Text Mining en Social Media

Máster en Big Data Analytics – Curso 2017/2018

UPV

Pablo Gomis Sellés

1. Introducción.

En este trabajo queremos distinguir el género de la persona que escribe un texto, así como la variedad de español utilizado. Tendremos que tratar estas dos vertientes de forma diferente, pues cada una tienes unas características distintas que nos permiten estudiarlas. Para esto utilizaremos distintos métodos de text mining como tfidf o algoritmos de inteligencia artificial, como support vector machines con diferentes kernels, con los que mejoraremos el accuracy.

El dataset que nos proporcionan tiene ya las particiones de training y test hechas. Cada muestra contiene un XML con 100 tweets de un mismo autor. El conjunto de training tiene 2800 muestras, y el de test, la mitad.

Para tratar la diferencia de género nos centraremos en las preposiciones, artículos y otras formas gramaticales que se ha demostrado que da información respecto al género de una persona. En el caso de la variedad, nos centraremos más en palabras específicas del país o región donde se habla.

Nuestro método de trabajo final fue probar muchas posibles soluciones hasta llegar a la que consideramos óptima(o que ya no podíamos mejorar mucho).

2. Pruebas variedad

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

Conclusión Variedad

Hemos llegado a un accuracy del 88% en la prueba 9, que consideraremos como óptima para nuestro caso. Cabe destacar la importancia de la cantidad de palabras que se escojan.

3. Pruebas género

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

Conclusión Género

Hemos llegado a un accuracy del 75 % utilizando los resultados de la prueba 3, que consideraremos como óptima para nuestro caso. Cabe destacar, al contrario que para la variedad, el poco interés que tiene ahora el coger cuantas más palabras mejor. Parece totalmente contraproducente.

Éramos totalmente conscientes de que se podría haber hecho mejor, utilizando otras técnicas, pero debido al tiempo limitado que teníamos decidimos enfocarlo del modo “prueba-error” y considero que no obtuvimos unos malos resultados.