Text Mining en Social Media
Máster Big Data Analytics - Curso 2016 / 2017
Universitat Politècnica de València

Detección de género y variedad de idioma en textos extraídos de la red social Twitter (datos *PAN-AP'17*)

Christian Ferrer Fas
chferfa@inf.upv.es

Abstract

A continuación, se describe un problema del mundo real consistente en detectar ciertas características acerca de los autores de textos publicados en redes sociales (concretamente su género y su variedad de idioma), y se detallan las diferentes acciones que se han llevado a cabo para proporcionar una solución al mismo.
La solución presentada se basa, grosso modo, en aplicar una serie de técnicas de *text mining* sobre el *dataset* proporcionado para generar un conjunto de características que permita, posteriormente, entrenar un modelo de *machine learning* que sea capaz de determinar, ante la llegada de un nuevo texto, ciertas características sobre el autor del mismo (su género y su variedad de idioma).
A lo largo de la tarea, se pone de relevancia la importancia que tienen ciertas decisiones iniciales sobre el resultado final (entendiendo este como el *accuracy* del modelo entrenado). Las decisiones más importantes se resumen en: aplicar los métodos de preprocesado más adecuados al *dataset* proporcionado, seleccionar el vocabulario más discriminante posible (el uso de pesos **tf-idf** ha marcado la diferencia en este aspecto) y seleccionar el método de *machine learning* más adecuado al tipo de características a estudiar (los candidatos más adecuados son **Support Vector Machine** y **Random Forest**).

1 Introducción

De forma general, el problema de **Author Profiling** consiste en detectar ciertas características del autor de un texto, a partir de características presentes en el propio texto, tales como el estilo de escritura, expresiones o palabras empleadas, etc. A continuación, se presentan dos problemas de **Author Profiling** a resolver sobre un mismo *dataset*. El *dataset* está extraído de la red social **Twitter** y los problemas a resolver consisten en:

- Determinar el género (hombre o mujer) del autor de cada *tweet* a partir del contenido del mismo.
- Determinar la variedad de idioma (argentina, chilena, colombiana, mexicana, peruana, venezolana o española) del autor de cada *tweet* a partir del contenido del mismo.

2 Dataset

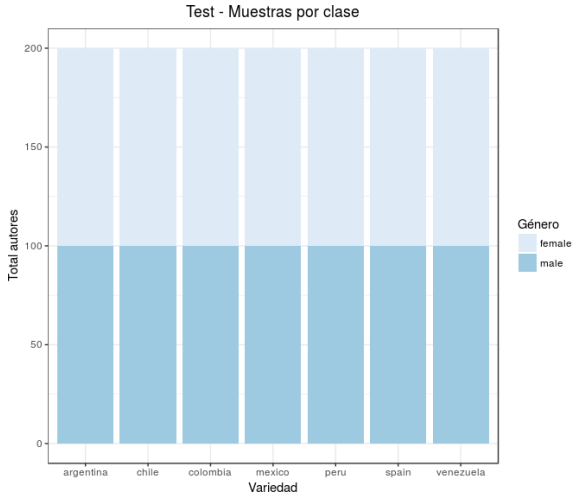
El *dataset* proporcionado cuenta con las siguientes características:

- Está separado en dos conjuntos: *training* y *test*.
- Cada muestra, tanto de *training* como de *test*, está constituida dentro de un fichero *XML* que contiene todos los textos de los *tweets* de un mismo autor. El nombre del fichero representa el identificador único de cada autor.
- Las etiquetas de cada muestra, tanto de género como de variedad, se encuentran en un fichero único y están asociadas a cada una de las muestras por medio del identificador único del autor.

En cuanto a las dimensiones del *dataset*, tanto en el conjunto de *training* como en el de *test*, cada una de las muestras está compuesta por el texto de 100 *tweets* diferentes de un mismo autor.
El conjunto de *training* contiene 2800 muestras, mientras que el conjunto de *test* contiene 1400 muestras (la mitad).

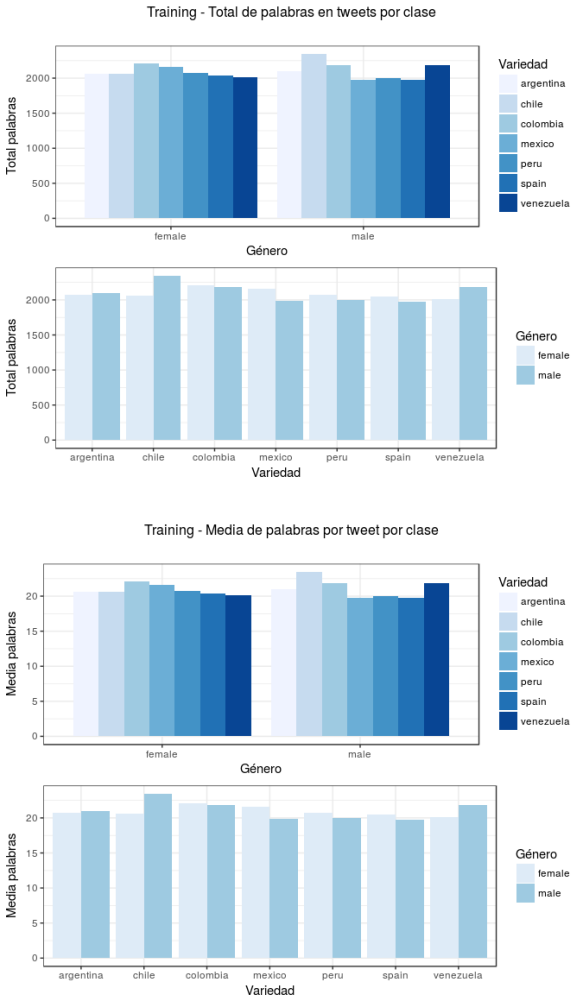
Las muestras están distribuidas de forma equitativa entre las dos características a estudiar: género y variedad. De este modo, el conjunto de *training* contiene 1400 muestras de cada género (200 muestras de cada variedad por cada género) y 400 muestras de cada variedad (200 muestras de cada género por cada variedad), como se puede observar en la siguiente gráfica:

De forma equivalente, el conjunto de *test* contiene 700 muestras de cada género (100 muestras de cada variedad por cada género) y 200 muestras de cada variedad (100 muestras de cada género por cada variedad), como se puede observar en la siguiente gráfica:



Por tanto, la distribución de las muestras está perfectamente equilibrada entre las dos características a estudiar.

Además, las muestras tienen características similares en cuanto a cantidad de palabras en total y cantidad de palabras por *tweet*. Este aspecto no está distribuido equitativamente de forma perfecta como los anteriores aspectos, pero igualmente está muy bien balanceado, como se puede apreciar en las siguientes gráficas del conjunto de *training*:



La distribución de palabras en los *tweets* del conjunto de *test* es muy similar, en términos relativos, a la que se ha mostrado para el conjunto de *training*.

El hecho de disponer de un *dataset* tan equilibrado entre las características a estudiar, reduce drásticamente el problema de desbalanceo que podría influir negativamente en las técnicas de *machine learning* a aplicar sobre el mismo.

3 Propuesta del alumno

Para resolver ambos problemas (detección del género y detección de la variedad) se ha decidido aplicar el mismo tipo de preprocesado sobre los datos, así como la misma técnica de *machine learning* para el aprendizaje del modelo, variando algunos aspectos mínimos de un problema a otro (el tamaño del vocabulario). Esta decisión se basa en el tiempo límite disponible para resolver los dos problemas, sin embargo, sería más idóneo

realizar un análisis por separado para cada caso, pues cada uno de ellos se podría abordar mejor aplicando un preprocesado diferente sobre los datos, dada la naturaleza de cada problema.
El proceso global consiste en los siguientes 3 pasos generales:

- Obtención del vocabulario a partir de las muestras de *training*:
  - Preprocesado de los textos de los *tweets* y separación por clases.
  - Obtención de las palabras más frecuentes para cada clase.
  - Obtención de los pesos **tf-idf** para las palabras más frecuentes con respecto a todas las clases.
  - Obtención las palabras con mayor peso **tf-idf**: este conjunto de palabras, de tamaño configurable, constituye el vocabulario.
- Obtención de las bolsas de palabras para los conjuntos de *training* y de *test*, mediante el vocabulario generado.
- Obtención del modelo de *machine learning* mediante **Random Forest**:
  - Preparación de los datos de las bolsas de palabras para el entrenamiento y el test del modelo.
  - Entrenamiento del modelo.
  - Test del modelo.
  - Evaluación del modelo.

El preprocesado de los datos incluye opciones para convertir todas las palabras a minúsculas, eliminar signos de puntuación, eliminar números, eliminar acentos, eliminar *stop words* (tanto obtenidas con la librería **tm** para idioma *es*, como proporcionadas adicionalmente) y normalizar espacios en blanco. Los resultados obtenidos son extraordinariamente variables en función del preprocesado que se aplique sobre los datos, por lo que estas opciones de preprocesado son altamente configurables para facilitar las sucesivas pruebas.
Se han evaluado dos métodos para seleccionar las palabras a incluir en el vocabulario: obtención de palabras más frecuentes (por medio de la función proporcionada **GenerateVocabulary**) y obtención de palabras con mayor peso **tf-idf**. El método finalmente empleado en la solución es, en realidad, una combinación de ambos métodos consistente en: obtener las n palabras más frecuentes de cada clase (conjunto de $n \cdot nclases$ palabras), calcular los pesos **tf-idf** del conjunto anterior y seleccionar las m palabras con mayor peso **tf-idf**.
El empleo del cálculo de pesos **tf-idf** (proporcionado por la librería **TidyText**) se basa en otorgar un valor mayor a una palabra cuanto más discriminante es (es decir, el valor será mayor cuánto más frecuente sea en su clase y menos en el resto de clases).
Una vez se dispone del vocabulario, se generan las bolsas de palabras para los conjuntos de *training* y de *test*. Este proceso consiste en medir las frecuencias de aparición en cada conjunto de las palabras incluidas en el vocabulario y seleccionar aquellas más frecuentes (junto a su frecuencia).
Los datos de las bolsas de palabras se adecúan para poder emplearlos en el aprendizaje del modelo de *machine learning* mediante **Random Forest** (proporcionado por la librería **caret**). El modelo entrenado es testeado con los datos de la bolsa de palabras de *test* y evaluado mediante una matriz de confusión, que proporciona el dato del *accuracy*.
Para llegar a este proceso, se han ido probando diferentes técnicas, siendo estas las que mejor resultado han ofrecido. Entre las técnicas descartadas, se encuentran principalmente las siguientes:

- Obtención de vocabulario basado en *n-gramas*, en lugar de palabras.
- Modelo de *machine learning* mediante **Support Vector Machine**. Aunque en la literatura se aconseja como el método más efectivo para este tipo de problemas, los resultados de las diferentes pruebas han resultado ser mejores en cuanto a *accuracy* empleando **Random Forest**.

4 Resultados experimentales

El *baseline* proporcinado es el siguiente, para la medida de *accuracy*:

Clase	***Accuracy***
Género	0.6643
Variedad	0.7721

Se han aplicado las siguientes configuraciones de preprocesado y vocabulario:

Preprocesado	Género	Variedad
Minúsculas	SI	SI
Puntuación	SI	SI
Números	SI	SI
Acentos	SI	SI
Stop words **tm**	SI	SI
Stop words adicionales	SI	SI
Espacios en blanco	SI	SI

Vocabulario	Género	Variedad
n (frecuencia)	1000	1000
m (tamaño vocabulario)	700	500

Con el proceso y la configuración descritos, se han obtenido los siguientes valores de accuracy:

Clase	*Accuracy*
Género	0.7343
Variedad	0.9157

A lo largo de todo el proceso, se han incluido mediciones de los tiempos para cada uno de los pasos computacionalmente más comlejos. A continuación, de indican estos tiempos (expresados en segundos):

Acción	Género	Variedad
Preprocesar y separar	12.5750	3.5552
Palabras frecuentes	12.3195	3.3954
Pesos **tf-idf**	0.0030	0.0050
Bolsa palabras *training*	169.8910	132.9490
Bolsa palabras *test*	88.8420	65.7310
Entrenar modelo	76.1590	62.2180
Testear modelo	0.2160	0.1840
Evaluar modelo	0.0030	0.0150

5 Conclusiones y trabajo futuro

Tras el trabajo realizado se puede observar que la detección de la variedad es un problema relativamente más sencillo de solucionar que la detección del género, a partir de textos publicados en **Twitter**. Uno de los posibles motivos puede ser que, debido a que existen palabras autóctonas o exclusivas en cada país (que se emplean en dicho país y no en el resto), la discriminación de la variedad usando como características el empleo de estas palabras ofrece poca probabilidad al error. Sin embargo, no parece existir una característica discriminante tan profunda entre las palabras que emplean hombres y mujeres (quizás el problema hubiera sido más sencillo si se tratara de detectar rangos de edad).

Ante la dificultad aparente de encontrar palabras que discriminen los géneros, parece razonable pensar que el uso de un vocabulario basado en *n-gramas* podría ofrecer nuevas posibilidades, ya que permitiría enfocar el aprendizaje del modelo a expresiones propias de cada gnero.

Otro aspecto que se ha podido observar de forma muy notable es la importancia que existe en el preprocesado de los datos, habiéndose mejorado considerablemente el *accuracy* en la detección de la variedad al quitar acentos en las palabras. Sin embargo, esto es otro ejemplo donde es posible que un preprocesado más concreto sobre los signos de puntuación, acentos, etc., podría permitir al modelo aprender características más discriminantes para la detección de la varidad, aprendiendo los estilos de escritura predominantes en cada país.

En cuanto a trabajo futuro, se consideran las siguientes posibilidades:

- Preprocesado de emoticonos, principalmente para la detección del género.
- Ampliar conjunto de *stop words* adicionales.
- Añadir información externa al vocabulario.
- Expansión de *URL*s, principalmente para la detección de la variedad *.
- *Tunning* de parámetros del proceso, desde las opciones de preprocesado hasta los tamaños del vocabulario.
- Métodos de *machine learning* alternativos (profundizar en la parametrización de **Support Vector Machine** y emplear otros métodos, como **Redes Neuronales Artificiales**).

References

Dr. Francisco Manuel Rangel Pardo y Dr. Paolo Rosso 2017 *Material asignatura Text Mining en Social Media*, Máster Big Data Analytics, curso 2016/2017 Universitat Politècnica de València

The Comprehensive R Archive Network https://cran.r-project.org

scikit-learn Machine Learning in Python http://scikit-learn.org

Wikipedia https://es.wikipedia.org/wiki/Tf-idf

*Esta funcionalidad ya se ha implementado en la solución, pero no se ha empleado puesto que, al tratarse de una ejecución secuencial, el tiempo necesario para su ejecución es muy elevado; por tanto, el trabajo futuro consistiría en realizar una implementación multihilo, o bien ejecutarlo de forma externa e incorporar los resultados como un *dataset* auxiliar.