.UBAfiuba
FACULTAD DE INGENIERÍA

75.06 /95.58 Organización de Datos
Dr. Ing. Rodríguez - 2°C 2023

Informe Final - Grupo 29

Integrantes	Padrón
Isidro Héctor Borthaburu	108901
María Delfina Cano Ros Langrehr	109338
Martín Wainwright	108211

TP2: Críticas Cinematográficas - Grupo XX

Introducción

En este trabajo práctico trabajamos con una compilación de críticas cinematográficas en español con el propósito de clasificarlas como positivas o negativas. El conjunto de datos tiene 7262 registros distribuidos en 3 columnas: *ID*, *review_es* (reseña en español) y sentimiento (indicando si la crítica es positiva o negativa).

Desarrollamos un conjunto de métodos y modelos de clasificación para analizar el sentimiento expresado en las reseñas.

Nos centramos en 5 modelos de clasificación, trabajando con la búsqueda de hiperparametros y verificando la predicción en la plataforma Kaggle. Estos métodos son: XG-Boost, Random Forest Classifier, Redes Neuronales, Bayes-Naive y un modelo de ensamble Voting, este último teniendo un XG-Boost, un Random Forest Classifier y un Bayes-Naive.

También realizamos un trabajo de preprocesamiento y normalización en el conjunto de los datos para poder optimizar a fondo las clasificaciones. Además, utilizamos “nltk” para poder pasar a lowercase, extraer palabras conectoras del dataset y tokenizar las reseñas. Para realizar este último, utilizamos el método *Bag of Words* en el que obtenemos un vector con todas las palabras de las reseñas. Las palabras no están ordenadas de ninguna forma y cada una puede aparecer más de una vez, incluso se cuentan todas las palabras.

Por último, tomamos como supuesto que el modelo de *Redes Neuronales* va a ser el que mejor predicción nos dé.

Cuadro de Resultados

Modelo	F1-Test	Presicion Test	Recall Test	Accuracy	Kaggle
XgBoost	0.9123	0.878	0.9495	0.9092	0.72087
Random Forest	0.84	0.86	0.81	0.83	0.71622
Red Neuronal	0.8485	0.94	0.77	0.86	0.742

Bayes Naive	0.8655	0.895	0.8258	0.8661	0.73076
Voting	0.8718	0.880	0.8593	0.8627	0.71835

Descripción de Modelos

Para comenzar, entrenamos un modelo simple de *XGBoost* el cuál consiste en construir distintos árboles de decisión donde cada nuevo árbol creado, corrige los errores de los árboles anteriores, y así se mejora constantemente las decisiones tomadas. Este modelo tiene regularización para poder evitar el sobreajuste de los datos, maneja automáticamente los datos faltante y ,además, utiliza la poda de árboles.

Para mejorar la predicción, buscamos hiper parámetros que mejoren nuestro modelo. Utilizamos distintos estimadores como: *frecuencia_máxima*, *frecuencia_minima*, *profundamente_maxima*, *learning_rate*, entre otros.
Luego, implementamos el modelo de *Random Forest* el cuál consiste en varios árboles que toman decisiones basándose en las características de los datos, por último, el modelo combina todas las decisiones tomadas y devuelve lo más “elegido” o “votado” por todos los árboles, y de esta forma tomar un decisión con mayor precisión. También, realizamos una búsqueda de hiperparámetros para este modelo utilizando el método Grid *Search*.

Continuando, implementamos un nuevo modelo para nosotros que fue el *Bayes Naive* el cuál utiliza el teorema de Bayes de probabilidad para calcular las probabilidades condicionales y las utiliza para clasificar las distintas reseñas. Este modelo asume que las características de entrada son independientes entre sí, dada la clase de salida.
Nuestro mejor modelo fue el de Redes Neuronales el cuál consiste en

En adición construimos, entrenar y evaluar un modelo de red neuronal para la clasificación binaria de sentimientos, agrupandolas en variables '. Se dividen los datos en conjuntos de entrenamiento y prueba, y luego se define un modelo secuencial que incluye capas de incrustación, aplanado y una capa densa con una unidad de salida y activación sigmoide. El modelo se compila con una función de pérdida binaria, un optimizador RMSprop y se monitoriza la precisión. Finalmente, el modelo se entrena con los datos de entrenamiento.

Por último, implementamos un modelo de ensamble que es un técnica la cuál combina múltiples modelos de aprendizaje para no depender solamente de uno y poder realizar una predicción más precisa. Decidimos utilizar un modelo de ensamble Híbrido llamado *Voting*. Para este, decidimos implementar tres distintos modelos de decisión que fueron: *XG-Boost, Random Fores y Bayes-Naive*. Este ensamble consiste en que se implementan estos modelos utilizando los mismos datos y cada uno toma una decisión, y finalizando se toma la predicción mayoritaria o ponderada.

Conclusiones generales

Para concluir, consideramos que el preprocesamiento y el análisis exploratorio de los datos fue completamente necesario ya que nos permitió “limpiar” el dataset, por ejemplo, que no haya valores nulos. También pudimos observar y analizar el conjunto de datos, implementar el modelo de *Bag of Words*, tokenizar el texto, entre otros aspectos. Creemos que este trabajo permitió que se mejore la performance de los modelos porque, como explicamos anteriormente, pudimos observar y analizar variables irrelevantes para el análisis, que todas los textos estén en minúscula, que todos los textos sean strings, y demás.

El modelo que tuvo el mejor desempeño en el *test* y en Kaggle fue la Red Neuronal, también este modelo fue uno de los más sencillos de entrenar y de los que más rápido se desarrollaron. Esta velocidad de aprendizaje es muy útil ya que se pueden hacer una mayor cantidad de pruebas.

Tareas Realizadas

Teniendo en cuenta que el trabajo práctico tuvo una duración de 9 (nueve) semanas, le pedimos a cada integrante que indique cuántas horas (en promedio) considera que dedicó semanalmente al TP

Integrante	Promedio Semanal (hs)
Isidro Hector Borthaburu	7
Maria Delfina Cano Ros Langrher	7
Martincito Wainwright	7