Informe TP2: Críticas Cinematográficas

Integrantes:
Agama Avila, Arely - 105829
Martinez, Selene Anahi - 100439
Meichtri, Melany - 102330

Denominamos true al set de (review, label) siendo label positivo o negativo según el dataset de entrenamiento.
Denominamos predicho al set de (review,label), siendo label positivo o negativo según la predicción dada por el modelo.
Notamos que la precisión, recall y accuracy dan iguales dado que:

Recall en multilabel utilizando average micro se calcula como la cantidad del resultado de la intersección de las coincidencias entre True y Predicho sobre la cantidad de elementos predichos, que va a ser la misma cantidad que los elementos del dataset:

$Cant(True \cap Predicho) \ / \ Cant(Predicho)$

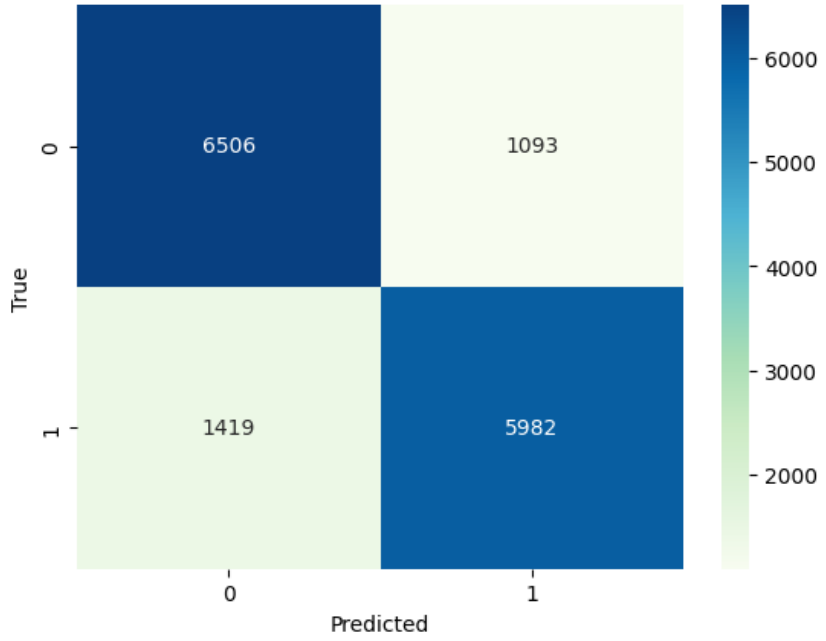
Precisión en multilabel utilizando average micro se calcula como la cantidad del resultado de la intersección de las coincidencias entre True y Predicho sobre la cantidad de elementos de True, que va a ser la misma cantidad que los elementos del dataset:

$Cant(True \cap Predicho) \ / \ Cant(True)$

Accuracy se calcula como la intersección entre los True y los Predichos sobre todos los elementos del dataset.

Como vemos los 3 producen el mismo resultado numérico

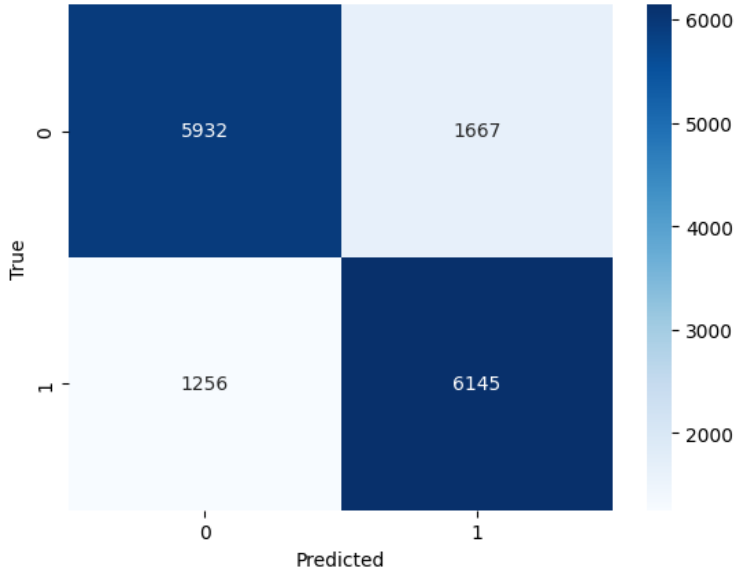
Bayes Naive:
Hiperparámetro escogido: alpha, el cual da 0.1
Vectorizer_max_features: ve la cantidad de palabras a usar en el vocabulario.
Vectorizer_stop_words:
Puntuación f1 score de validación cruzada: 0.8322260039899554



Score en Kaggle:

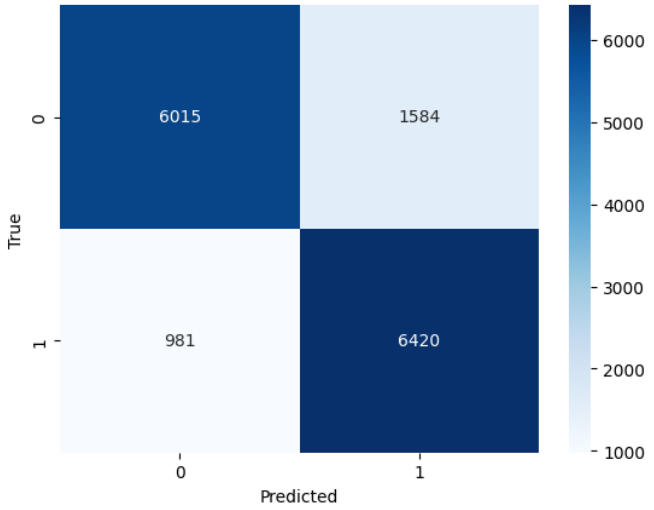
Random Forest:

.n_stimators: Para determinar la cantidad de árboles de decisión.
.max_depth: Para determinar la profundidad de cada árbol de decisión.
.min_samples_split: Mínimo de muestras para separar un nodo interno .
.min_samples_leaf: Mínimo de muestras requeridas para ser hoja.



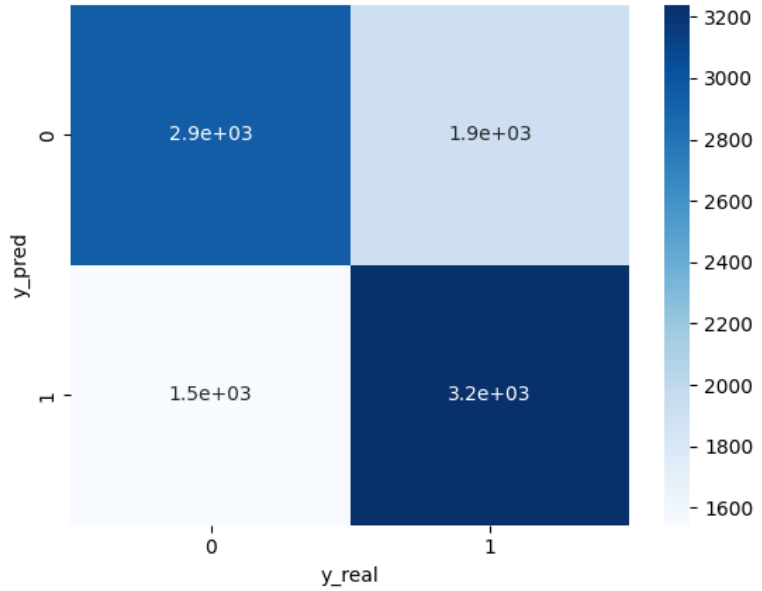
XGBoost:

n_estimators: Número de árboles que llevan a cabo el boosting
max_depth: profundidad de cada arbol de decision
learning_rate: cuanto aprende en cada iteracion, pasos chicos previenen el overfitting



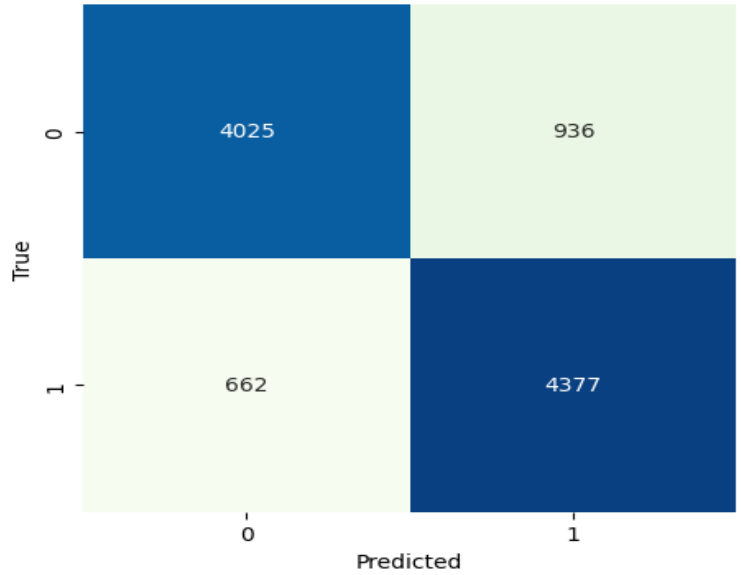
:

Red Neuronal:



The f1score countVectorizer is 0.6412784061429906
The accuracy countVectorizer is 0.6412784061429906
The precision countVectorizer is 0.6412784061429906
The recall countVectorizer is 0.6412784061429906

Ensamble Voting:



Accuracy: 0.8402

Score en Kaggle: 0.73773