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

Semana 6: (24 al 28 de marzo)

Tareas realizadas:

- ☑ Se hizo un recorte en el dataset, dejando únicamente dos categorías {positive y negative}.
- ☑ Se realizó la experimentación en los tres algoritmos, adaptándose para funcionar con las nuevas características del dataset. No obstante, los resultados fueron peor de los obtenidos anteriormente. Por lo cual **se optó por buscar más datos**, se encontró un dataset con más de 170 mil datos de tweets en español clasificados en positivo y negativo. Puede consultarse el dataset[1] en el repositorio.
- ☑ Se realiza el ajuste del dataset para adaptarlo a las necesidades del proyecto para posteriormente realizar la experimentación con el nuevo conjunto de datos.
- ☑ El día 25 de Abril, se realizó una reunión con mi compañera Paula Monge para la experimentación con los modelos. Se pueden consultar los modelos en el repositorio[2].
- ☑ El día 26 de abril, se realizó una reunión para probar los modelos con diferentes hiperparametros. A continuación, se muestran las estadísticas obtenidas según algunas combinaciones de hiperparametros:
  LSTM Bidireccional:

  Learning rate = 0.01
  Función de activación = sigmoid
  Epocas = 20
  Batch Size = 64
  Resultados:

	precision	recall	f1-score	support
0	0.77	0.81	0.79	18318
1	0.65	0.59	0.62	10912
accuracy			0.73	29230
macro avg	0.71	0.70	0.70	29230
weighted avg	0.72	0.73	0.72	29230

[1] Enlace al dataset:
https://github.com/Emmazch22/Paula_y_Emmanuel_CI0134/blob/main/Datasets/new_dataset_balanced.csv

[2] Enlace a los notebooks:
https://github.com/Emmazch22/Paula_y_Emmanuel_CI0134/tree/main/Notebooks

Learning rate = 0.001
Función de activación = sigmoid
Epocas = 20
Batch Size = 64

Resultados:

	precision	recall	f1-score	support
0	0.77	0.78	0.77	18318
1	0.62	0.60	0.61	10912
accuracy			0.71	29230
macro avg	0.69	0.69	0.69	29230
weighted avg	0.71	0.71	0.71	29230

Learning rate = 0.0001
Función de activación = sigmoid
Epocas = 20
Batch Size = 64

Resultados:

	precision	*recall*	*f1-score*	*support*
0	*0.77*	*0.81*	*0.79*	*18318*
1	*0.65*	*0.58*	*0.62*	*10912*
accuracy			*0.73*	*29230*
macro avg	*0.71*	*0.70*	*0.70*	*29230*
weighted avg	*0.72*	*0.73*	*0.72*	*29230*

Regresión Logística:
Para el modelo de regresión logística únicamente se utiliza un hiper parámetro, siendo la cantidad máxima de iteraciones.

max_iters = 1000
ngram_range = (1,2)
max_features = 1000
C = 1

Resultados:

	precision	recall	f1-score	support

```
              precision    recall  f1-score   support

           0       0.74      0.74      0.74     11111
           1       0.74      0.73      0.74     11032

    accuracy                           0.74     22143
   macro avg       0.74      0.74      0.74     22143
weighted avg       0.74      0.74      0.74     22143
```

max_iters: 4000
ngram_range = (1,2)
max_features = 50000
C = 1

Resultados:

```
              precision    recall  f1-score   support

           0       0.77      0.80      0.79     11111
           1       0.79      0.76      0.78     11032

    accuracy                           0.78     22143
   macro avg       0.78      0.78      0.78     22143
weighted avg       0.78      0.78      0.78     22143
```

max_iters = 4000
ngram_range = (1,2)
max_features = 50000
C = 0.01

Resultados:

```
              precision    recall  f1-score   support

           0       0.75      0.79      0.77     11111
           1       0.77      0.73      0.75     11032

    accuracy                           0.76     22143
   macro avg       0.76      0.76      0.76     22143
weighted avg       0.76      0.76      0.76     22143
```

max_iters = 1000
ngram_range = (1,3)
max_features = 50000
C = 1

Resultados:

```
              precision    recall  f1-score   support
```

```
              0       0.77      0.79      0.78     11111
              1       0.78      0.76      0.77     11032

    accuracy                            0.77     22143
   macro avg       0.77      0.77      0.77     22143
weighted avg       0.77      0.77      0.77     22143
```

Naïve Bayes:

alpha = 0.5
fit_prior = True

Resultados:

```
              precision    recall  f1-score   support

           0       0.71      0.87      0.78      8835
           1       0.83      0.65      0.73      8879

    accuracy                           0.76     17714
   macro avg       0.77      0.76      0.75     17714
weighted avg       0.77      0.76      0.75     17714
```

alpha = 1
fit_prior = True

Resultados:

```
              precision    recall  f1-score   support

           0       0.71      0.87      0.78      8835
           1       0.83      0.64      0.73      8879

    accuracy                           0.76     17714
   macro avg       0.77      0.76      0.75     17714
weighted avg       0.77      0.76      0.75     17714
```

Según los resultados obtenidos en los modelos, se puede observar como la utilización de un modelo u otro no afecta considerablemente en los resultados, dado que las estadísticas obtenidas no presentan variaciones importantes.

Estos resultados fueron obtenidos al clasificar textos de un dataset con más de 100 mil datos en español. Para contrastar estas estadísticas, se procederá a evaluar los resultados obtenidos con los mismos hiper parámetros pero con datos en inglés, obtenidos de otro dataset con las mismas características.

Para la próxima semana:

- Elaboración de la presentación para el avance 2.
- Correcciones del avance 1.

- Experimentación con datos en inglés para contrastar los resultados. Se plantea utilizar un dataset con la misma cantidad de datos y las mismas características en los textos, a excepción del lenguaje.