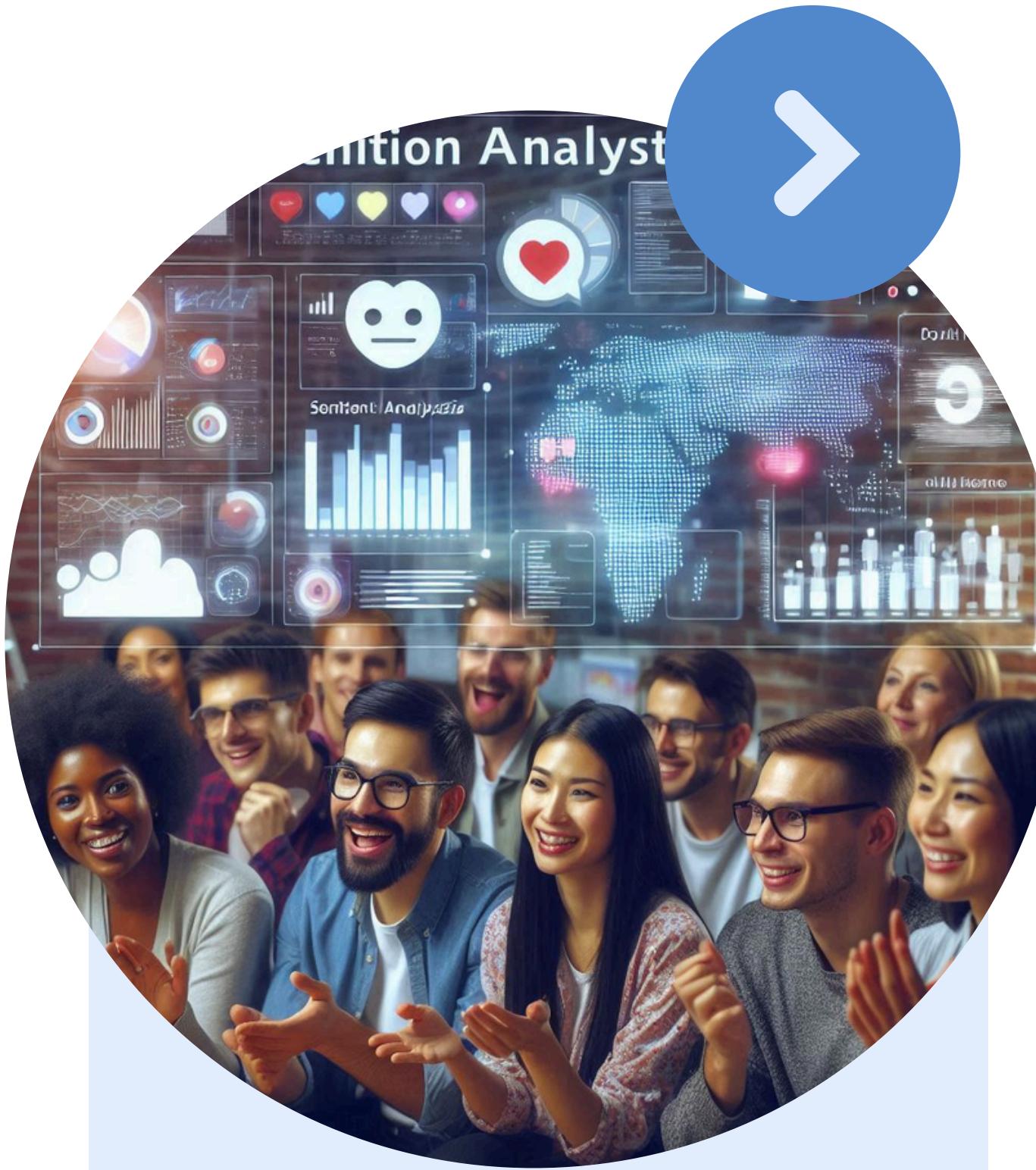


SENTIMENT ANALYSIS

BOOTCAMP XPERIENCE

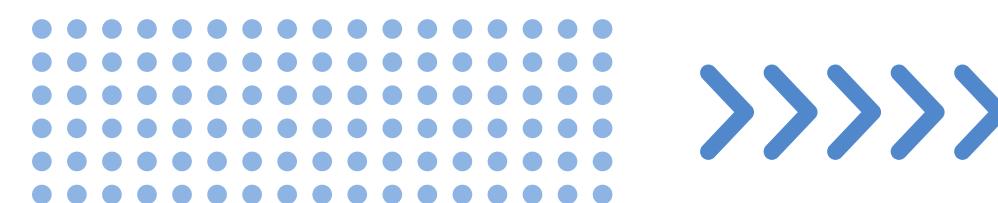
GRUPO 1





Introduction

El objetivo de este proyecto es realizar un análisis de sentimientos en los comentarios, clasificándolos en categorías de positivos o negativos, para comprender mejor las emociones y opiniones expresadas en los textos.



TRATAMIENTO DE DATOS

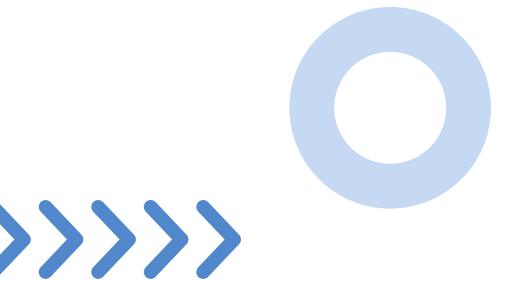
Se realizan varios tratamientos con la finalidad de limpiar nuestros comentarios y poder incluirlos en nuestro modelo de machine learning.



- Se aborda el tratamiento de hashtags, URLs, signos de puntuación, emoticones, número a texto, apóstrofes, mayúsculas, entre otros.
- Eliminación de stop words: Se abordó la eliminación de palabras comunes que no aportan significado relevante al análisis, como artículos, preposiciones y conjunciones.
- Stemming: Se aplicó un proceso de reducción de palabras a su raíz o forma base, eliminando sufijos y prefijos para normalizar términos con significados similares.

	comentario	nota	sentimiento	tratamiento_1	tratamiento_2	tratamiento_3
0	No doubt it has a great bass and to a great ex...	3	negativo	no doubt it has a great bass and to a great ex...	doubt great bass great extent noise cancellati...	doubt great bass great extent nois cancel dece...
1	This earphones are unreliable, i bought it be...	1	negativo	this earphones are unreliable i bought it befo...	earphones unreliable bought fifteen days meanw...	earphon unreli bought fifteen day meanwhig...
2	i bought itfor 999,I purchased it second time,...	4	positivo	i bought itfor nine hundred and ninety-nine i ...	bought itfor nine hundred ninety-nine purchas...	bought itfor nine hundr ninety-nin purchas sec...
3	Its sound quality is adorable. overall it was ...	1	negativo	its sound quality is adorable overall it was g...	sound quality adorable overall good two weeks ...	sound qualiti ador overal good two week stop w...
4	Its Awesome... Good sound quality & 8-9 hrs ba...	5	positivo	its awesome good sound quality eight nine hrs ...	awesome good sound quality eight nine hrs batt...	awesom good sound qualiti eight nine hr batter...
...
14332	Good	4	positivo	good	good	good
14333	An amazing product but a bit costly.	5	positivo	an amazing product but a bit costly	amazing product bit costly	amaz product bit costli
14334	Sound	1	negativo	sound	sound	sound
14335	the sound is good battery life is good but the...	5	positivo	the sound is good battery life is good but the...	sound good battery life good wire long good pr...	sound good batteri life good wire long good pr...
14336	M writing this review after using for almost 7...	1	negativo	m writing this review after using for almost s...	writing review using almost seven months stopp...	write review use almost seven month stop work ...

FEATURE ENGINEERING



En nuestro dataframe vectorizado, donde se contabilizan la frecuencia de las palabras, consideramos adicionar nuevas variables con la finalidad de poder mejorar nuestro modelo.

- **Conteo de palabras totales:** Número total de palabras en el texto.
- **Conteo de palabras únicas:** Cantidad de palabras distintas en el texto.
- **Palabras positivas, negativas y neutras:** Clasificación de palabras según su carga emocional.
- **Emoción del texto:** Identificación del sentimiento predominante (alegría, tristeza, etc.).
- **Polaridad:** Medida de si el texto es positivo, negativo o neutral.
- **Subjetividad:** Grado en que el texto expresa opiniones personales o hechos objetivos.

FEATURE ENGINEERING

	also	amaz	amazon	awesom	backup	bad	bass	batteri	best	better	...	subjectivity	neg_words	neu_words	pos_words	emotion_anger	emotion_fear	emotion_joy	emotion_love	emotion_sadness	emotion_surprise
0	0	0	0	0	0	0	1	1	0	0	...	0	14	29	6	1	0	0	0	0	0
1	2	0	0	0	0	0	1	0	0	0	...	0	0	31	9	0	0	1	0	0	0
2	0	0	0	2	0	0	0	2	0	0	...	0	0	42	12	0	0	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	2	12	2	0	0	1	0	0	0
4	0	0	2	1	0	1	0	1	0	1	...	0	5	16	7	0	0	0	0	1	0
...	
14315	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	1	0	0	0
14316	0	1	0	0	0	0	0	0	0	0	...	0	0	4	0	0	0	1	0	0	0
14317	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	1	0	0	0
14318	0	0	0	0	0	0	0	1	0	0	...	0	0	4	5	0	0	1	0	0	0
14319	0	0	0	0	0	0	0	0	0	0	...	0	4	23	0	1	0	0	0	0	0

14320 rows x 113 columns

ENTRENANDO EL MODELO

```
### Dividir el conjunto de datos en conjuntos de entrenamiento y prueba
from sklearn.model_selection import train_test_split

## standarización de matriz
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
matriz_dispersa_evaluaciones_scaler = scaler.fit_transform(matriz_dispersa_evaluaciones_new_new)

X_entrenamiento, X_prueba, y_entrenamiento, y_prueba = train_test_split(matriz_dispersa_evaluaciones_scaler, df.sentimiento, random_state=4978)
```

```
### Crear y entrenar el modelo de regresión logística
from sklearn.linear_model import LogisticRegression
regresion_logistica = LogisticRegression()
regresion_logistica.fit(X_entrenamiento, y_entrenamiento)
exactitud = regresion_logistica.score(X_prueba, y_prueba)
print(exactitud)
```

0.8108938547486033

```
## Verificando existencia de sobreajuste del modelo
## Evaluamos en los datos de entrenamiento, para ver que valor nos sale el accuracy.
```

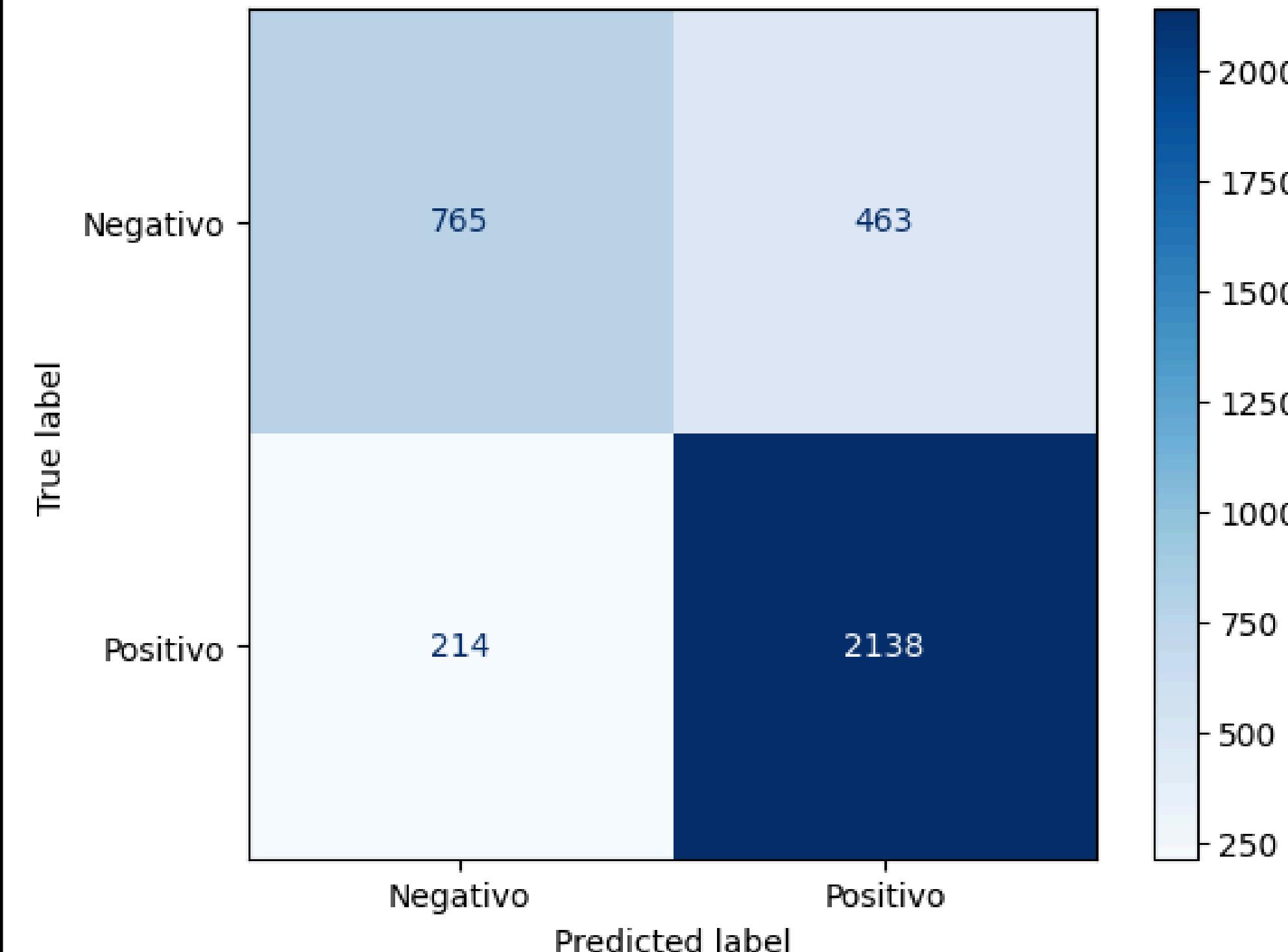
```
exactitud_train = regresion_logistica.score(X_entrenamiento, y_entrenamiento)
print(exactitud_train)
```

0.8189944134078212

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

# Calcular matriz de confusión
cm = confusion_matrix(y_prueba, regresion_logistica.predict(X_prueba))
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=["Negativo", "Positivo"])
disp.plot(cmap='Blues')
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x791f4e9f2c50>
```



CONCLUSIONES

- Balance de emociones: Mejorar el equilibrio entre emociones negativas y positivas ayudaría al modelo a generalizar mejor y aumentar su precisión.
- Ampliación de características: Agregar más características enriquecería el modelo, haciéndolo más robusto y mejorando su capacidad de clasificación.
- Explorar redes neuronales: Probar con redes neuronales podría mejorar el rendimiento y la capacidad de generalización, ya que no se ha observado sobreajuste hasta ahora.
- Más datos: Ampliar la base de datos permitiría mejorar la precisión del modelo y manejar una mayor variedad emocional en los textos.
- Manejo de emoticones: Es importante tratar los emoticones de manera adecuada, ya que se eliminaron en la fase de preprocesamiento, lo que dejó algunas líneas vacías. Esto puede impactar la calidad de los datos y el rendimiento del modelo, por lo que debería considerarse su inclusión o un manejo adecuado.
- Tuning de hiperparámetros: Experimentar con diferentes configuraciones de hiperparámetros podría mejorar aún más el rendimiento y optimizar el modelo.