CI-0134: Investigación en Ciencias de la Computación
Emmanuel Zúñiga - B98729

Semana 9: (15 al 19 de Mayo)

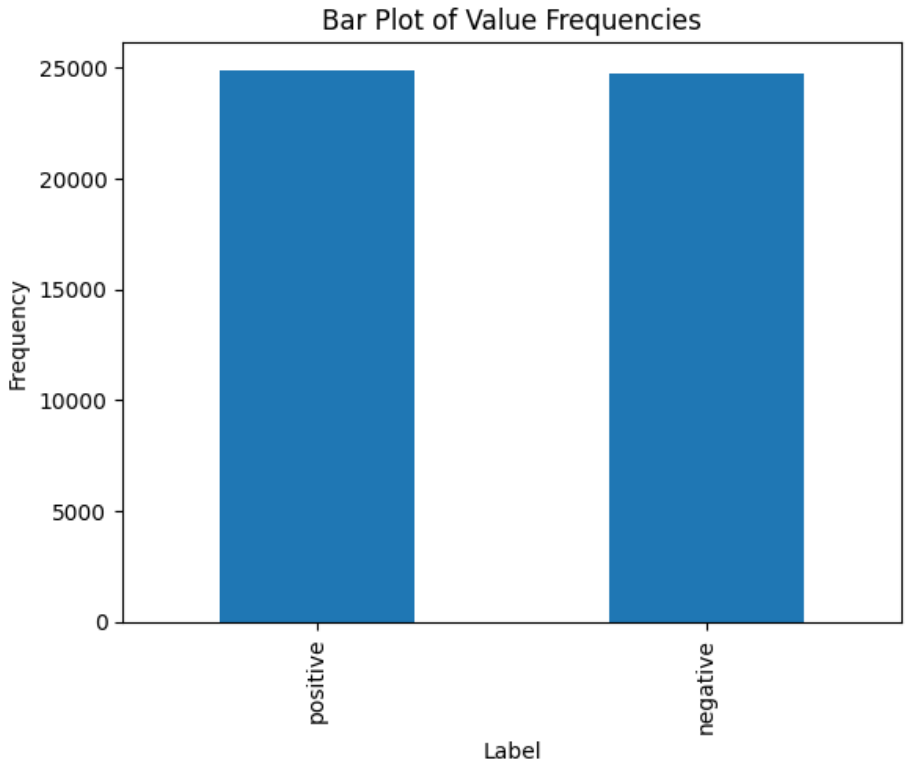
Tareas realizadas:

- ☑ Debido a la mala calidad del dataset seleccionado, se procede con la búsqueda de un nuevo dataset que se adecue a las necesidades del proyecto. Encontramos un dataset que contiene reseñas en inglés y traducción en español, clasificadas según la polaridad[1]. Por lo tanto, se procede a hacer el preprocesamiento de los datos y la experimentación.
- ☑ A continuación, se presenta un breve ejemplo de algunas entradas del dataset. En este caso, se muestra la entrada original y la entrada después del preprocesamiento:

Texto original en inglés	Texto original en español
Probably my all-time favorite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy or boring. It just never gets old, despite my having seen it some 15 or more times in the last 25 years.	Probablemente mi película favorita de todos los tiempos, una historia de desinterés, sacrificio y dedicación a una causa noble, pero no es predicada ni aburrida.Simplemente nunca envejece, a pesar de haberla visto unas 15 o más veces en los últimos 25 años.
Texto procesado en inglés	Texto procesado en español
probably time favorite movie story selflessness sacrifice dedication noble cause preachy boring never gets old despite seen 15 times last 25 years	probablemente pelicula favorita tiempos historia desinteresidad sacrificio dedicacion causa noble predicada aburrida simplemente nunca envejece pesar haberlo visto unas 15 mas veces ultimos 25 anos

Para el procesamiento de los datos se eliminan carácteres especiales, signos de puntuación, exclamación, interrogación; contracciones, stopwords, entre otros.
El notebook con el preprocesamiento de los datos se encuentra en la carpeta “notebooks” del repositorio.

[1] Enlace al dataset:
https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish



Igualmente existe un balanceo adecuado según las clases de los datos, por lo tanto, no se prevén problemas con respecto al dataset a la hora de ajustar los modelos.

☑ Se realiza la experimentación con el nuevo dataset, a continuación se presentan algunos resultados obtenidos con LSTM Bidireccional y Naïve Bayes, para ambos idiomas. Las clases están mapeadas como {1: positivo, 0: negativo}
LSTM Bidireccional:

Español:

```
410/410 [==============================] - 11s 26ms/step
Accuracy Score:  0.8144471594379963
Confusion Matrix:
 [[5608  929]
 [1501 5058]]
              precision    recall  f1-score   support

           0       0.79      0.86      0.82      6537
           1       0.84      0.77      0.81      6559

    accuracy                           0.81     13096
   macro avg       0.82      0.81      0.81     13096
weighted avg       0.82      0.81      0.81     13096
```

Inglés:

```
410/410 [==============================] - 11s 24ms/step
Accuracy Score:  0.8545357361026268
Confusion Matrix:
 [[5608  929]
 [ 976 5583]]
              precision    recall  f1-score   support

           0       0.85      0.86      0.85      6537
           1       0.86      0.85      0.85      6559

    accuracy                           0.85     13096
   macro avg       0.85      0.85      0.85     13096
weighted avg       0.85      0.85      0.85     13096
```

Naïve Bayes:

Español:

```
              precision    recall  f1-score   support

           0       0.84      0.86      0.85      3960
           1       0.86      0.84      0.85      3976

    accuracy                           0.85      7936
   macro avg       0.85      0.85      0.85      7936
weighted avg       0.85      0.85      0.85      7936
```

Inglés:

```
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      3960
           1       0.88      0.83      0.85      3976

    accuracy                           0.86      7936
   macro avg       0.86      0.86      0.86      7936
weighted avg       0.86      0.86      0.86      7936
```

Los cuatro resultados mostrados anteriormente serán analizados en conjunto a los resultados obtenidos en Regresión Logística y la revisión bibliográfica, para así poder dar respuesta a la pregunta de investigación planteada.

Los notebooks en cuestión donde se encuentran las implementaciones de los modelos se encuentran en la carpeta “notebooks” del repositorio. El dataset utilizado no pudo ser subido al repositorio debido a su tamaño.

Para la próxima semana:

- Experimentación con diferentes combinaciones de hiper parámetros.
- Elaboración de un reporte con los resultados.
- Lectura de trabajos relacionados.