Universidad Peruana de Ciencias Aplicadas

INFORME DEL TRABAJO FINAL

CURSO APLICACIONES DE DATA SCIENCE

Carrera de Ciencias de la Computación

Sección: CC54

Alumno	Código
Francesca Nicole Bances Torres	u202214090
Loana Colleen Rodríguez Matos	u202115571
Cesar Rafael Sánchez Garay	u202116752

2024

Índice

Descripción del caso de uso.

El crecimiento exponencial de las redes sociales ha generado un vasto caudal de datos textuales que contienen valiosa información sobre las opiniones, emociones y percepciones de los usuarios. Según el Digital Report (2023, como se cita en Santiago, 2023, p.11), se estima que existen más de 4.760 millones de usuarios de redes sociales en todo el mundo, lo que representa aproximadamente el 59,4% de la población global. Estas plataformas se han convertido en espacios donde millones de personas interactúan diariamente para compartir sus pensamientos sobre eventos sociales, productos y marcas, generando una cantidad enorme de datos. Por ejemplo, en Twitter se publican alrededor de 500 millones de tuits al día, mientras que en Facebook se registran más de 100 mil millones de mensajes diarios. Esta abundancia de información convierte estas interacciones en una fuente clave para el análisis de sentimientos.

Comprender estos sentimientos, ya sean positivos, negativos o neutrales, puede ofrecer una visión clara de la percepción pública y ayudar a las empresas o instituciones a adaptar sus estrategias de comunicación y marketing. De hecho, se estima que el 80% de los datos generados en redes sociales son no estructurados y contienen información potencialmente útil para las organizaciones. El mercado global de análisis de sentimientos está valorado en más de 6 mil millones de dólares y se espera que alcance los 10 mil millones para 2025, lo que demuestra la creciente importancia de esta herramienta en el mundo empresarial.

El proyecto del presente caso de uso es realizar un análisis de los sentimientos expresados en publicaciones de redes sociales como Twitter, Facebook e Instagram, con el fin de clasificar las emociones vinculadas a temas específicos o marcas. Este tipo de análisis, también conocido como minería de opiniones o sentiment analysis, busca no solo identificar qué emociones predominan, sino también cómo varían a lo largo del tiempo y cómo las diferentes plataformas pueden influir en la expresión emocional de los usuarios. Por ejemplo, estudios han demostrado que las opiniones negativas pueden difundirse un 70% más rápido que las positivas en las redes sociales, lo que resalta la necesidad de monitorear y responder rápidamente a los comentarios de los usuarios.

La fundamentación de este proyecto se basa en el hecho de que las redes sociales han transformado la manera en que las personas y las organizaciones interactúan, lo que ha incrementado la necesidad de monitorear en tiempo real la reputación de una marca o la respuesta emocional frente a determinados eventos. Se estima que el 90% de los consumidores leen comentarios en línea antes de realizar una compra, y que el 86% de ellos evitan comprar productos o servicios de empresas con reseñas negativas. En este contexto, se plantea utilizar algoritmos de clasificación, tales como Naive Bayes, Support Vector Machines (SVM) y redes neuronales, para clasificar los sentimientos asociados a las publicaciones. Además, se buscará comparar el rendimiento de estos modelos para identificar cuál es más adecuado en función de las características del dataset utilizado. Estudios previos han demostrado que los modelos basados en redes neuronales pueden alcanzar precisiones superiores al 90% en tareas de clasificación de sentimientos, lo que indica su potencial eficacia para este tipo de análisis.

Objetivos del Proyecto

- **Análisis de Sentimientos en Redes Sociales:** Realizar un análisis exhaustivo de los sentimientos expresados en publicaciones de plataformas como Twitter, Facebook e Instagram, enfocándose en temas específicos o marcas.
- **Clasificación de Emociones:** Utilizar algoritmos de clasificación como **Naive Bayes**, **Support Vector Machines (SVM)** y **redes neuronales** para categorizar las emociones en positivas, negativas o neutrales.
- **Comparación de Modelos:** Evaluar y comparar el rendimiento de los diferentes modelos de clasificación para determinar cuál es más eficaz según las características del conjunto de datos utilizado.
- **Monitoreo de Tendencias Temporales:** Identificar cómo varían las emociones a lo largo del tiempo y cómo las distintas plataformas influyen en la expresión emocional de los usuarios.
- **Apoyo a Estrategias Empresariales:** Proporcionar insights que permitan a empresas o instituciones adaptar sus estrategias de comunicación y marketing en función de la percepción pública obtenida del análisis.

Relevancia del Proyecto

Este proyecto es relevante porque entender y gestionar el sentimiento público en redes sociales es crucial para las organizaciones debido a la rápida difusión de opiniones, su influencia en las decisiones de los consumidores y el creciente valor del análisis de sentimientos en el mercado. La capacidad de analizar y responder a las emociones expresadas en línea permite a las empresas mejorar su reputación, adaptar sus estrategias y mantenerse competitivas en un entorno digital en constante cambio.

Preguntas a Responder

1. **¿Qué impacto tienen las variables contextuales (plataforma, número de 'likes' y 'retweets', país, y momento de publicación) en la predicción de los sentimientos expresados en las publicaciones de redes sociales?**

   Además de clasificar los sentimientos, se propone analizar cómo ciertas variables adicionales afectan el rendimiento predictivo del modelo. Este enfoque permitirá entender si el contexto de las publicaciones influye en la percepción emocional de los usuarios.

2. **¿Cuál es la relación entre el tipo de plataforma (Facebook, Instagram, Twitter) y el sentimiento predominante en las publicaciones?**

   Esta pregunta busca investigar si existen diferencias significativas en los sentimientos expresados por los usuarios en distintas plataformas y cómo estas diferencias pueden influir en la predicción.

Descripción del conjunto de datos (dataset).

Título: Social Media Sentiments Analysis Dataset

Origen: El dataset Social Media Sentiments Analysis Dataset ha sido extraído de Kaggle.

Características del Dataset:

- Número de Variables: 15
- Número de Observaciones: 732

Descripción de las variables:

Variable	Descripción	Tipo de variable
Text	Texto creado por los usuarios que expresa su opinión y sentimientos	Texto
Sentiment	Clasificación de las emociones en categorías.	Categórica
Timestamp	Información de fecha y hora	Fecha
User	Identificadores únicos de los usuarios que contribuyen	Categórica
Platform	Red social de donde proviene el contenido	Categórica
Hashtags	Identificador de temas y tendencias	Texto
Likes	Número de veces que la publicación fue marcada con "me gusta"	Numérica
Retweets	Número de veces que la publicación fue compartida o retuiteada	Numérica
Country	Origen geográfico de la publicación	Categórica
Year	Año de publicación	Numérica
Month	Mes de publicación	Numérica
Day	Día de publicación	Numérica
Hour	Hora de publicación	Numérica

Fuente: Elaboración propia

Análisis exploratorio de los datos (EDA).

Los datos recolectados deberán ser semiestructurados o no estructurados. Se debe incluir la descripción de las tareas de carga, inspección, preprocesamiento (o normalización en el caso de textos) y visualización de los datos.

- Carga de datos:

```
[15] data = pd.read_csv('sentimentdataset.csv')
     data.head()
```

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Likes	Country	Year	Month	Day	Hour
0	0	0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Nature #Park	15.0	30.0	USA	2023	1	15	12
1	1	1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #Morning	5.0	10.0	Canada	2023	1	15	8
2	2	2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #Workout	20.0	40.0	USA	2023	1	15	15
3	3	3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #Adventure	8.0	15.0	UK	2023	1	15	18
4	4	4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	ChefCook	Instagram	#Cooking #Food	12.0	25.0	Australia	2023	1	15	19

Fuente: Elaboración propia

- Inspección de datos

```
[11] data.shape
```

```
(732, 15)
```

Fuente: Elaboración propia

El dataset cuenta con 15 variables y 732 observaciones.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 732 entries, 0 to 731
Data columns (total 15 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Unnamed: 0.1  732 non-null    int64  
 1   Unnamed: 0    732 non-null    int64  
 2   Text          732 non-null    object 
 3   Sentiment     732 non-null    object 
 4   Timestamp     732 non-null    object 
 5   User          732 non-null    object 
 6   Platform      732 non-null    object 
 7   Hashtags      732 non-null    object 
 8   Retweets      732 non-null    float64
 9   Likes         732 non-null    float64
 10  Country       732 non-null    object 
 11  Year          732 non-null    int64  
 12  Month         732 non-null    int64  
 13  Day           732 non-null    int64  
 14  Hour          732 non-null    int64  
dtypes: float64(2), int64(6), object(7)
memory usage: 85.9+ KB
```

Fuente: Elaboración propia

```
data.describe()
```

	Unnamed: 0.1	Unnamed: 0	Retweets	Likes	Year	Month	Day	Hour
count	732.000000	732.000000	732.000000	732.000000	732.000000	732.000000	732.000000	732.000000
mean	366.464481	369.740437	21.508197	42.901639	2020.471311	6.122951	15.497268	15.521858
std	211.513936	212.428936	7.061286	14.089848	2.802285	3.411763	8.474553	4.113414
min	0.000000	0.000000	5.000000	10.000000	2010.000000	1.000000	1.000000	0.000000
25%	183.750000	185.750000	17.750000	34.750000	2019.000000	3.000000	9.000000	13.000000
50%	366.500000	370.500000	22.000000	43.000000	2021.000000	6.000000	15.000000	16.000000
75%	549.250000	553.250000	25.000000	50.000000	2023.000000	9.000000	22.000000	19.000000
max	732.000000	736.000000	40.000000	80.000000	2023.000000	12.000000	31.000000	23.000000

Fuente: Elaboración propia

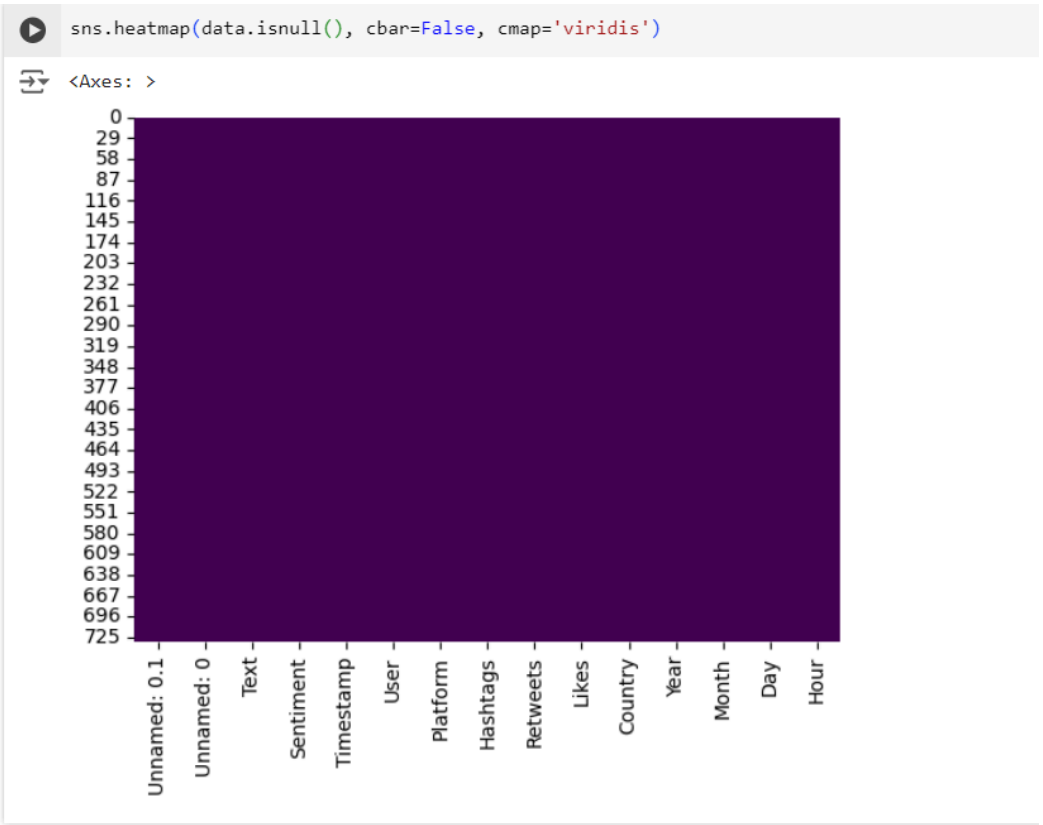
```
data.columns.values
```

```
array(['Unnamed: 0.1', 'Unnamed: 0', 'Text', 'Sentiment', 'Timestamp',
       'User', 'Platform', 'Hashtags', 'Retweets', 'Likes', 'Country',
       'Year', 'Month', 'Day', 'Hour'], dtype=object)
```

Fuente: Elaboración propia

Verificar valores nulos:

```
sns.heatmap(data.isnull(), cbar=False, cmap='viridis')
```



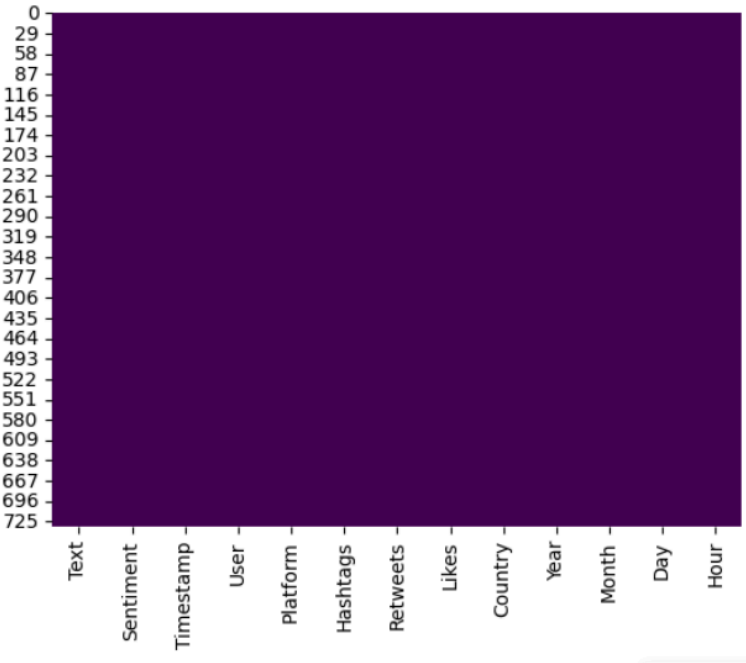
Fuente: Elaboración propia

No hay valores nulos en el dataset

Verificar valores Na:

```
sns.heatmap(data.isna(), cbar=False, cmap='viridis')
```

```
<Axes: >
```



Fuente: Elaboración propia

No hay valores vacíos

Verificación de datos duplicados:

```
[14] text_duplicate = data[data.duplicated(subset='Text')]
     text_duplicate
```

	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Likes	Country	Year	Month	Day	Hour
278	A compassionate rain, tears of empathy fallin...	Compassionate	2021-07-01 12:10:00	RainNurturer	Instagram	#Compassionate #TearsOfEmpathy	21.0	42.0	Canada	2021	7	1	12
279	Proudly scaling the peaks of achievement, a m...	Proud	2020-01-05 08:45:00	PeakConqueror	Twitter	#Proud #ScalingPeaks	23.0	46.0	USA	2020	1	5	8
280	Embraced by the hopeful dawn, a gardener sowi...	Hopeful	2022-07-17 06:15:00	DawnGardener	Instagram	#Hopeful #SeedsOfOptimism	14.0	28.0	UK	2022	7	17	6
281	A playful escapade in the carnival of life, c...	Playful	2018-08-22 17:20:00	CarnivalDreamer	Facebook	#Playful #CarnivalEscapade	24.0	48.0	Australia	2018	8	22	17
282	Floating on clouds of inspiration, an artist ...	Inspired	2021-12-08 14:30:00	SkyArtist	Twitter	#Inspired #CloudsOfCreativity	18.0	36.0	India	2021	12	8	14
283	Navigating the river of contentment, a serene...	Contentment	2019-04-27 09:50:00	RiverNavigator	Instagram	#Contentment #TranquilWaters	20.0	40.0	Canada	2019	4	27	9
284	With empathy as a lantern, wandering through ...	Empathetic	2023-09-03 21:40:00	LanternWanderer	Facebook	#Empathetic #LanternOfCompassion	16.0	32.0	USA	2023	9	3	21
285	A free spirit soaring on the wings of dreams,...	Free-spirited	2020-06-10 10:05:00	DreamSoarer	Twitter	#FreeSpirit #WingsOfDreams	22.0	44.0	UK	2020	6	10	10
286	Bathed in the golden hues of gratefulness, a ...	Grateful	2022-04-01 18:30:00	SunsetAdmirer	Instagram	#Grateful #GoldenHues	19.0	38.0	Australia	2022	4	1	18
287	Confident strides in the dance of life, a bal...	Confident	2021-01-15 13:00:00	DanceStrider	Facebook	#Confident #DanceOfLife	23.0	46.0	Canada	2021	1	15	13
288	Hopeful whispers of wind, carrying the promis...	Hopeful	2023-05-06 07:20:00	WindWhisperer	Twitter	#Hopeful #BrighterTomorrows	15.0	30.0	India	2023	5	6	7
289	Playfully juggling responsibilities, a circus...	Playful	2019-11-18 15:15:00	JugglingArtist	Instagram	#Playful #JugglingResponsibilities	25.0	50.0	USA	2019	11	18	15
290	Whispering tales of inspiration to the stars,...	Inspired	2020-08-29 20:45:00	StarStoryteller	Facebook	#Inspired #TalesToTheStars	14.0	28.0	UK	2020	8	29	20

Fuente: Elaboración propia

```
len(text_duplicate)
```

```
25
```

Fuente: Elaboración propia

En este caso tenemos 25 filas duplicadas, sus valores son los mismos en todas las columnas por lo que no aportan ningún valor adicional por ende las eliminaremos del dataset.

Limpieza de datos duplicados

```
[15] data = data.drop_duplicates(subset=['Text'], keep='first')

[16] len(data)

707
```

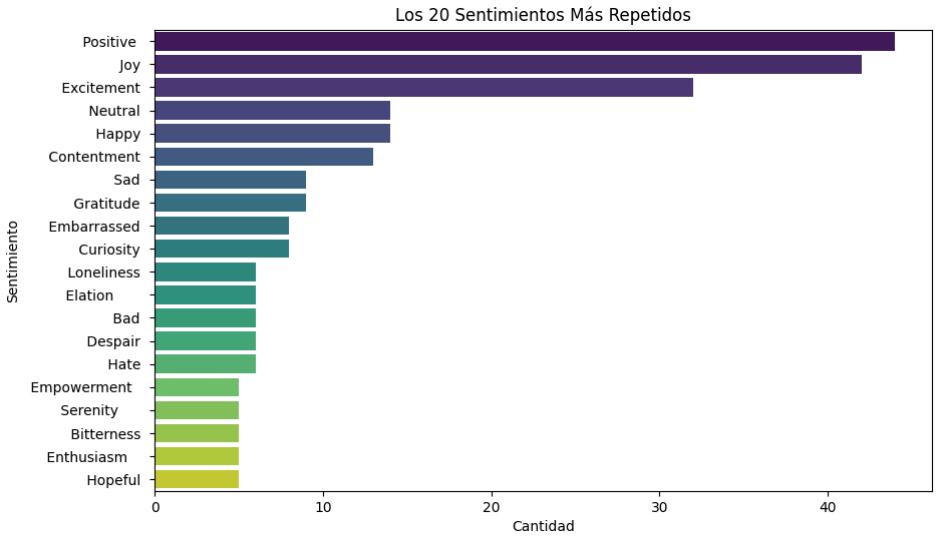
Fuente: Elaboración propia

Eliminamos los valores repetidos, quedándonos con 707 registros.

Visualización:

- Sentiment

```
[18] len(data['Sentiment'].unique())

278
```



Fuente: Elaboración propia

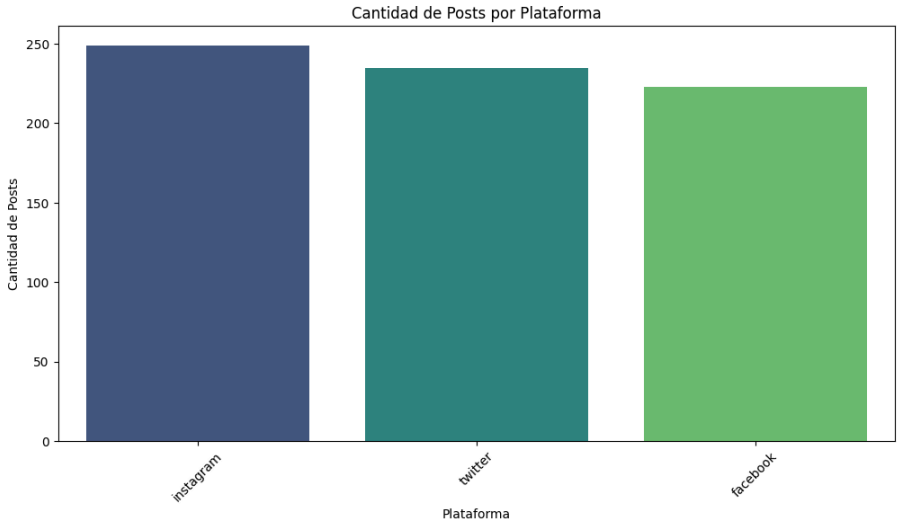
Al analizar el gráfico podemos ver que los sentimientos más repetidos son Positive, Joy, Excitement, Neutral y Happy.

- Plataformas más utilizadas

```
[20] # Limpiar los nombres de las plataformas
     data['Platform'] = data['Platform'].str.strip().str.lower()  # Eliminar espacios y convertir a minúsculas

     # Contar nuevamente la cantidad de posts por plataforma
     conteo_platforms = data['Platform'].value_counts()

     # Mostrar el conteo
     print(conteo_platforms)
```

```
Platform
instagram    249
twitter      235
facebook     223
Name: count, dtype: int64
```

Fuente: Elaboración propia



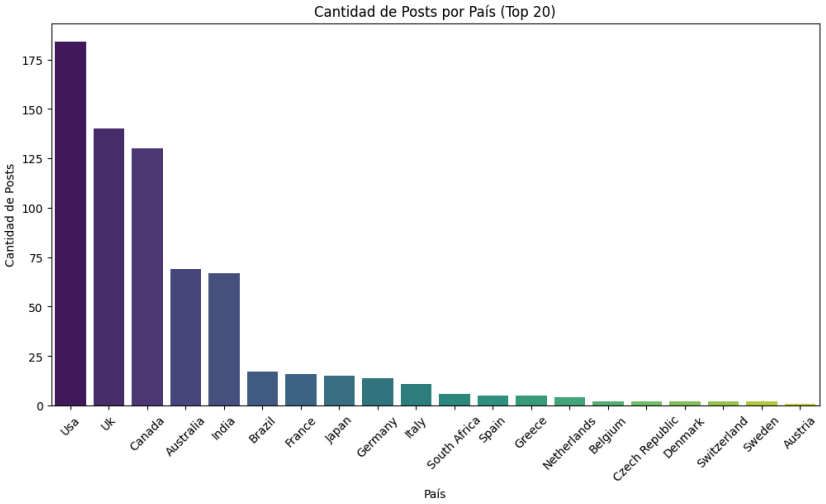
Fuente: Elaboración propia

En el gráfico podemos ver la cantidad de posts por país. Estados Unidos lidera con la mayor cantidad de posts, con más de 175, seguido por el Reino Unido con aproximadamente 150, y Canadá y Australia, ambos con alrededor de 125 posts.

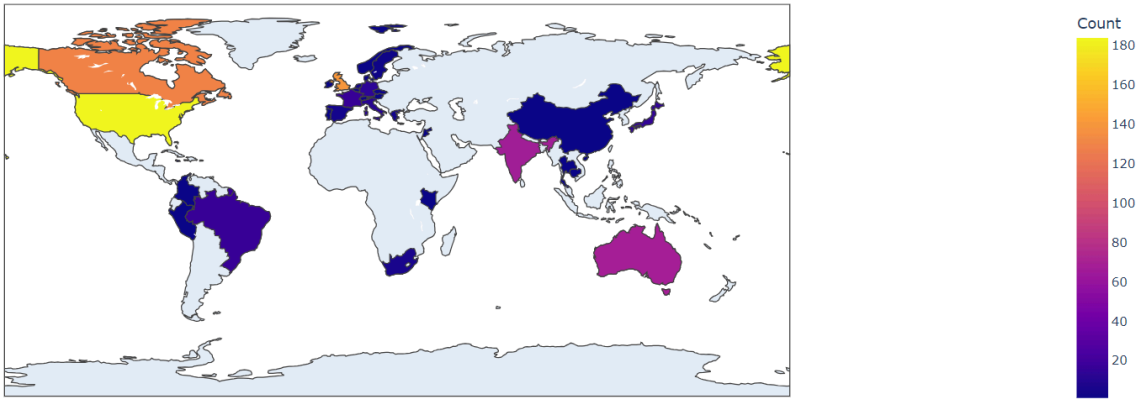
- Países más comunes

```
[22] # Limpiar los nombres de los países
     data['Country'] = data['Country'].str.strip().str.title()  # Eliminar espacios y poner en formato de título

     # Contar la cantidad de posts por país
     conteo_paises = data['Country'].value_counts()
```

Fuente: Elaboración propia



Fuente: Elaboración propia



Fuente: Elaboración propia

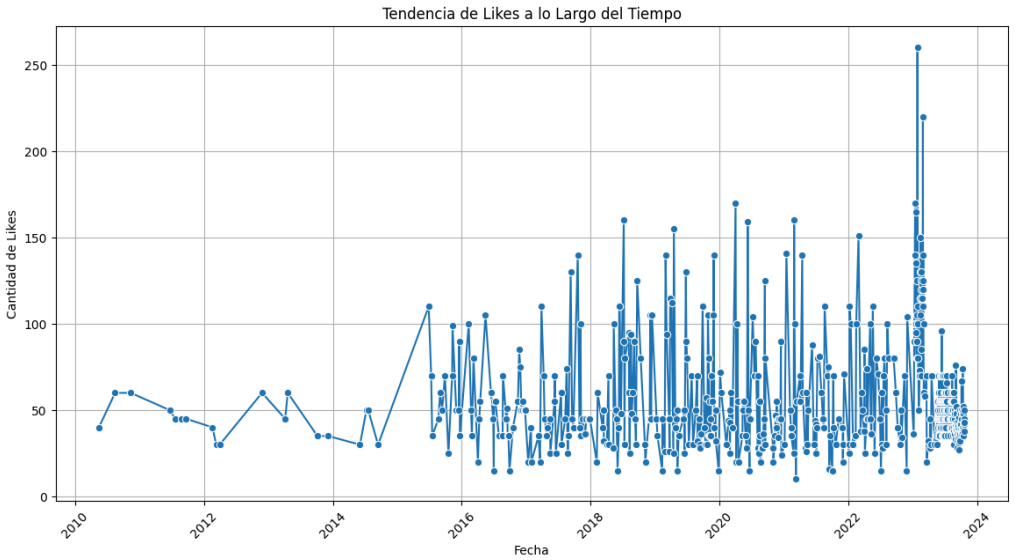
Los países con más posts son:

- USA: Con más de 175 posts, es el país más activo.
- UK: Con cantidades cercanas a 130 posts.
- Canadá: Con aproximadamente 125 posts.
- Australia e India: Presentan una disminución significativa con menos de 75 posts.

Después de los 5 primeros países, hay una caída significativa en la cantidad de posts por país. Los siguientes, como Brasil, Francia y Japón, tienen menos de 25 posts.

Países menos representados: Los últimos países en la lista, como Suecia, Suiza y Austria, tienen menos de 10 posts cada uno.

Tendencias de likes a lo largo del tiempo:



Fuente: Elaboración propia

Estabilidad inicial: Entre 2010 y 2015, los likes se mantienen relativamente constantes, con fluctuaciones entre 20 y 60, pero sin picos notables.

A partir de 2016, comienza a haber un aumento en la variabilidad de los "likes", con algunos picos que superan los 100.

A partir de 2018, el gráfico muestra una mayor cantidad de puntos de "likes", lo que sugiere una mayor frecuencia de interacciones. Los valores oscilan de manera más pronunciada, con picos más altos y más frecuentes.

A partir de 2020, hay un incremento en la cantidad máxima de "likes", con algunos picos que llegan a más de 200, especialmente hacia el final del gráfico (2024).

Normalización

- Tokenización y limpieza de palabras

```python
# Extraer la columna de texto
texts = data['Text'].fillna('')  # Asegúrate de usar la columna 'Text'

# Inicializar stop words en inglés
stop_words = set(stopwords.words('english'))

# Función de tokenización, eliminación de stop words y filtrado de palabras con menos de 2 letras
def tokenize_and_filter(text):
    if not isinstance(text, str):  # Asegurarse de que sea una cadena
        return []
    tokens = word_tokenize(text.lower())  # Tokenizar y convertir a minúsculas
    tokens_filtrados = [word for word in tokens if word.isalpha() and word not in stop_words and len(word) > 2]  # Eliminar stop words, palabras cortas y símbolos
    return tokens_filtrados

# Aplicar tokenización a la columna de texto
data['tokens'] = data['Text'].apply(tokenize_and_filter)

# Mostrar algunos ejemplos de los tokens
print(data[['Text', 'tokens']].head())
```

Fuente: Elaboración propia

```
                                                Text  \
0   Enjoying a beautiful day at the park!        ...
1   Traffic was terrible this morning.           ...
2   Just finished an amazing workout! 💪          ...
3   Excited about the upcoming weekend getaway!  ...
4   Trying out a new recipe for dinner tonight.  ...

                                  tokens
0       [enjoying, beautiful, day, park]
1           [traffic, terrible, morning]
2           [finished, amazing, workout]
3  [excited, upcoming, weekend, getaway]
4  [trying, new, recipe, dinner, tonight]
```

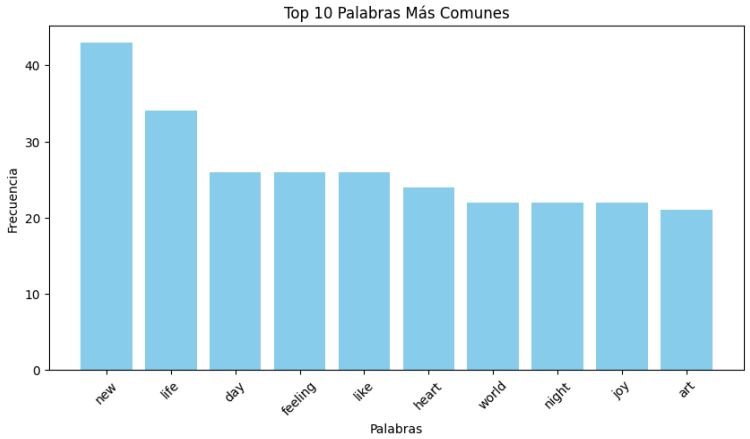
Fuente: Elaboración propia

Con esta función nos encargamos de tokenizar la columna Text, eliminamos las stop words y eliminamos las palabras con menos de 2 letras.

Frecuencia de palabras:

```python
# Combinar todos los tokens en una lista
all_tokens = [token for sublist in data['tokens'] for token in sublist]

# Calcular la frecuencia de cada palabra
freq_distribution = Counter(all_tokens)

# Mostrar las 10 palabras más comunes
print(freq_distribution.most_common(10))


# Gráfica de las 10 palabras más comunes
common_words = freq_distribution.most_common(10)
words, counts = zip(*common_words)  # Desempaquetar en dos listas

# Crear el gráfico de barras
plt.figure(figsize=(10, 5))
plt.bar(words, counts, color='skyblue')
plt.title('Top 10 Palabras Más Comunes')
plt.xlabel('Palabras')
plt.ylabel('Frecuencia')
plt.xticks(rotation=45)
plt.show()
```

Fuente: Elaboración propia



Fuente: Elaboración propia

Palabras más comunes: La palabra más frecuente es "new", con una frecuencia superior a 40, seguida de "life" con alrededor de 35. Esto sugiere que estos términos son los más recurrentes en el texto analizado.
Rango de frecuencias: El resto de las palabras, como "day", "feeling", "like", "heart", "world", "night", "joy" y "art", tienen frecuencias entre 20 y 30.
Las palabras más comunes parecen estar relacionadas con conceptos emocionales y existenciales como "vida", "sentimientos", "mundo", "corazón", lo que sugiere que el texto puede estar vinculado a sentimientos positivos.

Propuestas de Modelización.

1. **Naive Bayes:**

   El modelo **Naive Bayes** es uno de los métodos más utilizados en tareas de procesamiento de lenguaje natural (NLP), especialmente en la clasificación de texto, debido a su simplicidad, rapidez y efectividad en muchas situaciones. Este algoritmo pertenece a la familia de modelos probabilísticos basados en la **teoría de Bayes**, la cual utiliza la probabilidad condicional para realizar inferencias sobre la categoría a la que pertenece una instancia de datos, en este caso, un texto o una publicación en redes sociales.

   El principio fundamental del **Naive Bayes** es que las características del conjunto de datos (en este caso, las palabras o tokens) son independientes entre sí, lo que significa que la ocurrencia de una palabra no influye en la ocurrencia de otra dentro del mismo documento. Aunque esta suposición de independencia puede parecer irrealista en el procesamiento de lenguaje natural, en la práctica, Naive Bayes ha demostrado ser muy eficaz para problemas como la clasificación de sentimientos, donde las palabras relevantes suelen ser buenas indicadoras del sentimiento subyacente.

2. **Complement Naive Bayes:**

El Complement Naive Bayes es una variante del algoritmo Naive Bayes, diseñada para mejorar la clasificación en problemas de texto desbalanceados, donde una o más clases tienen una cantidad significativamente mayor de ejemplos que otras. El complemento de Naive Bayes ajusta la forma en que se calculan las probabilidades de las clases, haciendo que el modelo sea más robusto frente a datos desbalanceados, lo que lo convierte en una opción ideal para tareas de clasificación de texto en las que las clases no están distribuidas de manera uniforme.

3. **Modelo de Regresión Logística**

La **regresión logística** es un modelo estadístico utilizado comúnmente para problemas de clasificación binaria, aunque también puede extenderse a clasificación multiclase. Se basa en la idea de modelar la probabilidad de una clase en función de una combinación lineal de las características de entrada. A pesar de su nombre, la regresión logística no se utiliza para problemas de regresión, sino para clasificación, ya que su salida es la probabilidad de que una observación pertenezca a una clase específica, que luego se convierte en una clasificación utilizando un umbral (típicamente 0.5).

4. **Modelo Random Forest**

Random Forest es un algoritmo de aprendizaje supervisado basado en el ensamblaje de múltiples árboles de decisión. Cada árbol en el bosque se entrena utilizando un subconjunto aleatorio de los datos y las características, lo que ayuda a reducir el sobreajuste (overfitting) y mejora la capacidad de generalización del modelo. El modelo realiza predicciones promediando los resultados de todos los árboles, lo que suele dar lugar a una mayor precisión y estabilidad en comparación con un único árbol de decisión.

5. **Modelo Random Forest con SMOTE**

El **Random Forest** es un algoritmo de aprendizaje supervisado que crea múltiples árboles de decisión en un proceso de "ensamblaje" para mejorar la precisión y evitar el sobreajuste (overfitting). Sin embargo, en problemas de clasificación con clases desbalanceadas, este modelo puede inclinarse hacia la clase mayoritaria. Para resolver este problema, se utilizó **SMOTE** (Synthetic Minority Over-sampling Technique) antes de entrenar el modelo.

SMOTE es una técnica de sobremuestreo que genera nuevas muestras sintéticas para la clase minoritaria, en lugar de simplemente replicar las instancias existentes. Esto ayuda a equilibrar el conjunto de datos al crear ejemplos adicionales de la clase minoritaria, lo que permite que el modelo aprenda de una manera más equitativa entre todas las clases. En este caso, SMOTE se aplicó a los datos de entrenamiento representados por los vectores de características generados por el **TfidfVectorizer**.

6. **Modelo SVM con SMOTE**

El modelo **Support Vector Machine (SVM)** es uno de los enfoques más potentes para la clasificación, especialmente cuando se trata de problemas con márgenes de decisión claros entre clases. SVM busca encontrar el hiperplano óptimo que separa las diferentes clases en el

espacio de características. En este caso, se utilizó el **SVC** (Support Vector Classifier) con un **kernel lineal**, que es adecuado para datos que pueden ser linealmente separables. También es posible utilizar otros tipos de kernels (como 'poly' o 'rbf') dependiendo de la naturaleza de los datos y la complejidad de la frontera entre clases.

Al igual que con el modelo Random Forest, se aplicó la técnica **SMOTE** (Synthetic Minority Over-sampling Technique) para balancear el conjunto de entrenamiento. SMOTE genera nuevas instancias sintéticas de la clase minoritaria para que ambas clases (mayoritaria y minoritaria) estén equilibradas, lo que ayuda a evitar que el modelo se incline hacia la clase mayoritaria, mejorando así la capacidad de generalización.

Modelo	Descripción	Ventajas	Desventajas	Aplicación Típica
Naive Bayes	Modelo probabilístico basado en la teoría de Bayes, que asume que las características son independientes.	Simple, rápido, efectivo en clasificación de texto y análisis de sentimientos.	Supone independencia entre las características, lo que puede no ser realista.	Clasificación de texto, análisis de sentimientos.
Complement Naive Bayes	Variante de Naive Bayes, ajusta la probabilidad para mejorar la clasificación en datasets desbalanceados.	Mejora la precisión en clases desbalanceadas.	Puede no ser tan eficaz en datasets balanceados.	Clasificación de texto en problemas desbalanceados.
Regresión Logística	Modelo estadístico para clasificación binaria, basado en una combinación lineal de las características.	Intuitivo, fácil de implementar y comprender, efectivo en clasificación binaria.	Limitado en problemas con relaciones no lineales o multiclase sin ajustes.	Clasificación binaria y multiclase.
Random Forest	Algoritmo de ensamblaje de múltiples árboles de decisión, reduce el sobreajuste y mejora la precisión.	Robustez, precisión, reduce el sobreajuste, funciona bien con datos complejos.	Computacionalmente más costoso, menos interpretable.	Clasificación general, detección de anomalías.

Random Forest con SMOTE	Random Forest mejorado con SMOTE para balancear clases desbalanceadas generando instancias sintéticas.	Soluciona problemas de clases desbalanceadas, mejora la generalización.	Incrementa la complejidad y el costo computacional.	Clasificación con clases desbalanceadas.
SVM con SMOTE	Support Vector Machine con SMOTE para balancear clases desbalanceadas, busca el hiperplano óptimo de separación.	Potente en problemas con márgenes de decisión claros, SMOTE mejora la precisión en clases desbalanceadas.	Puede ser sensible a la elección del kernel, costoso computacionalmente.	Clasificación en problemas con márgenes claros y clases desbalanceadas.

Fuente: Elaboración propia

Monetización y Estimación

Monetización

El análisis de sentimientos en redes sociales ofrece múltiples oportunidades de monetización para las empresas y organizaciones:

- **Mejora de Estrategias de Marketing**: Al comprender las emociones y percepciones de los usuarios, las empresas pueden adaptar sus campañas de marketing para aumentar la efectividad y el retorno de inversión (ROI). Esto puede traducirse en un aumento de ventas y participación de mercado.
- **Gestión de la Reputación en Línea**: Las herramientas de análisis permiten a las empresas monitorear en tiempo real las opiniones de los consumidores, permitiendo una respuesta rápida a comentarios negativos y la potenciación de comentarios positivos, lo que puede mejorar la reputación de la marca y la lealtad del cliente.
- **Desarrollo de Productos y Servicios**: Al identificar tendencias y necesidades del mercado a través del análisis de sentimientos, las empresas pueden innovar y desarrollar productos o servicios que satisfagan mejor las demandas de los consumidores, generando nuevas fuentes de ingresos.
- **Segmentación de Audiencias**: El análisis permite segmentar a los clientes según sus emociones y opiniones, facilitando campañas de marketing más dirigidas y efectivas, optimizando los recursos invertidos.
- **Consultoría y Venta de Servicios de Análisis**: Las empresas especializadas pueden ofrecer servicios de análisis de sentimientos a otras organizaciones, creando un modelo de negocio basado en la venta de insights y asesoramiento estratégico.

Estimación

La implementación de un proyecto de análisis de sentimientos en redes sociales implica considerar varios aspectos en términos de recursos y costos:

- **Recopilación y Almacenamiento de Datos**:
  - **Costo de Acceso a Datos**: Dependiendo de las políticas de cada plataforma, puede haber costos asociados al acceso a las API de Twitter, Facebook e Instagram para la recolección de datos. El costo promedio de acceso a una API de Twitter puede oscilar entre $99 y $2,899 mensuales, dependiendo del volumen de datos requerido.
  - **Infraestructura de Almacenamiento**: Se requiere infraestructura para almacenar grandes volúmenes de datos textuales. Utilizar servicios en la nube como Amazon S3 puede costar entre $23 y $25 por terabyte al mes, dependiendo del nivel de uso y la redundancia.
- **Desarrollo y Entrenamiento de Modelos**:
  - **Recursos Humanos**: Se necesitan especialistas en ciencia de datos y aprendizaje automático para desarrollar, entrenar y optimizar los modelos. El salario promedio de un científico de datos es de $100,000 a $150,000 al año, y podría ser necesario contratar entre 1 y 3 especialistas para un proyecto de esta magnitud.
  - **Potencia Computacional**: El entrenamiento de modelos, especialmente redes neuronales profundas, requiere recursos computacionales significativos. Alquilar instancias de GPU en servicios como AWS puede costar entre $0.90 y $3.06 por hora, dependiendo del tipo de instancia utilizada. El costo total para entrenar modelos complejos podría oscilar entre $5,000 y $20,000, dependiendo del tiempo de entrenamiento.
- **Herramientas y Licencias**:
  - **Software**: Aunque existen herramientas y librerías de código abierto, algunas soluciones empresariales pueden requerir licencias pagadas, lo cual podría añadir entre $5,000 y $15,000 al presupuesto total, dependiendo de las herramientas necesarias.
- **Tiempo de Implementación**:
  - **Plazo Estimado**: Un proyecto de esta magnitud puede tomar entre 6 y 12 meses desde la recopilación de datos hasta la implementación y pruebas finales, con un costo asociado de salarios para el equipo involucrado.
- **Costos Adicionales**:
  - **Mantenimiento y Actualización**: Los modelos requieren actualización constante para mantener su precisión frente a nuevos datos y tendencias. Se debe considerar un costo mensual de mantenimiento de entre $2,000 y $5,000.
  - **Seguridad y Cumplimiento Normativo**: Es necesario garantizar la protección de datos y cumplir con regulaciones como el GDPR, lo que puede implicar costos legales y de implementación de medidas de seguridad, estimados en alrededor de $10,000 anuales.
- **Retorno de Inversión (ROI)**:
  - **Beneficios Potenciales**: Aunque los costos iniciales pueden ser significativos, el potencial aumento en ingresos por ventas, mejora en la satisfacción del cliente y optimización de estrategias puede generar un ROI positivo en el mediano plazo. Las empresas que implementan análisis de sentimientos suelen reportar un incremento en

ventas del 10% al 15%, lo que podría traducirse en cientos de miles de dólares anuales, dependiendo del tamaño del mercado.

Modelización

En la etapa de modelización, se utilizaron varios algoritmos principales para clasificar los sentimientos expresados en redes sociales, con un enfoque especial en el modelo **Support Vector Machines (SVM)** combinado con **SMOTE** (Synthetic Minority Over-sampling Technique), debido a su excelente rendimiento, logrando un **accuracy de 0.78**. Se seleccionaron tres modelos en total: **Naive Bayes**, **Support Vector Machines (SVM)** y **Redes Neuronales**. Cada uno tiene características particulares que los hacen adecuados para el análisis comparativo de la clasificación de sentimientos, pero fue el modelo SVM con SMOTE el que obtuvo los mejores resultados.

Naive Bayes fue considerado debido a su rapidez y simplicidad, siendo un modelo eficaz en tareas de clasificación de texto donde las palabras son buenos indicadores de la clase subyacente. Aunque la suposición de independencia de las características es simplista, Naive Bayes se ha mostrado eficaz para textos dispersos como los de las redes sociales.

El modelo **Support Vector Machines (SVM)**, elegido por su capacidad para manejar datos no lineales mediante funciones kernel, fue utilizado con un kernel lineal en este caso. SVM es particularmente útil en problemas de clasificación con muchas variables que no son linealmente separables, lo que lo hace adecuado para el análisis de sentimientos en textos de redes sociales.

Lo que diferencia al modelo **SVM** en este análisis es su combinación con **SMOTE**. SMOTE es una técnica de sobremuestreo que genera muestras sintéticas para la clase minoritaria, equilibrando las clases en el conjunto de datos. Esta técnica ayudó a mejorar significativamente el rendimiento de SVM en el conjunto de datos desbalanceado, lo que contribuyó a la obtención de un **accuracy de 0.78**, el más alto entre los modelos evaluados.

Finalmente, se emplearon **Redes Neuronales**, que son muy eficaces para aprender patrones complejos y capturar relaciones implícitas entre las palabras, pero no lograron superar al modelo SVM con SMOTE en cuanto a precisión en este caso particular. Para entrenar los modelos, se dividió el conjunto de datos en un conjunto de entrenamiento (80%) y uno de prueba (20%). Además, se utilizó **validación cruzada** para minimizar el sobreajuste y asegurar que los resultados fueran representativos. El rendimiento de los modelos fue evaluado utilizando varias métricas, como la **precisión** (accuracy), el **F1-score** y la **matriz de confusión**, lo que permitió evaluar tanto la exactitud general de la clasificación como la capacidad de los modelos para diferenciar entre las clases de sentimientos positivos, negativos y neutros.

Análisis de Resultados

Tras el entrenamiento y validación de los modelos, se realizó una comparación detallada de su desempeño utilizando métricas clave como precisión (accuracy), recall (exhaustividad), F1-score y el

reporte de clasificación. A continuación se presentan los resultados obtenidos de los tres modelos con mejor desempeño: Logistic Regression, Random Forest y SVM con SMOTE.

Modelo	Precisión (%)	Exhaustividad (%)	F1-Score (%)	AUC
Logistic Regression	60.56	58.00	59.00	-
Random Forest	77.51	77.00	77.00	0.78
SVM con SMOTE	78.99	79.00	79.00	0.80

Fuente: Elaboración propia

Logistic Regression:

El modelo de Logistic Regression presentó un accuracy de 60.56%, lo que lo posiciona como el de menor rendimiento entre los tres modelos. Aunque el modelo tuvo un desempeño moderado en términos de precisión y recall, los valores de F1-score fueron relativamente bajos, especialmente para la clase negativa, lo que indica que la capacidad del modelo para clasificar correctamente esa clase fue limitada. El modelo mostró un buen desempeño en la clase neutra, con una exhaustividad del 69%. Sin embargo, en general, Logistic Regression no logró capturar las complejidades del conjunto de datos de forma tan efectiva como los otros dos modelos.

- **Resultados del modelo Logistic Regression:**

```
Accuracy: 0.6056338028169014

Classification Report:
              precision    recall  f1-score   support

    negativo       0.55      0.50      0.52        24
      neutro       0.59      0.69      0.64        65
    positivo       0.66      0.55      0.60        53

    accuracy                           0.61       142
   macro avg       0.60      0.58      0.59       142
weighted avg       0.61      0.61      0.60       142
```

Fuente: Elaboración propia

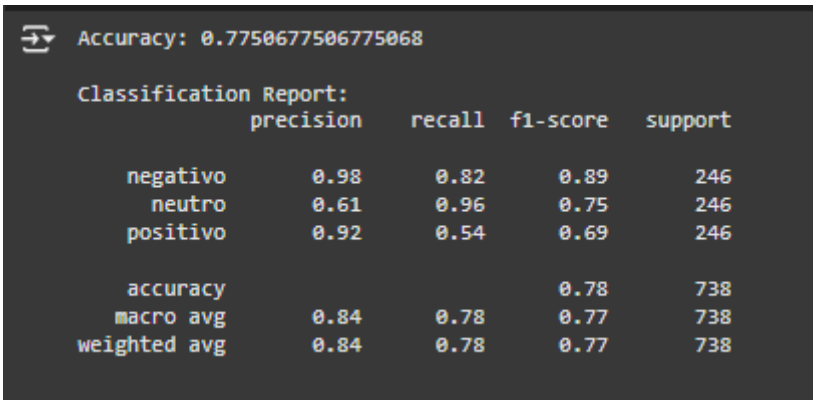
```
----------------------------------------
Texto: The app is convenient, but it has some bugs that need fixing. It crashes sometimes, and some of the features are difficult to use. I hope they release an update soon.
Predicción: Neutro
----------------------------------------
Texto: I had a great experience with this online store. The delivery was quick, and the product was exactly as described. I'll be shopping here again for sure.
Predicción: Positivo
----------------------------------------
Texto: I'm really not impressed with this service. It's slow, and the quality is not what I expected for the price. I would rather go elsewhere next time.
Predicción: Neutro
----------------------------------------
Texto: This book was such a pleasure to read! The characters were relatable, and the storyline was full of unexpected twists. I couldn't put it down and finished it in one sitting.
Predicción: Positivo
----------------------------------------
```

Fuente: Elaboración propia

Random Forest:

El modelo de Random Forest, por otro lado, obtuvo un accuracy de 77.51% en su evaluación con validación cruzada. Random Forest es conocido por su robustez al manejar conjuntos de datos con muchas características, y en este caso, pudo manejar el desbalance de clases gracias al muestreo balanceado con SMOTE. Random Forest mostró una excelente exhaustividad en la clase neutra (96%), lo que significa que fue muy efectivo para clasificar esta clase, pero las clases positiva y negativa tuvieron resultados más moderados. En general, el modelo alcanzó un buen balance entre precisión y recall, destacándose por su rendimiento en la clase neutra.

- **Resultados del modelo Random Forest:**

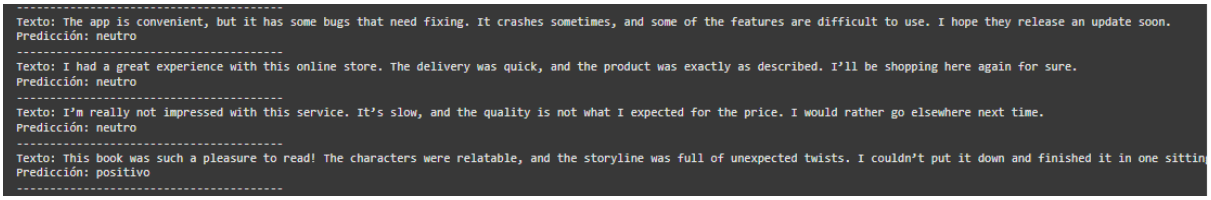


Accuracy: 0.7750677506775068

Classification Report:

	precision	recall	f1-score	support
negativo	0.98	0.82	0.89	246
neutro	0.61	0.96	0.75	246
positivo	0.92	0.54	0.69	246
accuracy			0.78	738
macro avg	0.84	0.78	0.77	738
weighted avg	0.84	0.78	0.77	738

Fuente: Elaboración propia



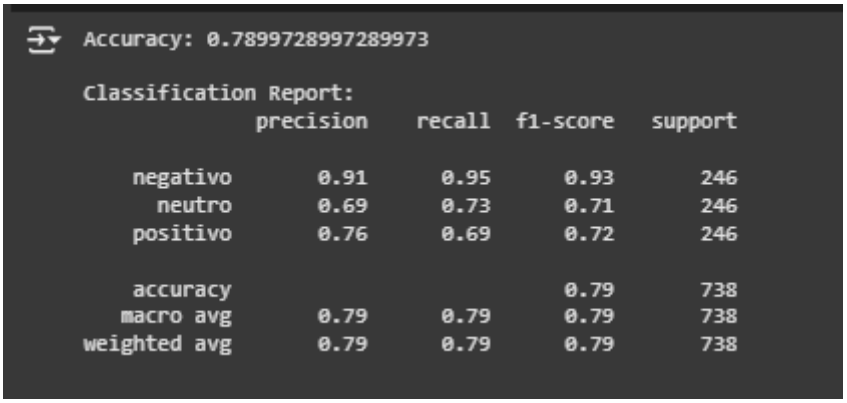
```
----------------------------------------
Texto: The app is convenient, but it has some bugs that need fixing. It crashes sometimes, and some of the features are difficult to use. I hope they release an update soon.
Predicción: neutro
----------------------------------------
Texto: I had a great experience with this online store. The delivery was quick, and the product was exactly as described. I'll be shopping here again for sure.
Predicción: neutro
----------------------------------------
Texto: I'm really not impressed with this service. It's slow, and the quality is not what I expected for the price. I would rather go elsewhere next time.
Predicción: neutro
----------------------------------------
Texto: This book was such a pleasure to read! The characters were relatable, and the storyline was full of unexpected twists. I couldn't put it down and finished it in one sittin
Predicción: positivo
----------------------------------------
```

Fuente: Elaboración propia

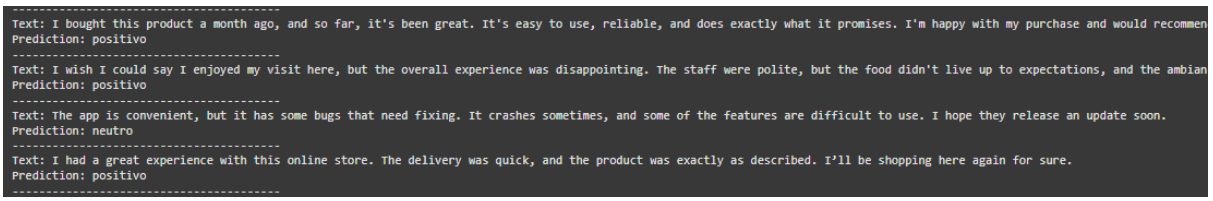
SVM con SMOTE:

El modelo que obtuvo el mejor desempeño fue el SVM con SMOTE, con un accuracy de 78.99%. Este modelo fue especialmente efectivo para clasificar las clases negativa y neutra, con una precisión de 91% para la clase negativa y un recall del 95%, lo que indica que fue muy eficaz para identificar los sentimientos negativos. SVM con SMOTE se benefició del balanceo de clases proporcionado por SMOTE, lo que permitió mejorar la capacidad del modelo para distinguir entre clases minoritarias y mayoritarias, sin sesgarse demasiado hacia la clase más frecuente. El rendimiento del modelo en la clase positiva fue también notable, con una precisión de 76% y un recall de 69%. En términos generales, el modelo SVM con SMOTE demostró ser el más equilibrado, con una F1-score de 79%, lo que lo convierte en el modelo más robusto para este problema.

- **Resultados del modelo SVM con SMOTE:**



Fuente: Elaboración propia



Fuente: Elaboración propia

Conclusiones

En este trabajo, se emplearon técnicas de balanceo de datos mediante SMOTE y modelos supervisados de aprendizaje automático (SVM, Random Forest y Logistic Regression) para la clasificación de sentimientos en redes sociales. Los resultados obtenidos fueron en general positivos, destacando el modelo SVM con SMOTE, que logró un accuracy del 78.99% y un F1-score del 79%, lo que demuestra su capacidad para manejar datos desbalanceados y clasificar sentimientos de manera efectiva. El modelo Random Forest también obtuvo un desempeño sólido con un accuracy del 77.51%, mostrando un buen manejo de las clases neutras, mientras que Logistic Regression, aunque eficiente en términos de tiempo de entrenamiento, presentó un rendimiento inferior con un accuracy del 60.56%, lo que limita su aplicación en este tipo de problemas complejos. Estos resultados destacan la importancia de combinar técnicas de preprocesamiento, como el balanceo de clases, con algoritmos adecuados para mejorar la precisión en tareas de clasificación de texto.

A pesar de los buenos resultados, los modelos tuvieron dificultades para diferenciar entre sentimientos sutiles o similares, como las clases positiva y neutra, lo que sugiere que enfoques más avanzados, como redes neuronales profundas o transformers, podrían mejorar aún más la capacidad de los modelos para captar relaciones complejas en los textos. En cuanto al trabajo futuro, se recomienda explorar y ajustar modelos de deep learning, ampliar el conjunto de datos para incluir más variabilidad y contextos, y considerar factores adicionales como los metadatos (plataforma de origen, horarios de publicación). Estas acciones podrían optimizar el desempeño y la generalización de los modelos, llevando a resultados más precisos y confiables. En resumen, este estudio demuestra el impacto positivo del balanceo de clases y el uso de algoritmos supervisados para abordar eficazmente el análisis de sentimientos en redes sociales, con un alto potencial para aplicaciones prácticas en áreas como marketing, atención al cliente y análisis de opiniones.

Referencias Bibliográficas

- **BrightLocal.** (2020). *Local Consumer Review Survey 2020*. Recuperado de https://www.brightlocal.com/research/local-consumer-review-survey/
- **Facebook Newsroom.** (2023). *Company Info*. Recuperado de https://about.fb.com/company-info/
- **Ferrara, E., & Yang, Z.** (2015). Measuring emotional contagion in social media. *PLoS ONE*, 10(11), e0142390. https://doi.org/10.1371/journal.pone.0142390
- **Giménes, S.** (2023). Redes Sociales, estado actual y tendencias 2023. OBS Business School. Recuperado de https://marketing.onlinebschool.es/Prensa/Informe%20OBS%20Tendencias%20Redes%20Sociales%202023.pdf
- **Gandomi, A., & Haider, M.** (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- **Internet Live Stats.** (2023). *Twitter Usage Statistics*. Recuperado de http://www.internetlivestats.com/twitter-statistics/
- **Kemp, S.** (2023). *Digital 2023: Global Overview Report*. DataReportal. Recuperado de https://datareportal.com/reports/digital-2023-global-overview-report
- **MarketsandMarkets.** (2020). *Sentiment Analysis Market by Component, Application, Deployment Mode, Organization Size, Vertical and Region - Global Forecast to 2025*. Recuperado de https://www.marketsandmarkets.com/Market-Reports/sentiment-analysis-market-932.html
- **Zhang, L., Wang, S., & Liu, B.** (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253. https://doi.org/10.1002/widm.1253