Text Mining II: Author Profiling

Víctor Núñez Monsálvez
vicnumon@alumni.upv.es

Abstract

El author profiling es el problema consistente en extraer rasgos del autor a partir de textos que haya escrito. En este caso, el estudio se centra en post publicados en redes sociales. Los rasgos bajo estudio son la edad y el género. Para realizar esta tarea, se emplea el concepto de sociolecto con el objetivo de escoger una serie de características a tener en cuenta, como el character flooding o el uso de comas, para clasificar correctamente cada autor en una clase de género y de edad. Además, se realiza una representación estadística del corpus de entrenamiento empleando una bolsa de palabras. Para cada uno de los autores se calcula el valor de la frecuencia para cada palabra de la bolsa de palabras y para cada una de la características. Estos valores se codifican, junto con el valor real para edad y género, en ficheros que se aplican a algoritmos de aprendizaje automático en la herramienta WEKA. Por último, el resultado de dichos algoritmos, el porcentaje de acierto en las clase tanto para edad como género, se compara con los resultados preliminares del baseline para concluir el grado de mejora que se consigue con la estrategia adopatada para abordar el problema de clasificar autor en cuanto edad y género.

1 Introducción

El tema tratado en este trabajo es el problema del Author Profiling ('evaluación por perfil del autor'). Se puede considerar la tarea del Author Profiling como una subtarea del Authorship Analysis ('análisis de autoría'), cuyo fin es el de clasificar textos en clases atendido al estilo del autor [1]. El objetivo del Author Profiling es el extraer rasgos del autor de uno o un conjunto de textos.

Algunos ejemplos comunes de rasgos extraídos son la edad, el género, si la lengua empleada es nativa o no, o el tipo de personalidad. En este caso, se ha centrado la atención en determinar la clase de edad o género a la cual pertenecen una colección de textos agrupados por autor.

La idea subyacente del Author Profiling es aprovechar el sociolecto, también llamado dialecto social o registro) empleado por una persona (el autor) al redactar un texto para determinar a qué grupo de edad y género

pertenece. El sociolecto hace referencia a la manera en la cual se comparte el lenguaje [1].

En cada situación, una misma persona emplea un registro determinado. Este registro tiene una parte pasiva y una parte activa. Para ser más precisos, hay una parte que se adquiere de manera pasiva mediante el contacto con la gente con la cual nos relacionamos habitualmente en nuestro entorno más cercano. Sin embargo, hay una compenente de aprendizaje activo de unas maneras de hablar y escribir que constituyen una elección consciente con el propósito de identificarnos con un grupo particular [2].

2 Dataset

Los datos empleados en este trabajo provienen del dataset que fue base de la tarea de Author Profiling de PAN-2013. Así pues, se ha tomado como dataset solamente el corpus de training (entrenamiento) en idioma español. Una serie de características globales del dataset son:

- Colección con miles de autores
- Obtenido de redes sociales
- Gran variedad de temas
- Gran dificultad para etiquetar la información
- Discriminar personas reales de chatbots
- Idioma: Español

Los criterios a la hora de seleccionar qué información se incluyó en el dataset fueron los siguientes:

- Posts agrupados por autor
- Se mantiene autores con pocos posts
- Se parte en chunks los autores con más de 1000 palabras
- Se equilibra por sexo
- Grupos de edad no equilibrados. La edad ha sido categorizada en tres clases: 10S (13-17), 20S (23-27), 30S (33-47).
- Partición aleatoria en tres datasets: Training, Early Bird (10%), Test (+20%). Se ha utilizado solo la parte de test en este trabajo.

Una exploración estadística de los datos revela que consta de 8160 registros distribuidos por edad y género como se muestra en la tabla 1.

Table 1: Distribución de casos del dataset

Edad	Género	Número Casos
10S	Hombre	144
10S	Mujer	144
20S	Hombre	2304
20S	Mujer	2304
30S	Hombre	1632
30S	Mujer	1632

De esta información se confirma los criterios de selección. Por tanto, el dataset está balanceado a nivel de sexo pero no a nivel de edad. Hay el mismo número de casos para cada sexo en cada rango de edades, mientras que hay diferente número de casos para las edades. Este desequilibrio en edad puede ser causa de sesgo hacia los rango 10S y 20S cuando se apliquen los algoritmos de aprendizaje.

La información de cada autor, es decir, la colección de sus post viene almacenada en un fichero XML. En dicho fichero, el autor está identificado por un código HASH-MD5. En un fichero adicional, denominado truth-es.txt se encuentra la equivalencia entre cada código de autor y los valores de edad y género. Esta información es necesaria cuando se apliquen los algoritmos de aprendizaje supervisado. Por lo tanto, se agrega al fichero (.arff) que se le pasa a WEKA, la herramienta de aprendizaje automático.

3 Propuesta del alumno

La solución propuesta consta de un BOW (bag-of-words o bolsa de palabras) y de una serie de características. A partir del BOW, las características y los valores reales de edad y género se obtiene un par de ficheros .arff (uno para edad y otro para género). Estos serán la entrada de la herramienta WEKA de aprendizaje automático. En el fichero true-es.txt, se asocia identificador HASH-MD5 de autor con los valores reales de género y edad. El formato es: id_autor:::género:::edad. Por lo tanto, se ha podido calcular el índcie discriminante porque se conocía los rasgos del autor que escribió cada fichero de posts. Por ejemplo, una línea de este fichero es 5684bc2274efb649d955f427036865a1:::female:::20s.

Así pues, se ha empleado un BOW pero se han escogido las palabras que lo componen de una manera diferente. En el baseline se tomaba las NTERM palabras más frecuentes en la colección de textos. NTERM es una variable que tomaba el valor 1000 por defecto.En la solución propuesta se ha escogido, en un primer momento, las palabras que aparecían más de 200 veces en total. Una vez hecho esto, se obtiene un BOW original de unas

800 palabras. Entonces, se procesa el fichero y se calcula un índice discriminante de género que relaciona la cantidad de veces aparece cada palabras seleccionada en textos escritos por un hombre con el total de veces. Este índice tiene un rango de valores entre 0 y 1. De esta forma, las palabras con un índice más próximo a 1 son consideradas como las palabras más discriminantes de género masculino. En otras palabras, son las más frecuentes en textos escritos por un autor de género masculino y, por tanto, las que más podrían ayudar a clasificar al autor como de este género. En el lado opuesto, las palabras del BOW original con un índice más próximo a 0 son las más proclives a clasificar al auto como de género femenino puesto que aparecen menos en textos escritos por autores de género masculino.Finalmente, se toma del BOW original las 200 palabras con mayor índice y las 200 con menor índice para formar un BOW procesado. Este será el conjunto de palabras que representarán al dataset tanto en la tarea de clasificar los autores por género como por edad. El BOW procesado contendrá 400 palabras con el número total de ocurrencias en la colección y su indíce discriminante. Así pues, el formato final del fichero será: palabra:::num_ocurrenciads:::índice.

Una vez se dispone del BOW procesado se le añade el valor para cada autor de una serie de 13 características que se supone ayudarán a mejorar el rendimiento de la clasificación de los autores en una clase de edad y un edad de género. La elección de estas características se basa en el concepto se sociolecto, es decir, en la manera de emplear el lenguaje. Las características son las siguientes:

- ratio-commas: Uso de comas.
- ratio-points: Uso de puntos.
- ratio-character-flooding: Character flooding.
- ratio-diff-word: Número de palabras diferentes entre el total de palabras.
- ratio-emoticon: Emoticionos. Se ha tomado del campo alt de la marca HTML para insertar imágenes (IMG). Generalmente es un texto alternativo a la imagen y se emplea ':)'.
- ratio-emphasis: Uso de marcas HTML de énfasis como EM o STRONG.
- ratio-laugh: Risas (jaja, jeje,...)
- ratio-numbers: Proporción de números que aparecen entre el total de palabras de los textos.
- ratio-accents: Uso de acentos.
- exclamation-flooding: Repetición de símbolos de exclamación (!!!, ???, ?!?!?).

- email: Indica si en el texto aparece o no una dirección de correo electrónico.
- uses-at: Indica si se usa la arroba (@) en en alguna palabra para indicar género indeterminado (p.ej. chic@s).
- url: En un valor binario que indica si aparece (1) o no (0) alguna dirección web en el mensaje. Indica que se ha añadido un enlace a otra página.

Los ratios son proporciones entre el número de veces que aparecen entre el total de palabras de los textos de un autor. En cuanto a las justificaciones a la hora de seleccionar estas características para representar a la colección de textos y ayudar a la clasificación de los autores por edad y género, se hicieron una serie de suposiciones en función del sociolecto de cada clase a predecir:

- Se supuso que los grupos de menor edad tienden a usar menos la puntuación y a enfatizar más. Así pues, se supuso que emplearían menos las comas (ratio-commas), los puntos (ratio-points) y los acentos (ratio-accents) y que tienden a alargar las palabras (character flooding), a emplear marcas de énfasis (ratio-emphasis), risas (ratio-laugh), muchas exclamaciones (exclamation-flooding) y emoticonos (ratio-emoticon) para describir sus emociones. Además, también suelen emplear un conjunto menos rico de vocabulario (ratio-diff-word) son menos estrictos a la hora de escribir según la norma y tienden a acortar las palabras o a variar su grafía con signos como la arroba (uses-at) o números (ratio-numbers).
- Se supuso que las mujeres tienden a expresar más sus emociones y estado de ánimo que los hombres. Por lo tanto, también tenderán más usar emoticonos (ratio-emoticon), alargar vocales (character-flooding), emplear muchas exclamaciones (exclamation-flooding) y compartir enlaces (url). También, tienden más a repetir ideas y, con ello, palabras, luego el vocabulario utilizado es menor (ratio-diff-word).

Por último, cabe describir la estructura de los archivos .arff que se emplean como entreda de los algoritmos de aprendizaje en la herramienta WEKA. Estos ficheros constan de una primera parte de atributos y una segunda parte de datos. En la parte de los atributos, aparecen las 400 palabras del BOW procesado con su tipo, las 13 características seleccionadas y 1 atributo de clase. Este último es el valor real de cada autor: la 'edad' para el fichero para clasificar al autor por edad y el género para clasificarlo por género. El valor real es necesario para el entrenamiento del algoritmo de aprendizaje supervisado. Es más recomendable para la tarea de Author Profiling usar aprendizaje supervisado puesto que es más sencillo evaluar su calidad [4].

A continuación, se muestra una serie de líneas del fichero .arff para la clasificación del autor por edad:

- Primero, se especifica el tipo de atributo y los atributos. En primer lugar, las 400 palabras del BOW procesado con su tipo

  @relation 'BOW'

  @attribute 'term-seiya' real

  @attribute 'term-as' real

  @attribute 'term-obra' real

  @attribute 'term-sonreir' real

  @attribute 'term-mexico' real

  ... (así hasta 400 palabras)

- Al final de la parte de atributos, se ha especificado las características seleccionadas y, finalmente, la clase a identificar con nombre 'class' y los valores que puede tomar.

  @attribute 'ratio-commas' real

  @attribute 'ratio-points' real

  @attribute 'ratio-character-flooding' real

  @attribute 'ratio-diff-word' real

  @attribute 'ratio-emoticon' real

  @attribute 'ratio-emphasis' real

  @attribute 'ratio-laugh' real

  @attribute 'ratio-numbers' real

  @attribute 'ratio-accents' real

  @attribute 'exclamation-flooding' real

  @attribute 'email' real

  @attribute 'uses-at' real

  @attribute 'url' real

  @attribute 'class' {10S, 20S, 30S}

- En segundo y último lugar, la parte de datos en la cual, para cada autor, se especifica el valor que tienen sus textos agrupados para todos los atributos: palabras del BOW, características seleccionadas y clase edad. Por ejemplo, la primera fila corresponde a un solo autor y tiene los siguientes valores

@data

0.0, 0.0, 0.0, 0.0, ... (así hasta 400 palabras), 0.0, 0.0546875, 0.16666666666666666, 0.0,0.5416666666666666,0.020833333333333332, 0.0026041666666666665, 0.0026041666666666665,0.0,0.0625,1.0,0.0,0.0, 0.0,20S

- El significado de las líneas anteriores para los textos del autor es que las 4 primeras palabras del BOW procesado ('seiya', 'as', 'obra', 'sonreir') no aparecen, los siguientes 13 números corresponden a los valores de la características seleccionadas y, por último, la edad real del autor está en el rango 20S (23-27 años).

En el repositorio [5], se encuentra todos los archivos del proyecto así como esta memoria del trabajo.

4 Resultados experimentales

Una vez se dispone de los ficheros .arff se carga uno en la herramienta WEKA y, en el apartado Classify, se selecciona los métodos de clasificación. En este trabajo se ha escogido dos para comparar los resultados del baseline con los de la solución propuesta.

Los metodos escogidos fueron Naïve Bayes y Random Tree. En ambos casos, se escogió la validación cruzada con 10 pliegues como opción de test. Los resultados para ambos algoritmos y ambos rasgos (edad y género) para el baseline se muestran en la tabla 2.

Table 2: Porcentaje de acierto para el baseline

	Género	Edad
Näive Bayes	54.88%	27.13%
Random Forest	58.97%	61.20%

Para la solución propuesta, se muestran los resultados en la tabla 3. Se puede observar como los resultados de la solución propuesta mejoran a los del baseline, especialmente en cuanto a la edad y género para el algoritmo Naïve Bayes.

Table 3: Porcentaje de acierto para la solución propuesta

	Género	Edad
Näive Bayes	58.96%	46.53%
Random Forest	60.44%	61.42%

5 Conclusiones y trabajo futuro

La BOW calculada introdujo como novedad el cálculo de un índice discriminante de género. Los resultados de género mejoraron, tanto por la introducción de esta nueva BOW como por la inclusión de la características. Aunque esta BOW se pensó inicialmente para mejorar la precisión (accuracy) en la clasificación del género, también se obtiene una gran mejora si se aplica al problema de la edad. Así pues, se decidió emplear la misma BOW para ambas tareas.

Los resultados entre algoritmos de aprendizaje son mejores para Random Forest, especialmente en el caso de la edad. Como ventaja del algoritmo Naïve Bayes se puede apuntar que es muy rápido y obtiene resultados similares para el caso de género.

El algoritmo de Random Forest es mucho más lento. Por ejemplo, para procesar en WEKA ambos algoritmos con los ficheros de 8160 autores y 414 atributos, se tarda unos 15 segundos con Naïve Bayes por los 7 minutos y medio que se tardó con Random Forest. En ambos casos, con validación cruzada con 10 pliegues. Con Näive Bayes se tarda medio segundo en calcular el modelo, mientras que con Random Forest se tarda unos 42 segundos.

Además, el consumo de memoria es mucho mayor en el caso de Random Forest. Si se ejecuta dos veces seguidas un algoritmo Random Forest, se consume la memoria y la herramienta WEKA se cierra si no se tiene la precaución de borrar del buffer de resultados los resultados de la primer ejecución.

Por lo tanto, Naïve Bayes sería más indicado para aplicaciones donde el tiempo de respuesta debiera ser bajo, como en tiempo real, y se tolerase cierto error. En tal caso, habría que tener en cuenta el peso de los errores de predicción. Así pues, sería capaz de calcular mientras se están generando los posts o refrescarse cada minuto. Random Forest estaría más indicado para tareas batch, donde la rapidez de predicción no fuera crítica pero si se requireise de una mayor precisión.

En un futuro, se podría plantear el aumentar el número de palabras en el BOW para tener más datos a la hora de entrenar el algoritmo y mejorar los resultados. Otra opción, aunque se perdiese autores sería no guardar los que tiene pocos textos. O todo lo contrario, ir actualizando el BOW conforme se obtengan más texto de los autores. Esta tarea se podría realizar una vez al día, cuando menos actividad se registre en la red social. De este modo, se podría ir mejorando progresivamente la capacidad del índice discriminante de género y, con ello, la del BOW.

Por último, atendiendo a las muchos ratios empleados se puede deducir que no va a mejorar mucho más las prestaciones si se introduce aún más nuevas características. La mejora podría ir más en la línea de mejorar el BOW, como se ha comentado anteriormente, o emplear otro enfoque. Otra manera de abordar la tarea del Author Profiling sería no emplear un BOW

calculado a partir de los textos sino diccionarios con expresiones coloquiales (slang). También, se podría complementar el BOW de la solución propuesta con diccionarios para algún aspecto en concreto del sociolecto, con el fin de mejorar las prestaciones en el cálculo de las características seleccionada. Otra opción, sería clasificar según n-gramas en vez de palabras completas para poder tratar mejor el tema de las terminaciones de género o las repeticiones de caracteres (character-flooding).

References

[1] Website oficial de la competición PAN-2013, sección de la tarea de Author Profiling http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html

[2] 'Sociolect', Wikipedia, artículo recuperado el 21 de julio de 2015, https://en.wikipedia.org/wiki/Sociolect

[3] 'Data Mining with WEKA', https://weka.waikato.ac.nz/explorer

[4] 'Author Profiling y La Vanguardia', Francisco Rangel, artículo recuperado el 21 de julio de 2015, http://www.kicorangel.com/2015/05/04/author-profiling-y-la-vanguardia/

[5] https://github.com/VictorBD/tm2_ap