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

Lorena Ponce Ruiz, Sergio Langarita Benítez, Luis Miguel Bartolin Arnau
lorena.ponru@gmail.com
sergiolangaritabenitez@gmail.com
luismibartolinarnau@gmail.com

Abstract

Dado un dataset de tweets de diferentes usuarios de twitter. Predecir si un autor es irónico o no. Para ello se han planteado diferentes modelos de predicción, y tratamientos de los datos.

1 Introducción

Dado un dataset con un conjunto de tweets 8400 se ha separado el dataset en dos partes: training y test. Se ha entrenado diferentes modelos con los tweets de la parte training y para validar el modelo se ha buscado predecir si los tweets de la parte de test son irónicos o no.

2 Dataset

Se ha proporcionado una carpeta 421 archivos, solo el archivo 'truth.txt' tiene un formato csv. Donde hay dos columnas, la primera columna representa el autor y en la segunda columna etiqueta si es un autor irónico o no. Respecto al resto de los archivos, cada archivo contiene 20 tweets de un autor, en formato xml. El nombre de estos archivos es el nombre del autor. Previamente, se habían anonimizado los nombres de los archivos para ocultar sus autores, y dentro de los archivos se habían anonimizado los hashtags, las menciones a otros usuarios y las urls.

3 Propuesta del alumno

Se ha creado un Bag of Words respecto al training y se ha realizado un árbol de decisión y random forest. Y con este Bag of Words se ha generado otro dataframe dibujando un Árbol de decisión y seleccionando las variables más relevantes. Y también se ha producido un tercer dataframe con una matriz de correlación de las variables entre sí y se han seleccionado las variables con más y con menos correlación respecto a la columna de resultado. Para estos tres dataframes se han propuesto el modelo lineal generalizado (glm) y regresión lineal (lm). Otro tratamiento que se ha realizado a los datos es la creación de una matriz Tf-idf (Term frequency - Inverse document frequency) y aplicar sobre esta matriz una red neuronal.

4 Resultados experimentales

Después de evaluar los diferentes modelos, se ha llegado a la conclusión que el random forest es el modelo que mejores resultados puede dar, así que se ha experimentado con diferente número de palabras en el vocabulario. Se ha probado hasta con, 15000 palabras, pero bajaban los resultados. Con 1000 palabras se ha llegado a 0.95 de accuracy y 0.9 de kappa, sobre el test.

5 Conclusiones y trabajo futuro

El mejor resultado obtenido ha sido mediante la creación de un Bag of Words y la aplicación de un random forest. Con respecto al trabajo futuro se puede centrar en la manipulación del dataset, dando más peso a elementos como los emoticonos, hastags, las menciones a otros usuarios y seleccionando diferentes palabras claves.