PAPER. Asignatura Text Mining en Social Media. Master en Big Data Analytics

Alberto García García
algar11@alumni.upv.es

Abstract

El presente proyecto trataba de abordar dos temas claramente diferenciados, por una parte se quería predecir cuando un tweet había sido escrito por un hombre o por una mujer y, por otro lado, se pretendía predecir el país de procedencia dentro del habla hispana. Para abordar ambas casuísticas se debía realizar un análisis a priori de que variables o atributos eran distintivos para poder diferenciar entre clases. Obviamente, no es lo mismo intentar trabajar en un problema de predicción del sexo, como en un problema de predicción de la procedencia, es por esto que se realizaron de proyectos completamente diferenciados.

Tratando la opción del sexo se debían establecer unas reglas en base a la forma de hablar, es decir, quien escribe tweets más largos, quien pone más adjetivos, quien escribe más preposiciones. En cambio, en la opción de la procedencia se debían analizar que palabras se utilizaban mucho en un país y, de ese modo, crear distinciones relevantes para poder diferenciarlos.

Así pues, para resolver el problema de Author Profiling ha sido necesario evaluar ciertas pautas de comportamiento que pueden ser muy relevantes, realizar estudios previos y analizar formas de escribir. Dentro de todos estos aspectos, englobando ambas problemáticas, tendrán mención honorífica las variables que engloban los adjetivos, las preposiciones, emoticonos y bolsas de palabras típicas en un ámbito que no están presentes en el otro. Con todos estos datos y entrenando los modelos adecuados, se podría llegar a predecir con una probabilidad del 70% cuando se trata de una clase u otra.

Introducción

El proyecto Author Profiling consta de dos apartados diferenciados, uno será la predicción del sexo y el otro la predicción del país de procedencia de un tweet escrito por una persona. Puesto que son problemas diferentes se han realizado metodologías de análisis y tratamiento completamente distintas y se han generado las agrupaciones correspondientes. El problema del sexo pretende discernir entre un tweet escrito por un hombre y una mujer, y por otro lado el país de procedencia (siempre teniendo en cuenta que serán países de habla Hispana), entre los que se encuentran Argentina, Chile, Colombia, Mexico, Peru, Spain y Venezuela.

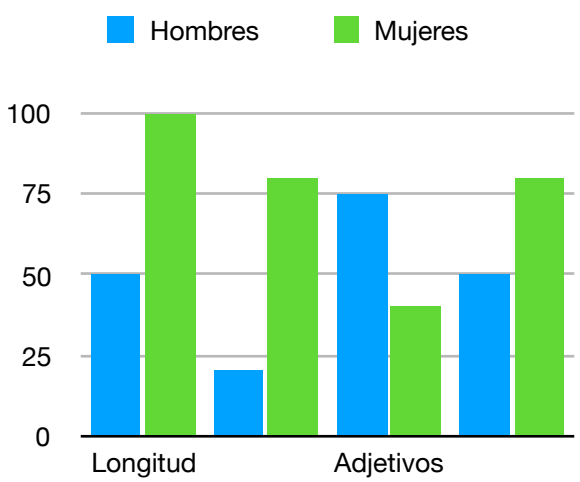
Dataset

Para abordar el proyecto de Author Profiling se disponía de un Dataset obtenido de Twitter de no un gran tamaño (aprox. 50 Mb), en el que se encontraba una colección de tweets de 300 autores, llegando a conseguir 100 tweets por autor, para poder abordar la predicción con un volumen considerable. En el dataset se han podido encontrar una gran variedad de temáticas de conversación, es por ello que el etiquetado de estos se hace especialmente complicado.

Analizando el Dataset se han podido encontrar muestras para entrenar y muestras para evaluar el modelo. Cada fichero muestra los 100 tweets referentes a un usuario en concreto, y cada tweet puede ser de una temática diferente, por lo que se ha conseguido un amplio abanico en la muestra para poder generar todas las casuísticas que se estimen oportunas para su evaluación.

Dentro de los aspectos relevantes que se analizaron en el dataset fueron:

- La longitud del tweet
- La cantidad de emoticonos
- La cantidad de adjetivos
- La cantidad de preposiciones



Por lo que se ha podido observar, los hombres escriben más adjetivos, pero los tweets son más cortos, ponen menos emoticonos y menos preposiciones, así pues, estas reglas se deberán tener en cuenta para el posterior algoritmo de predicción, ya que son propiedades claramente diferenciales.

Propuesta del alumno

En el presente proyecto Author Profiling, se debía abordar, a partir de un Dataset de twitter, el género y la variedad del lenguaje. Por cuestiones de tiempo sólo se implemento la solución de la predicción del sexo, pero se van a analizar las dos casuísticas:

- Sexo

Para llevar a cabo el análisis del sexo, se ha realizado un análisis exhaustivo, como ya se exponía en el apartado anterior, en el que se estimaba unas reglas del lenguaje que presuponían una distinción entre hombres y mujeres. Atributos tales como: Longitud del tweet, cantidad de emoticonos, cantidad de adjetivos masculinos, cantidad de adjetivos femeninos, cantidad de preposiciones y una bolsa de palabras con palabras que aparezcan en un sector y no en el otro y viceversa.

Para estimar todas estas variables como opciones relevantes a la hora de generar el algoritmo, se ha tenido en cuenta que, los hombres utilizan más adjetivos (aunque lo que se pretendía era establecer una regla en la que si el adjetivo terminaba en a/as o en o/os se catalogara como femenino o masculino, por otro lado, se ha observado como las mujeres empleaban más preposiciones, más emoticonos y sus tweets eran más largos.

Por último se pensó que podría ser de gran relevancia que se propusieran como bolsa de palabras aquellas que, aparezcan en un bloque y no en el otro, ya que nos ayudaran a discernir en el sexo.

- Variedad del lenguaje

En cuanto a la variedad del lenguaje como se mencionaba anteriormente no se ha podido implementar, pero se estimaba oportuno generar bolsas de palabras especificas por país, en el que la palabra que estuviera en un país no se repitiera en otro, es decir, fuera una muestra única. Por otro lado, se debería de haber analizado también la longitud de los tweets o la cantidad de emoticonos, para saber si globalmente algún país los emplea más que otro.

Resultados experimentales

Antes de de abordar los resultados reales obtenidos, se deberá de hablar de las hipótesis que se han ido generando, y como se ha llegado a la solución final.

Como ya se había descrito anteriormente, simplemente habrá resultados para la predicción del sexo, inicialmente se estimó como relevante, únicamente las variables de, longitud de tweet, cantidad de emoticonos, cantidad de preposiciones y cantidad de adjetivos masculinos y femeninos. Aún siendo variables muy relevantes y con mucha fuerza, se trataba de un modelo entrenado con muy pocos atributos para estimar, por lo que los resultados iniciales no fueron los esperados, la precisión obtenida rondaba el 60%, por debajo del baseline con simplemente la bolsa de palabras obtenida, claramente si se analizaba que únicamente se habían empleado 5 atributos, el resultado es fantástico, por lo que se siguió el proceso de análisis y se llego a la conclusión que se deberían incluir bolsas de palabra referentes a datos que aparecen en un sector y no en el otro.

Para ello se procedió a crear la bolsa de palabras y se incluyeron en el modelo, viendo que mejoraba sustancialmente hasta llegar a un 70% de precisión aproximadamente, lo que ya colocaba el resultado por encima del baseline esperado.

Cabe destacar que el mejor resultado ha sido obtenido con red neuronal, aunque se han entrenado otros modelos como, SVM, Random Forest o Decission Tree.

Las características del modelo entrenado han sido las siguientes:

- Longitud tweet
- Cantidad de emoticonos
- Cantidad de preposiciones
- Cantidad de Adjetivos maculinos y femeninos
- Bolsa de palabras masculina
- Bolsa de palabras femenina

En general, se han creado cerca de 2005 atributos a estimar, y con ello se ha obtenido dicha precisión cercana al 70%.

SVM	68,71
DT	60,71
RF	66,38
NN	72,7

Conclusiones y trabajo futuro

Teniendo en consideración el tiempo empleado en el desarrollo del proyecto, se puede estimar que los resultados son relativamente notables, ya que el resultado de la precisión obtenido para la predicción del sexo esta por encima del baseline a superar. Por contra, y por el mismo motivo, no se ha podido desarrollar la opción para la predicción de la variabilidad del lenguaje.

Como líneas de trabajo futuras cabe destacar dos ramas claramente diferenciadas:

- Por sexo:

Analizar comportamientos para establecer patrones de búsqueda de personas que hablen de si misma como si fuera del mismo sexo.

- Por variabilidad del lenguaje:

La primera opción a tener en cuenta es implementar el algoritmo con unas pocas variables e ir generando pruebas incrementando las variables de diferentes tipos, en concreto, longitud del tweet en las diversas zonas, cantidad de emoticonos, palabras clave de cada zona geográfica, entre otros aspectos relevantes.

Con todos estos datos se debe poder conseguir unos resultados bastante aceptables, que nos proporcionarán una precisión por encima del baseline.

Referencias

Rosso, Paolo y Rangel, Francisco. 2017-2108. Apuntes asignatura Text Mining en Social media. Autoritas