Asignatura Text Mining en Social Media. Máster en Big Data Analytics

Fátima Selma Juan
fatimaselju@gmail.com

Abstract

Como tarea evaluable de la asignatura de Text Mining en Social Media del programa del Máster en Big Data Analytics de la UPV, se propone realizar una práctica relativa a Author Profiling en Social Media. Este resumen muestra el marco del trabajo desarrollado y los resultados obtenidos.

La práctica propuesta se centra en abordar dos características concretas del autor: su género y la identificación de su variedad de idioma. La hipótesis de partida del estudio se basa en que los usuarios se expresan de manera distinta en función de sus propias características, como pueden ser sexo y variedad del idioma español en este caso. ¿Existe algún tipo de relación entre la forma en la que articulamos nuestro discurso y expresamos nuestras emociones, con nuestra edad o género?

Para este propósito, se dispone de un corpus o dataset obtenido de Twitter, con una colección de textos de miles de autores y cientos de tuits por autor, de una gran variedad de temas, etiquetados por sexo (2 clases) y por variedad de lenguaje (7 clases).

El presente documento mostrará la descripción del dataset, la propuesta desarrollada en el taller, así como el planteamiento de futuros desarrollos.

1 Introducción

Como primer paso, debemos introducir el concepto de Author Profiling. Recurriremos para ello a la propia introducción al problema expuesta por Francisco Rangel en su tesis doctoral:

"La posibilidad de conocer rasgos de una persona a partir únicamente de los textos que escribe se ha convertido en un área de gran interés denominada author profiling. Ser capaz de inferir de un usuario su sexo, edad, idioma nativo o los rasgos de su personalidad, simplemente analizando sus textos, abre todo un abanico de posibilidades desde el punto de vista forense, de la seguridad o del marketing.

Además, la proliferación de los medios sociales, que favorece nuevos modelos de comunicación y relación humana, potencia este abanico de posibilidades hasta cotas nunca antes vistas. La idiosincrasia inherente a estos medios sociales hace de ellos un entorno de comunicación especial, donde la libertad de expresión, la informalidad y la generación espontánea de temáticas y tendencias propician el acercamiento a la realidad diaria de las personas en su uso de la lengua. Sin embargo, esa misma idiosincrasia hace que en muchas ocasiones la aplicación de técnicas lingüísticas de análisis no sea posible, o sea extremadamente costoso."

En resumen, diferentes personas tienden a escribir de manera diferente. Además, el estilo de su escritura puede describir características del autor, como son edad, sexo, personalidad, idioma nativo, variedad del lenguaje, etc

"El author profiling estudia aspectos psicolingüísticos y sociológicos (cómo se usa el lenguaje y qué rasgos son compartidos por grupos similares) para tratar de determinar, a partir de los textos, aspectos personales de su autor como la edad, el sexo, el idioma nativo o los rasgos de su personalidad. El interés es evidente desde perspectivas como la forense, donde ser capaz de conocer el perfil lingüístico de un mensaje de texto sospechoso (lenguaje utilizado por cierto tipo de personas) e identificar características (lenguaje como evidencia), ciertamente podría ayudar a atribuir su autoría; o desde el punto de vista de la mercadotecnia, lo que proporcionaría a las compañías la capacidad de seg-

mentar su mercado en base, por ejemplo, al sexo, la edad o la región a la que pertenecen los usuarios que opinan de sus productos."

Como vemos, el proceso de perfilado del autor es de gran importancia: desde el punto de vista del marketing, las empresas pueden estar interesadas en conocer el perfil demográfico de sus seguidores en una red social para lograr una mejor segmentación del mercado; desde un punto de vista forense, determinar el perfil de una persona que escribió un "texto sospechoso" puede proporcionar informacin de fondo capaz.

La hipótesis en que se basa esta metodología parte de que los autores se expresan de manera distinta en función de ciertos rasgos de su persona. Y especialmente cuando hablamos de medios sociales donde no hay censura y prima la libertad de expresión.

En el caso que nos ocupa, nos centraremos en dos características concretas de los autores:

- Género: masculino, femenino
- Variedad del idioma: Argentina, Chile, Colombia, México, Perú, Spain, Venezuela

Se trata pues de un problema de clasificación supervisada, con una variable target binomial (sexo) y otras variable target multinomial (variedad de idioma). Es necesario, pues, disponer de un conjunto de datos de entrenamiento previamente clasificados por estas dos variables para poder entrenar los modelos.

2 Dataset

Se descarga del dataset PAN-AP´17, proporcionado por el profesor, que se construye de la siguiente forma:

1. Se recuperan tuits enmarcados en una región geográfica: longitud, latitud, radio
2. Se preseleccionan los usuarios únicos que han emitido tuits (filtrados por el idioma del perfil)
3. Se recuperan los timelines de los usuarios únicos
4. Se seleccionan los autores con más de 100 tuits (que no sean retuits) en:
    - el idioma correspondiente
    - con la localización geográfica esperada en su perfil
5. Se revisan manualmente los perfiles para asegurar el sexo
6. Se seleccionan 100 tuits por autor para la construcción del dataset final

La estructura es la siguiente:

- un par de ficheros training.txt y test.txt.
- El formato es: id, sexo, variedad.
- Un fichero json por autor: cada línea del fichero es un tuit en formato xml.

Exploraremos los aspectos que nos interesan del dataset, cualquier información que nos describa el conjunto de datos y aporte conocimiento como:

- Número de autores por clase (sexo y variedad del lenguaje).
- Número de tuits por autor.
- Número de tuits por clase (sexo y variedad del lenguaje).
- Número de palabras por documento / autor / clase.
- Distribución de palabras/documentos/autores por documento/autor/clase.
- Longitud media de los tuits, palabras, documentos, total y por clase.
- Distribución temporal de los tuits, tuit más antiguo, más nuevo, media, desviación,
- Palabras extrañas, frecuentes, comunes, etc...

El corpus de entrenamiento contiene 8.200 archivos XML, y el de test 1.400 archivos, cada uno perteneciente a un autor en particular, como conjunto de test. Cada archivo XML contiene los tuits publicados por cada autor.

Disponemos de 2.800 autores de los que:

- Sexo: 1.400 autores para cada clase, 2 clases.
- Variedad de idioma: 200 autores para cada clase, 7 clases.

Las probabilidades a priori para cada clasificador serían pues:

- Sexo: 0.5
- Variedad de idioma: 0.2

Uno de los primeros pasos del proceso consiste en cargar los tuits, desde los ficheros xml, asociados a cada autor etiquetado por sexo y variedad de idioma, en una tabla en R. Realizamos una vista de las primeras muestras del dataset y un análisis descriptivo básico.

3 Propuesta del alumno

Una vez se dispone del corpus etiquetado por género y variedad de idioma, se plantea como un modelo de clasificación supervisada.

Disponemos de dos tipos de clasificadores:

1. Clasificador binario para género
2. Clasificador de etiquetas múltiples o multinomial para la variedad de idioma

Inicialmente se proponen dos clasificadores independientes, sexo y variedad del idioma. Pero también se puede proponer un clasificador cruzado por género y variedad del lenguaje, de forma conjunta, con lo que el clasificador tendría 2x7=14 clases.

Centramos inicialmente nuestra atención en los siguientes aspectos:

- Bolsa de palabras: frecuencia de aparición. Entrenamos los modelos con diferentes tamaños de la bolsa de palabras. Nuestra hipótesis inicial es que el tamaño de la bolsa de palabras puede ser más relevante para el clasificador de variedad de idioma, y no lo es tanto para el género.
- Analizar el efecto que puede tener el tamaño de la bolsa de palabras en el rendimiento de los algoritmos de aprendizaje.
- Algoritmos de clasificación: compararemos distintos algoritmos como SVM, Regresión Logística, knn y Ranger (implementación rápida de Random Forests, particularmente adecuada para datos de alta dimensión).
- Ngramas: se propone el uso de estas estructuras que creemos mejorará la predicción, al menos se pretende probar con los bigramas, sobre todo en el clasificador de variedad de idioma.
- k-fold cross-validation: uso de cross-validation, se probará con diferentes valores del parámetro k.

A partir de un corpus dado, se obtienen las n palabras más frecuentes, aplicando distintos tipos de preprocesado como pasar a minúscula, eliminar números, palabras vacías, etc. Dado un vocabulario y un conjunto de datos, se obtiene su representación basada en bolsa de palabras por frecuencias relativas.

Adicionalmente se programan nuevas variables predictoras que consideramos que ayudarán a mejorar la predicción. En los resultados comprobaremos si esto es así. Las nuevas variables creadas son:

- N° total de palabras por tuit: nuestra hipótesis inicial es que las mujeres utilizan más palabras. Para la variedad de idioma consideramos que no será muy relevante.
- N° de palabras distintas por tuit: variedad del vocabulario de un autor, pensamos que puede ser un parámetro relevante en ambos clasificadores.
- Distancia de cada autor a cada clase: Se calculan el mínimo, máximo y media del n° total de palabras por tuit para cada clase del clasificador género y variedad de idioma. Se calculará la distancia de cada autor al mínimo, máximo y media de cada clase. Con esto, dispondremos de 6 variables adicionales por autor, 3 por cada clase. En el caso de la variedad de idioma serán 3x7=21 variables nuevas.

4 Resultados experimentales

La medida de evaluacin utilizada ha sido **accuracy**, calculada como el porcentaje de casos correctamente clasificados frente al total de casos.

A continuación mostramos los resultados obtenidos en cada ejecución, junto con los parámetros aplicados en cada caso:

n	Model	GENDER	VARIETY	JOIN
10	SVM	0,526	0,196	0,099
50	SVM	0,666	0,376	0,252
100	SVM	0,678	0,489	0,326
500	SVM	0,709	0,758	0,539
1000	SVM	0,659	0,782	NA
1000	**Ranger**	**0,7164**	**0,885**	**0,629**
100	RegLog	0,673	NA	NA
500	RegLog	0,706	NA	NA

Con estos parámetros, la combinación ganadora para ambos clasificadores es:

- bolsa de 1.000 palabras
- algoritmo Ranger
- Accuracy(Gender) = 0.716
  Accuracy(Variety) = 0.885

A partir de esta combinación ganadora, procedemos a añadir nuevas variables al modelo para intentar mejorar el ajuste:

- Con la variable n° total de palabras por tuit, para el clasificador Sexo, se obtiene un accuracy de **0,719**, algo superior al anterior y mejora sensiblemente la preddiccin.
- Con las variables de distancia al mínimo, máximo y media de cada clase, para el clasificador Sexo, se obtiene un accuracy de **0,723**, algo superior de nuevo con lo que se vuelve a incrementar la predicción.

Por lo tanto, podemos concluir que si se consideran nuevas variables derivadas, sí se mejora el grado de ajuste del modelo.

En cuanto al tiempo de ejecución de los algoritmos, se observa un creciminento exponencial del tiempo de procesamiento en función del tamaño de la bolsa de palabras:

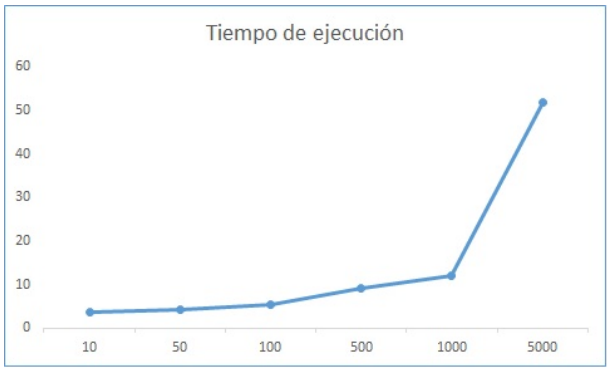

Figure 1: Tiempo y bolsa de palabras

5 Conclusiones y trabajo futuro

Análisis de los resultados obtenidos:

- El mejor accuracy se obtiene con una bolsa de 1.000 palabras y el algoritmo Ranger, para ambos clasificadores.
- Se observa que el incremento de la bolsa de palabras tiene un efecto un importante en los resultados obtenidos. Este efecto es más acusado en el caso del clasificador, como podemos ver en el siguiente gráfico:

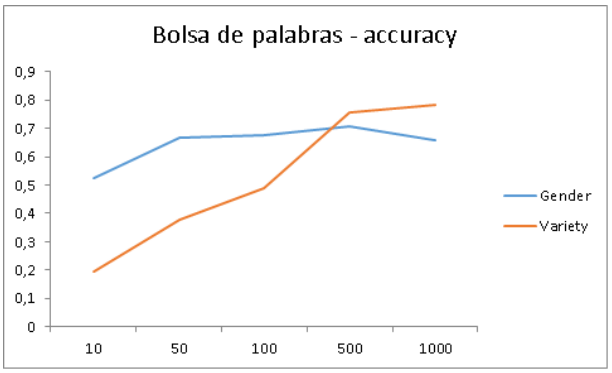

Figure 2: Accuracy Algoritmo SVM

- El clasificador conjunto presenta peores ajustes.
- Añadir nuevas variables derivadas que describen las clases incrementa el ajuste del modelo.

Algunas ideas de futuro para mejorar los resultados:

- Diccionarios: uso de diccionarios, términos fuera del diccionario
- Bolsa de palabras por clase
- Análisis semántico: sintagmas
- Uso de term frequency-inverse document frequency (tf-idf), frecuencia de término y frecuencia inversa de documento.
- Ngramas: el uso de estas estructuras creemos que debe mejorar la prediccin, sobre todo en el clasificador de variedad de idioma.
- Algoritmos de clasificación: probar diferentes Kernel en SVM y otros algoritmos de clasificación como ExtraTree Clasifier, Naïve Bayes o Redes Neuronales.

- Añadir información de uso de símbolos o emoticonos.
- Análisis de subperfiles: estudiantes, amas de casa, etc...
- Equilibrar el corpus por clase, siguiendo una distribucin realista del uso de Twitter por cada clase.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.