Text Mining en Social Media. Master Big Data

Pablo Julio Sáez
pabjusae@gmail.com

Abstract

La tarea realizada para la asignatura de text mining consiste en la creación de algoritmo para resolver un problema de *author profiling*, el cual debe distinguir a partir de los datos dados entre el género del autor y, por otro lado, diferenciar la variedad del idioma. Los datos consisten en un *dataset* que contiene tweets de usuarios de distinto género e idioma.

El objetivo es abarcar todo el proceso de creación de un algoritmo clasificador, desde la limpieza y tratamiento de los tweets a la selección y entranamiento del algoritmo empleando el lenguaje de programación R.

1 Introducción

Author profiling es una rama de investigación dentro del área del procesamiento del lenguaje natural, y fuertemente relacionada con técnicas de aprendizaje automático o *machine learning*. Ésta estudia la posibilidad de extraer características sobre el autor de un texto, como por ejemplo la detección de su edad e incluso de su ideología política.

Esta rama es de gran utilidad en el ámbito empresarial y público, ya que permite clasificar el *feedback* recibido de clientes y usuarios acerca de productos o servicios. Del mismo modo, también es útil para conocer los perfiles de personas en determinadas redes, con el fin de mejorar campañas de *marketing*.

2 Dataset

El dataset proporcionado, conocido como PAN-AP17, consiste en un conjunto formado por tweets que provienen de miles de autores, de distinto género y lengua. Para el entrenamiento de los algoritmos emplearemos una pequeña parte del *dataset* nombrado anteriormente.

Ésta consiste en un *subset* con 300 autores por género y lenguaje, estando este último dividido en las siguientes variantes del español:

- Argentina
- Chile
- Colombia
- México
- Perú
- Venezuela
- España

Cada uno de los autores cuenta con 100 tweets. El entrenamiento de los algoritmos empleará 200 autores con sus respectivos tweets, y el resto serán empleados para el test y la verificación de su precisión. El *dataset* debe ser contruido a partir de unos ficheros en formato JSON, cada fichero corresponde a un autor con sus correspondientes tweets. Por otro lado, se proporcionan dos ficheros con los identificadores de cada autor etiquetados de la siguiente forma: id:::sexo:::variedad.

3 Propuesta del alumno

Nuestro equipo planteó el problema dividiendo el *dataset* en dos, uno para crear un algoritmo que identificara por género y otro para detectar la variedad del español. Antes de comenzar a entrenar el algoritmo hicimos un preprocesado de los datos. Éste consistió en:

- Eliminación de símbolos de puntuación y espacios en blanco. Para que no se detecten como palabras.

- Eliminación de números, ya que junto con los signos de puntuación, pensamos que no son de relevancia a la hora de identificar ni por género ni por variedad.

- Conversión a minúsculas de todas las palabras. De modo, que se detecten como las mismas palabras aquellas que presentan distinta forma.

- Eliminación de *stopwords*. Éstas son palabras que pensamos que no influyen a la hora de identificar al autor, del mismo modo que el uso de números o los signos de puntuación. En este caso se hizo un análisis visual de las palabras más comunes del *dataset* y se eliminaron en la mayoría de los casos palabras sin semántica y adverbios. Destacar que palabras frecuentes como servicios de internet (youtube, twitter ...) se dejaron, ya que en un principio fueron eliminadas pero al entrenar algoritmos, pero en pruebas posteriores se mostraron se relevantes a la hora de mejorar los resultados en el algoritmo encargado de identificar el género. Algunas de las *stopwords* eliminadas fueron: si, q, gracias, hoy, ser, día, mejor.

Una vez tenemos los datos finales ya limpios, comenzamos a entrenar algoritmos buscando mejorar el baseline dado. Los algoritmos usados para la deteccin fueron:

- Support vector machine.

- Cross validation.

Cada uno de los algoritmos se realizó con y sin stopwords, para ver su importancia a la hora del entrenamiento.

4 Resultados experimentales

Antes de comenzar a comentar los resultados de los algoritmos nombrados anteriormente, cabe destacar que el *baseline* a mejorar era del 0.7721 en el caso de la variedad, del cuál voy a comentar los algoritmos empleados a continuación.

Con el empleo de las máquinas de vectores de soporte o SVM el resultado fue de 0.7721 sin stopwords y de 0.7693 con éstas. Por otro lado, al emplear *cross validation* obtuvimos unos valores de 0.7721 y 0.7729, sin eliminar *stopwords* y eliminándolas respectivamente.

5 Conclusiones y trabajo futuro

Los resultados anteriores son similares al *baseline*, y en el caso del uso de *cross validation* junto con la eliminacin de las *stopwords* ligeramente superior.

Observamos que la eliminación de las *stopwords* no muestra un incremento de la precisión en el resultado, por lo que una posible mejora sería revisarlas, con el fin de obtener las más determinantes.

Por otro lado, la falta de tiempo debido a problemas técnicos con las librerías de R, nos limitó a emplear únicamente dos algoritmos a diferencia de los algoritmos empleados en la detección del género. Por ello, como trabajo futuro se podrían emplear los algoritmos que mejores resultados nos dieron en el caso del género, que fueron *random forest* y *mixture discriminant analysis*.

References

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