-El modelo que se ha usado ha sido

- Sexo (Gender): 73,43
- Varidad (Variety): 86,79

Tabla resultado con todos lo valores intermedios mientras se iban probando los cambios:

Porcentajes Accuracy por modelos		
Modelo	Género	Variedad
SVM	0,6743	0,845
Naive Bayes con CV	0,6653	0,835
Random forest	0,7343	0,8679



Como se puede observa, en la tabla y en el gráfico se muestran los valores de accuracy para cada modelo que hemos usado con nuestro modelo de training final, Random Forest, que es el que más nos ha dado tampoco se va mucho del Suport Vector Machine.

5. Conclusiones y trabajo futuro

La mayor conclusión a la que se ha llegado, es que con una buena investigación dependiendo de lo que vayamos a clasificar, extrayendo palabras clave y formando con ellas las bolsas de palabras, se puede mejorar el porcentaje de acierto.

En el problema que se ha descrito aquí, para nuestro caso, aumentar la bolsa de palabras nos ha

salido mejor, aunque al hacerlo se ha ralentizado el tiempo de cómputo, que bien, si no es algo urgente, puede ser aceptable.

En un futuro se ha pensado en hacer otras mejoras para la limpieza de datos, entre ellas:

- Hacer una investigación más profunda, separando los factores, es decir, crear una bolsa de palabras única para cada factor, género o variedad, para que el modelo sea más real y por lo tanto se clasifique mejor.
- Crear una nueva función, que en el procesado de los datos lo que hará es quitar todas las palabras que se formen con un solo carácter, con ello conseguimos eliminar las abreviaciones como: 'q', 'x', 'd', etc...
- Probar más algoritmos de predicción, para probar si otros, aparte de los ya probados, conseguimos una mejora en el accuracy.