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

Teresa Benlloch García
tebengar@gmail.com

Abstract

Este trabajo forma parte de la evaluación de la asignatura Text Mining in Social Media, impartida por Paolo Roso y Francisco Rangel. Partiendo de que cada individuo tiene una forma de expresarse y un lenguaje característico, la tarea nos propone identificar el género y la variedad del idioma del autor. Para ello disponemos de un dataset, proporcionado por el profesor, con tweets de 300 autores y 100 tweets por autor. Nos proporciona un par de ficheros, training y test con los que abordaremos el problema de clasificación planteado.

1 Introducción

Las redes sociales se han convertido en uno de los canales de gestión de las relaciones empresa-cliente, de conocimiento de la opinión pública... Las redes sociales necesitan algoritmos de Text Mining para poder realizar búsquedas de palabras clave, clasificación y agrupamiento. El conocimiento obtenido de las redes sociales, en nuestro caso Twitter, ha resultado ser muy valioso para las empresas. Nuestro problema se centra en un problema de clasificación en el que a partir de una muestra extraída de Twitter, mediante los modelos estadsticos, determinaremos el género de los usuarios (masculino/femenino) y la variedad del idioma (Espaa, Venezuela, México, Perú, Chile, Argentina, Colombia)

2 Dataset

Descargamos el dataset PAN-AP'17:
https://s3.amazonaws.com/autoritas.pan/pan-ap17-bigdata.zip

El dataset se estructura de la siguiente forma:
Proceso de construcción:

1. Se recuperan tuits enmarcados en una región geográfica
   - longitud, latitud, radio

2. Se preseleccionan los usuarios únicos que han emitido tuits (filtrados por idioma del perfil)
3. Se recuperan los timelines de los usuarios únicos
4. Se seleccionan los autores con más de 100 tuits (que no sean retuits) en:
   - el idioma correspondiente
   - con la localización geográfica esperada en su perfil

5. Se revisan manualmente los perfiles para asegurar el sexo
6. Se seleccionan 100 tuits por autor para la construcción del dataset final

Etiquetado del sexo:
Basado en diccionario de nombres propios

- Mujeres: 24.429
- Hombres: 25.949

Características:

- Obtenido de Twitter.
- Colección de miles de autores.
- Cientos de tuits por autor.
- Gran variedad de temas.
- No un gran tamao: 54Mb descomprimido.
- Dificultad para etiquetar la información:
  - Variedad del lenguaje por posicin geográfica.
  - Sexo por nombre (mujeres, hombres)
  - Posibilidad de personas reales vs. robots (chatbots)?

Formato de los ficheros:
Una vez descomprimido el dataset, la estructura es la siguiente:

- Un par de ficheros de verdad: training.txt y test.txt. El formato es:
  - id: sexo / variedad
- Un fichero .json por autor:
  - Cada línea del fichero un tuit en formato xml

Exploración:

- Nmero de autores por clase (sexo y variedad del lenguaje).
- Nmero de tuits por autor.
- Nmero de tuits por clase.
- Nmero de palabras por documento / autor / clase.
- Distribucin de palabras/documentos/autores por documento/autor/clase
- Longitud media de tuits, palabras, documentos...por clase.
- Distribucin temporal de los tuits, tuit ms antiguo, ms nuevo, media, desviacin
- Palabras extraas, frecuentes, comunes
- cualquier informacin que nos describa el dataset y aporte conocimiento nuevo y valuable.

3 Propuesta del alumno

Centramos el problema en determinar el género de los usuarios de twitter:

Partimos de una bolsa de palabras en la que disponemos la frecuencia de aparición de cada una de las palabras.
Obtenemos las palabras mas frecuentes (Variamos el tamaño de la bolsa para certificar su relevancia al aplicar los algoritmos de clasificación.) y, con el fin de mejorar los resultados, creamos 6 nuevas variables como el número total de palabras por tweet y su distancia a su mínimo, máximo y media

Finalmente aplicamos distintos algortimos de clasificación como Ranger (implementación de Random Forest), VSM, y Regresin Logstica.

n	Model	Gender
10	SVM	0,526
50	SVM	0,666
100	SVM	0,678
100	RL	0,673
500	SVM	0,709
500	RL	0,706
1000	SVM	0,659
1000	Ranger	0,7164

4 Resultados experimentales

Utilizando como medida de evaluación el accuracy, observamos que el mejor accuracy se obtiene con una bolsa de 1.000 palabras y el algoritmo Ranger. Vemos los resultados obtenidos en la siguiente tabla:

5 Conclusiones y trabajo futuro

El mejor resultado se obtiene con una bolsa de 1000 palabras y el algoritmo Ranger. Con estos parámetros y añadiendo las nuevas variables creadas se mejoran los resultados (0,719 con el número total de palabras y 0,723 con las distancias al mínimo, máximo y media) por lo que podemos concluir que la creación de nuevas variables mejora el ajuste del modelo.
Como ideas de futuro para mejorar los resultados:

- Análisis de pronombres y proposiciones
- Otros clasificadores