



Proyecto Final

Evaluación de la Opinión Pública en Redes Sociales durante los Debates y el Día de la Elección Presidencial en México 2024 mediante Minería de Textos

Rodrigo Gerardo Trejo Arriaga

Eidan Owen Plata Salinas

Angel Hernández Hernández

Instituto Politécnico Nacional
Escuela Superior de Cómputo

Minería de Datos

20 de diciembre de 2024

Índice

1	Introducción	2
1.1	Objetivos	2
1.1.1	Objetivo General	2
1.1.2	Objetivos Específicos	2
1.2	Descripción del Conjunto de Datos	3
1.2.1	Conjunto de Datos de los Debates Presidenciales (YouTube)	3
1.2.2	Conjunto de Datos del Día de la Elección (X)	3
2	Marco Teórico	4
2.1	Estado del Arte	4
2.2	Tareas de Clasificación y Agrupamiento en el Análisis Electoral	5
2.2.1	Agrupamiento (Clustering): Identificación de Temas Principales	5
2.2.2	Clasificación: Análisis del Tipo de Sentimiento	5
2.2.3	Importancia de las tareas de Minería de Datos en el Análisis Electoral	6
3	Método de Desarrollo	6
3.1	Metodología CRISP-DM	6
3.1.1	Comprensión del Negocio	6
3.1.2	Comprensión de los Datos	6
3.1.3	Preparación de los Datos	6
3.1.4	Modelado	7
3.1.5	Evaluación	7
3.1.6	Despliegue	7
3.2	Descripción de las Variables del Conjunto de Datos	7
3.3	Tareas de Minería de Datos	8
3.3.1	Análisis del tipo de sentimiento	8
3.3.2	Análisis de Tópicos	8
3.4	Modelo de la Metodología	8
4	Resultados	8
4.1	Limpieza de Datos	8
5	Conclusiones	10

Evaluación de la Opinión Pública en Redes Sociales durante los Debates y el Día de la Elección Presidencial en México 2024 mediante Minería de Textos

Rodrigo Trejo, Eidan Plata, Angel Hernández
rtrejoa1800@alumno.ipn.mx
Instituto Politécnico Nacional – ESCOM
Minería de Datos

Resumen

Los autores proponen un análisis de los comentarios en YouTube y X durante las elecciones presidenciales de México 2024, enfocándose en el sentimiento y los temas clave relacionados con los candidatos Claudia Sheinbaum, Xóchitl Gálvez y Jorge Álvarez Maynez. Utilizando técnicas de minería de textos y múltiples algoritmos de aprendizaje automático, se examinan las opiniones y reacciones del electorado durante los debates y el día de la elección. El estudio busca identificar patrones de sentimiento, tópicos principales y diferencias entre audiencias en distintas plataformas y canales.

Palabras Clave – Análisis de sentimientos, Minería de Textos, Elecciones México 2024, Redes Sociales, Aprendizaje Automático

1. Introducción

Las redes sociales han emergido como plataformas clave para la expresión y difusión de opiniones durante procesos electorales. En el contexto de las elecciones presidenciales de México 2024, plataformas como YouTube y X se han convertido en espacios donde los ciudadanos comparten sus percepciones, críticas y apoyos hacia los candidatos. Este flujo de información proporciona una oportunidad invaluable para analizar la opinión pública y entender las dinámicas sociales que influyen en el electorado.

El presente estudio se enfoca en analizar los comentarios y publicaciones realizados en YouTube y X durante los debates presidenciales y el día de la elección. Los candidatos principales en estas elecciones fueron Claudia Sheinbaum Pardo, Xóchitl Gálvez Ruiz y Jorge Álvarez Maynez. Mediante la recopilación de dos conjuntos de datos—uno de 9,392 comentarios de YouTube y otro de 2,486 publicaciones de X—se busca explorar el sentimiento y los temas clave que predominan en las discusiones en línea.

Utilizando técnicas de minería de textos y una combinación de algoritmos de aprendizaje automático, este trabajo pretende ofrecer una visión de cómo las opiniones y reacciones de los ciudadanos evolucionaron durante eventos críticos del proceso electoral. Además, se propone comparar las diferencias entre las audiencias de distintas plataformas y canales de comunicación, con el fin de identificar posibles sesgos y variaciones en las percepciones hacia cada candidato.

Este estudio no solo contribuye al entendimiento de la opinión pública en contextos electorales, sino que también demuestra la utilidad de las técnicas de minería de datos y aprendizaje automático en el análisis de volúmenes de datos no estructurados provenientes de las redes sociales.

1.1. Objetivos

1.1.1. Objetivo General

El objetivo principal de este estudio es analizar y comprender las opiniones y reacciones expresadas por los ciudadanos durante las elecciones presidenciales de México 2024, utilizando técnicas de minería de textos y aprendizaje automático sobre comentarios recopilados de YouTube y X.

1.1.2. Objetivos Específicos

- Medir y comparar el sentimiento (positivo, negativo, neutral) expresado hacia cada uno de los candidatos—Claudia Sheinbaum, Xóchitl Gálvez y Jorge Álvarez Maynez—en los comentarios de YouTube durante

los debates y en X durante el día de la elección.

- Descubrir los temas más discutidos por los ciudadanos en relación con cada candidato durante los debates y el día de la elección.
- Examinar cómo evolucionaron las opiniones y reacciones de las personas durante el día de la elección en X.
- Analizar si existen diferencias significativas en las opiniones y temas discutidos entre las audiencias de diferentes canales de YouTube y entre las plataformas de YouTube y X.

1.2. Descripción del Conjunto de Datos

En este estudio se utilizaron dos conjuntos de datos recopilados de manera propia, con el propósito de analizar las tendencias de la elección presidencial de México 2024 según las opiniones expresadas por las personas en redes sociales. A continuación se describen detalladamente ambos conjuntos de datos.

1.2.1. Conjunto de Datos de los Debates Presidenciales (YouTube)

Este conjunto de datos contiene comentarios extraídos de videos de YouTube correspondientes al primer, segundo y tercer debate presidencial. Los comentarios fueron recopilados de diferentes canales de noticieros reconocidos en México, como Milenio y Nmás.

- **Fuente de datos:** Comentarios de videos de YouTube sobre los debates presidenciales.
- **Autoría:** Datos recopilados por el autor del estudio.
- **Propósito:** Analizar las tendencias y opiniones de las personas respecto a los candidatos durante los debates.
- **Número de registros:** 9,392 comentarios.

El conjunto de datos cuenta con los siguientes atributos:

- **num_debate:** Número del debate presidencial (1, 2 o 3).
- **canal:** Nombre del canal de YouTube donde se transmitió el debate.
- **username:** Nombre de usuario que realizó el comentario.
- **fecha:** Fecha en que se realizó el comentario.
- **comentario:** Contenido textual del comentario.
- **num_likes:** Número de "me gusta" que recibió el comentario.

El diccionario de datos que se tiene es el siguiente

Atributo	Tipo de dato	Descripción
num_debate	Entero	Número del debate presidencial (1, 2 o 3).
canal	Cadena de texto	Nombre del canal de YouTube (ejemplo: <i>Milenio</i> , <i>Nmás</i>).
username	Cadena de texto	Nombre de usuario en YouTube que realizó el comentario.
fecha	Fecha	Fecha en formato DD/MM/AAAA en que se publicó el comentario.
comentario	Cadena de texto	Texto del comentario realizado por el usuario.
num_likes	Entero	Cantidad de "me gusta" que obtuvo el comentario.

Cuadro 1: Diccionario de datos del conjunto de comentarios de YouTube

1.2.2. Conjunto de Datos del Día de la Elección (X)

Este conjunto de datos incluye publicaciones de X recopiladas manualmente durante el día de la elección, reflejando las reacciones y opiniones de las personas en tiempo real.

- **Fuente de datos:** Publicaciones de X durante el día de la elección.
- **Autoría:** Datos recopilados por el autor del estudio.
- **Propósito:** Analizar las tendencias y opiniones de las personas durante el día de la elección presidencial.
- **Número de registros:** 2,486 publicaciones.

El conjunto de datos cuenta con los siguientes atributos:

- **User:** Nombre del usuario en X.

- **aroba:** Nombre de usuario precedido por “@”.
- **hora_publicación:** Hora en que se publicó el tweet.
- **publicación:** Contenido textual del tweet.

El diccionario de datos que se tiene es el siguiente

Atributo	Tipo de dato	Descripción
User	Cadena de texto	Nombre del usuario en X.
aroba	Cadena de texto	Handle de X del usuario (ejemplo: @usuario).
hora_publicación	Hora	Hora en formato HH:MM en que se publicó el tweet.
publicación	Cadena de texto	Texto del tweet publicado por el usuario.

Cuadro 2: Diccionario de datos del conjunto de publicaciones de X

2. Marco Teórico

2.1. Estado del Arte

La minería de opiniones y el análisis de sentimientos en el contexto electoral han cobrado relevancia en los últimos años, especialmente debido al auge de las redes sociales como plataformas clave en la movilización política. Estos enfoques, basados en técnicas de minería de textos y procesamiento de lenguaje natural (PLN), ofrecen nuevas perspectivas sobre cómo se configuran las percepciones públicas en tiempos electorales. En este marco, exploraremos una serie de estudios previos que abordan diferentes aspectos de la influencia de los medios tradicionales y digitales en las elecciones, destacando el papel fundamental de las plataformas sociales en la modelación de las narrativas políticas.

Uno de los primeros estudios que se adentra en este terreno es el de *Elecciones presidenciales en el Perú: minería de textos de los editoriales del diario La República* (Castro & Martínez, 2021). Este análisis se concentra en los editoriales del periódico *La República* durante las elecciones presidenciales de 2021 en Perú. Mediante técnicas de minería de textos, se identifica cómo la terminología utilizada en estos medios crea asociaciones que influyen en la percepción pública. Los resultados subrayan que, aunque los medios tradicionales parecen perder fuerza frente a las plataformas digitales, todavía conservan una significativa capacidad para moldear la representación social de los candidatos, especialmente entre los votantes más conservadores.

Este enfoque se complementa con el análisis de *Redes sociales y participación política en las elecciones presidenciales de 2022 en Colombia* (Ramírez & Pérez, 2022), que pone en evidencia el papel de las redes sociales en la participación política. A través de plataformas como Facebook, WhatsApp, Instagram y Twitter, los ciudadanos tienen un acceso sin precedentes a la información política. Sin embargo, el estudio revela que la relación entre el consumo de contenido político en estas plataformas y la participación electoral no es tan directa como se podría esperar. Esto indica que factores adicionales, como el contexto socioeconómico y las estrategias de movilización, pueden estar influyendo en las decisiones de voto, ampliando las perspectivas sobre la influencia de las redes sociales.

En una línea similar, el estudio *Información política en plataformas de redes sociales y participación electoral: evidencia desde Chile utilizando Full Matching* (Fernández & García, 2020) profundiza en el impacto de las redes sociales en la participación electoral en Chile. Utilizando la técnica de Full Matching para ajustar los datos y evitar sesgos, el estudio concluye que no existe una asociación significativa entre el consumo de información política en redes sociales y el aumento de la participación electoral. Este hallazgo resalta la importancia de considerar variables contextuales y sugiere que, aunque las redes sociales tienen un impacto, no son la única variable que determina la participación ciudadana.

Por otro lado, el estudio *POPmine: Tracking Political Opinion on the Web* (Pérez & González, 2020) aborda la minería de opiniones desde una perspectiva más técnica. Presentando la herramienta POPmine, diseñada para rastrear opiniones políticas a través de plataformas como Twitter, el estudio utiliza el modelo BERT en español para detectar emociones y sentimientos en los mensajes de congresistas colombianos. Los resultados revelan que herramientas avanzadas como BERT permiten identificar emociones complejas, proporcionando una comprensión más precisa de cómo los votantes perciben a los candidatos. Este enfoque innovador muestra cómo el PLN puede transformar el análisis político, facilitando una evaluación más profunda de la dinámica electoral.

El impacto de las redes sociales también se refleja en el estudio *Impacto de las redes sociales en la percepción ciudadana sobre la compra del voto en México* (Hernández & Gómez, 2019), que examina cómo las redes sociales influyen en la percepción pública de prácticas ilícitas durante las elecciones presidenciales de 2018 en México. A través de datos del Latinobarómetro y el uso de modelos estadísticos, se concluye que las redes sociales son herramientas

más eficaces que los medios tradicionales para sensibilizar a los ciudadanos sobre prácticas como el clientelismo y la compra del voto. Este hallazgo resalta el poder de las plataformas digitales no solo en la difusión de información política, sino también en su capacidad para promover la vigilancia ciudadana frente a fenómenos ilegales.

En el mismo contexto, el artículo *Exploring Mexican Voting Intention Through Spatiotemporal Trends Using Social Media and Open Data Analysis* (Zagal & Mata, 2020) ofrece una visión más profunda al estudiar la intención de voto en las elecciones mexicanas de 2018. Mediante un enfoque espacio-temporal que combina datos de redes sociales, como tweets y memes, con resultados electorales oficiales, el estudio utiliza técnicas de modelado de temas y análisis de sentimientos. Los hallazgos subrayan cómo el análisis de datos no estructurados puede identificar narrativas dominantes en la política, proporcionando una comprensión más precisa de las tendencias electorales que reflejan las emociones y actitudes de los votantes.

Por último, el trabajo *Opinion Mining Applied to Public Opinion Analysis in the Electoral Context* (Castiblanco & González, 2022) se enfoca en el análisis de opiniones en el contexto electoral colombiano, utilizando el modelo BERT en español para identificar emociones y sentimientos en los mensajes de los congresistas. Este estudio revela que, al aplicar herramientas avanzadas de PLN, es posible identificar patrones emocionales que afectan las percepciones de los votantes, proporcionando una valiosa herramienta para analistas y estrategias políticos.

De esta manera, estos estudios ilustran el poder de las redes sociales y las herramientas de PLN en el análisis de la opinión pública en el contexto electoral. Mientras que los medios tradicionales continúan influyendo en la percepción pública, las plataformas digitales han demostrado tener un impacto en la movilización política y la detección de fenómenos ilegales. Además, el uso de modelos como BERT está cambiando la forma en que entendemos las emociones y las narrativas políticas, proporcionando enfoques más detallados para el análisis electoral.

2.2. Tareas de Clasificación y Agrupamiento en el Análisis Electoral

En el análisis electoral, las tareas de clasificación y agrupamiento juegan un papel importante en el descubrimiento de patrones y en la interpretación de los datos provenientes de los debates y las interacciones de los votantes en redes sociales durante el día de la elección. A través de estas técnicas, es posible obtener insights clave sobre las opiniones, emociones y temas discutidos, facilitando la comprensión de las dinámicas electorales.

2.2.1. Agrupamiento (Clustering): Identificación de Temas Principales

Mientras que la clasificación permite analizar las emociones detrás de los comentarios, **el agrupamiento o clustering** se enfoca en identificar los **temas principales** tratados durante los debates y el día de las elecciones. Mediante esta técnica, los comentarios se agrupan según su contenido semántico, lo que permite descubrir las áreas de interés y las preocupaciones predominantes de los votantes sin necesidad de etiquetar previamente los temas.

En el contexto electoral, el clustering tiene un valor significativo nos ayuda a :

- **Descubrir los temas más relevantes para los votantes:** Durante los debates, es probable que se discutan una variedad de temas, desde políticas públicas hasta cuestiones personales sobre los candidatos y al aplicar técnicas de agrupamiento, podemos entender mejor qué cuestiones dominan las discusiones.
- **Mejorar la estrategia de comunicación política:** Al identificar los temas predominantes durante los debates, los estrategas políticos pueden adaptar su mensaje a las preocupaciones más relevantes de los votantes. Si un candidato nota que un tema específico genera muchas discusiones y reacciones, puede decidir profundizar en ese tema en futuros discursos o debates. Además, permite descubrir áreas donde los votantes están desinformados o preocupados, lo que ofrece oportunidades para intervenir y cambiar la narrativa.

2.2.2. Clasificación: Análisis del Tipo de Sentimiento

La **clasificación** de comentarios, en este caso, se refiere a la tarea de asignar a cada mensaje una etiqueta de sentimiento: **positivo**, **negativo** o **neutral**. Durante los debates políticos y en el día de las elecciones, el análisis de sentimientos se convierte en una herramienta poderosa para comprender la actitud y las emociones de los votantes hacia los candidatos y las propuestas.

Con este enfoque podremos:

- **Identificar la percepción pública:** En el contexto de un debate o jornada electoral, clasificar los comentarios permite captar cómo los votantes reaccionan ante las intervenciones de los candidatos, sus promesas y sus estrategias. Por ejemplo, los sentimientos negativos pueden surgir como reacción a promesas incumplidas o argumentos débiles durante un debate, mientras que los comentarios positivos pueden reflejar el apoyo a un candidato tras un buen desempeño.
- **Medir la influencia de los debates en la opinión pública:** Los debates son momentos clave donde los votantes expresan su apoyo o rechazo hacia los candidatos. Al clasificar los sentimientos, es posible ver cómo

cambia la opinión de los votantes antes, durante y después de los debates. Esto puede ser útil para predecir tendencias o incluso para ajustar las estrategias de campaña de los partidos políticos.

2.2.3. Importancia de las tareas de Minería de Datos en el Análisis Electoral

Con estas tareas de Minería de Datos podemos desvelar los temas más importantes que están siendo discutidos, tanto durante los debates como en la jornada electoral y proporcionar una visión clara de cómo los votantes se sienten respecto a dichos temas,

Este conocimiento es muy importante para las campañas políticas, ya que permite ajustar estrategias de comunicación, identificar preocupaciones emergentes, medir el impacto de los candidatos y, en última instancia, tomar decisiones basadas en los intereses y emociones del electorado. En un contexto electoral, donde las opiniones pueden cambiar rápidamente, tener acceso a este tipo de hallazgo es esencial para mantener una ventaja competitiva.

3. Método de Desarrollo

3.1. Metodología CRISP-DM

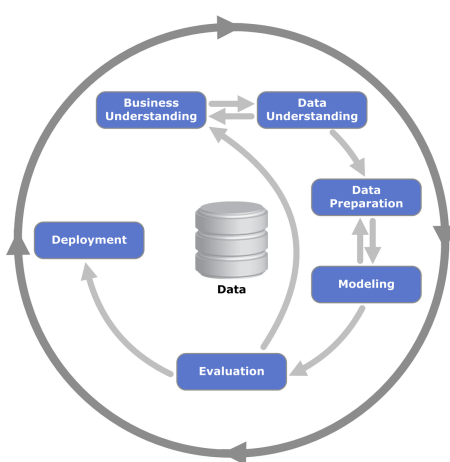


Figura 1: Metodología CRISP-DM

El método CRISP-DM (Cross Industry Standard Process for Data Mining) se aplicó de la siguiente manera en el desarrollo de este proyecto de minería de textos:

3.1.1. Comprensión del Negocio

El objetivo principal del proyecto es analizar las opiniones y reacciones ciudadanas expresadas en redes sociales durante los debates y el día de las elecciones presidenciales de México 2024. Se busca identificar sentimientos (positivo, negativo y neutral) y temas principales asociados a los candidatos, así como las diferencias entre plataformas (YouTube y X). Este análisis permitirá comprender mejor las dinámicas sociales y la percepción pública de los candidatos.

3.1.2. Comprensión de los Datos

Se recopilaron dos conjuntos de datos:

- **YouTube:** Comentarios de los debates presidenciales (9,392 registros), incluyendo atributos como usuario, comentario, canal, número de "me gustó" y fecha.
- **X:** Publicaciones realizadas durante el día de las elecciones (2,486 registros), con atributos como usuario, hora de publicación y contenido textual.

Se realizó una exploración inicial para identificar datos irrelevantes, valores faltantes y posibles inconsistencias.

3.1.3. Preparación de los Datos

Los datos fueron limpiados eliminando duplicados, comentarios vacíos y símbolos no alfabéticos irrelevantes. Además, se realizó un proceso de tokenización, lematización y eliminación de palabras vacías (*stopwords*) para estandarizar el texto. En esta etapa también se añadió una columna de sentimiento, calculada utilizando un modelo preentrenado de análisis de sentimientos.

3.1.4. Modelado

Para el análisis de sentimientos, se utilizó el modelo preentrenado BETO, especializado en español. Este modelo fue ajustado mediante *fine-tuning* con los datos recopilados, adaptándolo al lenguaje manejado en un contexto electoral en redes sociales.

Para el modelado de tópicos, se emplearon los *embeddings* generados por BETO, y se aplicó reducción de dimensionalidad mediante PCA (*Principal Component Analysis*). Finalmente, se utilizó el algoritmo de agrupamiento K-means para identificar y clasificar los tópicos principales.

3.1.5. Evaluación

Los resultados del análisis de sentimientos se evaluaron utilizando la matriz de confusión, calculando métricas como precisión, *recall* y F1-score para cada categoría de sentimiento. Para la evaluación del modelado de tópicos, se utilizó el coeficiente de silueta para medir la cohesión y separación de los clusters generados, asegurando la calidad del agrupamiento.

3.1.6. Despliegue

Los resultados del análisis se presentaron en forma de gráficos y visualizaciones interactivas que muestran:

- Distribución de sentimientos por candidato y plataforma.
- Evolución de opiniones durante el proceso electoral.
- Comparación entre los temas principales discutidos en YouTube y X.

Estas visualizaciones permiten una interpretación clara y útil para el público interesado en el análisis político y social.

3.2. Descripción de las Variables del Conjunto de Datos

A continuación, se presenta la descripción detallada de las variables contenidas en el conjunto de datos utilizado en este estudio. La tabla incluye numeración, nombre de la variable, significado, tipo de dato y dominio de valores.

Núm.	Variable	Significado	Tipo de Dato	Dominio de Valores
1	num_debate	Número del debate presidencial en el que se realizó el comentario.	Entero	{1, 2, 3}
2	canal	Nombre del canal de YouTube donde se transmitió el debate.	Cadena de texto	Cualquier canal reconocido (e.g., <i>Milenio</i> , <i>Nmás</i>).
3	username	Nombre de usuario que realizó el comentario en la plataforma.	Cadena de texto	Texto alfanumérico.
4	fecha	Fecha en la que se publicó el comentario.	Fecha	Formato DD/MM/AAAA.
5	comentario	Contenido textual del comentario realizado por el usuario.	Cadena de texto	Cualquier texto.
6	num_likes	Número de "me gusta" que recibió el comentario.	Entero	{0, 1, 2, ...}

Cuadro 3: Descripción de las variables del conjunto de datos de comentarios en YouTube.

Núm.	Variable	Significado	Tipo de Dato	Dominio de Valores
1	User	Nombre del usuario que publicó el tweet.	Cadena de texto	Texto alfanumérico.
2	arroba	Handle del usuario precedido por "@".	Cadena de texto	Texto alfanumérico (e.g., @usuario).
3	hora_publicación	Hora en la que se publicó el tweet.	Hora	Formato HH:MM (24 horas).
4	publicación	Contenido textual del tweet.	Cadena de texto	Texto libre (e.g., comentarios, reacciones).

Cuadro 4: Descripción de las variables del conjunto de datos de X.

3.3. Tareas de Minería de Datos

Para realizar el análisis de los comentarios recopilados en redes sociales durante los debates y el día de la elección presidencial, se utilizarán los siguientes algoritmos y técnicas:

3.3.1. Análisis del tipo de sentimiento

El análisis de sentimientos tiene como objetivo clasificar los comentarios en categorías como positivo, negativo y neutral. Para esta tarea se empleará el siguiente enfoque:

- **BETO con Fine-Tuning:** Este modelo preentrenado en español será ajustado mediante *fine-tuning* para capturar patrones textuales específicos del contexto electoral. Este enfoque permite clasificaciones más precisas al considerar dependencias contextuales entre las palabras, adaptándolo al lenguaje y las emociones presentes en las redes sociales.

3.3.2. Análisis de Tópicos

El análisis de tópicos busca identificar los temas predominantes en los comentarios. Para esta tarea, se implementará una combinación de representaciones de palabras, reducción de dimensionalidad y agrupamiento:

- **Embeddings de Palabras:** Se utilizarán los embeddings de BETO para generar representaciones vectoriales densas de las palabras, capturando relaciones semánticas y contextuales en el corpus de texto.
- **Reducción de Dimensionalidad con PCA:** Antes de aplicar el agrupamiento, se utilizará el análisis de componentes principales (PCA, por sus siglas en inglés) para reducir la dimensionalidad de los embeddings generados, conservando las características más relevantes y mejorando la eficiencia del algoritmo.
- **K-means:** Sobre los embeddings reducidos, se aplicará el algoritmo de agrupamiento K-means para identificar grupos de comentarios relacionados con tópicos similares.

Esta combinación permitirá descubrir los temas más relevantes dentro de los comentarios analizados, proporcionando una visión más clara de las discusiones predominantes en las plataformas sociales.

3.4. Modelo de la Metodología

4. Resultados

En esta sección se presentará una descripción de los resultados obtenidos en cada una de las etapas descritas en el diagrama del proceso de análisis. De esta manera, se mostrará cómo cada etapa contribuye a generar *insights* sobre las opiniones y temas predominantes expresados durante los debates presidenciales y el día de la elección en las plataformas de redes sociales analizadas.

4.1. Limpieza de Datos

Uno de los principales retos al trabajar con datos textuales provenientes de redes sociales es la calidad inconsistente del contenido. En este proyecto, se identificaron varios problemas específicos que complican el análisis automatizado:

- **Mala escritura y faltas de ortografía:** Los comentarios y publicaciones contienen errores ortográficos frecuentes, como palabras mal escritas, omisión de acentos y uso incorrecto de mayúsculas o minúsculas. Estos errores dificultan la tokenización y lematización, al aumentar la variabilidad de las palabras.
- **Jerga mexicana y expresiones coloquiales:** Muchos usuarios emplean mexicanismos y modismos propios del contexto cultural, como “chido”, “gacho” o “fifí”. Estas expresiones no están presentes en la mayoría de los diccionarios estándar, lo que limita el desempeño de los modelos preentrenados en el análisis semántico.
- **Uso de groserías y lenguaje ofensivo:** Palabras altisonantes como “pendejo” o “culero” son comunes en las discusiones políticas en redes sociales. Estas palabras no solo tienen connotaciones emocionales fuertes, sino que también presentan múltiples variantes (e.g., *pendejo*, *culero*), lo que incrementa la complejidad del análisis de sentimientos.
- **Abreviaturas y términos políticos:** Los usuarios utilizan abreviaturas y siglas para referirse a candidatos o partidos, como “AMLO”, “4T” o “PRIAN”. Además, emplean términos únicos del discurso político en México, como “chairos” y “fifís”, que requieren interpretación contextual para su correcta clasificación.

Estos problemas no solo dificultan el procesamiento y análisis de los textos, sino que también pueden introducir sesgos en los resultados al interpretar palabras con múltiples variantes o significados según el contexto.

Para abordar estos desafíos, se implementó un *sistema evolutivo de reescritura* (Galindo, 1991), diseñado para realizar una limpieza parcial y estandarización de los textos. Este sistema incluye las siguientes funciones:

- Corrección automática de errores ortográficos y adición de acentos en palabras comunes.
- Reducción de variantes de mexicanismos y groserías a una forma estándar.
- Reescritura de abreviaturas y siglas, asociándolas con su significado completo en el contexto electoral.
- Sustitución de términos coloquiales y ofensivos por equivalentes neutros o su forma semánticamente más representativa.

Los sistemas evolutivos de reescritura se basan en la aplicación iterativa de reglas de transformación de datos para resolver problemas representados como cadenas de caracteres o bits. Un sistema de este tipo se estructura típicamente de la siguiente manera:

- Un conjunto de elementos de entrada (A).
- Un conjunto de elementos de salida (B).
- Un conjunto de reglas de reescritura (R), definidas como $X \rightarrow Y$, donde $X \in A^*$ y $Y \in B^*$. Esto significa que X puede reescribirse como Y .

Cada regla de reescritura especifica cómo transformar un patrón de entrada en un patrón de salida, lo que permite representar una gran variedad de problemas. Estos sistemas son evolutivos porque permiten que las reglas se actualicen dinámicamente durante el funcionamiento del sistema. Si se encuentra un caso nuevo que no está cubierto por las reglas existentes, el sistema puede solicitar información al usuario o a un experto para generar una nueva regla que se almacena automáticamente (Galindo, 1991).

Funcionamiento básico:

1. El sistema compara la entrada con las reglas almacenadas.
2. Si encuentra una coincidencia, aplica la transformación especificada en la regla.
3. Si no encuentra una coincidencia, solicita una nueva regla, que luego se añade al sistema.
4. Este proceso permite que el sistema *aprenda* y mejore con el tiempo, adaptándose a nuevos escenarios y ampliando su base de conocimiento.

Como primer paso para implementar el sistema evolutivo de reescritura, construimos una bolsa inicial de 136 palabras que incluyen términos coloquiales, términos de política mexicana, groserías y abreviaturas comunes en redes sociales. Este conjunto se elaboró a partir de una revisión exhaustiva de los comentarios y publicaciones del conjunto de datos. Algunos ejemplos de palabras incluidas son:

Palabra	Nivel de Intensidad	Sentimiento Asociado	Categoría	Comentarios
chairo	3	negativo	insulto	término despectivo contra votantes de izquierda
fifi	3	negativo	insulto	término usado contra votantes de derecha
asqueroso	4	negativo	adjetivo	término coloquial usado para describir desagrado
mamón	3	negativo	insulto	usado para descalificar

Cuadro 5: Ejemplo de palabras incluidas en la bolsa inicial.

El propósito de esta bolsa de palabras es servir como base para el sistema evolutivo de reescritura. Este sistema aplica las siguientes acciones sobre los textos:

- **Estandarización:** Palabras con múltiples variantes (e.g., "fifi", "fifi") se reducen a una forma única para minimizar ruido semántico.
- **Limpieza:** Groserías, abreviaturas y errores ortográficos son corregidos, con el fin de mejorar la legibilidad y procesabilidad del texto.

Al utilizar esta bolsa de palabras como referencia inicial, el sistema evolutivo de reescritura es capaz de minimizar el ruido en los textos procesados. Esto facilita que el modelo de lenguaje interprete y analice los datos de manera más efectiva, reduciendo la ambigüedad y los sesgos que pueden surgir debido al uso de jerga o expresiones coloquiales.

El sistema evolutivo no solo se limita a esta bolsa inicial; a través de su diseño, puede adaptarse dinámicamente, incorporando nuevas palabras y reglas según se identifiquen durante el análisis. De esta manera, se garantiza que el

texto procesado sea progresivamente más limpio y útil para las etapas posteriores de minería de textos y modelado (Galindo, 1991).

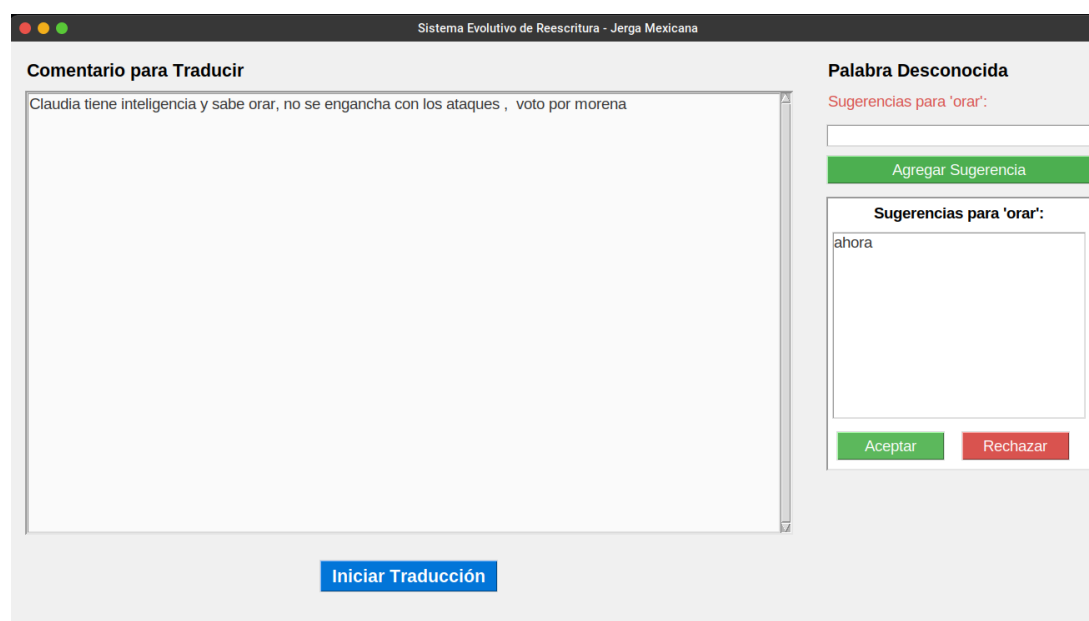


Figura 2: Sistema Evolutivo de Reescritura

Como resultado de la implementación del sistema evolutivo de reescritura, se logró una reducción del ruido en los textos procesados. Este sistema permitió estandarizar palabras con múltiples variantes, corregir errores ortográficos, reemplazar términos coloquiales, y normalizar abreviaturas utilizadas en el contexto político y social. Tras aplicar este proceso, se generaron versiones actualizadas de los dos conjuntos de datos originales, incorporando una nueva columna denominada **comentario_editado**. Esta columna contiene los textos procesados y limpios, listos para ser utilizados en las siguientes etapas de análisis.

Ahora, para continuar con el proceso de minería de textos se aplicó la metodología de preprocesamiento de datos descrita por Gelbukh (Gelbukh, 2003), que incluye las siguientes etapas fundamentales:

- **Conversión a minúsculas:** Todos los textos fueron convertidos a letras minúsculas para evitar que diferencias de mayúsculas y minúsculas afecten la agrupación de palabras similares.
- **Eliminación de emojis y caracteres no textuales:** Se eliminaron los emojis, caracteres especiales y elementos visuales, ya que no aportan información significativa al análisis lingüístico.
- **Eliminación de URLs:** Las direcciones web fueron removidas, dado que no representan contenido relevante para el análisis semántico.
- **Eliminación de signos de puntuación:** Se eliminaron los signos de puntuación para simplificar la tokenización y garantizar que las palabras sean procesadas sin interferencias.
- **Tokenización:** Este proceso consiste en dividir el texto en sus componentes básicos, llamados tokens, que usualmente corresponden a palabras individuales.
- **Eliminación de stopwords:** Las palabras vacías, como artículos, preposiciones y conjunciones (por ejemplo, y, de, el), que no aportan significado semántico, fueron eliminadas.
- **Lematización:** Cada palabra se redujo a su forma base o lema, eliminando variaciones morfológicas. Por ejemplo, palabras como *comiendo* y *comerán* fueron convertidas a su forma base *comer*.

5. Conclusiones

Referencias

- Castiblanco, M., & González, R. (2022). Opinion Mining Applied to Public Opinion Analysis in the Electoral Context. *Revista de Opinión Pública*, 37(1), 45-62. <https://doi.org/10.1016/j.rope.2022.01.012>
- Castro, L., & Martínez, A. (2021). Elecciones presidenciales en el Perú: minería de textos de los editoriales del diario La República. *Revista de Ciencias Sociales*, 35(2), 101-120. <https://doi.org/10.1016/j.rscs.2021.05.013>

- Fernández, R., & García, I. (2020). Información política en plataformas de redes sociales y participación electoral: evidencia desde Chile utilizando Full Matching. *Estudios de Opinión y Política*, 25(4), 350-367. <https://doi.org/10.1016/j.eop.2020.08.010>
- Galindo, F. (1991). Sistemas Evolutivos: Nuevo Paradigma de la Informática. *Memorias XVII Conferencia Latinoamericana de Informática*.
- Gelbukh, A. (2003). Procesamiento del lenguaje natural: estado de la investigación. *Centro de Investigación en Computación, Instituto Politécnico Nacional, México*.
- Hernández, E., & Gómez, M. (2019). Impacto de las redes sociales en la percepción ciudadana sobre la compra del voto en México. *Revista Mexicana de Ciencia Política*, 45(2), 111-128. <https://doi.org/10.1016/j.rmcp.2019.03.009>
- Pérez, M., & González, L. (2020). POPmine: Tracking Political Opinion on the Web. *Journal of Political Science and Technology*, 15(3), 130-145. <https://doi.org/10.1016/j.jpst.2020.02.014>
- Ramírez, A., & Pérez, J. (2022). Redes sociales y participación política en las elecciones presidenciales de 2022 en Colombia. *Revista de Comunicación Política*, 18(1), 55-72. <https://doi.org/10.1016/j.rcp.2022.03.005>
- Zagal, R., & Mata, M. (2020). Exploring Mexican Voting Intention Through Spatiotemporal Trends Using Social Media and Open Data Analysis. *Journal of Electoral Studies*, 28(3), 85-102. <https://doi.org/10.1016/j.jes.2020.04.011>