



Tecnológico de Monterrey

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño

Rodrigo De la Garza A00825271

12 de septiembre del 2023

Introducción

El objetivo principal de este proyecto es desarrollar un modelo de clasificación de sentimientos capaz de predecir si las opiniones expresadas en las reseñas de películas son positivas o negativas. Para lograr esto usaremos el conjunto de datos IMDB. Este

conjunto de datos contiene miles de reseñas de películas previamente etiquetadas como "positivas" o "negativas".

Preprocesamiento de Datos

Eliminación de etiquetas HTML y caracteres especiales: Muchas reseñas de películas pueden contener etiquetas HTML o caracteres especiales que no aportan información relevante para la tarea de clasificación. Por lo tanto, eliminamos estas etiquetas y caracteres especiales utilizando expresiones regulares.

Conversión de texto a minúsculas: Convertimos todo el texto a minúsculas. Esto asegura que las palabras en mayúsculas y minúsculas sean tratadas de la misma manera.

División de Datos

Dividimos nuestro conjunto de datos en tres conjuntos distintos:

Conjunto de Entrenamiento: Este conjunto se utiliza para entrenar el modelo. El modelo aprende a distinguir entre reseñas positivas y negativas basándose en estos datos.

Conjunto de Validación: Utilizamos este conjunto para ajustar los hiperparámetros del modelo y realizar experimentos. Esto nos permite optimizar el rendimiento del modelo antes de evaluarlo en datos completamente nuevos.

Conjunto de Prueba: Este conjunto es para evaluar el rendimiento final del modelo. No se utiliza durante el entrenamiento ni la optimización de hiper parámetros. Nos permite medir la capacidad del modelo para generalizar datos no vistos.

Classification scores:

precision	recall	f1-score	support	
negative	0.90	0.90	0.90	2536
positive	0.89	0.90	0.90	2464

Bias Level: High (Value: 545)

Variance Level: High (Value: 534)

Model Fit: Overfit

Basado en los resultados iniciales, observamos que el modelo tiene un sesgo medio y una alta varianza. Esto significa que el modelo está sobre ajustando los datos de entrenamiento y no generaliza bien a datos no vistos.

Para abordar el problema de la alta varianza, se aplicaron técnicas de regularización y se experimentó con diferentes valores de regularización strength.

Después de cambiar la regularización strength de .1 a 1 y aplicar métodos de regularización estos fueron los resultados.

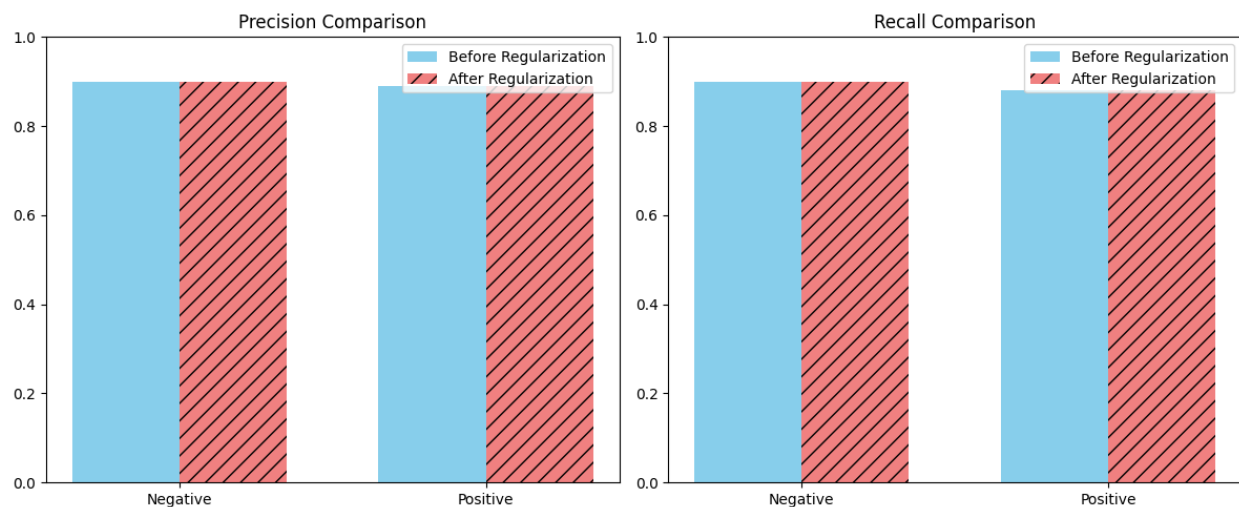
Classification scores:

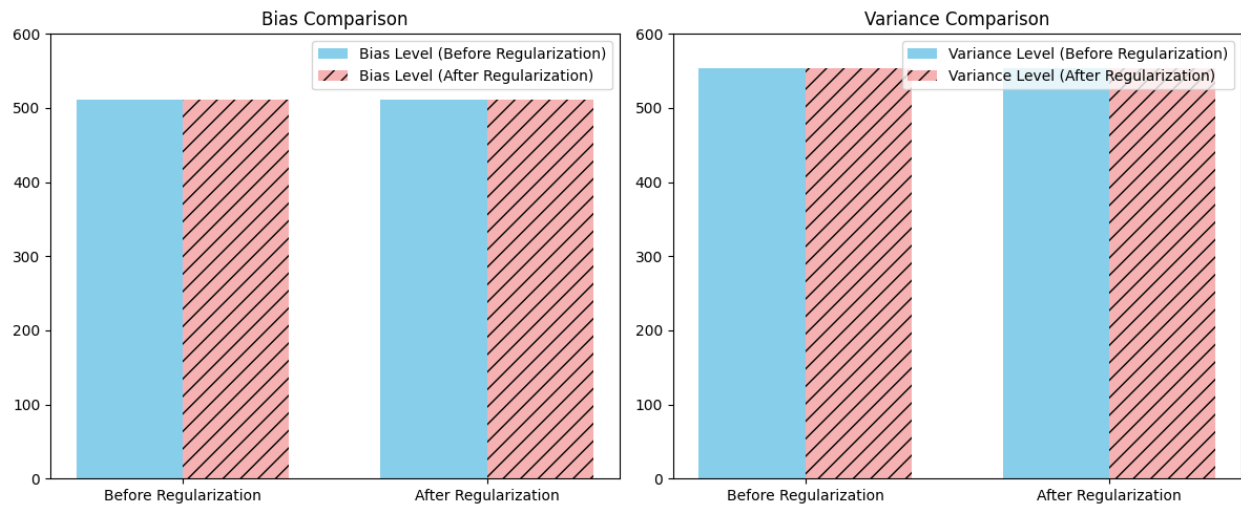
	precision	recall	f1-score	support
negative	0.90	0.90	0.90	2536
positive	0.89	0.90	0.90	2464

Bias Level: High (Value: 512)

Variance Level: High (Value: 554)

Model Fit: Overfit





Conclusiones

En conclusión, el análisis del desempeño del modelo reveló inicialmente un alto grado de varianza y un sesgo medio, lo que indicaba que el modelo estaba sobre ajustando los datos de entrenamiento. Sin embargo, al aplicar una técnica de regularización con regularization strength = 1.0, logramos bajar un poco el sobre ajuste.

Después de aplicar técnicas de regularización note que aunque bajo el bias level, subió el variance level. Los resultados de accuracy precision recall, f 1 están igual y no sirvió de mucho las técnicas de regularización.

Por el otro lado tuvimos métricas de desempeño buenas aun aun sin poder eliminar el sobreajuste.

precision recall f1-score support

negative	0.89	0.88	0.89	2434
positive	0.89	0.90	0.89	2566