Proyecto 1.1

Análisis de sentimientos de películas

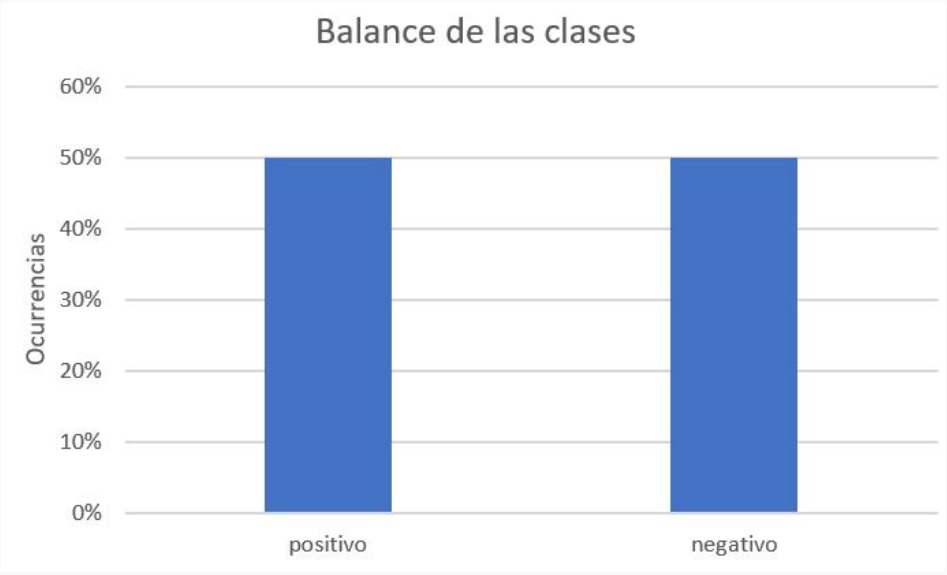
Santiago Campo - 201921995
Camilo Falla - 201821059
Pablo Pastrana - 201822920

Entendimiento de los datos
Clases Balanceadas

Encontramos que las clases estaban perfectamente balanceadas.

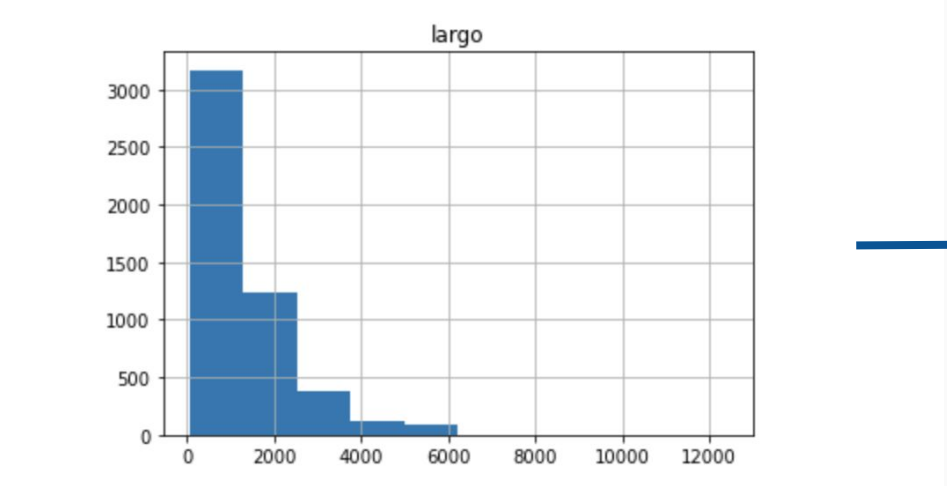
```
df["sentimiento"].value_counts(dropna = False, normalize = True)
```

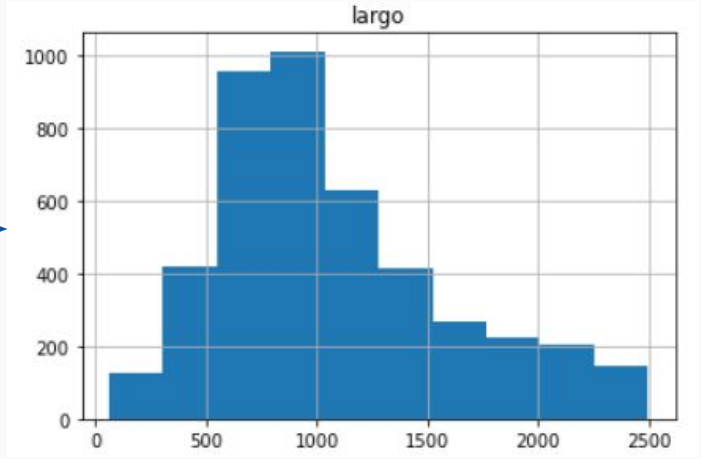
```
positivo    0.5
negativo    0.5
Name: sentimiento, dtype: float64
```



Entendimiento de los datos

Longitud Cadenas



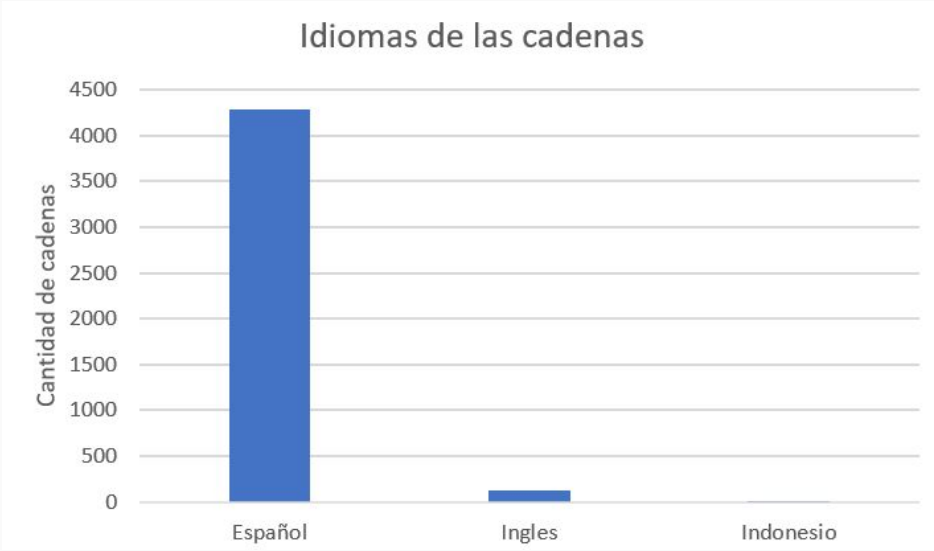


Eliminamos las cadenas que tenían más de 2500 caracteres.

Entendimiento de los datos
Idioma Cadenas

Eliminamos las cadenas que no estaban en español.

```
es    4281
en     118
id       1
Name: idioma, dtype: int64
```



Pre-Procesamiento
Limpieza y Lemmatización

- Eliminar caracteres especiales
- Eliminar caracteres singulares (a, y, o)
- Eliminar espacios innecesarios
- Eliminar stop-words
- Convertir todo a minúscula

Lemmatización: Reducir un conjunto de palabras a solo una representación dependiendo de distintas formas morfológicas.

WordNetLemmatizer()

Pre-Procesamiento
Bag of Words

En este modelo, se considera cada documento como un conjunto no ordenado de palabras y se cuenta el número de veces que aparece cada palabra en el documento.

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Pre-Procesamiento
Tfid Transformer

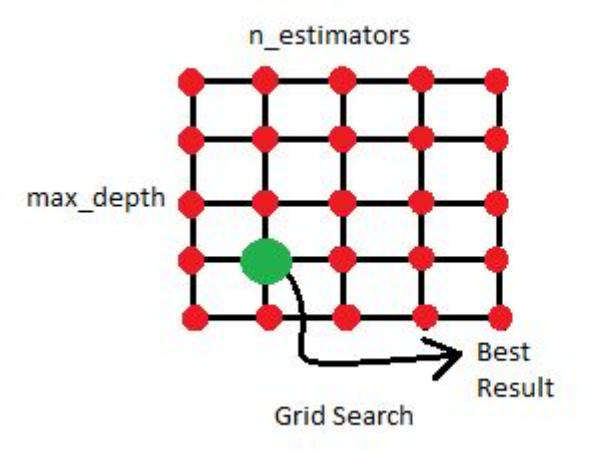
El modelo TF-IDF asigna un peso numérico a cada palabra en un documento, de acuerdo con la frecuencia de esa palabra en el documento y la frecuencia de esa palabra en todo el corpus de documentos.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Algoritmos Utilizados

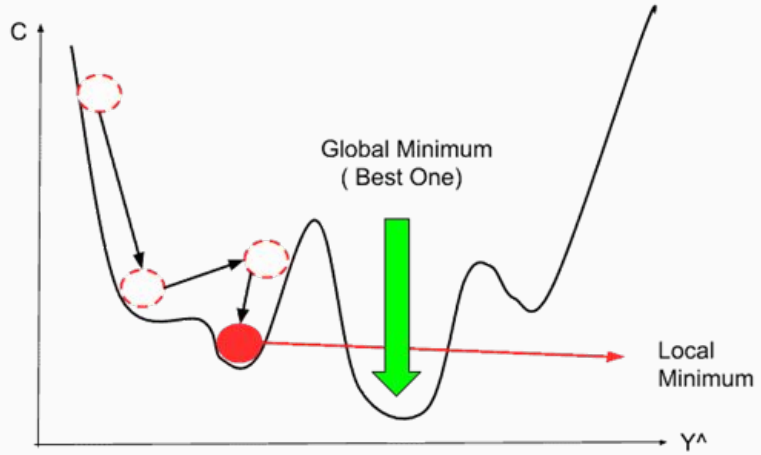
GridSearch - Todos los modelos

GridSearch es una técnica para encontrar los mejores hiper-parámetros para un modelo de Machine Learning.



Stochastic Gradient Descent (SGD)

Es un tipo de clasificador lineal que utiliza el descenso de gradiente estocástico para optimizar la función de pérdida y encontrar los mejores pesos para las características del modelo.



Stochastic Gradient Descent (SGD)

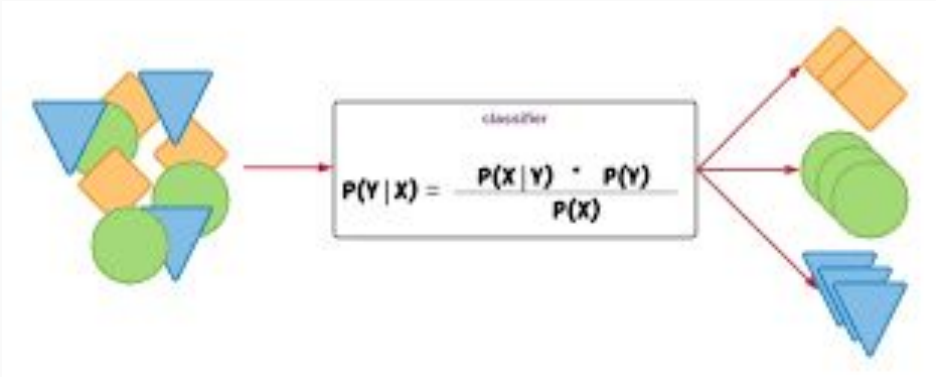
	precision	recall	f1-score	support
negativo	0.85	0.87	0.86	424
positivo	0.87	0.85	0.86	433
accuracy			0.86	857
macro avg	0.86	0.86	0.86	857
weighted avg	0.86	0.86	0.86	857

Accuracy: 0.86

	Actual Pos	Actual Neg
Pred Pos	370	54
Pred Neg	66	367

Multinomial Naive Bayes

El modelo Naive Bayes clasifica los textos en categorías utilizando una función de probabilidad condicional que mide la probabilidad de que un texto pertenezca a una determinada categoría dado el conjunto de características del texto.



Multinomial Naive Bayes

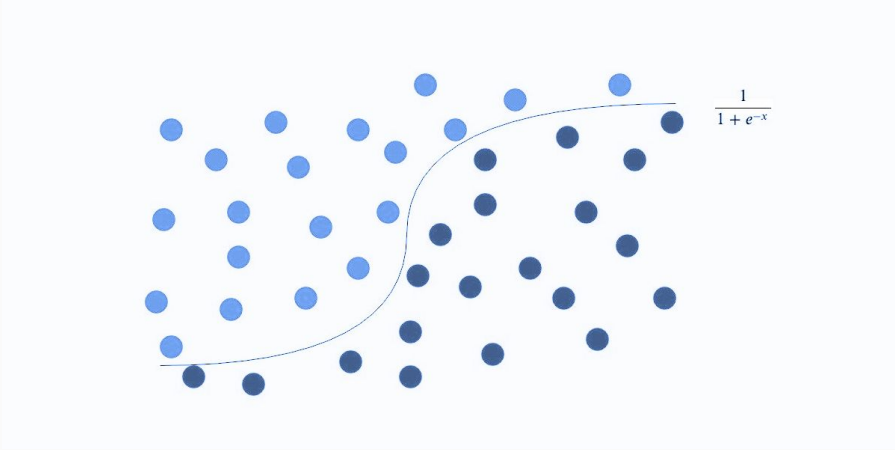
	precision	recall	f1-score	support
negativo	0.82	0.87	0.84	424
positivo	0.86	0.82	0.84	433
accuracy			0.84	857
macro avg	0.84	0.84	0.84	857
weighted avg	0.84	0.84	0.84	857

Accuracy: 0.8425

	Actual Pos	Actual Neg
Pred Pos	367	57
Pred Neg	78	355

Regresión Logística

La regresión logística utiliza una función sigmoide para transformar una entrada lineal en una salida entre 0 y 1, que representa la probabilidad de que una instancia pertenezca a una de las categorías.



Regresión Logística

	precision	recall	f1-score	support
negativo	0.86	0.83	0.84	424
positivo	0.84	0.86	0.85	433
accuracy			0.85	857
macro avg	0.85	0.85	0.85	857
weighted avg	0.85	0.85	0.85	857

Accuracy: 0.84597

	Actual Pos	Actual Neg
Pred Pos	351	73
Pred Neg	59	374

Modelo seleccionado:

Stochastic Gradient Descent (SGD) Classifier

- loss: hinge
- penalty: l2
- alpha: 0.001
- max_iter:17

	precision	recall	f1-score	support
negativo	0.85	0.87	0.86	424
positivo	0.87	0.85	0.86	433
accuracy			0.86	857
macro avg	0.86	0.86	0.86	857
weighted avg	0.86	0.86	0.86	857

Accuracy: 0.86

Impacto en el negocio

Los resultados encontrados tendrán un gran impacto en el negocio.

1. Mejorar la experiencia del usuario al mostrarle comentarios buenos o malos dependiendo de la preferencia que cada usuario tenga. De este modo, la satisfacción general de los usuarios aumentará, lo cual será bastante beneficioso para la rentabilidad general de la organización.

2. Además, la empresa también podrá hacer uso de la categorización automática de los comentarios para saber qué series están teniendo mayor aceptación y de esta manera, poder mantenerlas en oferta por un mayor tiempo.

Gracias!