Text Mining en Social Media. Master Big Data 2016-2017

Silvia Aroca Ortega
silvia.aroca@gmail.com

Abstract

Este artículo expone la solución propuesta al problema planteado en la asignatura Text Mining en Social Media. La tarea ha consistido en la clasificación de tweets por variedad y sexo utilizando criterios basados en Author Profiling y técnicas de Machine Learning para deducir el género y el país de origen. A continuación veremos las estrategias utilizadas y otros posibles planteamientos a ser utilizados con el fin de mejorar el resultado.

1. Introducción

El objeto de estudio de esta tarea es averiguar el perfil del autor de un texto mediante los criterios de clasificación de Sexo (hombre/mujer) y Variedad lingüística (país) de cada autor. Nos enfrentamos a una tarea de *Author Profiling*, un campo de investigación muy reciente que permite identificar rasgos personales, como la edad, el sexo, la personalidad, el idioma nativo o la variedad regional del idioma de los autores de los textos a partir de su forma de escribir. Tarea que se dificulta cuando nos encontramos con textos cortos y espontáneos como los de las redes sociales e incluso si los idiomas son parecidos entre sí. Nuestra tarea se basa en trabajar sobre una base de datos dada, que recoge un corpus compuesto por un listado de usuarios de Twitter con hasta 30.000 tweets de cada uno, todos ellos pertenecientes a países de habla hispana (Argentina, Chile, Colombia, México, Perú, España y Venezuela). Evaluando dicho corpus mediante distintas técnicas basadas el estudio del lenguaje natural y aplicando tres métodos diferentes de algoritmos de clasificación, SVM (Support Vector Machine), Naive Bayes con Cross-Validation y Random Forest, obteniendo así los resultados que veremos más adelante.

2. Dataset

Para el análisis se dispone de un Dataset, proporcionado por la empresa Autoritas, denominado *pan-ap17-bigdata*, que contiene 30.000 tweets de variedad lingüística español (ES) divididos entre 300 autores identificados mediante una secuencia alfanumérica para mantener su anonimato, de los cuales se asignan 200 autores para realizar el entrenamiento (Train) y 100 para el Test. Tal y como presenta la *Figura 1*.

Figura 1: Datos de análisis.

El Dataset consta de 2 carpetas (Train y Test), dentro de cada una de ellas se encuentran unos ficheros en formato XML, 2.800 y 1.400 respectivamente, donde cada fichero es un autor con sus tweets asociados junto con un fichero ”truth.txt”que contiene la lista de títulos de los ficheros xml, el género y país al que corresponden.

Hay que decir que este Dataset con el que vamos a trabajar es un subconjunto de un Dataset más amplio,(*Figura 2*), que recoge distintos idiomas según su situación geográfica. Y que la arquitectura utilizada para la recopilación de datos sigue el esquema representado en la *Figura 3*.

Figura 2: Dataset original.

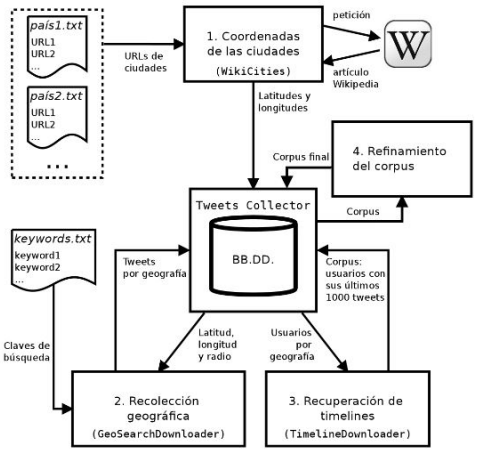

Figura 3: Arquitectura del sistema para la recolección de tweets y usuarios.

Proceso de construcción del Dataset:

- Recuperación de tweets enmarcados en una región geográfica (longitud, latitud, radio).
- Preseleccionan de los usuarios únicos que han emitido tweets, filtrados por idioma del perfil.
- Recuperación los timelines de los usuarios únicos.
- Selección de los autores con más de 100 tweets, que no sean retweets, en el idioma correspondiente o con la localización geográfica esperada en su perfil.
- Revisión manual de los perfiles para asegurar el sexo.
- Y selección de 100 tweets por autor para la construcción del Dataset final.

Llegados a este punto se preparan los datos para su explotación, utilizando el *lenguaje R* para el tratamiento de los datos.

3. Propuesta del alumno

En primer lugar antes de centrarnos en una clasificación por sexo o variedad, hemos realizado una limpieza de datos general aplicada a los dos tipos de clasificación. Hemos generado un vocabulario utilizado las funciones que se han proporcionado para recoger los tweets, limpiando el corpus de signos de puntuación, números, espacios en blanco, stopwords (palabras vacías) y términos más frecuentes, a esa limpieza le hemos añadido la eliminación de los acentos mediante la función de R:
chartr("áéíóú","aeiou",corpus.preprocessed)
Una vez generado el vocabulario creamos una bolsa de palabras (n=1000) con nuestra limpieza inicial y lanzamos la función. Como resultado obtenemos un gráfico con las palabras que más se repiten, pero como no se puede apreciar, reducimos la bolsa a n=100 y relanzamos para poder visualizarlas tal y como se ve en el gráfico siguiente:

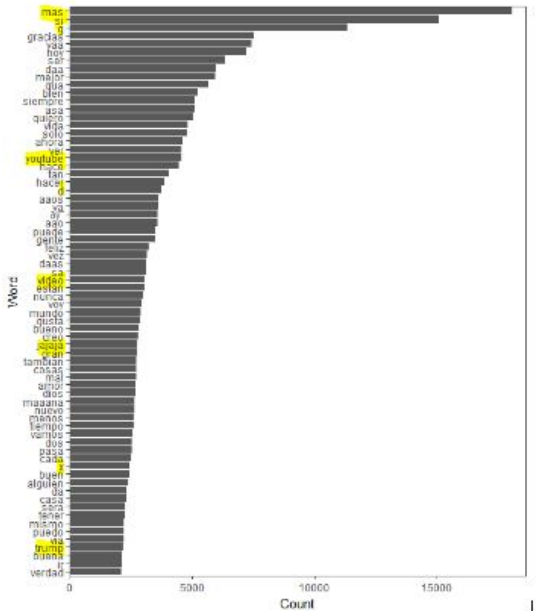

Figura 4: Las 100 palabras más frecuentes.

Como hemos dicho, este gráfico muestra las palabras más utilizadas en los tweets y nos permite ajustar el modelo eliminando las palabras comunes a todas las variedades del español, independientemente del género. Las palabras seleccionadas por el grupo a eliminar son:

- *youtube*
- *video*
- *trump*
- *mas*
- Abreviaciones de escritura: *q , x y d*
- Onomatopeyas: *jajaja*

Para ello utilizamos la siguiente función de R:
swlist=c("youtube", "video", "q", "si","x", "jajaja", "trump", "d", "mas")
y generamos un nuevo vocabulario.

A partir de este momento aplicamos los tres modelos algorítmicos de clasificación seleccionados por el grupo, al conjunto de *training*, por una parte para sexo y por otra para variedad.

```
bow_training_gen <- GenerateBoW(path_training, vocabulary, n, class="gender")
bow_test_gen <- GenerateBoW(path_test, vocabulary, n, class="gender")

bow_training_var <- GenerateBoW(path_training, vocabulary, n, class="variety")
bow_test_var <- GenerateBoW(path_test, vocabulary, n, class="variety")
```

Figura 5: Generación de los datasets de Training y Test para Sexo y Variedad.

Los modelos seleccionados para el estudio son los siguientes:

SVM (Support Vector Machine)
Naive Bayes con validación cruzada
Random Forest

Los resultados iniciales obtenidos, con la limpieza previa, la eliminación de las palabras más frecuentes y comunes a todas las variedades, y una bolsa de palabras de n=1.000 mejoramos un poco los resultados de *Accuracy* proporcionados en clase *(Figura 6 y 7)*:

VARIETY	GENDER
0.7721	0.6643

Figura 6: Accuracy proporcionado.

VARIETY	GENDER
0.845	0.6743

Figura 7: Accuracy n=1.000.

Con el fin de obtener mejores resultados, decidimos aumentar la bolsa de palabras de n=1.000 a n=2.000

4. Resultados experimentales

Tras el aumento de la bolsa de palabras obtenemos una mejora de los resultados en todos los métodos utilizados, tal y como muestra la siguiente tabla y gráfico,*(Figura 8 y 9)*. En ella se muestra los valores obtenidos con nuestro *training* final, con el que hemos obteniendo mejores resultados que con la bolsa inicial de 1.000 palabras. Destacando el mejor resultado obtenido utilizando el algoritmo de clasificación de *Random Forest*.

Total Accuracy	Column Labels		
Row Labels	Genero	Variedad	Grand Total
RandomForest	0,7343	0,8679	1,6022
SVM	0,6743	0,845	1,5193
Naive Bayes con CV	0,6653	0,835	1,5003
Grand Total	**2,0739**	**2,5479**	**4,6218**

Figura 8: Accuracy n=2.000.

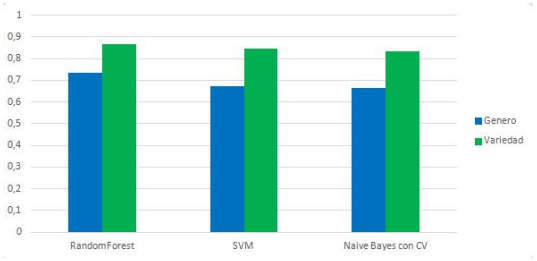

Figura 9: Gráfico de resultados.

5. Conclusiones y trabajo futuro

La conclusión a la que hemos llegado es que con una mayor limpieza de los datos y principalmente un aumento de la bolsa de palabras podemos obtener mejores resultados.

En un trabajo futuro se podría definir una limpieza más exhaustiva, tanto por *Variedad* como por *Sexo*, en esta ocasión por cuestión de tiempo solo hemos podido realizar una limpieza genérica a las 2 clasificaciones juntas, aunque consideramos en el planteamiento previo de la actividad otras acciones como:

- Discretizar entre palabras más utilizadas comúnmente por mujeres que por hombres y viceversa.
- Diferencias lingüísticas dentro de la Variedad:
  - El uso de distintas modalidades de *voseo* característico del Cono Sur, especialmente de Argentina y Uruguay, Centroamérica y ciertas zonas de Colombia y Venezuela. El mismo es inexistente en España.
  - Uso diferente de diminutivos, los terminados en *-illo, -ete e -ín* son propios de España, mientras que en paises como Venezuela y Colombia este diminutivo se usa solo en las palabras terminadas en *-te, -ta y -to*.
  - El uso del sistema pronominal para la segunda persona del plural, en España se

diferencia entre ”vosotros”(confianza) y üstedes”(respeto) y sus respectivas formas verbales y pronominales mientas que en Latinoamérica solo se usa üstedes”, sin diferenciar entre la confianza y el respeto en el plural.

- En Latinoamérica se prefiere la perífrasis de futuro *ir a + infinitivo*, y en España se usa comparativamente más la conjugación del futuro.
- Uso de *arcaísmos* como ”pararse”(Latinoamérica) por ”ponerse de pie”(España) y de *marinerismos* como ”virar”por ”girarz ”dar la vuelta.º ”doblar”.

- Estilística en la puntuación dentro de la Variedad:
  - En España se usan preferentemente las comillas «latinas», al igual que en francés (« »), mientras que en Latinoamérica se utilizan las comillas dobles (”) o simples (’) como las inglesas, sin embargo, no hay variaciones normativas respecto a su empleo.
  - En algunos países de Latinoamérica, especialmente en México, se emplea el punto como separador decimal en lugar de la coma, del mismo modo que en inglés.

- Crear nuevas funciones para el procesado de datos, como por ejemplo eliminar palabras de una única consonante.

- Probar con nuevos algoritmos de clasificación.

Finalmente, y si se dispone de tiempo, ir realizando ampliaciones de la bolsa de palabras , ya que como hemos detectado proporcionan un mejora notable en los resultados, aunque posee el inconveniente de necesitar un mayor tiempo de procesado.

References

Apuntes asignatura:*Text Mining en Social Media* Paolo Rosso y Francisco Rangel.