Esquema de paper. Asignatura Text Mining en Social Media. Master Big Data

Mari Carmen Sevilla Solera
emsisiem@gmail.com

Abstract

Author Profiling se emplea para intentar inferir la edad, sexo, idioma nativo o personalidad de una persona a través del análisis de sus textos. Nuestro objetivo es el de identicar automáticamente la edad y el idioma de usuarios del medio social Twitter.

En este paper proponemos una aproximación a esta tarea para determinar el sexo y el idioma de individuos que escriben en Twitter utilizando para ello Machine Learning. El idioma a inferir es español, la mayoría de estudios se han realizado sobre lengua inglesa.

La tarea se enmarca en el ámbito de Big Data debido a la gran variedad de temas o conversaciones posibles y en cuanto al volumen aunque el texto de los tuits puede ser corto se convierte, si juntamos todos los comentarios de ese autor, en un texto mucho más largo para cada autor. Esta tarea es planteada por Autoritas Consulting.

1 Introducción

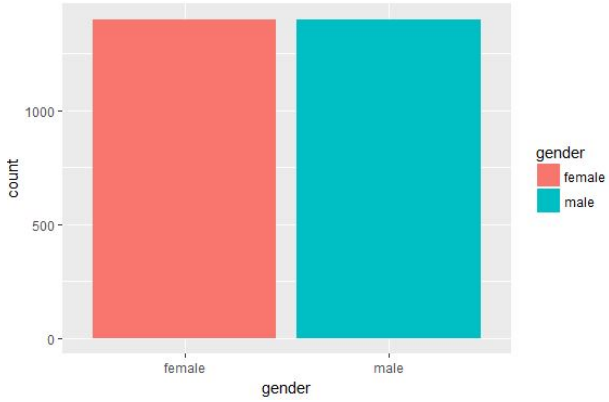
Dentro del área de Author profiling vamos a inferir el sexo y el idioma nativo de los autores de tuits. Esto que parece magia se consigue con Machine Learning, hipótesis e ideas sobre cómo clasificar los tuits según las dos clases antes comentadas, sexo e idioma nativo del autor del tuit.
Utilizaremos métodos supervisados ya que son más fáciles de evaluar su calidad. Como base nos proporcionan unos valores de accuracy a batir.
Volviendo a los métodos supervisados, esto significa tener un conjunto de datos de entrenamiento que permitan aprender con autores etiquetados con su sexo e idioma nativo para permitirnos evaluar la precisión que hayamos obtenido. Vamos a esquematizar el aprendizaje:

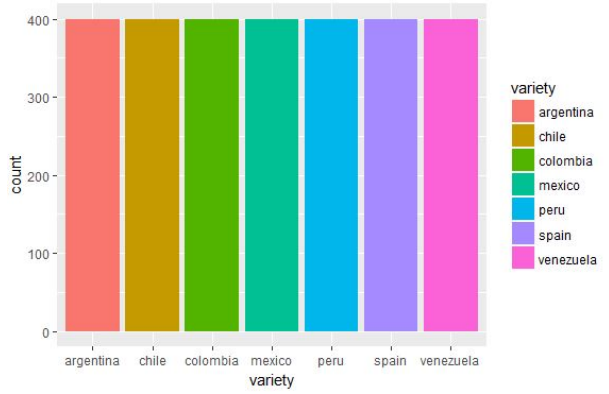
- Vamos a aprender un modelo a partir de la representación del training
- Vamos a evaluar el modelo
- Vamos a predecir con la representación del test.
- Vamos a utilizar la medida accuracy

Como accuracy entendemos la proporción del total de predicciones correctas, obteniéndose de la siguiente manera:
Accuracy = (True Positive+True Negative) / (True Positive+ True Negative + False Positive + False Negative).
Nos proporcionan código en R de partida para el estudio. En dicho código tenemos como clasificador las máquinas de vectores de soporte y como modelo de representación bolsa de palabras. La bolsa de palabras propuesta tiene una longitud de 1000 palabras que serán las más frecuentes sobre el dataset.
Presentaremos y llevaremos a cabo varias hipótesis para conseguir esa clasificación de los tuits. Intentaremos conseguir resultados mejorados a los base dados. Estos resultados serán presentados como la accuracy calculada como el porcentaje de casos acertados frente al total de casos.
Además haremos un estudio de las hipótesis planteadas que sean comparables y competitivas entre ellas y entre el resto de grupos de analistas.

2 Dataset

El corpus es proporcionado por la empresa Autoritas Consulting.
Este corpus está constituido por ficheros xml, un fichero por autor y con su contenido del tuit dentro del tag document.





Mostramos unas gráficas con la distribución de los datos por clases de genero y variedad según el fichero truth.txt.

Estos posts se agrupan por autor y se han eliminado los retuits en los ficheros proporcionados.

Para la clasificación tenemos un fichero truth.txt donde tenemos el autor que es el nombre del fichero xml y su etiqueta de sexo e idioma.

Los ficheros con los corpus de train y test los tenemos en distintas carpetas. Para los ficheros del corpus de test eliminaremos la etiqueta para que sea el modelo entrenado con los datos de train el que haga la predicción.

Ya que tenemos datos para entrenar y evaluar con los de test no es necesario realizar un proceso de dividir los datos entregados.

Los ficheros xml los cargaremos en un dataset en el que se realiza un preprocesado que consiste en pasar el texto a minúsculas, eliminar los números, palabras vacías, signos de puntuación y acentos.

3 Propuesta del alumno

Para nuestra propuesta trabajaremos con modelos separados para clasificar por sexo y por idioma ya que se mejoran los resultados en lugar de si lo hiciéramos de forma conjunta.

Seguimos con la estrategia de bolsa de palabras. Por cada una de las palabras calculamos su frecuencia de aparición en el corpus. Nosotros creamos un vector con este resultado que denominamos diccionario. De estas palabras sólo nos quedamos con las 1000 primeras siendo el orden la frecuencia de aparición, es decir, nos quedamos con las más frecuentes.

No se tiene en cuenta la semántica de las palabras ni su uso en el contexto, ya que una misma palabra dicha en el contexto de un hombre puede tener una connotación o significado diferente a lo dicho por una mujer. Esto queda fuera del estudio realizado.

La propuesta o hipótesis con la que mejor resultado obtuvimos fue en obtener una bolsa de palabras de mujeres, una bolsa de palabras de hombres y juntarlas en una única. Al obtener las palabras exclusivas obtuvimos una mejora en el modelo. Este estudio lo hicimos para ambas clasificaciones de sexo y de idioma.

Sobre esta bolsa de palabras de la unión de las exclusivas de cada sexo o variedad de idioma si añadíamos más palabras con cierta exclusividad en cada clase mejoramos de nuevo el resultado.

En la bolsa de palabras de la unión de las exclusivas por clase añadimos algunas palabras. Estas fueron seleccionadas por la guía de estudios que afirman que las mujeres utilizan más pronombres,negaciones, presente en tiempos verbales que los hombres, y que los hombres emplean más adjetivos, determinantes y modificadores.

Nos quedamos con una reducida bolsa, exactamente 417 palabras. Queríamos complementar esta lista pero no seguimos con el estudio.

4 Resultados experimentales

Elegimos un modelo de regresión lineal y de árboles pues tanto en la clasificación del sexo como del lenguaje nativo nos enfrentamos a clasificación por variable discreta.

La intuición nos dice que separemos el vocabulario de las mujeres y de los hombres para la clasificación de sexo y lo mismo para las distintas clases de idioma. Así conseguir palabras significativas en la clasificación.

Realizamos varias hipótesis y fuimos anotando el resultado de accuracy. Pasamos a detallar estos resultados.

Las dos primeras hipótesis que nos planteamos fue pensar que si teníamos una bolsa de palabras más frecuentes para mujeres clasificaríamos mejor a

las mujeres, y por tanto a los hombres. Lo mismo para el caso contrario.
Hipótesis 1: Bolsa de palabras más frecuentes para mujeres
Accuracy: 0,67
Hipótesis 2: Bolsa de palabras más frecuentes para hombres
Accuracy: 0,68
Hipótesis 3: Bolsa formada por la unión de las bolsas de palabras exclusivas de hombres y mujeres. Obtenemos una bolsa de 379 palabras.
Accuracy: 0,6957
Hipótesis 4: Bolsa formada en la hipótesis 3 más la intersección de las bolsas de las hipótesis 1 y 2.
Accuracy: 0,6907
Hipótesis 5: Bolsa formada por la hipótesis 3 más una lista reducida de palabras
Accuracy: 0,6979
Para la clasificación de los autores de los tuits por su idioma se ha seguido las siguientes hipótesis.
Hipótesis 1: Bolsa de palabras más repetidas en cada país y exclusivas por país
Accuracy: 0,8393
Hipótesis 2: Utilizando la misma bolsa de palabras que en hipótesis 1 pero con clasificación randomForest de 50 árboles
Accuracy: 0,8507
Hipótesis 3: Bolsa de palabras con las 100 más frecuentes de las exclusivas de cada país.
Accuracy: 0,7271

5 Conclusiones y trabajo futuro

Las palabras utilizadas más frecuentemente por una de las clases, ya sea sexo (hombre, mujer), o variedad en el idioma nativo (Colombia, Argentina, España, Venezuela, Peru, Chile, Mexico) cuando son comparadas con las otras clases son buenas características para el modelo.
La cantidad de variables o características también es un punto muy importante para tener modelos eficientes y no solo en accuracy si no en tiempo o rendimiento, así que, es primordial la selección de un buen número de variables. Además con nuestras hipótesis nos dimos cuenta que reduciendo el número de palabras de nuestro diccionario mejorábamos el resultado.
Por tanto, nuestro trabajo futuro iría encaminado a la selección de las palabras más significativas para la clasificación en cada clase y en encontrar el número ideal de ellas para conseguir modelos eficientes.
Como modelos de clasificación utilizamos el de máquinas de vector soporte y randomForest para el idioma. Quisimos probar con más modelos pero tuvimos problemas con la herramienta RStudio y en futuros trabajos sabemos de la importancia de probar con distintos modelos y así comparar resultados.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.