Master en Big Data Analytics

Text Mining Social Media

Autor:

Orlando Landaeta Leal

Junio 2022

Abstract

Detección de expresiones irónicas en mensajes de Twitter, pruebas realizadas aplicando algunos modelos de predicción.

Introducción

A través de un dataset dividido en 2 partes: entrenamiento y pruebas, se realizaron algunas pruebas de predicción, los dataset son un grupo de 360 archivos XML en donde cada uno contiene un conjunto de mensajes de Twitter, con estos datos se busca detectar si los tweets de la parte de test son de sarcasmo o no.

Dataset

El dataset consiste en un conjunto de 600 archivos XML, 420 para el training y 180 para los tests, adicionalmente entre los archivos de entrenamiento se encuentra el archivo truth.txt, este archivo contiene 2 columnas indicando el autor en la primera columna y en la segunda indica si el usuario es sarcástico en sus mensajes, en los archivos se encuentran anonimizados los nombres de usuarios, las direcciones URL y los hashtags de los mensajes.

Propuesta del alumno

Antes de realizar las pruebas se realizaron varios ensayos, se encontraron algunas palabras y letras como "amp", "d", ""x", "gt", "rt", "st", "th" entre otros, estos posiblemente no ayudarían en los resultados al verle ningún significado, razón por la que se procedió a realizar algunos cambios en el código eliminándolas de la bolsa para evitarlas en los resultados de las pruebas, esta lista de palabras se encuentra la línea 201 del código.

SW = c("amp","d","k","m","r","s","u","w","x","gt","rt","st","th","vs", "youtube", "video", "jajaja")

Se ha realizado un Bag of Words con respecto al training, para realizar las pruebas los parámetros establecidos fueron los siguientes:

Número de palabras en el vocabulario = 5000
Número de pliegues en cross-validation = 10
Número de repeticiones en cross-validation = 3

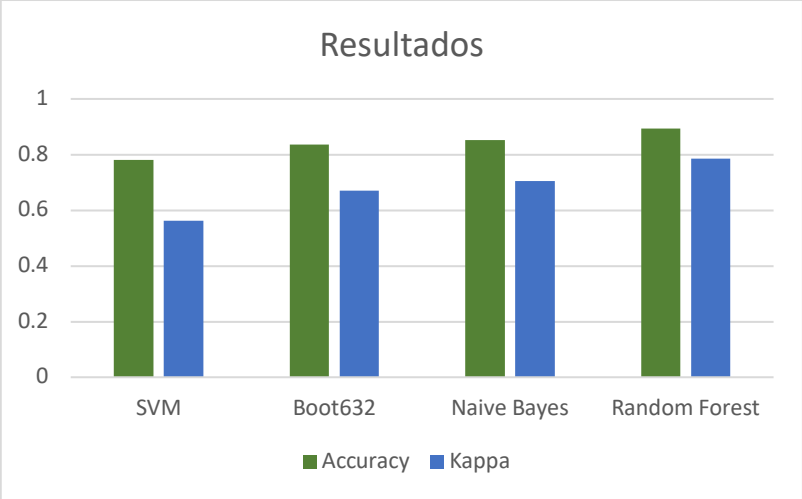
Resultados experimentales

Mientras se iban haciendo cambios en los datos, y se iba buscando limpiar el dataset, se han hecho distintas pruebas para observando cómo se obtenían diferentes accuracy de la clasificación.

Al hacer los cambios que se han especificado en el apartado anterior, y probando entre distintos modelos para la predicción, se han obtenido los siguientes valores de accuracy y kappa:

Porcentajes Accuracy por modelos

Modelo	Accuracy	Kappa
SVM	0.7809524	0.5619048
Boot632	0.8363339	0.6714948
Naive Bayes	0.8523810	0.7047619
Random Forest	0.8928571	0.7857143



Como se puede observar, en la tabla se muestran los valores obtenidos de accuracy y kappa para cada uno de los modelos aplicados, obteniendo mejores resultados con el Random Forest, las funciones para estas pruebas se encuentran entre las líneas 217 y 239 del código.

Conclusion y trabajo a futuro

Entre las pruebas realizadas los mejores resultados se consiguieron al aplicar Random Forest, para un trabajo futuro en la detección de sarcasmos en los mensajes, se podría ampliar la lista de palabras que podrían no aportar mejoras en los resultados, de igual manera una lista de posibles palabras que los puedan mejorar, también se podría buscar la manera de interpretar sarcasmos en los emoticonos.