Práctica con PAN-AP 17.
Asignatura Text Mining en Social Media.
Master Big Data

Elena Tejadillos
elena.tejadillos@gmail.com

Abstract

El objetivo de esta práctica es conseguir un método que partiendo de un tuit, nos permita conocer el género y el país de procedencia del autor del tuit. Para ello se tiene como fuente inicial un conjunto de tuits. De ellos se extraerán una lista de 2000 palabras, las que más frequentemente aparecen en los tuits, de las cuales se descartarán las stopwords (las palabras que se consideren no relevantes). Estos tuits son procesados y agrupados de manera balanceada en 2 datasets, uno de training y otro de test, en los cuales se estudiará el género y el país de origen de los autores de los tuits. Se aplican 3 modelos diferentes (Suppert Vector Machine, Naive Bayes con validación cruzada y RandomForest) para ver con cual de todos obtiene un mejor resultado. Finalmente se selecciona el RandomForest como mejor modelo ya que con él se obtiene un .ªccuracy"más alto, tanto para género como para variedad. Como objetivo final, se deberán superar los umbrales dados. El resultado a superar para reconocer el género es *0.6643* y para reconocer el país es *0.7721*

1. Introducción

.*Aauthor profiling"* es el área de estudio que da la posibilidad de conocer rasgos de una persona a partir de los textos que escribe. En el caso concreto de esta práctica, se quiere averiguar el sexo y el país de origen de los autores de un conjunto de tuits. Para realizar el estudio, se ha proporcionado un conjunto de 3200 tuits, que forman parte del dataset PAN-AP'17. Estos tuits se agruparán en dos subconjuntos, uno utilizado para el *"training"* de los modelos y el otro para el *"test"*.

Cada dataset se estudiará de manera separada para predecir el género y y el país de origen de los autores de los tuits.

2. Dataset

Se proporciona la información necesaria para el dataset PAN-AP'17 a través de un fichero zip. Este fichero contiene los tuits de 300 autores distintos: 200 autores para training y 100 autores para test. Al descomprimirlo, se generan 2 carpetas llamadas *"training"* y *"test"*.

En la carpeta *"training"* se almacenarán 2800 ficheros xml que contienen los tuits a analizar y un fichero "truth.txt"que contiene la lista de títulos de los ficheros xml y el género y páis a que corresponden.

La carpeta "testçontiene la misma información pero con tan solo 1400 ficheros xml. Estos tuits son los se van a clasificar dependiendo del género (male/female) o de proceder de uno de los 7 paises:

Argentina	Chile
Colombia	España
Mexico	Perú
Venezuela	

Al cargar toda la información en el programa, se tendrán 2 dataset, uno para training y otro para test. Estos datasets se han creado de manera que los datos estén balanceados, lo que significa que habrá el mismo número de tuits para hombres y para mujeres así como para cada uno de los países incluidos en la muestra.

Una vez que ambos datasets están cargados, se hará una limpieza de datos eliminando los signos de puntuación de los textos de los tuits.

TUITS	TRAINING		TEST	
	MALE	FEMALE	MALE	FEMALE
ARGENTINA	200	200	100	100
CHILE	200	200	100	100
COLOMBIA	200	200	100	100
ESPAÑA	200	200	100	100
MEXICO	200	200	100	100
PERÚ	200	200	100	100
VENEZUELA	200	200	100	100
TOTAL	1400	1400	700	700
	2800		1400	

Figura 1: Distribución del Dataset original.

3. Propuesta del alumno

Lo primero que se realiza es obtener la lista de las 1.000 palabras más frecuentes que aparecen tanto en el corpus de training como en el de test.

Partiendo de esta lista de 1.000 palabras y tras ordenarlas de mayor a menor frequencia, se seleccionan las 100 primeras para estudiarlas más en profundidad.

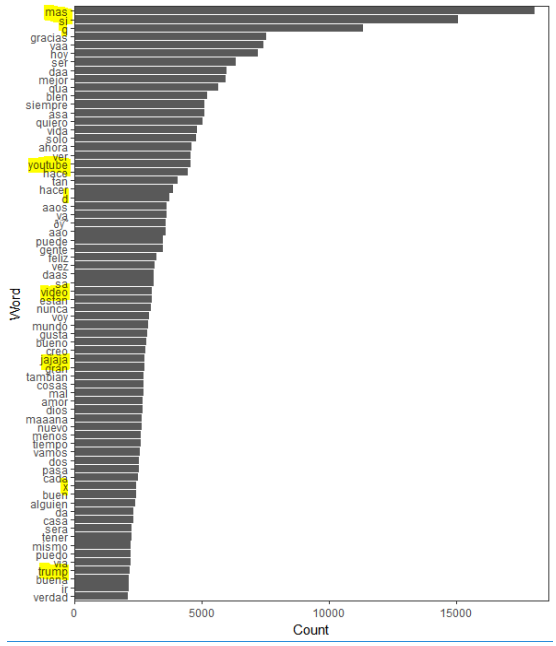

Figura 2: Lista de las 100 palabras más fecuentes.

Se decide excluir alguna de estas palabras a las que se denominarán *stopwords*. Las *stopwords* son palabras vacías , un listado de términos (preposiciones, determinantes, etc.) considerados de escaso valor semántico, que cuando se identifican en un documento se eliminan.

La supresión de todas estas palabras evita los problemas de ruido y supone un considerable ahorro de recursos, ya que aunque se trata de un número relativamente reducido de palabras tienen una elevada tasa de frecuencia.

Tras un análisis inicial, se decide incluir como *stopwords* las siguientes palabras:

- La palabra más utiliza y por ello, considerada como con escaso valor semántico ***"mas"***
- Palabras de una sola letra ***"d"*** o ***"q"*** (preposición a o abreviación)
- Onomatopeyas universales (***"jajaja"***)
- Nombres propios o genéricos (***"Trump"*** o ***"video"***)

Las palabras seleccionadas como *stopwords* están más orientadas a la variedad que al género. Se buscó en internet si estas palabras se dirían de igual manera en todos los países. Queda pendiente el estudio de qué tipo de palabras son más utilizadas por autores masculinos y cuales por autores femeninos.

A continuación, se aplican diferentes modelos para comprobar con cual de ellos se obtiene un mejor resultado (un mayor *Accuracy*. Los modelos seleccionados son:

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

De todos ellos, Naive Bayes es el que peor resultado obtiene, y Random Forest es del que se obtiene un mayor *Accuracy*.

Aunque se mejora levemente el umbral dado por el profesor, se decide aplicar como propuesta de mejora aumentar la bolsa de palabras de 1000 a 2000. Se vuelve a generar de nuevo la bolsa con 2000 palabras eliminando las *stopwords* y aplicando de nuevo los modelos.

De esta forma se consigue mejorar el resultados de todos los modelos tanto para la predicción de género como la de variedad.

4. Resultados experimentales

Tras aplicar los 3 modelos elegidos con una bolsa de 2000 palabras, el mejor resultado obtenido tanto para género como para variedad es el generado por el modelo de Random Forest.

Los resultados son los siguientes:

- Género: 0.7313
- Variedad: 0.8679

Tal y como puede observarse en la siguiente gráfica, los resultados obtenidos por los tres métodos superan, aunque sea levemente, el umbral inicial proporcionado. Pero el Random Forest destaca sobre los tres.

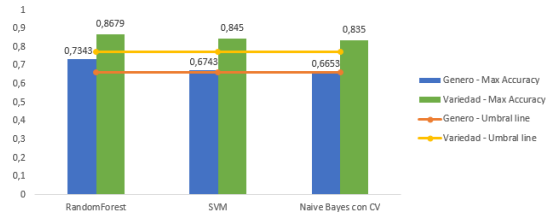

Figura 3: Comparativa de resultados obtenidos y umbrales.

5. Conclusiones y trabajo futuro

Las concluiones de la práctica han sido que el mejor modelo que podemos utilizar es Random Forest. Pero lo que más ha ayudado a conseguir una mayor mejora del resultado de la predicción, ha sido el aumento de 1000 a 2000 palabras en la bolsa de palabras a utilizar.

Pero este aumento en el número de palabras también ha conllevado a un aumento considerable en el tiempo cálculo de los modelos.

El estudio que se ha hecho de las palabras ha sido muy básico y muy orientado a la variedad, por lo que quedaría pendiente hacer un estudio más en profundidad de las palabras que puedan ser significativas o no por ser utilizadas por los hombres.

También se podría aplicar otros algoritmos como el TF-IDF (*Term frequency – Inverse document frequency*, basada en pesos que permite mejorar la selección de *stopwords* o como el de S-stemmer, que permite reducir las formas plurales al singular y así calcular la frequencia de una palabra esté en singular o plural.

References

Mari Vallez y Rafael Pedraza-Jimenez (Universitat Pompeu Fabra) 2007. *El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. https://www.upf.edu/hipertextnet/numero-5/pln.html*