



UNIVERSIDAD PARAGUAYO ALEMANA

OPEN DATA & BIG DATA

**Proyecto de Big Data y Machine
Learning**

Análisis de Sentimiento en Twitter

Recomendación de Películas

Sebastian Ovelar

San Lorenzo, Paraguay - Junio 2024

Resumen

El objetivo de este proyecto es analizar el sentimiento de los tweets de clientes y hacer una predicción de su satisfacción. Se utiliza un conjunto de datos que incluyen calificaciones de sentimiento, positivas , negativas o neutrales. El objetivo es hacer una limpieza y analizar dichos tweets, extrayendo información sobre la polaridad y subjetividad de los tweets. Además utilizaremos biblioteca. También el proyecto realiza un modelo de aprendizaje automático y evalúa esos resultados con métricas .

Este proyecto también analiza la recomendación de películas, esto es independiente a los tweets. También se usarán herramientas de clasificación de sentimiento, polaridad y subjetividad así crear un sistema de recomendaciones de películas y evaluarlo con métricas.

Compararemos los proyectos, el éxito de uno y el fracaso del otro para comentar el aprendizaje de las mismas.

Datos del proyecto

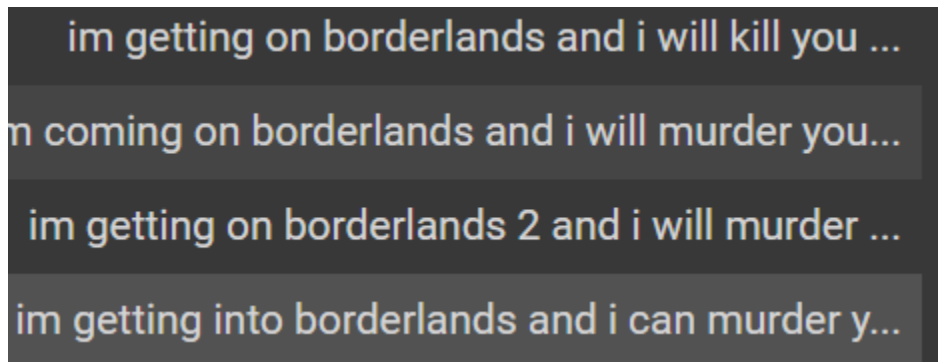
Los datos de proyecto utilizados provienen de Kaggle. Para el analysis de twitter se usa el dataset

[Twitter Sentiment Analysis \(kaggle.com\)](https://www.kaggle.com/datasets/rohitkumar9001/twitter-sentiment-analysis) con mas de 74 mil entradas y para las películas se usa el dataset [Movie Reviews Dataset: 10k+ Scraped Data \(kaggle.com\)](https://www.kaggle.com/datasets/rohitkumar9001/movie-reviews-dataset) que si bien tiene muchos errores, se hizo el esfuerzo para compilar algunas informaciones.

1.Análisis de Sentimiento en Twitter

1.1 Exploración de Datos (EDA) y Procesamiento de Datos

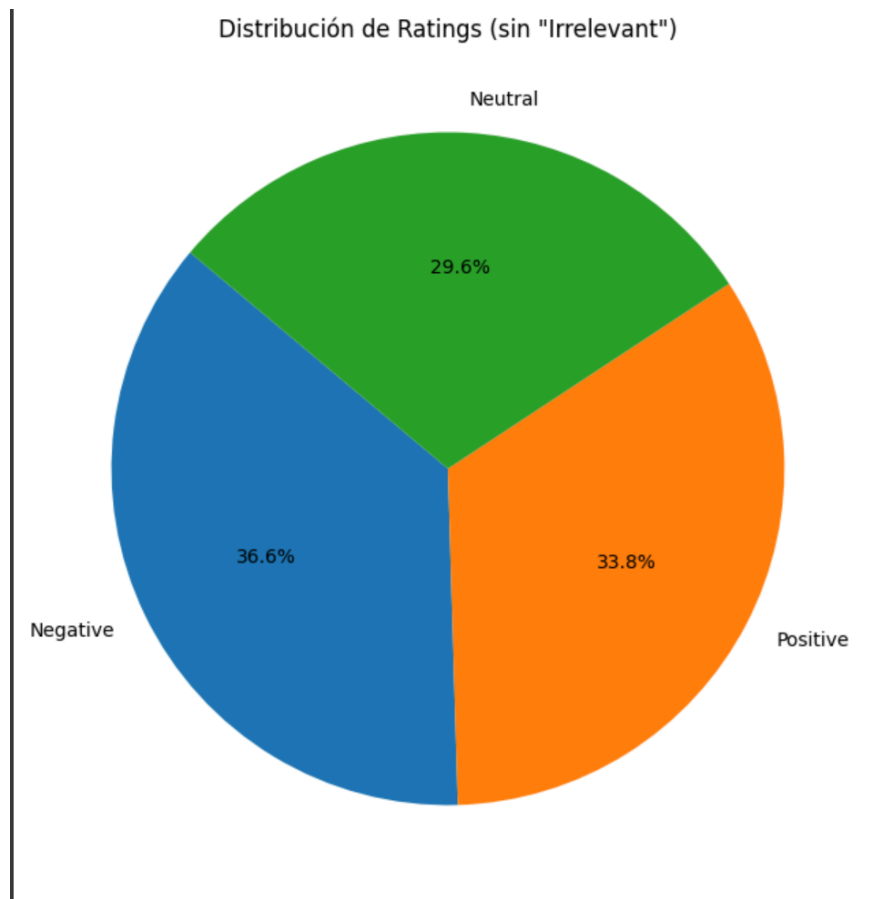
En este proyecto se utilizaron librerías como Pandas , Seaborn , Matplotlib y otros mas para la exploracion y el procesamiento. Primero se verifica la presencia de valores nulos en una columna de opiniones y se eliminan las filas correspondientes. Luego se creó una función para limpiar el texto, eliminando caracteres especiales y convirtiendo todo a minúscula.



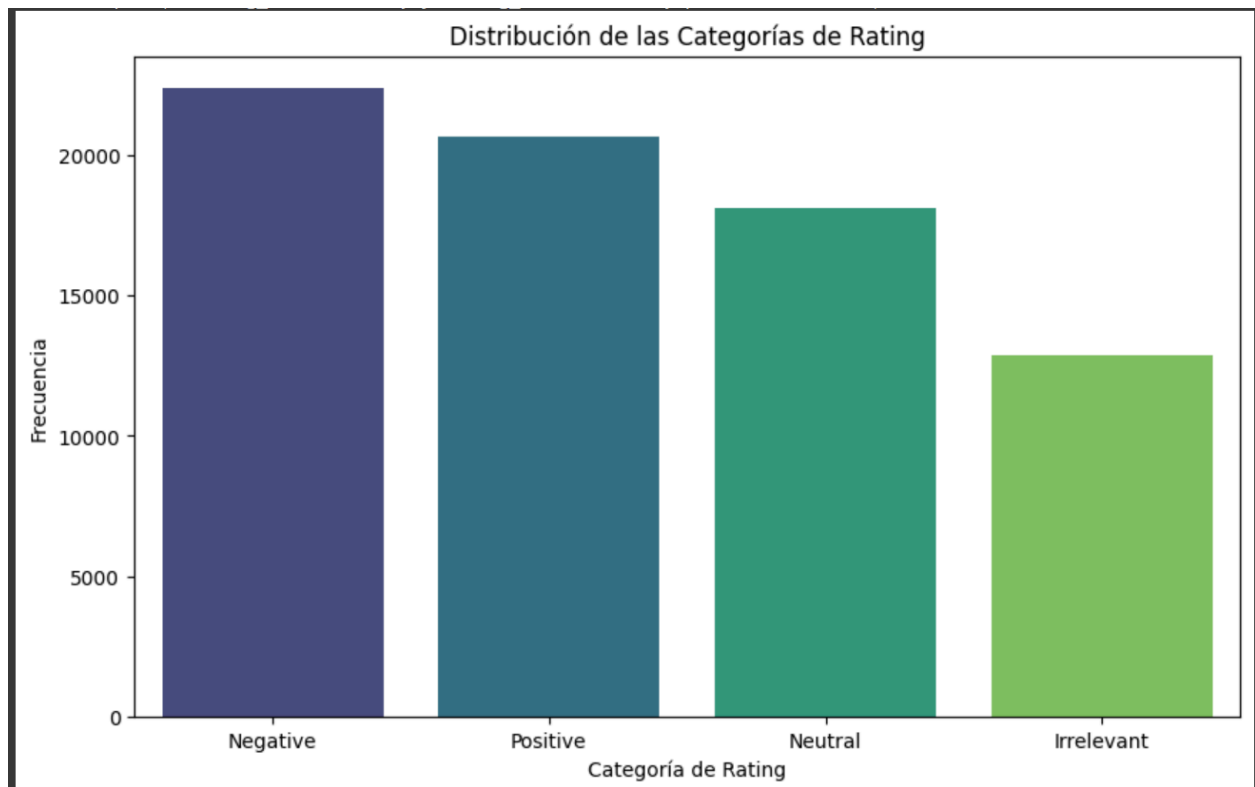
im getting on borderlands and i will kill you ...
n coming on borderlands and i will murder you...
im getting on borderlands 2 and i will murder ...
im getting into borderlands and i can murder y...

El gráfico a continuación muestra la distribución de los ratings excluyendo datos irrelevantes. Se puede ver que excluyendo los datos irrelevantes, los ratings son casi

parejos, solo predominan los negativos por un margen muy pequeño.



Si incluimos los ratings, de todos modos hay más negativos que positivos, hay muchos neutros y muchos irrelevantes pero no a la escala de los otros tres. Sin embargo, esto no quiere decir que todos los juegos y títulos sean más malos que buenos.

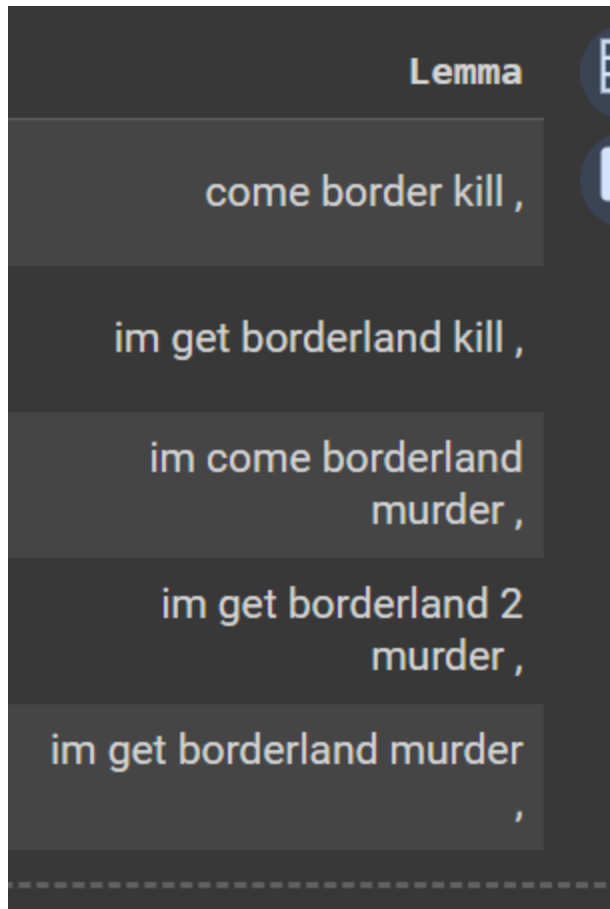


1.2 Extracción de Características

Podemos ver en la siguiente imagen como utilizando NLKT, se crea el Pos Tagging. Que adjunta cada palabra a su grupo (Sustantivo, Verbo, etc)

clean opinion	POS tagged
m coming to the borders and I will kill you...	[(coming, v), (borders, n), (kill, v), (, , None)]
n getting on borderlands and i will kill you ...	[(im, n), (getting, v), (borderlands, n), (kil...
ming on borderlands and i will murder you...	[(im, n), (coming, v), (borderlands, n), (murd...
getting on borderlands 2 and i will murder ...	[(im, n), (getting, v), (borderlands, n), (2, ...
etting into borderlands and i can murder y...	[(im, n), (getting, v), (borderlands, n), (mur...

Luego lemmatization, que transforma las palabras guardadas de Pos Tagging a su versión más simple



Se intenta buscar hashtags relevantes pero no se encuentra ninguno

		opinion	num_hashtags
0	I am coming to the borders and I will kill you...		0
1	im getting on borderlands and i will kill you ...		0
2	im coming on borderlands and i will murder you...		0
3	im getting on borderlands 2 and i will murder ...		0
4	im getting into borderlands and i can murder y...		0

Podemos luego ver de los ratings cuantos son positivos, negativos , neutros o irrelevantes

	rating	
	Negative	22358
	Positive	20654
	Neutral	18108
	Irrelevant	12875

1.3 Entrenamiento del Modelo

Se utilizaron (SVM) y Naive Bayes para la clasificación de las opiniones en categorías. Luego se evalúan los modelos creados utilizando métricas como recall, precisión, exactitud y F1-Score

Podemos decir que el modelo tiene una precisión del 83% aproximadamente, destacando en la neutralidad y en especial lo negativo. En general para todas las métricas esta bien balanceado sin mucha diferencia

SVM:				
	precision	recall	f1-score	support
Irrelevant	0.84	0.76	0.80	3900
Negative	0.83	0.88	0.85	6709
Neutral	0.86	0.78	0.82	5418
Positive	0.79	0.86	0.82	6172
accuracy			0.83	22199
macro avg	0.83	0.82	0.82	22199
weighted avg	0.83	0.83	0.83	22199
Accuracy: 0.8269291409522952				

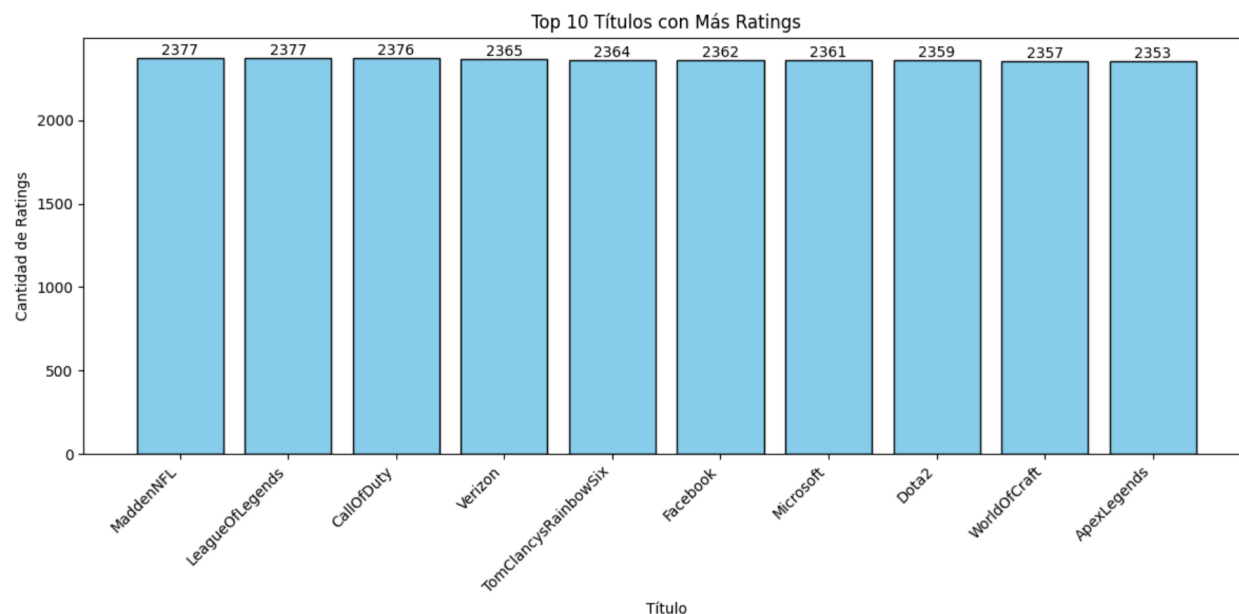
Por otro lado, Naive Bayes solo consiguió el 70% que no es un número muy alto a comparación del anterior. Se destaca que los comentarios negativos son verdaderamente negativos con su recall al 90%

Naive Bayes:				
	precision	recall	f1-score	support
Irrelevant	0.96	0.36	0.52	3900
Negative	0.63	0.90	0.75	6709
Neutral	0.84	0.60	0.70	5418
Positive	0.69	0.82	0.75	6172
accuracy			0.71	22199
macro avg	0.78	0.67	0.68	22199
weighted avg	0.76	0.71	0.70	22199
Accuracy: 0.7097166539033289				

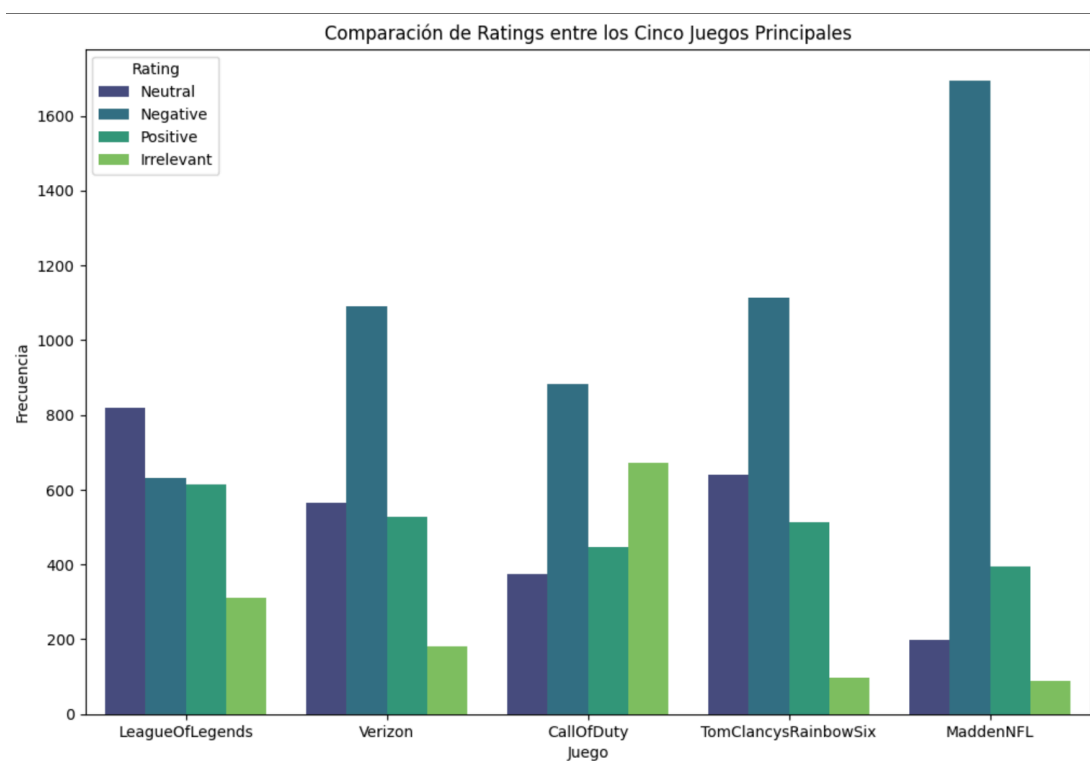
1.4 Visualización de Resultados

A continuación veremos algunos graficos de los resultados del Análisis del sentimiento:

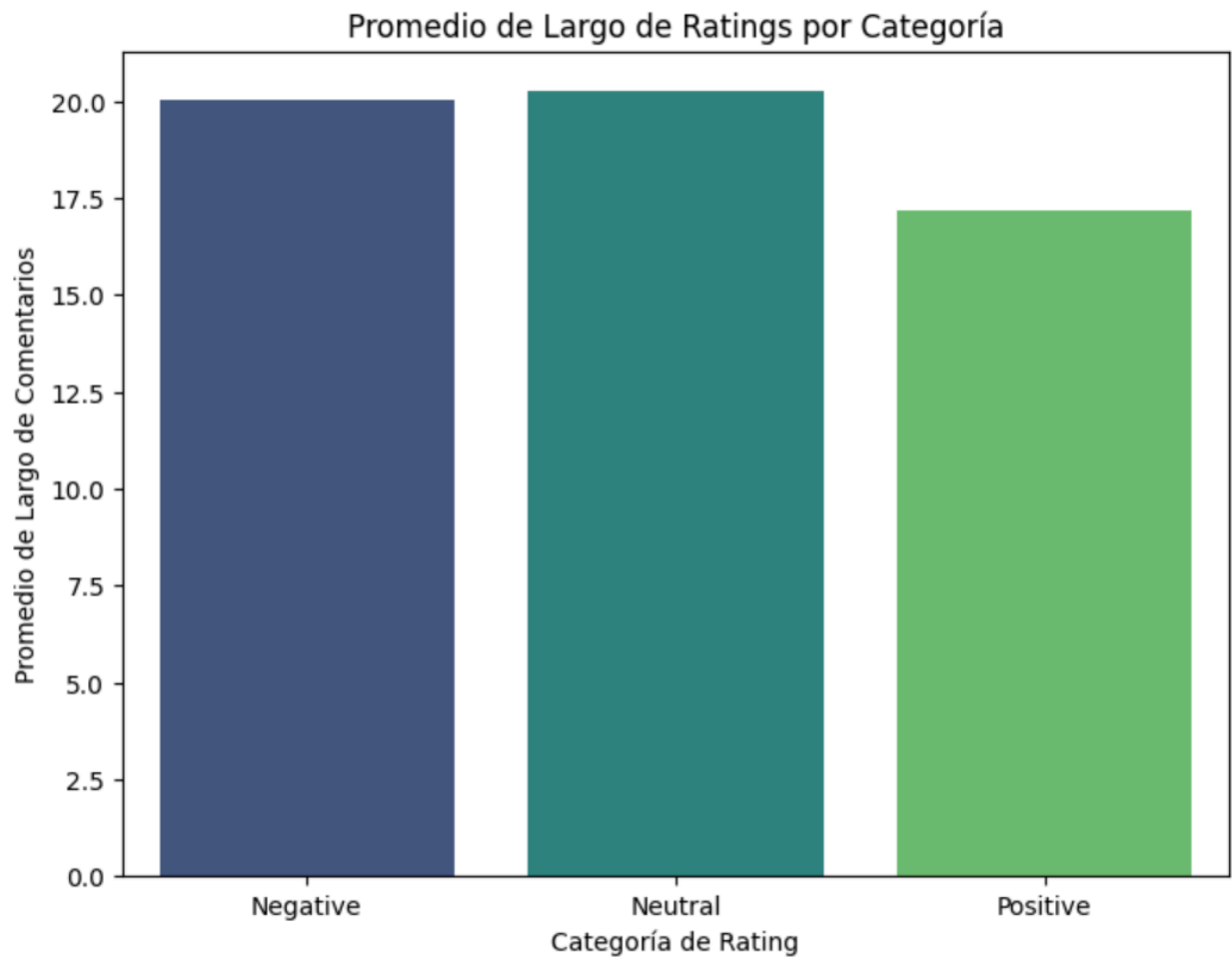
Aquí vemos que Madden tiene la mayor cantidad de ratings, independientemente si son positivas o no.



Podemos ver más a profundidad, que los 5 mayores juegos/títulos con más ratings son en lo general negativos, esto da a entender que mientras más reviews tengas , puede ser más por lo negativo que lo positivo



Por último vemos el largor de los reviews por categoría, destacando que mientras menos positivo es, más largo será el tweet.



Ahora que mostré el que salio un exito, quiero mostrar el que salio un fallo

2.Recomendación de Películas:

2.1 Exploración de Datos (EDA) y Procesamiento de Datos

Haciendo exactamente lo mismo que con el dataset anterior pero con este de

[Movie Reviews Dataset: 10k+ Scraped Data \(kaggle.com\)](#). Es un dataset con mas de 10 mil entradas de reseñas de películas

2.2 Extracción de Características

Primero limpio el texto de mayúsculas y minúsculas como el anterior. Luego hago el pos tagging similar al anterior para que me de las palabras categorizadas

```
POS tagged
[(review, n), (may, None), (contain,
v), (spoi...
[(youâ, v), (?, None), (?, None),
(never, r), ...
[(Puss, n), (Boots, n), (:, None),
(Pussy-Vers...
[(leave, v), (donkey, n), (outside,
None), (lâ...
[(Watch, v), (fun, v), (film, n),
(Twitter, n)...
```

Luego para la lematización de la misma, guardando las palabras en su modo mas sencillo

Lemma
review may contain spoiler .
youâ ? ? never swim ocean course pool seem d...
Puss Boots : Pussy-Verse
leave donkey outside lâ ? ? sad
Watch fun film Twitter tell itâ ? ? overrate

Como se podrá ver, el error del dataset que contiene mucho errores hizo que el resto del análisis no salga tan preciso

De todos modos intento hacer la polaridad, subjetividad y el análisis los cuales me dan resultados

Polarity	Subjectivity	Analysis
0.00	0.000	Neutral
0.00	0.400	Neutral
0.00	0.000	Neutral
-0.25	0.525	Negative
0.30	0.200	Positive

Entonces podemos ver que hay muchas reseñas positivas pero muchas de ellas neutras

```
Analysis
Positive      1512
Neutral       1281
Negative        767
Name: count, dtype: int64
```

2.3 Cálculo de Similitud:

Utilizando la librería scikit-learn se calcula la similitud de coseno entre usuarios en base a sus calificaciones de películas

```
Matriz de Similitud entre Usuarios:
User name          #1 gizmo fan  24framesofnick  AJ Ford  \
User name
#1 gizmo fan          1.000000          0.015448          0.0
24framesofnick        0.015448          1.000000          0.0
AJ Ford               0.000000          0.000000          1.0
ASYA                  0.000000          0.000000          0.0
Aaron Michael         0.000000          0.000000          0.0
```

Con esta información intentamos hacer el sistema de recomendaciones, y hacemos un testeo para ver qué sale

```
Película: Eraserhead  
Año de Lanzamiento: 19 6 6  
Calificación del Usuario Similar: 2.0  
Similitud con el Usuario Similar: 0.5714
```

```
Película: Eternals  
Año de Lanzamiento: 2021  
Calificación del Usuario Similar: 4.0  
Similitud con el Usuario Similar: 0.5714
```

```
Película: Sonic the Hedgehog  
Año de Lanzamiento: 2020  
Calificación del Usuario Similar: 2.0  
Similitud con el Usuario Similar: 0.5714
```

```
Película: Venom: Let There Be Carnage  
Año de Lanzamiento: 2021  
Calificación del Usuario Similar: 3.0  
Similitud con el Usuario Similar: 0.5714
```

```
Película: After Yang  
Año de Lanzamiento: 2021  
Calificación del Usuario Similar: 5.0  
Similitud con el Usuario Similar: 0.1839
```

2.4 Evaluación del Sistema

Haciendo el mismo proceso que con el conjunto anterior, intentamos ver si el modelado es verdaderamente bueno o no. Esta vez haciendo un promedio de ambos

```
Precisión promedio: 0  
Recall promedio: 0
```

Como pueden ver, el sistema es un total fracaso, significa que no funciona y que desde la limpieza hasta el modelado están totalmente erróneos

Conclusiones

A pesar de fallar con un sistema para la recomendación de películas, se podría decir que se tuvo un éxito para el análisis de sentimiento twitter. Mediante ambos análisis se ha aprendido a cómo limpiar los datos, juntar las palabras importantes, ver la polaridad, subjetividad y el análisis. Todo con el fin del aprendizaje y para mejorar en futuros proyectos.