Text Mining en Social Media
Máster en Big Data Analytics – Curso 2016 / 2017
UPV

Fermín Leal Payá
fermin.leal.paya@gmail.com

Abstract

Muchas veces se intentan desarrollar ciertas campañas que dependen del género o variedad de las personas que las van a recibir. Por este tipo de desarrollos o muchos otros casos de estudio (I+D), se propone crear un algoritmo de detección de género y variedad sobre frases, en este caso sobre Twitter. Para este menester vamos a intentar aplicar una serie de técnicas de text mining sobre el dataset proporcionado para generar ciertas características que nos permitan segmentar los autores de las frases según género y país.

Usaremos técnicas como extractores de características por frecuencia y peso de las palabras (tfidf) además de algoritmos de inteligencia artificial, que en este caso serán sopport vector machines con diferentes kernels, que nos permitirán mejorar el accuracy.

1. Introducción

Mirándolo desde una perspectiva global y generalista, el problema de detección de género y variedad, consiste en detectar ciertas características del autor de un texto, a partir del texto en sí. Los dos problemas tendrán que ser tratados de diferente manera, ya que para detectar si es género tendremos que limpiar los datos de manera distinta a si queremos detectar variedad.

Estas dos tareas se podrían hacer con el mismo pre-procesado pero el resultado sería peor por lo que optaremos por darle tratamientos diferentes.

2. Dataset

Veamos las características del dataset que tenemos para desarrollar la tarea:

- El dataset está separado en dos conjuntos: training y test
- Las muestras vienen formateadas en XML que contienen todos los textos de los tweets por autor. El nombre del fichero es el identificador del autor.
- El fichero etiquetado tendrá el identificador y la etiqueta correspondiente. Esto se aplica tanto a variedad como a género.

En cuanto a las dimensiones del dataset, tanto en el conjunto de training como en el de test, cada una de las muestras está compuesta por el texto de 100 tweets diferentes de un mismo autor.

El conjunto de training contiene 2800 muestras, mientras que el conjunto de test contiene 1400 muestras (la mitad).Las muestras están distribuidas de forma equitativa entre las dos características a estudiar: género y variedad.

3. Propuesta del alumno

Para resolver estos problemas primero de todo veremos el preprocesado de cada una de las partes.

Para género dejaremos toda la información en cuanto a preposiciones, artículos, etc.. ya que la forma de usar este tipo de lenguaje estadísticamente da información del género de la persona.

En cambio para ver la variedad quitaremos toda esta información para dejarnos solo las palabras típicas de cada país, es decir, lo que más peso tiene a la hora de distinguir variedad.

Como en la práctica teníamos un tiempo limitado nos hemos centrado en hacer un bastantes pruebas (ensayo y error) para llegar a la solución más óptima posible con el tiempo que disponemos.

4.Pruebas

4.1 Variedad

A continuación vamos a ver todas las pruebas que se hicieron para detectar la variedad y la solución final.

Prueba 1

- Bolsa 100 palabras
- SVM (Lineal)

Overall Statistics

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

Kappa : 0.4433
Mcnemar's Test P-Value : 0.0008919

Prueba 2

- Bolsa 100 palabras
- SVM (Radial)

Accuracy : 0.5193
95% CI : (0.4927, 0.5458)
No Information Rate : 0.1429
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4392
Mcnemar's Test P-Value : 1.491e-06

Prueba 3

- Bolsa 100 palabras
- Naive Bayes

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

Kappa : 0.325
Mcnemar's Test P-Value : < 2.2e-16

Prueba 4

- Bolsa 200 palabras
- SVM (Lineal)

Accuracy : 0.6329
95% CI : (0.607, 0.6582)
No Information Rate : 0.1429
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5717
Mcnemar's Test P-Value : 9.389e-05

Prueba 5

- Bolsa 200 palabras
- SVM (Radial)

Accuracy : 0.6757
95% CI : (0.6505, 0.7002)
No Information Rate : 0.1429
P-Value [Acc > NIR] : < 2e-16

Kappa : 0.6217

Prueba 6

- Bolsa 200 palabras
- KNN

Accuracy : 0.43

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

Kappa : 0.335
Mcnemar's Test P-Value : 0.001508

Prueba 7

- Bolsa 500 palabras
- SVM (Radial)

Accuracy : 0.7864
95% CI : (0.764, 0.8076)
No Information Rate : 0.1429
P-Value [Acc > NIR] : < 2e-16

Kappa : 0.7508
Mcnemar's Test P-Value : 0.02969

Prueba 8

- Bolsa 500 palabras
- SVM (Radial)
- Ajustes de sigma
- Ajustes de centroide

grid <- expand.grid(sigma = c(.00001, .0001, .001, .005), C = c(1.5))

Accuracy : 0.7893
95% CI : (0.767, 0.8104)
No Information Rate : 0.1429
P-Value [Acc > NIR] : <2e-16

Kappa : 0.7542
Mcnemar's Test P-Value : 0.2251

Prueba 9

- Bolsa 5000 palabras
- SVM (Radial)
- Ajustes de sigma
- Ajustes de centroide

Accuracy : 0.8879
95% CI : (0.849, 0.8852)
No Information Rate : 0.1429
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8658
Mcnemar's Test P-Value : 0.0002838

4.1.2 Conclusión Variedad

Como vemos para variedad hemos podido ajustar el accuracy a un 88%, y para esto hemos usado los parámetros de la prueba 9, donde los parámetros óptimos eran :

- Bolsa 5000 palabras
- SVM (Radial)
- Ajustes de sigma (1e-04)
- Ajustes de centroide (1,50)

Como vemos en este problema la cantidad de palabras usadas en la bolsa es muy importante ya que es un tema de variedad. Como veremos más adelante para género esto no importa tanto.

4.2 Género

A continuación vamos a ver todas las pruebas que se hicieron para detectar el género y la solución final.

Prueba 1

- Bolsa 1000 palabras
- SVM (Linear)
- Con signos de puntuación
- Sin preposiciones

Accuracy : 66%
Kappa : 0.33

Prueba 2

- Bolsa 1000 palabras

- SVM (Linear)
- Con signos de puntuación
- Con preposiciones

Accuracy : 67%
Kappa : 0.35

Prueba 3

- Bolsa 1000 palabras
- SVM (Radial)
- Con signos de puntuación
- Con preposiciones

Accuracy : 75%
Kappa : 0.50

Prueba 4

- Bolsa 1000 palabras
- Naive Bayes
- Con signos de puntuación
- Con preposiciones

Accuracy : 67%
Kappa : 0.34

Prueba 5

- Bolsa 1000 palabras
- NN + PCA
- Con signos de puntuación
- Con preposiciones

Accuracy : 73%
Kappa : 0.42

Prueba 6

- Bolsa 1000 palabras
- RF
- Con signos de puntuación
- Con preposiciones

Accuracy : 69%
Kappa : 0.40

Prueba 7

- Bolsa 5000 palabras
- SVM (Radial)
- Con signos de puntuación
- Con preposiciones

Accuracy : 74%
Kappa : 0.49

4.2.2 Conclusión Género

Como vemos para género hemos podido ajustar el accuracy a un 75%, y para esto hemos usado los parámetros de la prueba 3, donde los parámetros óptimos eran :

- Bolsa 1000 palabras
- SVM (Radial)
- Con signos de puntuación
- Con preposiciones

Como vemos para este caso el uso de una mayor bolsa de palabras es contraproducente ya que conseguimos peores resultados, tanto en tiempo como en precisión.

5. Conclusiones y trabajo futuro

Como vemos cada problema tiene que tratarse de forma diferente. Además de la bolsa de palabras de podría haber ajustado mucho mejor aplicando otras técnicas que sumen información al dataset.