Memoria para la asignatura Text Mining en Social Media. Master Big Data UPV 2018

Joan Buigues Romera
joanbuiguesromera@gmail.com

Abstract

El Author Profilling es una disciplina que nos permite clasificar rasgos identificativos de una persona en base a un texto escrito, sin contar con ninguna informacin adicional. En este caso vamos a tratar de pronosticar la variedad del lenguaje nativo de unos tweets determinados, pertenecientes al corpus del PAN-AP 2017, una competición internacional sobre perfilado de autores. Nos centraremos en crear diccionarios locales de modismos para cada una de las 7 variedades de lenguaje que tenemos identificadas y, aplicando diversos algoritmos de machine learning, trataremos de encontrar el mejor índice de acierto en este problema de clasificación.

1 Introducción

La capacidad de categorizar un texto en base a su contenido, sin contar con ninguna información adicional del autor, resulta muy interesante y útil desde el punto de vista de varias disciplinas. Estas tareas de Author Profiling pueden servirnos para identificar aspectos del autor como la edad, el género, el lenguaje nativo y/o el tipo de personalidad.

En este caso práctico de Author Profiling, vamos a usar como referencia datos de PAN-AP 2017. Se trata de una competición a nivel mundial donde el objetivo es resolver un problema de clasificación para identificar género y variedad del lenguaje de los autores.

Nos centraremos en un corpus reducido de tweets escritos por autores de 7 países distintos de habla hispana, de los cuales deberemos conseguir clasificar a qué país pertenecen los autores y de qué gnero son, en base a lo que han escrito en cada tweet.

Podemos abordar el problema de distintas maneras, aunque contamos con una bolsa de palabras que utilizaremos para que el modelo aprenda, contrastando los textos a clasificar con textos ya clasificados correctamente. Además, enriqueceremos esta bolsa de palabras con modismos locales para probar el desempenyo de distintos modelos de machine learning y quedarnos con el que mejor fiabilidad nos ofrezca.

2 Dataset

El conjunto de datos del que disponemos está dividido en conjunto de test y conjunto de training, en una proporción de 30% - 70%, respectivamente. El formato de presentación se trata de archivos individuales en XML, uno por cada autor, nombrados con el identificador proporcionado por la API de Twitter. Y cada fichero contiene 100 tweets del mismo autor, además del propio identificador del autor y la etiqueta de variedad y la de género.

3 Propuesta del alumno

Después de ejecutar el modelo de base con una bolsa de 100 palabras, hemos visto que el accuracy de la clase variedad tenía más margen de mejora, concretamente 33'83%, frente al 73'75% de accuracy para la clase género. Por tanto, decidimos centrarnos en mejorar la clasificación del aspecto variedad.

- N GENDER VARIETY JOINT TIME
- 10 0.5875 0.2608 0.1442 3.62m
- 50 0.6850 0.3167 0.2142 4.32m
- 100 0.7375 0.3383 0.2525 5.36m
- 500 0.7358 0.5717 0.4175 9.16m
- 1000 0.6983 0.6167 0.4325 12.11m
- 5000 0.7550 0.7275 0.5517 51.81m

- 10000 IMPOSSIBLE, RSTUDIO CRASHES

Para ello, recurrimos a buscar en foros especializados los modismos locales más usados en cada variedad a clasificar: chileno, mexicano, peruano, venezolano, colombiano, argentino y espanyol.

Creamos un diccionario por cada una de las 7 variedades con las palabras más usadas en cada una de ellas. Para ello, hemos tenido que aplicar un pequenyo preproceso, eliminando modismos duplicados. El criterio que seguimos para eliminar estos duplicados fue una comparativa visual del número de variedades donde aparecían modismos iguales, eliminando en cada caso el de la lista con menor número de palabras.

Somos conscientes que esta criba de datos no es la mejor, pero dadas las limitaciones de tiempo, nos pareció la mejor manera para resolver este problema de los modismos duplicados.

Además, tuvimos que hacer una pequenya transformación en estos diccionarios, cambiando todos los modismos para que empezaran por minúscula y adaptando estos nuevos datasets para que las variables tuvieran el mismo nombre y contaran con el mismo número de columnas.

Posteriormente, establecimos la semilla para que los diferentes resultados fueran totalmente comparables.

4 Resultados experimentales

Con estos antecedentes, procedemos a aplicar algunos de los algoritmos de machine learning más contrastados a nuestro vocabulario de 100 palabras.

Comenzaremos con Support Vector Machine, con el que obtenemos los siguientes datos:

- Accuracy : 0.5229
- 95% CI : (0.4963, 0.5493)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

Y a continuacin, probaremos con Random Forest, con el que obtenemos una ligera mejora:

- Accuracy : 0.5393
- 95% CI : (0.5128, 0.5656)
- No Information Rate : 0.1429
- P-Value [Acc ¿ NIR] : <2.2e-16

Tambin probamos con K-Nearest Neighbors, aunque el accuracy empeora considerablemente:

- Accuracy : 0.345
- 95% CI : (0.3201, 0.3706)
- No Information Rate : 0.1429
- P-Value [Acc ¿ NIR] : <2.2e-16

Y con Naive Bayes recuperamos nivel de accuracy, pero seguimos sin llegar al nivel de Random Forest:

- Accuracy : 0.4214
- 95% CI : (0.3954, 0.4478)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

Probamos con el Nearest Neighbour:

- Accuracy : 0.43
- 95% CI : (0.4039, 0.4564)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

Y, por ltimo, probamos un Classification Tree C5.0:

- Accuracy : 0.5071
- 95% CI : (0.4806, 0.5337)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

Como conclusión a esta comparativa de modelos, el mejor accuracy de 53,93% lo obtenemos con el algoritmo Random Forest.

Probamos a modelar con el baseline de 100 palabras contra la lista individual de Chile, pero mediante la matriz de confusión vemos que existe sobreajuste, es decir, clasifica correctamente para Chile pero falla bastante con el resto de países.

- Prediction argentina chile colombia mexico peru spain venezuela
- argentina 74 14 10 7 13 5 1
- chile 9 99 13 6 7 6 6

- colombia 4 3 9 3 5 2 5
- mexico 17 13 1 27 10 3 2
- peru 5 6 1 1 14 0 3
- spain 78 55 131 134 117 165 134
- venezuela 13 10 35 22 34 19 49

Y los datos de accuracy:

- Accuracy : 0.3121
- 95% CI : (0.2879, 0.3371)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

Ahora procederemos a ampliar la lista de 100 palabras, agregando las listas de todas la variedades. Modelamos con Random Forest el baseline de 500 palabras contra la lista conjunta.

Finalmente incrementamos el número de palabras hasta llegar a cerca de las 1000, con 500 de la bolsa de palabras inicial ms 416 de los diccionarios con todas las variedades, y de esta manera llegamos a conseguir con Random Forest un accuracy del 87,14%.

- Accuracy : 0.8714
- 95% CI : (0.8528, 0.8885)
- No Information Rate : 0.1429
- P-Value [Acc <NIR] : <2.2e-16

5 Conclusiones y trabajo futuro

Otra idea que sería interesante a la hora de clasificar la variedad, se basaría en algo similar a los diccionarios locales que hemos planteado pero utilizando nombres de personajes famosos y políticos de cada país. De esta manera, crearíamos listas para cada uno de los 7 países y las utilizaríamos para tratar de identificar los tweets en base a las menciones a que se realicen.

Complementariamente trabajaríamos el aspecto género. Ya que, como comentamos inicialmente, hemos planteado el problema de clasificación únicamente para el aspecto variedad, omitiendo el género ya que su accuracy inicial era más alto. Pero podríamos trabajarlo, planteando hipótesis como las siguientes, que tendríamos que contrastar:

- Los tweets más largos están escritos por mujeres.
- Los tweets con más número de emoticonos están escritos por mujeres.