Author profiling in Social Media. Máster en Big Data Analytics

Lorena Muñoz Albors
lomuoal@masters.upv.es

Abstract. En este documento se hace una breve descripción del proyecto final de la asignatura de *Text Mining in Social Media* del Master en *Big Data*. Dicho proyecto se ha basado en la utilización del *Machine Learning* como herramienta para distinguir el género y la procedencia demográfica de los autores de los tweets que han sido escogidos como muestra de datos. A partir de un *baseline* de 66% de precisión en predicción de género y un 77% en diversidad, se generan diferentes modelos, siendo el mejor de ellos el que alcanzaba un 70,5% de precisión en género y un 80,7% en diversidad.

1 Introducción

A pesar del impacto significativo de las redes sociales, observamos una falta de información sobre quienes crean los contenidos. Con *Machine Learning* ahora es posible crear un perfil de autores basado en sus escritos que nos permite conocer las características de los mismos, es lo que se conoce como *Author Profiling*.

Hoy en día, *Author Profiling* es de creciente importancia en una variedad de áreas, incluyendo marketing, seguridad y forense. Por ejemplo, hay muchas empresas interesadas en analizar si a la gente le gusta o no les gusta sus productos. Y con esa información, obtienen una primera respuesta de sus clientes y, por lo tanto, pueden adaptar su estrategia con respecto a este producto.

En este proyecto nos hemos centrado en analizar el género y la procedencia geográfica de los autores de los tweets escogidos como muestra, de entre las múltiples variables que se podrían haber analizado. Dada la nula correlación entre las variables estudiadas, se van a analizar como si de dos problemas distintos se tratasen, dando lugar, por tanto, a diferentes modelos.

2 *Dataset*

El *dataset* utilizado en este proyecto ha sido facilitado por PAN, la cual es una organización que se dedica al análisis de los textos y que ha creado una comunidad internacional para ello.

La fuente de información de la que se ha extraído la muestra de datos ha sido Twitter, escogiendo datos de 2017 y recopilando un total de 50.378 autores, siendo 24.429 mujeres y, el resto, 25.949, hombres. Así, de manera porcentual, las mujeres representan un 51,50% de los autores seleccionados, mientras que los hombres, un 48,5%.

De dicha muestra también conviene destacar que son cuatro los idiomas utilizados por los autores, encontrándose el español, el portugués, el inglés, y el árabe. Dado que en este proyecto nos centramos en la comunidad hispanohablante, debemos descartar los autores que utilizan el resto de idiomas. Por lo que, tras este descarte inicial, los autores pueden proceder de los siguientes países: Argentina, Chile, Colombia, México, Perú, España y, por último, Venezuela.

3 Propuesta: pre-procesamiento y modelos

En la muestra de datos, a simple vista, se observan un conjunto de elementos que podrían dificultar el análisis de texto. Por ello, se ha procedido a realizar una limpieza de los mismos mediante la eliminación de enlaces, emoticonos y signos de puntuación y de exclamación.

Como elementos que también podrían obstaculizar el análisis, encontramos los acentos y las palabras con vocales repetidas, en cuyos casos, se ha procedido a la sustitución de la vocal por otra que no se encuentre en ninguno de los supuestos anteriores.

Posteriormente, se ha realizado un pre-procesamiento de texto, para obtener un conjunto de datos con el que los algoritmos de *Machine Learning* pueda operar y obtener buenos resultados.

La técnica de pre-procesamiento de texto recibe el nombre de *Stemming* y consiste en mantener tan solo las raíces de las palabras para poder encontrar, a través de ellas, las familias de palabras.

Así obtenemos una *Bag of Words* o lista de raíces de palabras más frecuentes que vamos a poder utilizar para conocer el género y la procedencia geográfica de nuestros autores.

Por último, de los tres algoritmos de *Machine Learning*, que explicamos a continuación, elegiremos el mejor como resultado final:

- **Random Forest**:

  Es una combinación de árboles predictores, dependiendo cada uno de ellos de un vector aleatorio.

  Se trata de uno de los algoritmos con mayor certeza, dado que para una muestra de datos muy grande, da lugar a un clasificador muy certero, ya que cuenta las proximidades entre los pares de casos, localizando los valores atípicos o dando vistas interesantes de los mismos.

  Pero en el caso de que los datos contengan atributos correlacionados, los grupos más pequeños resultarán favorecidos respecto a los grupos grandes.

- **Regresión por mínimos cuadrados**:

  Dados un conjunto de pares ordenados, consiste en encontrar la función continua dentro de una familia de funciones, la cual debe aproximarse a los datos, utilizando para ello el criterio de mínimo error cuadrático.

  Esta técnica exige, además, que los errores de medida estén distribuidos de manera aleatoria.

- ***Support Vector Machines* (SVM)**:

  El modelo que se obtiene con esta técnica representa puntos de muestra en el espacio, separando las clases a dos espacios, lo más amplios que sean posibles, mediante un hiperplano de separación llamado vector entre los dos puntos o vector soporte.

  Las muestras, según el espacio en el que se encuentren, pertenecerán a una clase u otra, y esta es la base de este algoritmo que trata de encontrar el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo.

  Así, se permite clasificar a los puntos del vector según se encuentren en un lado u otro del hiperplano.

4 Resultados experimentales

Una vez ya hemos aplicado las técnicas ya explicadas, hemos superado el *baseline* propuesto para el estudio: 66,43% en distinción de género y 77,21% en procedencia geográfica.

- **Género**:

  Aplicando los tres algoritmos, ya explicados, hemos podido extraer las siguientes conclusiones:

  - Se ha alcanzado un **70,5%** de aciertos sobre el total usando ***Random Forest***.
  - Se ha alcanzado una precisión de **67,3%** utilizando la técnica de **Regresión por mínimos cuadrados**.
  - Se ha obtenido un **66,71%** de aciertos sobre el total con la técnica de **SVM**.

- **Procedencia geográfica**:

  Dado el escaso tiempo, en esta variable solo hemos podido aplicar un modelo, el de ***Random Forest***, elegido por su alta eficiencia y sencillez. Con esta técnica hemos obtenido un **80,7%** de aciertos sobre el total de los supuestos.

5 Conclusiones y trabajo futuro

Tras el estudio realizado y ya explicado, y teniendo en cuenta los resultados finales obtenidos para ambas variables, se puede llegar fácilmente a la conclusión de que el mejor algoritmo ha resultado ser el *Random Forest* y, por tanto, el elegido para el presente estudio.

En cuanto a los aspectos a mejorar en el futuro, convendría aplicar los diferentes algoritmos para la variable de procedencia geográfica y así poder comprobar, de manera certera, qué algoritmo da mejor resultado y eficiencia para dicha variable.

Otro de los aspectos que se plantean es la extracción de más características acerca del autor con la elaboración de un *web crawler*.

Por último, dado que ambas variables han sido tratadas como problemas no relacionados y distintos, y como propuesta para obtener un resultado en la variable de género que pudiese ser utilizado no solo en la comunidad hispanohablante, se podra eliminar, tan sólo para dicha variable, la segregación inicial que se ha realizado sobre los autores que no utilizan la lengua española. De este modo se podría obtener un resultado de la variable de género que podría ser usado globalmente, dada la muestra de datos que se habría escogido.

References

JKhalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX* , 6, 98-106.

Dichiu, D., & Rancea, I. (2016). Using Machine Learning Algorithms for Author Profiling In Social Media. *In CLEF (Working Notes)* (pp. 858-863).

Mechti, S., Jaoua, M., Belguith, L. H., & Faiz, R. (2014). Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media. *Proceedings of the PAN@ CLEF*, Sheffield, England.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26-32.

PAN 2017, Author Profiling `http://pan.webis.de/clef17/pan17-web/author-profiling.html`.