“Tarea 3: Clasificación de sentimientos de películas - Baseline”

Rodolfo Jesús Cruz Rebollar

A01368326

Grupo 101

Accuracy de features basados en Bag-of-Words (BOW):

	LR	SVM	RF	GBM
a)1000, 50, 20	0.69	0.6866	0.7116	0.6883
b)1000, 100, 100	0.785	0.7633	0.7866	0.7583
c)1000, 200, 100	0.7933	0.7866	0.7683	0.7333

Accuracy de features basados en TF-IDF (TFIDVectorizer() de scikit-learn):

	LR	SVM	RF	GBM
a)1000, 50, 20	0.715	0.7183	0.6816	0.6683
b)1000, 100, 100	0.795	0.8016	0.7666	0.7566
c)1000, 200, 100	0.8033	0.8183	0.7266	0.7333

En la tabla de resultados para BOW (Bag-of-Words), se observa que el modelo de regresión logística entrenado con Nsamp igual a 1000, maxtokens igual a 200 y maxtokenlen igual a 100 es el que tiene el mayor porcentaje de accuracy (exactitud) de todos los modelos generados, siendo éste del 79.33%, por lo que en Bag-of-Words, el mejor modelo fue el de regresión logística del inciso c. Además, en la tabla de resultados de TF-IDF, se aprecia que el modelo de SVM (Support Vector Machine) entrenado con Nsamp igual a 1000, maxtokens igual a 200 y maxtokenlen igual a 100, resultó ser el que tiene el mayor porcentaje de exactitud de todos los modelos generados con el vectorizador TF-IDF, teniendo dicho modelo un porcentaje de exactitud del 81.83%, motivo por el cual, al usar el vectorizador TF-IDF, el mejor modelo fue el de SVM del inciso c.

Por último, el modelo con el porcentaje de exactitud más alto de todos (considerando tanto los de BOW como los de TF-IDF), es el de SVM generado con el vectorizador TF-IDF y entrenado bajo las condiciones del inciso c (Nsamp = 1000, maxtokens = 200, maxtokenlen = 100), por lo que éste modelo resulta ser el más adecuado para clasificar los sentimientos de las películas de la base de datos original con la mayor exactitud y el menor margen de error posibles.