UNIVERSIDAD DE GUADALAJARA

MAESTRÍA EN CIENCIAS EN BIOINGENIERÍA Y CÓMPUTO INTELIGENTE

Análisis de datos

Proyecto de final

Análisis de datos en sentimientos y tendencias de opinión en los debates presidenciales de México 2024

Alumno:

IBQ. Rigoberto Rincón Ballesteros

Docente:

Dr. Germán Andrés Preciat

Descripción

Este proyecto analiza los tres debates presidenciales de México 2024 más vistos en YouTube, utilizando técnicas de procesamiento de lenguaje natural (NLP), análisis de sentimientos y análisis estadístico. La información se recolecta, procesa y visualiza mediante una combinación de Python, SQL y Power BI.

Objetivo general

Identificar tendencias, variación de sentimientos y patrones discursivos entre los diferentes debates, generando conocimiento relevante para el contexto político y social.

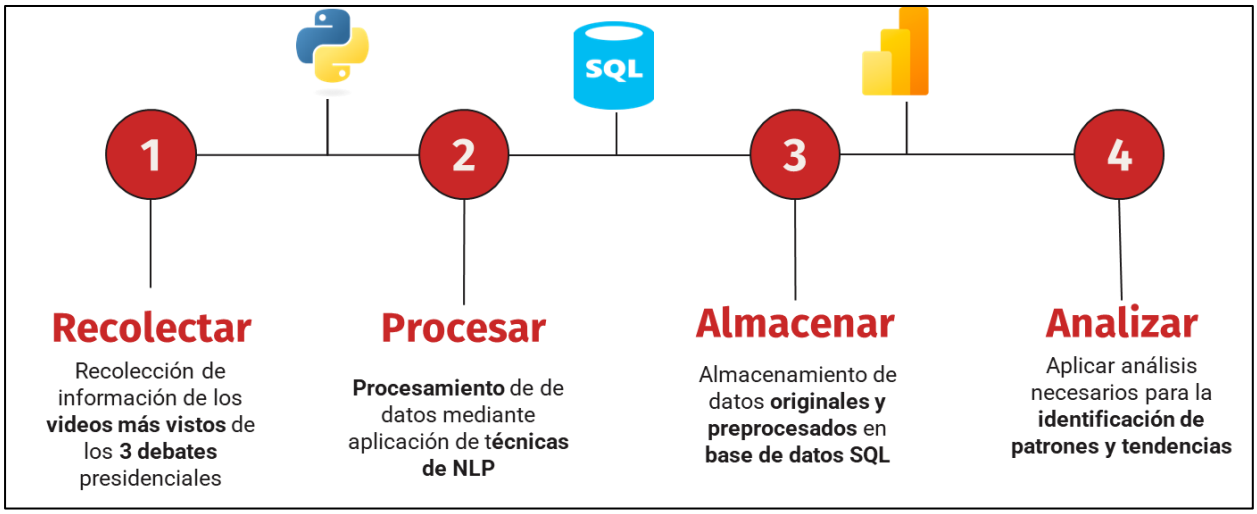
Objetivos específicos

1. Recolectar información de los videos más vistos de los tres debates presidenciales y los comentarios publicados.
2. Preprocesar los datos recolectados mediante técnicas de procesamiento de lenguaje natural (NLP) para extraer información relevante.
3. Estructurar los datos originales y derivados en una base de datos organizada.
4. Realizar un análisis estadístico para la interpretación de las tendencias y patrones encontrados en los datos.
5. Desarrollar un informe interactivo que presente los resultados.

Preguntas de investigación

En este proyecto de análisis de datos se plantearon una serie de 8 preguntas de investigación, las cuales fueron de mi interés abordar con base a la información obtenida durante un período comprendido desde la publicación del video del primer debate hasta la fecha del 20 de octubre de 2024.

Metodología seguida



Recolección:

La recolección de datos se realizó en dos etapas. En la primera etapa, se accedió directamente a la plataforma de YouTube, donde, mediante un filtro de búsqueda, se seleccionó el video más visto para cada uno de los tres debates presidenciales. Además, se descartaron aquellos videos cuya duración era mayor a la esperada, como en los casos donde algunos noticieros añadían un programa breve al final del debate para discutir los eventos sucedidos. De los videos seleccionados, se recopiló el enlace para su uso en la etapa posterior.

En la segunda etapa, se utilizó el ID de los videos para extraer la información en un entorno de desarrollo con Python, utilizando la API de YouTube V3, ofrecida gratuitamente por Google Cloud Products (GCP).

Primer debate: https://www.youtube.com/watch?v=kZauclTWv00&t=1051s

Segundo debate: https://www.youtube.com/watch?v=0osEeTQLk3Q

Tercer debate: https://www.youtube.com/watch?v=DEbALmrsZs8

Procesamiento:

Para el procesamiento de los datos, se utilizó Python para manejar los formatos de fecha y el texto en los comentarios. En el caso de las fechas, se aplicó una corrección para estandarizarlas a un formato más adecuado de *datetime*. Por otro lado, para el texto, se emplearon diversas técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés), tales como:

- Conversión de todo el texto a minúsculas.
- Eliminación de caracteres especiales como hashtags (#), menciones (@), números y otros símbolos ($!).
- Corrección ortográfica: se creó un diccionario con correcciones ortográficas típicas de usuarios mexicanos para restaurar la forma original de las palabras.

- Lematización: se aplicó este proceso para reducir el ruido en el texto, de modo que las variaciones de palabras, como las conjugaciones verbales, se normalizaran.
- Eliminación de palabras de relleno (*stopwords*): se creó un diccionario en español con palabras que sirven principalmente como conectores y que añaden ruido a las frases, lo que podría afectar el rendimiento en análisis posteriores.

Además, se llevó a cabo un análisis de sentimiento con el fin de clasificar la percepción en los comentarios como negativa, positiva o neutra. Para esto, se utilizó un modelo de análisis de sentimiento disponible gratuitamente en la plataforma HuggingFace, llamado "robertuito", el cual fue entrenado con alrededor de 5,000 tweets en español.

Almacenamiento:

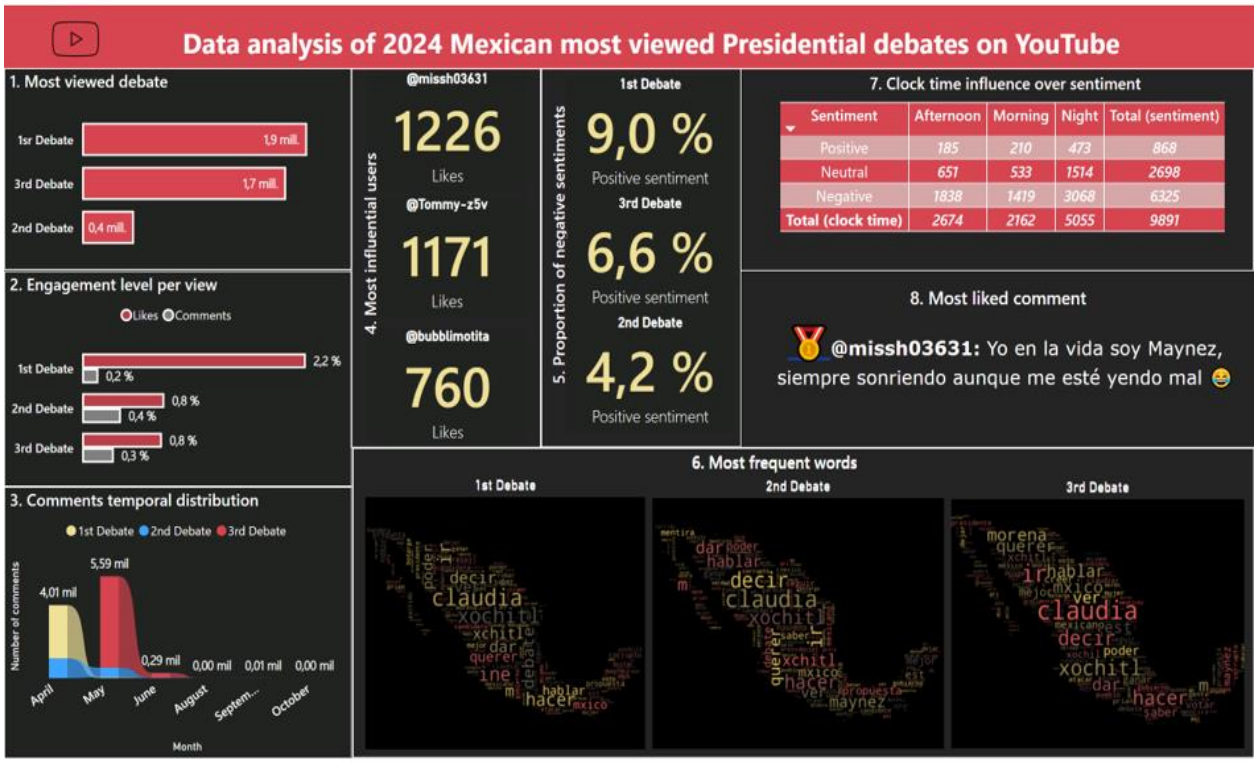
Con el fin de almacenar la información y garantizar un acceso eficiente, se optó por crear una base de datos relacional para conservar tanto los datos originales como los procesados. Para ello, se utilizó SQLite, por su facilidad de uso y adaptabilidad en procesos que no requieren grandes cantidades de memoria. La base de datos se generó como un archivo con extensión ".db".

Análisis:

Finalmente, para responder las preguntas de investigación planteadas, se decidió utilizar el software Power BI, con el propósito de facilitar la generación de visualizaciones y crear un panel de control que resumiera el análisis. En este sentido, almacenar la información en una base de datos SQL fue muy útil, ya que permitió crear consultas específicas (queries) para extraer los datos y cargarlos directamente en Power BI.

Resultados

Después de ejecutar la metodología, mostramos como resultado final el siguiente panel de control donde podemos observar en resumen los hallazgos encontrados luego de responder las preguntas de investigación planteadas.

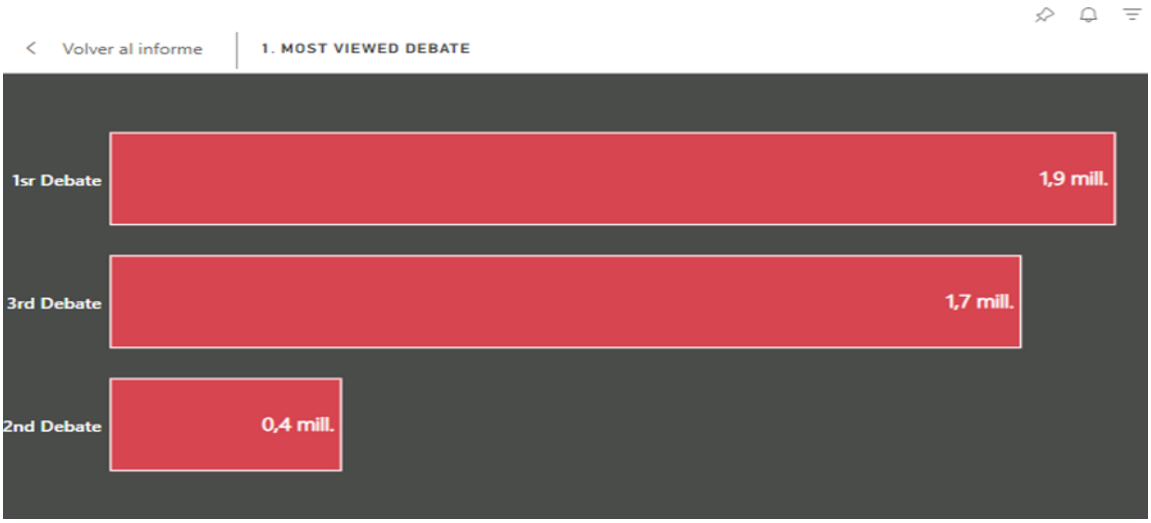


Conclusiones

A continuación, se mencionan algunas de las conclusiones derivadas de los hallazgos encontrados en cada uno de los resultados.

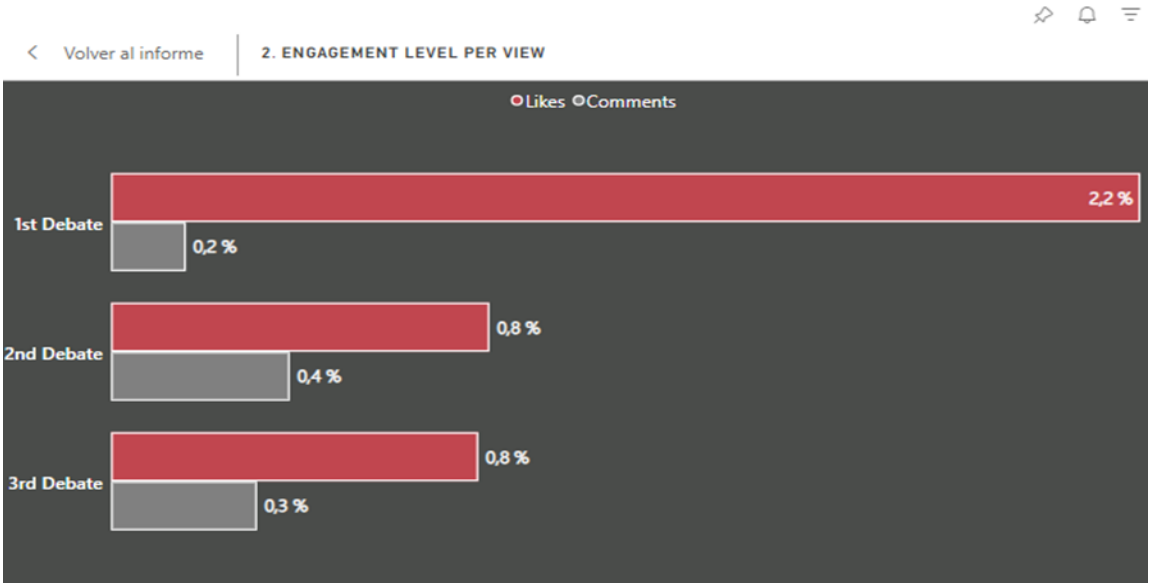
1. ¿Cuál fue el debate con mayor cantidad de visualizaciones durante el periodo de análisis?

Conclusión: El primer debate capturó significativamente más atención, debido posiblemente a la expectativa inicial de los votantes, el interés disminuye a medida que avanza el tiempo.



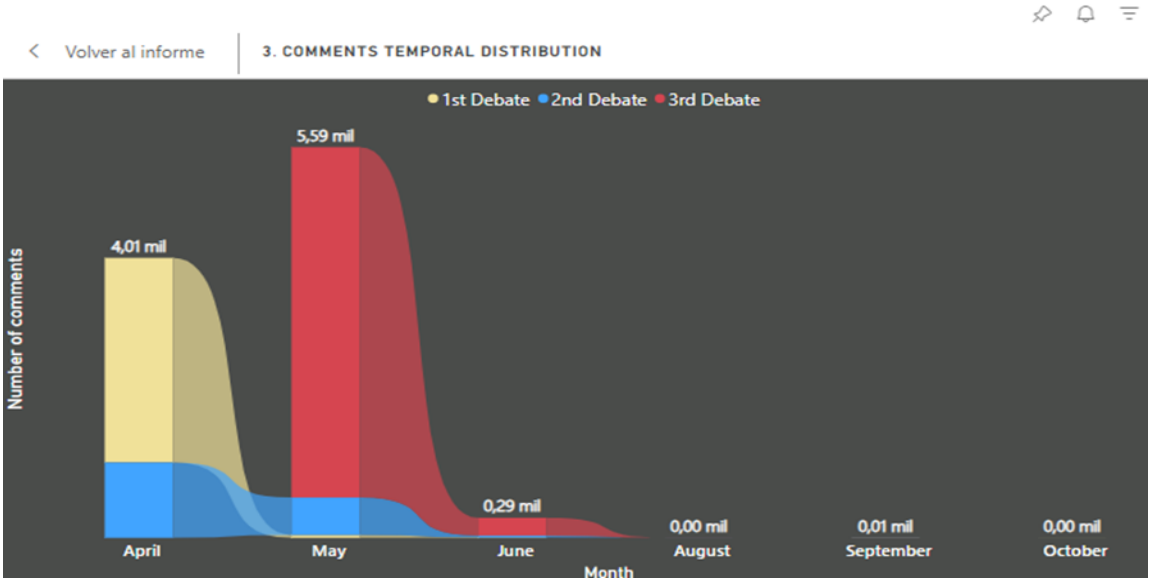
2. ¿Qué nivel de engagement (número de *likes* y comentarios) registró cada debate en el último durante el período de análisis?

Conclusión: A pesar de la repercusión del tercer debate en términos de visualizaciones, el nivel de engagement fue bajo. Esto podría indicar que, aunque la audiencia estaba presente, el debate no motivó a la interacción tanto como el primero.



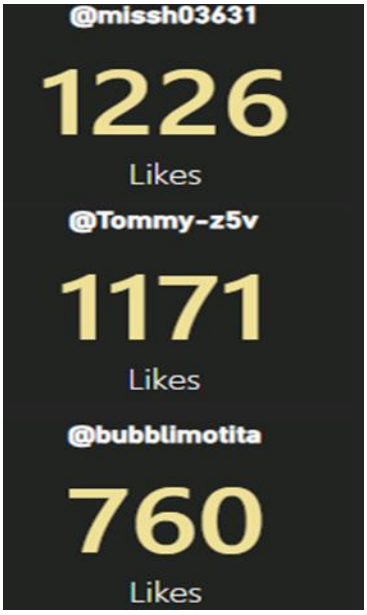
3. ¿Cómo se distribuyeron los comentarios en cada debate a lo largo del período de análisis, considerando intervalos de 1 mes?

Conclusión: La distribución temporal de los comentarios indican un interés momentáneo, en especial el primer y tercer debate. Esto podría ser aprovechado por los políticos para maximizar su popularidad entre los votantes.



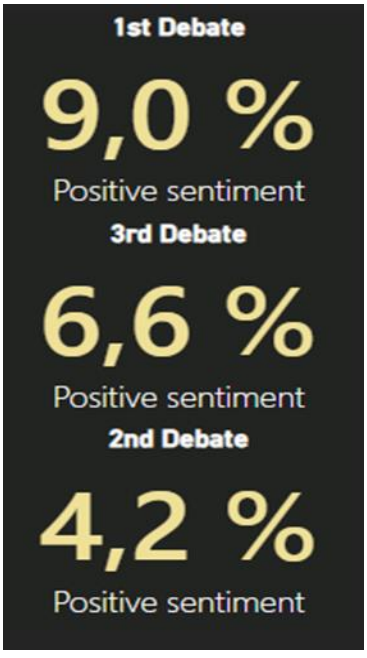
4. ¿Quiénes fueron los tres usuarios más influyentes, en términos de interacciones (*likes*), en la sección de comentarios de los debates durante el período de análisis?

Conclusión: Se identificaron los tres usuarios más influyentes en términos de likes en sus comentarios, las campañas podrían considerar involucrar directamente a estos usuarios o seguir el tipo de discurso que manejan como estrategia política.



5. ¿Qué porcentaje de comentarios expresaron una sentimiento positivo en cada debate, analizados durante el período de análisis?

Conclusión: El porcentaje de comentarios positivos disminuye en debates posteriores. Esto podría señalar la insatisfacción de los usuarios frente a los candidatos o sus propuestas, especialmente en el segundo debate.



6. ¿Cuál fue la palabra más mencionada en los comentarios de cada debate analizado?

Conclusión: Las palabras más mencionadas como “Claudia” o “Xóchilt” sugiere que el dialogo en los comentarios se centró más sobre personajes individuales más que en políticas o propuestas específicas

Debate 1:

Debate 2:

Debate 3:

7. ¿Cómo influye la hora del día en el sentimiento de los comentarios publicados en los debates?

Conclusión: Comentarios positivos y negativos tienen mayor recurrencia por la noche, sin embargo, se puede identificar una tendencia a mayor un mayor número de comentarios positivos por la mañana en comparación con la tarde. Podría ser que la hora del día influya en la percepción de los usuarios.

Sentiment	Afternoon	Morning	Night	Total (sentiment)
Negative	1838	1419	3068	6325
Neutral	651	533	1514	2698
Positive	185	210	473	868
Total (clock time)	2674	2162	5055	9891

8. ¿Cuál fue el comentario con mayor número de interacciónes (*likes*) en todos los debates analizados?

Conclusión: El humor puede ser una herramienta de conexión entre los usuarios y los temas políticos.