Author profiling

Luis Miguel Montes Novella
lmontesnovella@gmail.com

Abstract

En este artículo se describe cómo se ha aproximado la tarea de *author profiling* para la identificación de género y país en un conjunto de textos extraídos de *Twitter*. Para llevar a cabo esta tarea se han extraído una serie de características del texto que se cree que pueden ayudar a diferenciar entre géneros o países de procedencia. Las características extraídas son el uso de emoticonos, el uso de ciertas características en el estilo de escritura como el *character flooding*, la inclusión en los textos de números, direcciones *web*, correos electrónicos, etc. Además de estas características se ha obtenido un *BoW (Bag of Words)* con las palabras que ayudan a diferenciar mejor entre géneros y países. Para el procesamiento de los textos se ha usado el lenguaje de programación *Python* y para el entrenamiento y evaluación de los modelos de clasificación se ha usado *R*. Se han usado dos modelos, uno entrenado mediante *Naive Bayes* y otro que hace uso de árboles de clasificación.

1 Introducción

El *author profiling* es una tarea que consiste en la identificación de tipos o clases de autores a partir de textos escritos mediante la búsqueda de patrones comunes en el estilo de escritura. Esta clasificación de autores en diferentes grupos tiene numerosas aplicaciones como por ejemplo la identificación de personas sobre las que dirigir las campañas publicitarias o políticas. En concreto en este trabajo se han seguido dos aproximaciones, la detección de género y la detección de variedad del lenguaje en países hispanohablantes. Para la primera aproximación tenemos 3 posibilidades, detectar si se trata de un hombre, una mujer o bien no se ha podido identificar el género (se cree que muchos de estos perfiles pueden pertenecer a diferentes instituciones). En cuanto a la segunda aproximación tenemos 7 posibles países, Argentina (AR), Chile (CL), Colombia (CO), España (ES), México (MX), Perú (PE) y Venezuela (VE).

2 Dataset

El *dataset HispaTweets* (1) contiene *tweets* de 4550 autores hispanohablantes, se compone de una serie de ficheros, uno por autor, con todos los *tweets* que este ha escrito. Este *dataset* contiene un total de 4550 ficheros donde tenemos 650 autores de cada país. En la tabla 1 se puede ver la distribución de los autores en función de las clases a las que pertenecen, como se puede observar la distribución en cuanto a país es equilibrada pero en lo referente al género el número de muestras de cada clase es muy diferente, en este caso se ve como la mayor parte de los *tweets* pertenecen al género desconocido, es decir, no se han podido clasificar.

	Mujer	**Hombre**	**Desc**	**TOTAL**
AR	150	198	302	**650**
CL	126	285	239	**650**
CO	136	249	265	**650**
ES	148	249	253	**650**
MX	111	265	274	**650**
PE	155	259	236	**650**
VE	98	258	294	**650**
TOTAL	**924**	**1763**	**1863**	**4550**

Table 1: Distribución de los totales por clase

3 Propuesta del alumno

En este apartado se describen las características extraídas del texto y las variables obtenidas a partir de éstas para entrenar diferentes modelos mediante el algoritmo *Naive Bayes* y árboles de clasificación. Para la extracción de características del texto se ha usado el lenguaje de programación *Python* y para entrenar y validar los modelos de clasificación el lenguaje de análisis de datos *R*.

3.1 Información extraída del texto

A la hora de procesar el dataset se han seguido los siguientes pasos. En primer lugar se han procesado los ficheros para extraer la información necesaria y guardarla de una forma más compacta, para ello se han creado dos ficheros, uno de entrenamiento y otro de test. En estos ficheros por cada autor aparece su género, país y los textos de todos sus *tweets*. Con este preproceso se ha conseguido reducir el tamaño del dataset original desde 13 GB hasta aproximadamente 350 MB. Una vez realizado este preproceso se han usado una serie de scripts en *Python* para obtener un vocabulario y una representación de las muestras de entrenamiento y test para poder entrenar diversos modelos de clasificación. En los siguientes subapartados se describen las características extraídas de los textos.

3.1.1 Bag of Words

Un *Bag of Words, BoW* o Bolsa de Palabras es un conjunto de palabras que se usa para representar un texto, normalmente se escogen palabras de uso común o que ayudan a discriminar entre las diferentes clases que hay que diferenciar.

En este caso se ha modificado la forma de obtener el *BoW* para intentar mejorar los resultados en ambas tareas, para ello se han procesado los textos y se han *tokenizado* mediante el uso de la clase *TweetTokenizer* de la librería *NLTK* (2). De estos *tokens* se han eliminado los caracteres extraños (como por ejemplo los emoticonos), se han convertido todos sus caracteres a minúsculas y se ha creado un vocabulario con los *tokens* que aparecen más de 500 veces. Para cada palabra de las de este vocabulario se han obtenido las frecuencias de aparición por cada clase diferenciando entre género y país. Para las frecuencias de país y las frecuencias de género de cada palabra se ha calculado la desviación típica. Por tanto por cada palabra tenemos su frecuencia absoluta por cada clase, la desviación típica de las frecuencias por género y la desviación típica de las frecuencias por país. De esta forma al ordenar las palabras por desviación típica de forma decreciente es posible ver cuáles son las palabras que más ayudan a discriminar entre las clases, esto es porque la desviación típica será mayor cuanto mayores sean las frecuencias y mayor diferencia haya entre ellas. Una vez ordenadas estas palabras se han obtenido dos *BoWs* con 1000 palabras, uno para género y otro para país.

3.1.2 Uso de emoticonos

En las redes sociales es muy frecuente el uso de emoticonos, su uso podría estar muy relacionado con el género ya que se usan frecuentemente para expresar sentimientos y en este aspecto suele haber muchas diferencias entre hombres y mujeres. Dado que el texto se encuentra codificado en Unicode se han identificado los códigos correspondientes con estos emoticonos durante el procesamiento de los textos para contar el número de emoticonos que aparecen.

3.1.3 Hashtags y menciones

Dado que *Twitter* es una red social donde abunda el uso de *hashtags* y menciones se han identificado estas características para ver si su uso puede ayudar a diferenciar entre géneros o países.

3.1.4 Risas

La frecuencia del uso de risas en los mensajes de *chat* podría ayudar a detectar el género de una persona, sobretodo para diferenciar los perfiles institucionales ya que el lenguaje usado por ellos debería ser más formal. Para ello se ha realizado el parseo del texto buscando si los *tokens* que aparecen se corresponden con risas. Se ha considerado una risa toda cadena de texto que contiene en su interior una subcadena como *jaj*, *jej*, *jij*, *joj* o *juj*. Esta detección se ha hecho mediante el uso de expresiones regulares sobre cada uno de los *tokens*.

3.1.5 Character flooding

El *character flooding* es un estilo de escritura que consiste en el alargamiento de palabras duplicando letras o añadiendo letras que no existen como por ejemplo la o en *hooola* o la h en *tengohh*. Se cree que esta característica puede ayudar a diferenciar entre géneros ya que se puede considerar una forma de expresar sentimientos y estados de ánimo.

Se ha usado un algoritmo para detectar *character flooding* a nivel de palabras. Para ello se ha considerado que una palabra tiene *character flooding* siempre que contiene más de dos vocales seguidas o dos caracteres seguidos que no sean c, l, r o vocal. Se han excluido estos caracteres porque en castellano pueden aparecer juntos.

3.1.6 Otros

Además de todo lo anterior también se ha procesado el texto con expresiones regulares en busca de direcciones *web*, correos electrónicos y números para ver si la aparición de éstos en un texto puede ayudar a diferenciar entre los diferentes géneros o países.

3.2 Modelos empleados

Una vez obtenidos los ficheros de entrenamiento y test a partir de los textos se han escogido los modelos de aprendizaje automático a usar para la tarea de clasificación. Se ha decidido escoger dos tipos de modelos diferentes. En primer lugar se ha escogido un modelo de clasificación que emplea la técnica de *Naive Bayes* debido a la simplicidad de este tipo de modelos y la rapidez con la que se pueden entrenar y evaluar. En segundo lugar se ha escogido un modelo que hace uso de árboles de clasificación, este modelo se ha escogido debido a que los árboles de clasificación generalmente escogen mejor las variables que más ayudan a discriminar entre clases y los modelos resultantes suelen ser más fáciles de interpretar por los humanos. Para el entrenamiento y evaluación de los modelos se ha empleado el lenguaje de programación *R* con las librerías *e1071 (Naive Bayes)* y *rpart* (árboles de clasificación). En los modelos se han usado las siguientes variables basadas en las características extraídas de los textos:

- ***BoW***: *Bag of Words* con 1000 palabras obtenido con el proceso del apartado 3.1.1. Para cada palabra se indica la frecuencia relativa de aparición dentro de cada texto. Hay que tener en cuenta que los *BoWs* usados para diferenciar entre géneros y países son diferentes.
- **ratio de emoticonos**: número de emoticonos en el texto entre el total de palabras del texto.
- **ratio de hashtags**: número de hashtags en el texto entre el total de palabras del texto.
- **ratio de menciones**: número de menciones en el texto entre el total de palabras del texto.
- **ratio de risas**: número de risas entre el total de palabras del texto.
- **ratio de *character flooding***: número de palabras con *character flooding* entre el total de palabras del texto.
- **ratio de uso de direcciones *web***: número de direcciones *web* en el texto entre el total de palabras del texto.
- **ratio de números**: cantidad de veces que aparecen números en un texto entre el total de palabras del texto.
- **ratio de uso de direcciones de correo electrónico**: número de direcciones de correo electrónico entre el total de palabras del texto.

4 Resultados

En la tabla 2 se pueden observar los resultados obtenidos, en la primera fila se muestran los resultados para el algoritmo *Naive Bayes* haciendo uso del *BoW*, se puede observar que los resultados obtenidos para el género son pésimos ya que prácticamente coinciden con la probabilidad de acertar la clase al azar. Sin embargo para el país se obtienen buenos resultados. En la segunda fila se pueden ver los resultados que se obtienen cuando no se usa *BoW*, en el caso del género mejoran pero en el caso del país empeoran considerablemente. Por último al usar árboles de clasificación con todas las variables los resultados en la clasificación de género mejoran y los de país se aproximan a los obtenidos con el modelo entrenado mediante *Naive Bayes*. Por tanto, para la clasificación de *tweets* por género funcionan mejor los árboles de clasificación y para la identificación de la variedad del lenguaje funciona mejor el algoritmo de *Naive Bayes* haciendo uso del *BoW* y del resto de características extraídas. Como conclusión cabe destacar que para todas las tareas no funcionan igual de bien todos los métodos y que por más características que se añadan al dataset no siempre van a mejorar los resultados ya que eso depende en gran medida de la capacidad del algoritmo de aprendizaje usado para escoger las variables que más ayudan a discriminar entre las diferentes clases.

	Género	País
Naive Bayes con BoW	33.68 %	**85.16 %**
Naive Bayes sin BoW	40.05 %	20.88 %
Arboles de clasificación	**48.02 %**	83.46 %

Table 2: Resultados obtenidos

5 Conclusiones y trabajo futuro

Tras observar los resultados obtenidos se puede concluir que aunque las tareas de clasificación parezcan similares no siempre funciona bien el mismo tipo de modelo ya que para género *Naive Bayes* no ha funcionado bien mientras que para país se han obtenido unos buenos resultados. También cabe destacar la dificultad de la tarea de clasificación en el caso del género debido a la existencia de un género desconocido, un autor perteneciente a este género puede ser hombre o mujer, también puede darse el caso de que haya *tweets* mezclados tanto de hombres como de mujeres si se trata de un perfil institucional. Al ser el género desconocido la clase mayoritaria influye de manera negativa en los resultados.

Tras analizar los *BoWs* obtenidos para género y país se puede apreciar como sobretodo en el caso de los paises abundan nombres de ciudades, políticos, palabras características de esos países, etc. En el caso del género también se aprecia como se han seleccionado algunas palabras relacionadas con sentimientos, determinantes, etc.

En cuanto a posibles mejoras se puede intentar mejorar la generación del *BoW* y extraer otras características mediante el uso de diccionarios, por ejemplo se podría usar un diccionario con palabras en inglés para detectar el uso de este idioma en los textos en castellano ya que es algo bastante frecuente. Otra mejora sería la eliminación de algunos *tokens* del *BoW* como por ejemplo los símbolos, signos de puntuación o números ya que incluirlos en el *BoW* puede causar un sobreajuste del algoritmo. Por otra parte si se extrae como característica la frecuencia de uso de estos símbolos se podrían obtener mejores resultados. A parte de todo esto se podría realizar un mayor análisis de *tokens* que aparecen entre los textos como por ejemplo las direcciones web o de correo electrónico, si de estas direcciones se analiza el dominio esto podría ayudar a identificar los países. También se podría añadir información adicional sobre los autores extrayendo su descripción y añadiéndola a sus *tweets* o intentando analizar el género de las palabras que emplean a la hora de describirse.

References

[1] s3.amazonaws.com/cosmos.datasets/hispatweets.zip

[2] www.nltk.org

[3] github.com/lmontes/text_mining_hispatweets