Paper de la asignatura Text Mining en Social Media.

David Sánchez Pérez
david.sanchez@realrent.es

Abstract

El Author Profilling es un método de análisis en plena evolución que consiste en extraer patrones de textos para tratar de proporcionarnos un perfil lo más exacto posible del autor, como pueda ser su edad, género, nacionalidad...etc.

En este paper vamos a tratar de pronosticar la variedad del lenguaje nativo de unos tweets determinados.

Los datos son sacados del PAN-AP 2017, que se trata de una competición a nivel mundial donde el objetivo es resolver un problema que varía cada año la clase de clasificación que hay que hacer, por ejemplo, puede ser solo de genero, solo de variedad, de edad, combinacion de variables...

1. Introducción

Para realizar este caso práctico de Author Profilling nos proporcionan:

Un dataset sacado del PAN-AP 2017, un corpus reducido de tweets escritos por autores de siete países distintos de habla hispana, de los cuales deberemos conseguir clasificar a qué país pertenecen los autores y de qu género son, en base a lo que han escrito en cada tweet.

Un código en R que utilizaremos como base, dicho código genera una bolsa de palabras, las cuales las extrae del corpus, predice el genero y la variedad utilizando técnicas de machine learning. En los siguientes puntos explicaremos como hemos modificado el código para obtener mejores resultados en variedad ya que a priori tenía un margen de mejora mayor.

Para mejorar estos resultados nos hemos centrado en:

Hacer una bolsa de palabras que contuviera una parte de palabras extraídas del corpus y otra que fuera de modismos o palabras más usadas de cada país y usar otro modelo de machine learning.

Para la tarea contamos con cinco horas para primero examinar nuestro dataset, y posteriormente plantear cómo vamos abordar el problema y llevarlo a cabo con las pruebas necesarias. Y finalmente comprobar los resultados obtenidos y aportar nuestras conclusiones al respecto.

2. Dataset

El dataset que utilizamos en la tarea como ya hemos comentado, son los mismos que los del PAN-AP 2017, y se trata de un conjunto de archivos en formato XML, cada uno perteneciente a un individuo en concreto con cien tweets publicado por dicho individuo.

Para el training serán 2800 archivos que contienen 100 tweets, total 280 mil tweets(unos 34 MB de texto) para analizar y entrenar con nuestros modelos para poder realizar una correcta clasificación.

Y para el test serían 140 mil tweets.

Las muestras estarán distribuidas de forma que podamos estudiar tanto el g/'enero como la variedad.

Para unir cada archivo individual en uno solo que es lo que necesitamos a la hora de trabajar, debemos hacer una función que lea primero todos los archivos del tipo XML, y posteriormente los metamos en una variable todos juntos de la siguiente manera:

```
files = list.files(pattern="*.xml")
  corpus.raw <- NULL
  i <- 0
  for (file in files) {
    xmlfile <- xmlTreeParse(file, useInternalNodes = TRUE)
    corpus.raw <- c(corpus.raw, xpathApply
                    (xmlfile, "//document", function(x) xmlValue(x)))
    i <- i + 1
    if (verbose) print(paste(i, " ", file))
  }
```

3. Propuesta del alumno

Una vez explorado el dataset, ejecutamos el modelo que se nos ha proporcionado en el código, con una bolsa que se ha extraído del propio corpus, en base a las palabras más frecuentes y con un modelo predictivo de Super Vector Machine, el resultado que obtenemos es el siguiente:

Tabla 1. Resultados SVM.

N	GENDER	VARIETY	TIME
10	0.5875	0.2608	3.62m
50	0.6850	0.3167	4.32m
100	0.7375	0.3383	5.36m
500	0.7358	0.5717	9.16m
1000	0.6983	0.6167	12.11m
5000	0.7550	0.6167	51.81m

Como observamos el accuracy para el género es relativamente alto, en cambio para la variedad del lenguaje es muy bajo, por ello nos hizo plantearnos en centrarnos en el problema de la variedad y tratar de mejorarla dado el tiempo del que disponíamos.

La primera idea es centrarnos en la bolsa de palabras que tenemos, al tratarse de las palabras más frecuentes de los tweets y que se trata del mismo idioma, el castellano, aunque hayan 7 países distintos observamos que las palabras son de lo más común en todos los países y muy complicado que sirvan para diferenciar de qué país proviene cada uno. Por ello decidimos que debemos mejorar la bolsa incluyendo palabras típicas de cada país.

Para obtener estas palabras recurrimos a internet para encontrar los modismos o palabras más usadas en los países que tenemos que son: México, Colombia, Venezuela, Perú, Chile, Argentina y Espaa. Una vez obtenidas las palabras, las guardamos por separado en un csv para cada país y de esta manera poder cargarlo a nuestro dataset para generar la bolsa de palabras.

Pero no solo nos basta con guardarlas en csv conforme las encontramos, hemos tenido que aplicar un pequeo preproceso, eliminando modismos duplicados y también eliminando aquellos que se trataban de expresiones combinadas, quedándonos solo con palabras únicas.

Además, tuvimos que hacer una pequea 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.

Pero para ver que vamos bien encaminados en nuestra idea, hacemos la comprobacin con uno de nuestros particulares diccionarios para ver si el uso de este mejora la clasificación de los autores de dicho país. Por ejemplo lo hacemos con Chile.

Tabla 2. Usando 100 palabras de Vocabulary.

STATITICS	CLASS
	CHILE
Sensitivity	0.49500
Specificity	0.88500
Pos Pred Value	0.41772
Neg Pred Value	0.91316
Prevalence	0.14286
Detection Rate	0.07071
Detection Prevalence	0.16929
Balanced Accuracy	0.69000

Accuracy : 0.5229

Tabla 3. Usando 102 palabras de Chile

STATITICS	CLASS
	CHILE
Sensitivity	0.49500
Specificity	0.96083
Pos Pred Value	0.67808
Neg Pred Value	0.91946
Prevalence	0.14286
Detection Rate	0.07071
Detection Prevalence	0.10429
Balanced Accuracy	0.72792

Accuracy : 0.3121

Comprobamos que efectivamente el accuracy general disminuye a la hora de clasificar a las 7 países pero en el caso de Chile mejora individualmente al usar el diccionario con los modismos que hemos obtenido de este país. Por tanto, decidimos seguir adelante con nuestra idea de unir todos los diccionarios que hemos buscado para que sea nuestra bolsa de palabras conjunta y mejore el nivel de clasificación del modelo. Para crear la lista conjunta simplemente unimos cada lista.

```
venezolano<-read.table("~/listas/venezolano.csv", quote="\"")
colnames(venezolano)<-c("WORD")
venezolano\$FREQ<-0
peruano<-read.table("~/listas/peruano.csv",quote="\"")
colnames(peruano)<-c("WORD")
peruano\$FREQ<-0
argentino<-read.table("~/listas/argentino.csv", quote="\"")
colnames(argentino)<-c("WORD")
argentino\$FREQ<-0
chileno<-read.table("~/listas/chileno.csv",quote="\"")
colnames(chileno)<-c("WORD")
chileno\$FREQ<-0
colombiano<-read.table("~/listas/colombiano.csv", quote="\"")
colnames(colombiano)<-c("WORD")
colombiano\$FREQ<-0
mexicano<-read.table("~/listas/mexicano.csv",quote="\"")
colnames(mexicano)<-c("WORD")
mexicano\$FREQ<-0
espannol<-read.table("~/listas/espannol.csv", quote="\"")
colnames(espannol)<-c("WORD")
espannol\$FREQ<-0
vocabularios7<-rbind(venezolano,peruano,argentino,
chileno,colombiano,mexicano,espannol)
```

Conseguimos tener un resultado positivo con la bolsa de palabras, pasamos a probar distintos modelos de machine learning para ver si hay alguno que nos ofrezca mejores resultados.

4. Resultados experimentales

Una vez vomprobado que las listas de los modismos si que mejoran la clasificación para cada país decidimos comprobar de varios modelos de machine learning, cual nos aporta un mayor accuracy para el vocabulario dado usando solo 100 palabras.

- Usamos primero Support Vector Machine, que es el modelo que tenemos cuando se nos proporciona el modelo. 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.

- A continuación usamos Random Forest:

  Accuracy : 0.5393.

  95 % CI : (0.5128, 0.5656).

  No Information Rate : 0.1429.

  P-Value [Acc ¿NIR] : ¡2.2e-16.

- Ahora usamos K-Nearest Neighbors:

  Accuracy : 0.345.

  95 % CI : (0.3201, 0.3706).

  No Information Rate : 0.1429.

  P-Value [Acc ¿NIR] : ¡2.2e-16.

- Usando Naive Bayes:

  Accuracy : 0.4214

  95 % CI : (0.3954, 0.4478)

  No Information Rate : 0.1429

  P-Value [Acc ¿NIR] : ¡2.2e-16

- Usando Nearest Neighbour:

  Accuracy : 0.43

  95 % CI : (0.4039, 0.4564)

  No Information Rate : 0.1429

  P-Value [Acc ¿NIR] : ¡2.2e-16

- Usando 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 podemos observar, el modelo que mejor resultados nos ha dado de base usando únicamente 100 palabras del vocabulario base, es el Random Forest con un accuracy de 53,93 %.

Una vez elegido el modelo,le aplicamos nuestra bolsa de palabras que consta de 416 palabras únicas propias de cada pas mas 500 palabras del vocabulario inicial, 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] : ¡2e-16

En definitiva, hemos conseguido mejorar la clasificación de la variedad del lenguaje desde un 33 % hasta un 87 % enriqueciendo la bolsa de palabras y el número de palabras usadas, como del modelo de machine learning utilizado para la casificación.

5. Conclusiones y trabajo futuro

Otra idea que sera interesante a la hora de clasificar la variedad, se basara en algo similar a los diccionarios locales que hemos planteado pero utilizando nombres de personajes famosos y polticos de cada pas. De esta manera, crearamos listas para cada uno de los 7 pases y las utilizaramos para tratar de identificar los tweets en base a las menciones a que se realicen.

Complementariamente trabajaramos el aspecto gnero. Ya que, como comentamos inicialmente, hemos planteado el problema de clasificacin nicamente para el aspecto variedad, omitiendo el gnero ya que su accuracy inicial era ms alto. Pero podramos trabajarlo, planteando hiptesis como las siguientes, que tendramos que contrastar:

Los tweets ms largos estn escritos por mujeres. Los tweets con ms nmero de emoticonos estn escritos por mujeres.