PROYECTO ETL

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA

ANALISIS ALGORITMOS

MAIKOL VERGARA

VICTOR PEÑARANDA

JULIANA CASTRO

VALENTINA LEON

JUAN PERES

LUKAS RADRIGUEZ

JUAN DAVID BARAJAS

ARQUITECTURA

El sistema se compone de varios módulos que trabajan juntos para extraer, transformar, cargar y analizar los datos de comentarios relacionados con Gustavo Petro en redes sociales. La arquitectura del sistema incluye los siguientes componentes:

Extracción de Datos (Scraping)

- Herramienta Utilizada: UIPATH.
- Descripción: UIPATH se utiliza para automatizar el scraping de comentarios desde diferentes redes sociales. Los datos extraídos se almacenan temporalmente en archivos XLSX en Google Drive.

Almacenamiento temporal

- Plataforma: Google Drive.
- Descripción: Los archivos XLSX generados por UIPATH se almacenan en Google Drive para su posterior procesamiento.

Transformación y Carga (ETL)

- **Lenguaje de Programación**: Python.
- **Librerías Utilizadas**: pandas, numpy, transformers, tqdm, emoji, matplotlib, seaborn, wordcloud.
- **Descripción**: Los archivos XLSX se cargan en Google Colab, donde se realizan las transformaciones necesarias para limpiar y preprocesar los datos. Posteriormente, los datos transformados se almacenan en una base de datos adecuada para su análisis.

Análisis sentimiento

- **Modelo de Análisis**: nlptown/bert-base-multilingual-uncased-sentiment.
- **Descripción**: Se utiliza un modelo preentrenado de análisis de sentimientos para clasificar los comentarios en diferentes categorías emocionales.

Visualización y Reporte

- **Herramientas de Visualización**: matplotlib, seaborn, wordcloud.
- **Descripción**: Se generan gráficos y visualizaciones para interpretar y comunicar los resultados del análisis de sentimientos.

TECNOLOGIAS UTILIZADAS

- **UIPATH**: Herramienta de automatización utilizada para realizar el scraping de comentarios en redes sociales.
- **Google Drive**: Plataforma de almacenamiento utilizada para guardar los archivos XLSX generados por UIPATH.
- **Google Colab**: Entorno de programación en Python utilizado para el procesamiento y análisis de los datos.
- **Python**: Lenguaje de programación utilizado para la transformación y análisis de los datos.
- Librerías utilizadas
    - **pandas**: Para la manipulación y análisis de datos.
    - **numpy**: Para operaciones numéricas.
    - **transformers**: Para el análisis de sentimientos.
    - **tqdm**: Para mostrar el progreso del procesamiento.
    - **emoji**: Para el preprocesamiento de texto.
    - **matplotlib, seaborn**: Para la visualización de datos.
    - **wordcloud**: Para la generación de nubes de palabras.

ETL

Extracción

- **Proceso**: Utilizando UIPATH, se extraen los comentarios de redes sociales y se guardan en archivos XLSX en Google Drive.
- **Automatización del Scraping**:
    - **Configuración de Flujos de Trabajo**: Utilizar UIPATH Studio para configurar flujos de trabajo que automatizan la navegación y extracción de comentarios de las páginas web de las redes sociales.
    - **Interacción con el Navegador**: Emplear actividades específicas de UIPATH para interactuar con elementos de las páginas web, como hacer clic en botones, desplazarse por la página, y extraer contenido.
    - **Manejo de Captchas y Autenticación**: Implementar soluciones para manejar captchas y procesos de autenticación que puedan surgir durante el scraping.
    - **Gestión de Errores**: Configurar manejadores de errores para garantizar que el proceso de scraping se recupere automáticamente de fallos comunes, como la desconexión de la red o cambios en la estructura de la página web.

Transformación

- **Proceso**: En Google Colab, se cargan los archivos XLSX y se realizan las siguientes transformaciones:
    - Eliminación de duplicados.
    - Manejo de datos faltantes.
    - Normalización del texto (eliminación de emojis, caracteres especiales, conversión a minúsculas).
    - Corrección de errores ortográficos comunes.

- Código transformación:

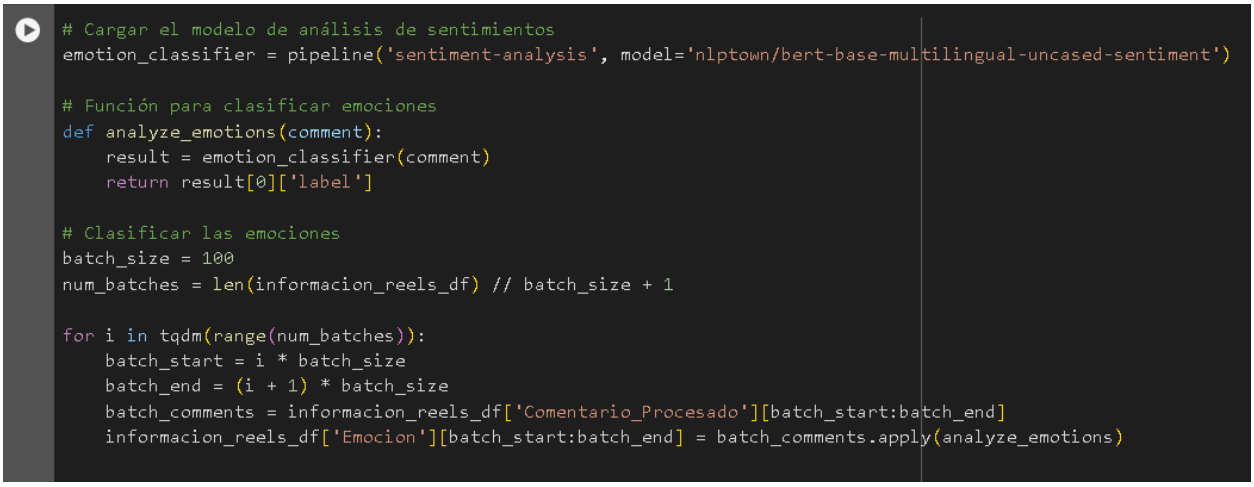
```python
import pandas as pd
import re
from transformers import pipeline
from tqdm import tqdm

# Cargar los archivos XLSX desde Google Drive
file_path = '/content/drive/MyDrive/Scraping Tiktok Data.xlsx'
informacion_reels_df = pd.read_excel(file_path, sheet_name='Informacion_Reels')

# Función de preprocesamiento de texto
def preprocess_text(text):
    emoji_pattern = re.compile(
        "["
        u"\U0001F600-\U0001F64F"
        u"\U0001F300-\U0001F5FF"
        u"\U0001F680-\U0001F6FF"
        u"\U0001F1E0-\U0001F1FF"
        u"\u200d"
        "]+", flags=re.UNICODE)
    text = emoji_pattern.sub(r'', text)
    text = re.sub(r'[^\w\s]', '', text)
    text = text.lower()
    text = re.sub(r'boto', 'voto', text)
    return text

# Limpiar los datos
informacion_reels_df = informacion_reels_df.dropna(subset=['Comentario'])
informacion_reels_df['Comentario'] = informacion_reels_df['Comentario'].astype(str)
informacion_reels_df['Comentario_Procesado'] = informacion_reels_df['Comentario'].apply(preprocess_text)
```

Carga

- **Proceso**: Los datos preprocesados se almacenan en una base de datos adecuada (por ejemplo, PostgreSQL) para su análisis posterior.

- **Razón selección:**
    - **Idioma**: Este modelo está entrenado para manejar múltiples idiomas, incluyendo el español, lo cual es crucial para analizar comentarios en este idioma.
    - **Precisión**: Proporciona una clasificación detallada y precisa de las emociones, que incluye categorías como positivo, negativo y neutral, así como subcategorías más específicas.
    - **Preentrenado**: Al ser un modelo preentrenado, permite una implementación rápida y eficiente sin necesidad de entrenar un modelo desde cero.
    - **Base Multilingual**: La capacidad de analizar comentarios en varios idiomas aumenta la versatilidad y aplicabilidad del modelo en diferentes contextos y plataformas de redes sociales.

- Código carga:



```
# Cargar el modelo de análisis de sentimientos
emotion_classifier = pipeline('sentiment-analysis', model='nlptown/bert-base-multilingual-uncased-sentiment')

# Función para clasificar emociones
def analyze_emotions(comment):
    result = emotion_classifier(comment)
    return result[0]['label']

# Clasificar las emociones
batch_size = 100
num_batches = len(informacion_reels_df) // batch_size + 1

for i in tqdm(range(num_batches)):
    batch_start = i * batch_size
    batch_end = (i + 1) * batch_size
    batch_comments = informacion_reels_df['Comentario_Procesado'][batch_start:batch_end]
    informacion_reels_df['Emocion'][batch_start:batch_end] = batch_comments.apply(analyze_emotions)
```

Implementación

- Proceso: Aplicar el modelo de análisis de sentimientos a los comentarios preprocesados para clasificar las emociones.
- Detalles: El modelo nlptown/bert-base-multilingual-uncased-sentiment se integra en un pipeline de análisis de sentimientos de transformers, permitiendo procesar y clasificar grandes volúmenes de comentarios de manera eficiente.

ANALISIS SENTIMIENTO

- **Modelo Utilizado**: nlptown/bert-base-multilingual-uncased-sentiment.
- **Razón de la Selección**:

- **Preentrenamiento**: El modelo ha sido preentrenado en una gran cantidad de datos multilingües, lo que permite una comprensión profunda de contextos y emociones en diferentes idiomas.
- **Fine-tuning**: Aunque el modelo es preentrenado, se puede ajustar (fine-tune) para mejorar su rendimiento en conjuntos de datos específicos si es necesario.
- **Escalabilidad**: La arquitectura de transformers y la capacidad de procesamiento por lotes permiten que el modelo escale eficientemente a grandes volúmenes de datos.
- **Precision y Recall**: El modelo proporciona métricas detalladas de precisión y recall, lo que permite evaluar y mejorar continuamente el rendimiento del análisis de sentimientos.

- **Validación**: Validar la precisión del modelo mediante la revisión manual de una muestra de comentarios para asegurar que las clasificaciones de emociones sean coherentes y precisas.
- **Almacenamiento**: Guardar los resultados del análisis en una base de datos estructurada, permitiendo consultas y análisis posteriores de los datos de emociones.

RESULTADOS

Porcentajes emociones

	Emocion	Porcentaje
0	neutral	82.903210
1	fear	5.729833
2	joy	4.600191
3	anger	2.421525
4	surprise	2.072857
5	disgust	1.635512
6	sadness	0.636872

Neutralidad: La mayoría de los comentarios se clasificaron como neutrales, lo que sugiere una tendencia hacia la comunicación informativa más que emocional. Esta alta proporción de comentarios neutrales puede deberse a la naturaleza informativa de muchos comentarios o a la falta de contexto emocional claro.

Miedo y Alegría Significativos: Los sentimientos de miedo y alegría son los más destacados entre las emociones no neutrales, lo que puede reflejar divisiones en la percepción pública. Esta dualidad sugiere que hay una notable polarización en las reacciones de las personas.

Presencia de Emociones Negativas: Aunque en menor proporción, las emociones negativas como enojo, disgusto y tristeza están presentes. Estas emociones negativas son importantes de abordar para mejorar la percepción pública y abordar las preocupaciones subyacentes de los usuarios.

Emocion Cuenta	anger	disgust	fear	joy	neutral	sadness	surprise
dejate_llevarr	2.015589	0.953185	4.696421	6.235417	82.802959	1.176587	2.119843
gustavopetrooficial	2.490073	1.730234	5.919645	4.405690	82.875027	0.573312	2.006019

Neutralidad:

- **Cuenta: dejate_llevarr**: La mayoría de los comentarios (82.80%) son neutrales, lo que sugiere que los seguidores de esta cuenta tienden a hacer comentarios informativos o sin una carga emocional clara.
- **Cuenta: gustavopetrooficial**: Similarmente, un alto porcentaje de comentarios (82.87%) son neutrales, indicando una tendencia general hacia la neutralidad en ambas cuentas.

Miedo y Alegría significativos:

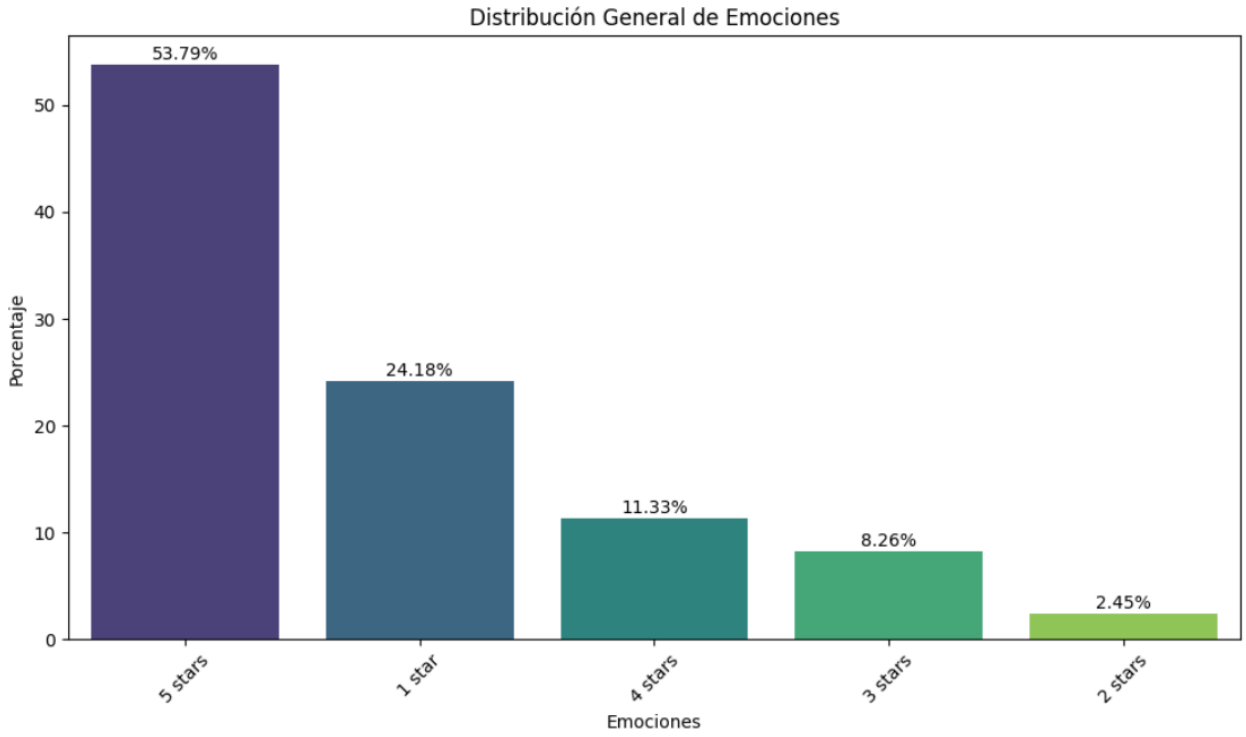
- **Miedo**: Los comentarios que expresan miedo son ligeramente más altos en la cuenta de gustavopetrooficial (5.91%) comparado con dejate_llevarr (4.69%), lo que podría indicar una mayor preocupación o incertidumbre entre los seguidores de la cuenta oficial.
- **Alegría**: La cuenta dejate_llevarr tiene un mayor porcentaje de comentarios alegres (6.23%) en comparación con gustavopetrooficial (4.40%), sugiriendo que los seguidores de la primera son más propensos a expresar sentimientos positivos.

Presencia de Emociones Negativas:

- **Enojo**: La cuenta gustavopetrooficial tiene un mayor porcentaje de comentarios de enojo (2.49%) comparado con dejate_llevarr (2.01%). Este ligero aumento podría reflejar frustraciones específicas de los seguidores de la cuenta oficial.
- **Disgusto**: El disgusto es más prominente en la cuenta oficial (1.73%) frente a dejate_llevarr (0.95%), lo cual puede indicar un descontento mayor entre los seguidores de la cuenta oficial.

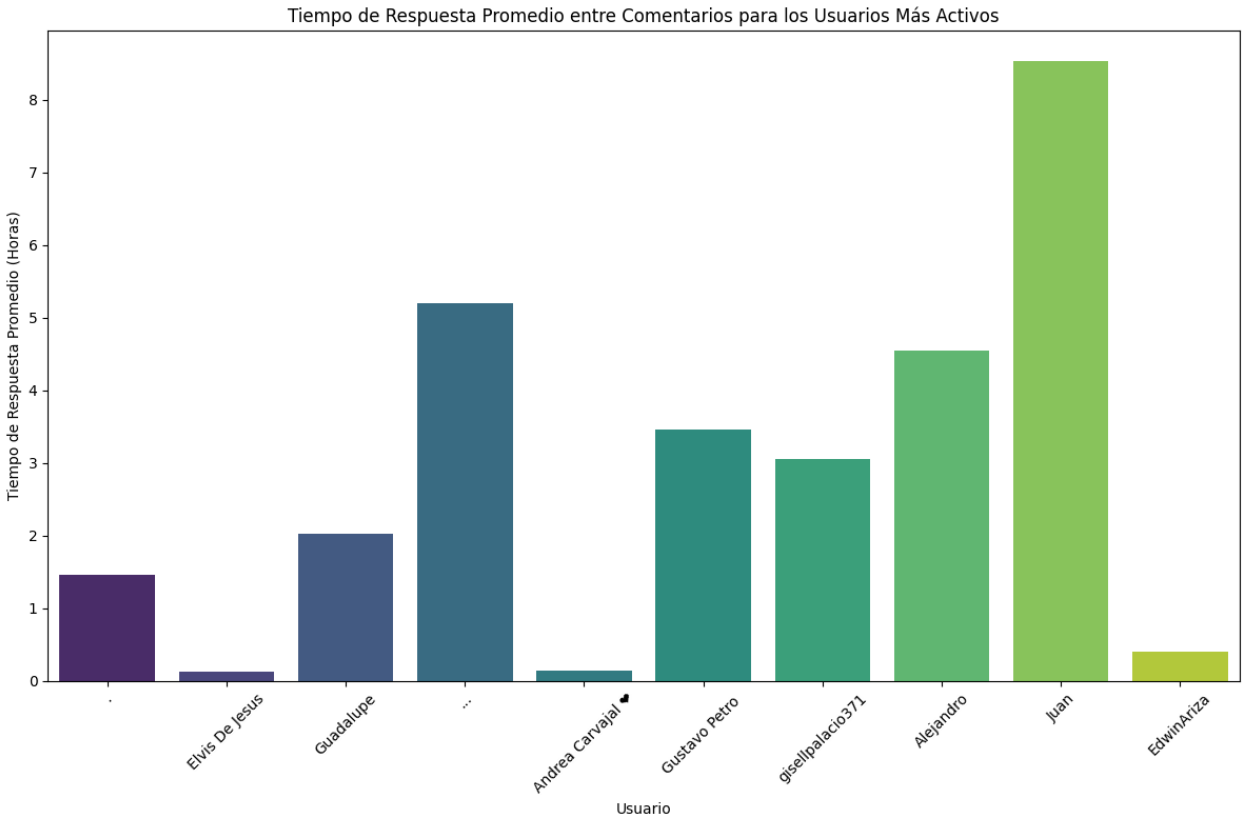
- **Tristeza**: Aunque en menor proporción, la tristeza es más evidente en dejate_llevarr (1.17%) comparado con gustavopetrooficial (0.57%), lo que podría reflejar momentos o eventos específicos que impactaron a los seguidores de esta cuenta.

Distribución estrellas



- Este gráfico de barras muestra la distribución porcentual de los comentarios clasificados por estrellas, de 1 a 5 estrellas. La mayor parte de los comentarios (53.79%) son extremadamente positivos (5 estrellas). Los comentarios extremadamente negativos (1 estrella) representan el 24.18% del total. Las demás clasificaciones (2, 3 y 4 estrellas) tienen una distribución menor, con 2.45%, 8.26%, y 11.33% respectivamente.
- En conclusión, se puede decir que los comentarios tienden a ser polarizados, con una mayoría significativa de comentarios muy positivos y un número considerable de comentarios muy negativos. Esto sugiere que la opinión de los usuarios sobre el presidente es bastante dividida, con sentimientos fuertes en ambos extremos del espectro.

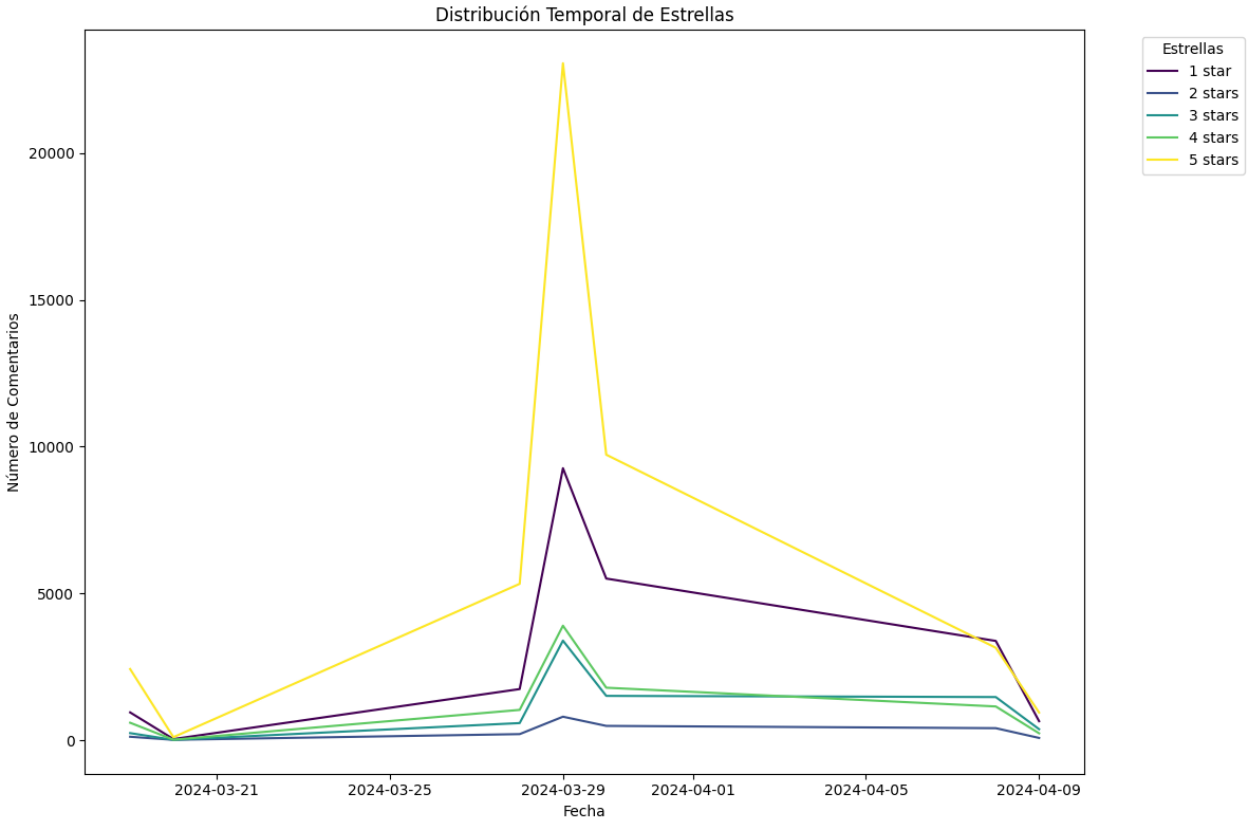
Tiempo de Respuesta Promedio entre Comentarios para los Usuarios Más Activos



Este gráfico de barras muestra el tiempo de respuesta promedio entre comentarios para los usuarios más activos en la cuenta del presidente. Se observan grandes variaciones en el tiempo de respuesta promedio, que varía desde menos de una hora hasta más de ocho horas.

En conclusión, se puede decir que algunos usuarios tienden a interactuar rápidamente, mientras que otros toman más tiempo para responder. La distribución del tiempo de respuesta puede ayudar a identificar patrones de comportamiento y niveles de compromiso de los usuarios más activos.

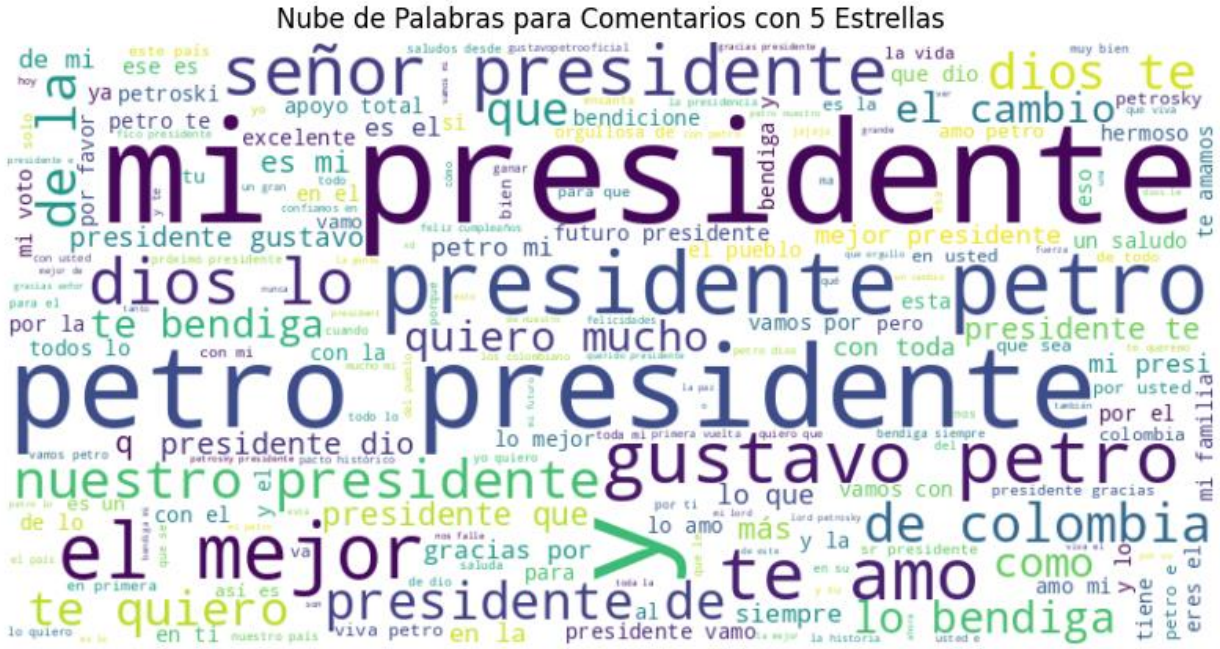
Distribución Temporal de Estrellas



Este gráfico de líneas ilustra la fluctuación de las clasificaciones de estrellas a lo largo del tiempo. Se destacan picos significativos en la actividad de comentarios, especialmente en la clasificación de 5 estrellas.

En conclusión, los picos de actividad pueden correlacionarse con eventos específicos o publicaciones importantes en la cuenta del presidente. El alto número de comentarios positivos en ciertos momentos sugiere una respuesta favorable del público durante esos períodos.

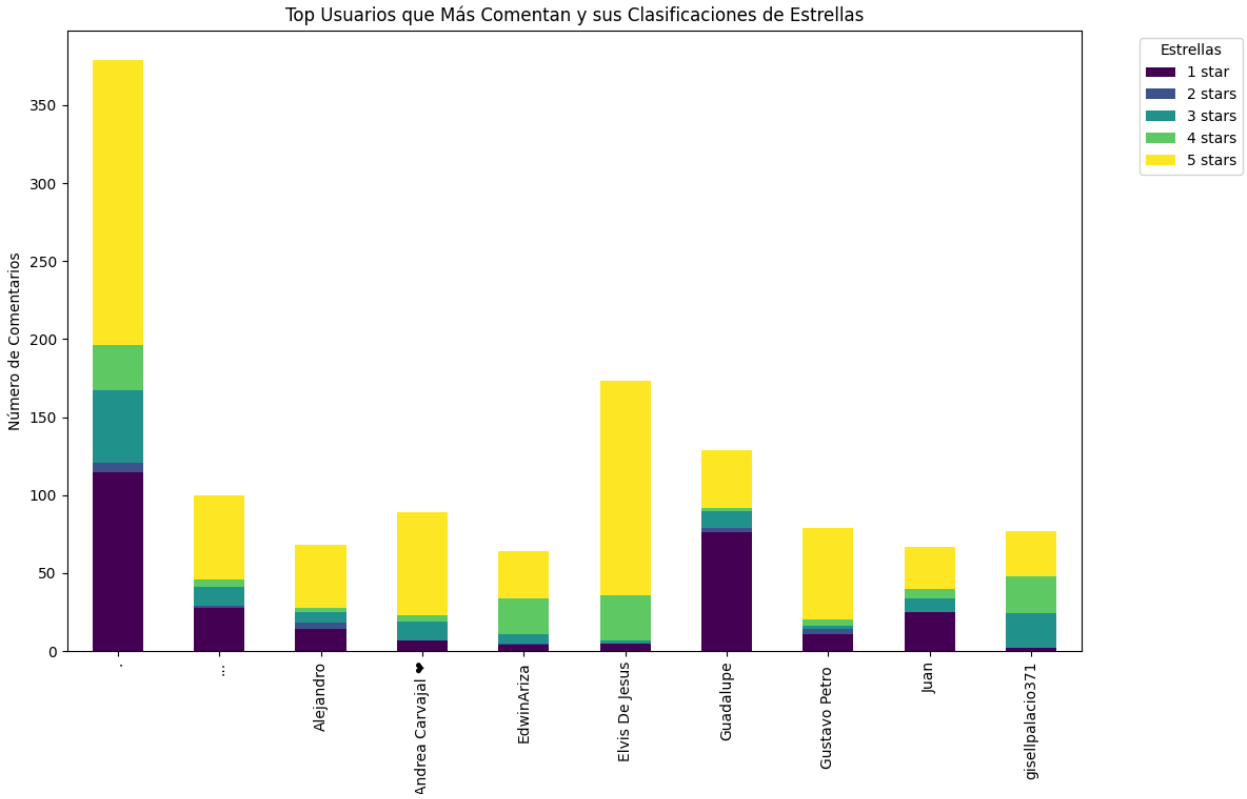
Nube de Palabras para Comentarios con 5 Estrellas



Esta nube de palabras resalta las palabras más frecuentes en los comentarios clasificados con 5 estrellas. Términos como "presidente", "mi presidente", "petro", "gustavo", y "te bendiga" son prominentes.

En resumen, los comentarios muy positivos tienden a ser elogiosos y muestran un fuerte apoyo hacia el presidente. Las palabras más comunes reflejan respeto y cariño hacia la figura presidencial.

Top Usuarios que Más Comentan y sus Clasificaciones de Estrellas



Este gráfico de barras apiladas muestra la distribución de estrellas en los comentarios realizados por los usuarios más activos. Se observa que algunos usuarios tienen una mayor cantidad de comentarios positivos (5 estrellas), mientras que otros tienen una mezcla de clasificaciones.

En conclusión, la diversidad en las clasificaciones de estrellas entre los usuarios más activos sugiere diferentes niveles de satisfacción y opiniones variadas sobre el presidente. Identificar estos usuarios y sus patrones de comentarios puede ayudar a entender mejor la dinámica de interacción en la cuenta.

CONCLUSIÓN

El análisis de los comentarios en la cuenta del presidente revela una polarización significativa en las opiniones de los usuarios, con una mayoría de comentarios extremadamente positivos y una cantidad considerable de comentarios extremadamente negativos. Los patrones de tiempo de respuesta y la distribución temporal de los comentarios también proporcionan información valiosa sobre el comportamiento y el compromiso de los usuarios. Este tipo de análisis puede ser útil para diseñar estrategias de comunicación más efectivas y para comprender mejor las percepciones públicas sobre el presidente en las redes sociales.