Text Mining en Social Media
Detección automática de género y variedad

Gabriel Piles González
gabpigon@upvnet.upv.es

Abstract

En el siguiente documento se detalla el trabajo realizado para obtener unos pronósticos de género y de procedencia de un conjunto de usuarios de twitter usando solo sus correspondientes tweets como entrada.

El objetivo de este trabajo consiste en enfrentarse a un problema real de análisis de datos y se ha primado el cómo se aborda el problema a los resultados obtenidos.

La forma de abordar esta tarea ha sido probar variaciones del código dado como punto de partida donde se usaba una bolsa de palabras con las **n** palabras más frecuentes y se usaba el clasificador **SVM** para dar un pronóstico.

En los siguientes apartados se detallarán todas las pruebas hechas y se explicará el porqué de su éxito o su fracaso. Se deben destacar tres resultados obtenidos. El primero es que quitar las mayúsculas y las tildes en los tweets da buenos resultados para la procedencia de los autores pero da malos resultados para el pronóstico del género. El segundo resultado destacable es que usar una bolsa de palabras con las palabras más discriminantes, mejora los resultados para la detección del género y da muy buenos resultados para determinar la procedencia de los autores de los tweets. El último resultado destacable es que el agoritmo **Random Forest** es el que mejores resultados obtuvo.

1 Introducción

En este documento se va a detallar el trabajo realizado para pronosticar de la mejor forma posible el género y la procedencia de los autores de un conjunto de tweets previamente etiquetados.

El pronóstico del género y la procedencia de autores de tweets entra dentro del ámbito de problemas de Author Profiling que consiste en predecir qué usuario hay detrás de un texto. Entre los diferentes parámetros de los usuarios que tienen sentido predecir se encuentran: el género, el país de procedencia, la edad o la ideología del usuario, entre otros.

Los tweets que se han hecho servir en esta tarea han sido seleccionados para minimizar los sesgos. Para cada género y para cada variedad se tiene exactamente el mísmo número de tweets, y los usuarios han sido seleccionados para tener una muestra representativa de tweets de cada usuario. Además los tweets etiquetados se han dividido en un set de training y un set de test para poder obtener una aproximación del ajuste en las pruebas que se hagan.

Como punto de partida se disponía de un código que realizaba un pronóstico usando una bolsa de palabras cogiendo las **n** palabras más frecuentes y el clasificador **Support Vector Machine** para realizar los pronósticos. Esta aproximación se ha tomado como **Base Line** a superar. Sobre este punto de partida se han realizado diferentes variaciones para conseguir un mejor ajuste. En los siguientes puntos se detalla cada variación por separado.

2 Stop Words

La primera prueba realizada para mejorar el resultado del **Base Line** fue quitar palabras que se creían no importantes del vocabulario de palabras más repetidas. Se realiazaron varias pruebas al respecto para pronósticar el género del autor del conjunto de tweets de test y consistentemente los resultados empeoraban al quitar palabras.

Es posible que las palabras que se pensaba que no aportaban información sobre el género del autor de los tweets en realidad sí que lo hiciera. Por poner un ejemplo, en un principio se pensó que la palabra "q" que estaba muy alta en la lista de palabras más usadas no serviría para decidir si se trataba de un hombre o de una mujer. No obstante, dado los resultados obtenidos quedó claro que esta palabra en concreto sirve para discernir el género, muy probablemente porque hay un género que la usa más que el otro.

Visto los resultados con el género, se dejó de lado esta vía para mejorar los resultados y en las posteriores pruebas no se quitó ninguna de las palabras del vocabulario.

3 Diferentes rangos de vocabulario

Una idea propuesta era probar a coger como bolsa de palabras las **n** palabras más usadas descartando las **m** primeras. Con un par de pruebas se vio que esta técnica no mejora los resultados del **Base Line**. Con el tiempo justo del que se disponía, no se consideró adecuado emplear tiempo para descubrir los motivos detrás de este comportamiento, pero puede tener sentido que las palabras que aparecen más, sean las que más información aporten.

4 Tildes

Otra opción explorada fue la de no tener en cuenta las tildes, y que se contara en la bolsa de palabras la misma palabra con y sin tilde. Los resultados obtenidos con esta idea han sido en un principio sorprendentes, ya que esta modificación conseguía consistetemente mejores resultados para predecir la variedad pero no el género. Las diferencias en los ajustes eran bajas, del orden de la segunda posición detrás de la coma decimal pero el tiempo extra añadido para hacer este ajuste es muy bajo para que se considerara dejarlo fuera para pronosticar la variedad.

5 Vocabulario de bigramas

Se programó un algoritmo para construir una bolsa de palabras con los **n** bigramas más frecuentes, pero los resultados obtenidos con este método fueron peores que el **Base Line**. Probando diferentes **n** se consiguió llegar a un ajuste de **62%**, un resultado muy pobre considerando que el **Base Line** está en el orden del **66%**.

Este resultado se puede reproducir ejecutando el script del siguiente enlace:

`https://github.com/gabriel-piles/text_mining/blob/master/my-pan-ap17_bigrams.r`

Los resultados obtenidos fueron más bien inesperados, ya que a priori, podía tener sentido que los bigramas aportaran más información que las palabras sueltas. Por otro lado, es probable que el problema de este mal resultado sea que la frecuencia de los bigramas más frecuentes es considerablemente más baja que de las palabras, y habría que coger muchos más bigramas que palabras para mejorar el resultado. Con el tiempo del que disponiamos, no tenía sentido entrenar un modelo con semejante número de columnas.

6 SVM vs Random Forest

Otro resultado sorprendente fue que el algoritmo **Random Forest** obtenía mejores resultados que el clasificador **SVM**. En todas las pruebas realizadas este resultado se repetía consistentemente. A continuación se muestran dos comparaciones de ajustes del set de test con **Random Forest** y **SVM**.

- Pronóstico para el género
- Algoritmo **tf-idf**
- 300 palabras más discriminantes

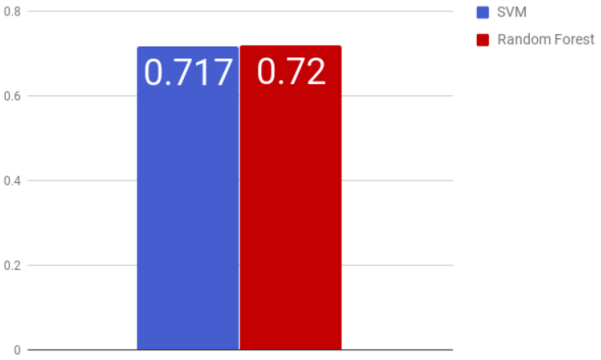

Figure 1: Ajuste de los tweets de test con SVM y Random Forest para género

- Pronóstico para la variedad
- Algoritmo **tf-idf**
- 500 palabras más discriminantes

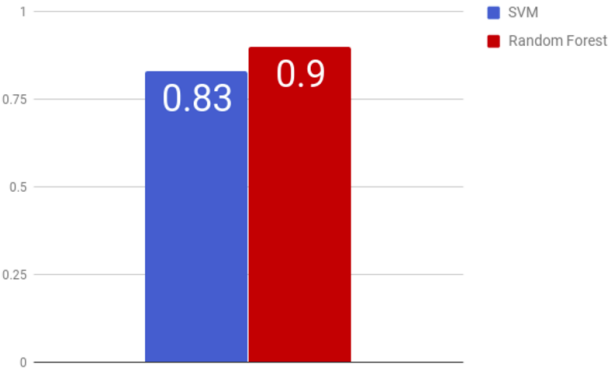

Figure 2: Ajuste de los tweets de test con SVM y Random Forest para la variedad

No obstante, como otros grupos demostraron, el clasificador **SVM** no obtiene los mejores resultados con los parámetros por defecto, y si se ajustan estos parámetros correctamente es posible obtener mejores resultados que el clasificador **Random Forest**. Estas pruebas se dejan para trabajo futuro.

7 TF-IDF

El método que mejores resultados obtuvo fue crear una bolsa de palabras con las **n** palabras más discriminantes obtenidas con el algoritmo **tf-idf** de la librería **TidyText** para **R**. A continuación se muestran las palabras más discriminantes para el género y para la variedad:

	WORD	FREQ	gender	tf	idf	tf_idf
1	gol	474	male	0.0011256795	0.6931472	0.0007802616
2	vs	430	male	0.0010211860	0.6931472	0.0007078322
3	retweeted	418	male	0.0009926878	0.6931472	0.0006880788
4	whatsapp	397	female	0.0009578359	0.6931472	0.0006639213
5	sola	375	female	0.0009047568	0.6931472	0.0006271297
6	llorar	357	female	0.0008613285	0.6931472	0.0005970274

Figure 3: Palabras más disciminantes para el género

	WORD	FREQ	variety	tf	idf	tf_idf
1	peru	1052	peru	0.008241543	1.94591	0.016037302
2	vos	949	argentina	0.007451261	1.94591	0.014499484
3	bogota	854	colombia	0.006586355	1.94591	0.012816456
4	elcooperante	869	venezuela	0.006378356	1.94591	0.012411708
5	nicolasmaduro	723	venezuela	0.005306734	1.94591	0.010326427
6	epn	584	mexico	0.004691554	1.94591	0.009129343

Figure 4: Palabras más disciminantes para la variedad

Es destacable que muchas palabras son intuitivamente discriminantes pero otras no lo son en absoluto. Palabras como “retweeted” o “whatsapp” no se entiende por qué consiguen discriminar el género.

8 Mejores resultados

Los mejores resultados obtenidos se han conseguido con las siguientes configuraciones:

Para el género:

- No quitar mayúsculas ni tildes
- Algoritmo **tf-idf**
- 700 palabras más discriminantes
- **Random Forest**.

$$\text{Ajuste} = 73.43\%$$

Este resultado se puede reproducir ejecutando el script del siguiente enlace:

`https://github.com/gabriel-piles/text_mining/blob/master/my-pan-ap17_tf_idf_genre.r`

Para la variedad:

- Quitanr mayúsculas y tildes
- Algoritmo **tf-idf**
- 500 palabras más discriminantes
- **Random Forest**.

$$\text{Ajuste} = 91.57\%$$

Este resultado se puede reproducir ejecutando el script del siguiente enlace:

`https://github.com/gabriel-piles/text_mining/blob/master/my-pan-ap17_tf_idf_variety.r`

Usando la técnica de **tf-idf** se ha obtenido un buen ajuste para la variedad pero no tanto para el género. Como se dijo en el transcurso de la tarea, para predecir el género pesa más cómo se dicen las cosas y para pronósticar la variedad importa el qué se dice.

9 Conclusiones y trabajo futuro

Con el tiempo dado para realizar este trabajo, los resultados obtenidos han sido muy satisfactorios. Se han hecho muchas pruebas, que era el objetivo de esta tarea, se han descartado ideas y se han obtenido buenos resultados con otras.

La intuición ha fallado para muchas de las pruebas realizadas, y es ésto mismo la lección que más se quiere remarcar en este trabajo: **no te fies de tú intuición en todos los casos y haz la prueba.**

Para un trabajo futuro se destacarían los siguientes puntos:

- Añadir a las matrices de train y test columnas con información como el número medio de carácteres por tweet.
- Construir una bolsa de palabras juntando las palabras más discriminantes y los bigramas más discriminantes.
- Ajustar correctamente los parámetros del clasificador **SVM**.
- Probar otros clasificadores como las **Redes Neuronales**.
- Estudiar bibliografía sobre **Author Profilin**.