

Análisis de comentarios en TikTok mediante técnicas de PLN para detectar tendencias electorales.

Analysis of Comments on TikTok Using NLP Techniques to Detect Electoral Trends.

Jorge Emiliano Mora Herrera¹, Luz Ahide Gallardo Felipe², Hadassah Alejandra Hernandez García¹, Adrian Corona Pacheco¹, Maria Isabel Rocha Vargas¹, José Adrian Rodriguez Gonzalez¹ y Angel Díaz Pacheco^{1*}

¹ División de Ingenierías, Campus Irapuato-Salamanca, Universidad de Guanajuato, México.

² División de Ciencias Naturales y Exactas, Universidad de Guanajuato, México.

*angel.diaz@ugto.mx

Resumen

En México, el análisis de tendencias electorales se basa principalmente en encuestas telefónicas dirigidas a ciudadanos con servicio de línea fija. A pesar de la validación comprobada de tales métodos, este enfoque de muestreo, además de ser económicamente y logísticamente pesado, no logra reflejar con precisión la realidad actual de la mayoría de los votantes, quienes son predominantemente usuarios activos de plataformas de redes sociales. Dada la amplia disponibilidad de vastos repositorios de datos de acceso gratuito y la aplicación de técnicas modernas de inteligencia artificial, surge la oportunidad de obtener percepciones significativas sobre las preferencias electorales que conducen a junio de 2024. Comparar estos hallazgos con los resultados finales del proceso electoral se erige como un medio para fortalecer la validez de los métodos alternativos empleados en dichos procesos.

Palabras clave: Elecciones, TikTok, inteligencia artificial, deep learning, NLP.

Introducción

En el año 2018, México experimentó uno de los momentos más significativos de su historia contemporánea al elegir un gobierno de orientación izquierdista, desencadenando una rápida transformación a través de una variedad de ámbitos, instituciones y legislaciones. A medida que nos aproximamos a las elecciones de 2024, surge una interrogante de importancia crucial: ¿Estará el pueblo mexicano satisfecho con su elección en 2024?

Históricamente, esta pregunta ha encontrado respuesta a través de encuestas telefónicas (Arts et al., 2008; Crown, 2010), un método que no logra adaptarse plenamente al panorama actual del país. En vista del creciente involucramiento digital en nuestra sociedad, resulta evidente que la información compartida libremente a través de plataformas como Facebook, Twitter (ahora X) y otras redes sociales puede ofrecer un medio más eficaz para evaluar la opinión pública en una nación con una población mayoritariamente joven.

Aunque las redes sociales han sido ampliamente utilizadas para medir diversos constructos en varias disciplinas (Deng et al., 2019; Sirasapalli & Malla, 2023), su aplicación para evaluar la satisfacción electoral en elecciones presidenciales ha sido relativamente poco explorada. Es relevante destacar que la investigación en ciencias sociales en México sigue confiando en gran medida en encuestas in situ como método principal. Por consiguiente, la adopción de una metodología alternativa que capitalice las vastas colecciones de datos generadas diariamente tiene un potencial significativo para proporcionar hallazgos de gran utilidad e interés.

El propósito de este estudio es descubrir el sentimiento colectivo del pueblo mexicano respecto al proceso electoral de 2024 mediante la adopción de diversas técnicas de inteligencia artificial y el análisis exhaustivo de datos provenientes de las redes sociales. Este análisis permitirá identificar grupos de opiniones que reflejen las diversas voces de la ciudadanía. Al sumergirse en el discurso digital, se podrá obtener una



comprensión integral de las percepciones predominantes, preocupaciones y aspiraciones del público, incluso en regiones con acceso limitado a Internet. Este enfoque capacita a los responsables de políticas públicas, investigadores y partes interesadas para evaluar de manera exhaustiva el pulso de la nación, fomentando la toma de decisiones informadas y promoviendo un discurso democrático más inclusivo y receptivo.

Este documento está organizado de la siguiente forma: en la sección 2 se presenta una revisión de los trabajos relacionados al tema bajo investigación, la sección 3 presenta la metodología seguida para la recopilación y etiquetado de datos, la sección 4 presenta los análisis preliminares en los conjuntos de datos. La sección 5 presenta la Discusión y la sección 6 presenta las conclusiones y el trabajo futuro.

Trabajo relacionado

Tradicionalmente, las encuestas han sido la piedra angular de los análisis de tendencias en los procesos electorales a nivel mundial (Pasek, 2015). Si bien su eficacia para medir una amplia gama de constructos en las ciencias sociales no se cuestiona, resulta innegable el alto costo, tanto logístico como económico, asociado con estos procedimientos.

En un esfuerzo por mitigar la complejidad logística inherente al proceso y alcanzar muestras más representativas, se ha sugerido el empleo de encuestas telefónicas. No obstante, a pesar de las mejoras en el procedimiento, estas encuestas han experimentado una disminución en los niveles de respuesta (Keeter et al., 2017), y su fiabilidad se ve comprometida por su propensión al sesgo, dado que la disponibilidad de servicio fijo varía entre los ciudadanos (Jorge y Miro, 2021).

Es de suma relevancia destacar la contribución del estudio realizado por Brito y sus colegas (2021), en el cual llevaron a cabo una exhaustiva revisión sistemática que abarcó 83 investigaciones centradas en el empleo del análisis de redes sociales para el análisis de tendencias y la prospectiva en diversos procesos electorales en los Estados Unidos. A pesar de identificar resultados poco confiables en los métodos alternativos basados en procesamiento del lenguaje natural (NLP), también destacaron la distinción de ciertas plataformas de redes sociales en relación con otras para estos procesos. Lamentablemente, este estudio aún no aborda el uso de TikTok; no obstante, según datos del portal [statista.com](https://es.statista.com), el índice de participación en México asciende aproximadamente al 76%¹.

Considerando la conveniencia logística y la pertinencia de los métodos fundados en el Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) para la cuantificación de constructos en plataformas de redes sociales, tales como la percepción del destino en el ámbito turístico (Díaz-Pacheco et al., 2024; Guerrero-Rodríguez et al., 2024) o la satisfacción de los turistas (Díaz-Pacheco et al., 2023), se puede inferir que el análisis de las tendencias políticas en comentarios difundidos a través de Tik Tok puede ser abordado con eficacia mediante dichas técnicas, las cuales han demostrado un notable desempeño en contextos de análisis análogos. Abajo se presentan algunos de los trabajos que más influyeron en el presente estudio.

Dharani et al., (2023) exploraron el potencial de los datos de Twitter para monitorear el sentimiento público y predecir las preferencias de voto. Utilizando técnicas de procesamiento del lenguaje natural, categorizaron tweets sobre programas gubernamentales y emplearon aprendizaje supervisado para predecir resultados electorales. Métodos no supervisados identificaron tendencias y patrones. A pesar de desafíos como el sarcasmo y la jerga, sus hallazgos indican la viabilidad de utilizar datos de Twitter para análisis político. Abogan por un preprocesamiento y análisis exhaustivos de datos para superar estos obstáculos. El estudio sugiere que las campañas políticas podrían aprovechar Twitter para evaluar la opinión pública y atraer votantes, resaltando la importancia de las redes sociales en el discurso político moderno.

Bhadauria et al., (2024) contribuyeron al campo del análisis de sentimientos dentro del ámbito electoral mediante un examen exhaustivo de la predicción de tendencias de sentimiento previas a las elecciones. Su estudio utiliza metodologías de aprendizaje automático de vanguardia, con un énfasis específico en evaluar la efectividad de un modelo de red neuronal profunda en contraste con modelos convencionales como Naive Bayes y Regresión Logística. Aprovechando un conjunto de datos enriquecido con índices de corrupción y

¹ <https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/>

puntajes de popularidad como características, su modelo de red neuronal profunda muestra un rendimiento prometedor, alcanzando una impresionante tasa de precisión del 79.0% en la predicción de sentimientos.

En consonancia con lo anteriormente expuesto, la presente iniciativa postula la implementación de una amalgama de técnicas de Procesamiento del Lenguaje Natural con el propósito de monitorear las tendencias electorales, así como los intereses primordiales y las inquietudes del electorado en el marco del proceso electoral presidencial que habrá de tener lugar en junio de 2024.

Recolección de datos

Con el propósito de adquirir las observaciones pertinentes para nuestro estudio, hemos empleado la herramienta denominada *tiktok-comment-scraper*². Mediante su utilización, se ha conseguido la recopilación de **50,000** comentarios, los cuales fueron descargados de forma semanal, provenientes de los 10 videos que ostentan el mayor número de interacciones, derivados de la búsqueda efectuada con las palabras clave "*elecciones, 2024, México, presidenciales*".

Ante la imperativa exigencia de contar con un formato idóneo para el análisis de los datos, uno de los colaboradores del equipo efectuó una modificación sustancial al repositorio, orientada a la generación de archivos en formato CSV en lugar de JSON. Consecuentemente, se procedió a la ejecución de las labores concernientes al preprocesamiento y análisis de los datos.

Técnicas para el análisis de los datos

Preprocesamiento

Como parte de los análisis llevados a cabo, se realizó el pre procesamiento de los datos. Aunque la información se obtuvo de la forma más estructurada posible, es necesario realizar algunas acciones antes de poder utilizarla como entrada de los algoritmos de aprendizaje automático. Las etapas seguidas en el pre-procesamiento son:

- Convertir el texto de mayúsculas a minúsculas.
- Eliminación de palabras que no aportan información (artículos y preposiciones).
- Eliminación de signos de puntuación.
- Reemplazo de dígitos con el carácter "d".
- Reducción de la variabilidad léxica.
- Eliminación de las palabras que aparecen menos de 10 veces en el corpus.

Con lo anterior se limpió el conjunto de datos y se obtuvo un formato más adecuado para ser analizado con otros algoritmos. Como ejemplo para el lector, dicho pretratamiento permitió pasar de comentarios de la forma:

"se tuvo que retirar y cumplir con su trabajo de gobernador, y ahora el candidato es un conservador expriista"

A comentarios limpios y sin ruido como se muestra a continuación:

"tuvo retirar cumplir trabajo gobernador ahora candidato conservador expriista"

Modelado de tópicos

El modelado de tópicos se refiere a una técnica utilizada para la clasificación no supervisada de documentos, similar al agrupamiento en el análisis de datos numéricos, donde se discernirá grupos inherentes de

² <https://github.com/moraxh/tiktok-comment-scraper>

elementos, denominados tópicos, incluso en ausencia de criterios claros. Es importante destacar que los documentos pueden estar asociados con múltiples tópicos, asemejándose al concepto de agrupamiento difuso, donde cada dato pertenece a más de un clúster (Kherwa & Bansal, 2018). La importancia del modelado de tópicos radica en su capacidad para organizar, comprender, buscar y resumir automáticamente vastos repositorios electrónicos. A través de este método, se pueden descubrir temas latentes dentro del corpus, los documentos pueden ser categorizados según estos temas, y la clasificación resultante puede ser utilizada para estructurar, encapsular y buscar en los documentos. Por ejemplo, si un documento está categorizado bajo temas como comida, perros y salud, un usuario que busca "comida para perros" puede encontrarlo relevante debido a su cobertura de estos temas, evitando así la necesidad de revisar todo el documento para evaluar su relevancia. En consecuencia, al anotar documentos basados en los temas inferidos mediante el modelado, se puede optimizar la eficiencia del proceso de búsqueda.

Similitud de textos

La similitud de textos sirve como un concepto fundamental en el ámbito del procesamiento del lenguaje natural (PLN), sustentando diversas tareas cruciales como la recuperación de información, la respuesta automática a preguntas, la traducción automática, los sistemas de diálogo y el emparejamiento de documentos. En ella se abarca la medición de la semejanza entre segmentos de texto, delineada por las similitudes compartidas entre ellos. La evaluación de la similitud de textos engloba múltiples dimensiones, incluyendo la distancia de longitud, la distancia de distribución y la distancia semántica, cada una contribuyendo a una comprensión matizada del parecido textual. Además, los enfoques de representación de textos se diversifican en representaciones basadas en cadenas de caracteres, basadas en corpus, textos de un solo significado, textos de múltiples significados y representaciones basadas en estructuras de gráficos, cada uno ofreciendo percepciones únicas sobre las características textuales (Wang & Dong, 2020). En las aplicaciones contemporáneas de PLN, la similitud de textos adquiere una importancia primordial debido a su utilidad multifacética. Más allá de simplemente medir la semejanza semántica, se esfuerza por comprender las propiedades semánticas más amplias que subyacen a las palabras. Por ejemplo, mientras que las palabras 'Rey' y 'hombre' pueden exhibir una asociación cercana, no se consideran semánticamente similares; por el contrario, 'Rey' y 'Reina' comparten similitud semántica. Esta distinción subraya la necesidad de un análisis semántico en la evaluación precisa de la similitud de textos. Es importante destacar que las relaciones semánticas, incluida la similitud, se cuantifican en términos de distancia semántica, donde una distancia semántica más alta denota una relación más débil entre los términos.

En el marco de la presente investigación, se hizo uso de la similitud de textos con el propósito de generar segmentaciones en el corpus de datos y discernir los fragmentos pertinentes a cada candidato específico. Tal segmentación resulta imperativa para la realización efectiva de los análisis subsecuentes.

Análisis de Sentimientos

El análisis de sentimientos (SA) emerge como un aspecto fundamental en el ámbito de la clasificación de textos, especialmente en el contexto de la vasta cantidad de contenido generado en línea por autores aficionados sobre diversos temas. El SA sirve para extraer y amalgamar los sentimientos de los usuarios hacia una entidad objetivo, proporcionando información invaluable sobre la percepción pública. El aprovechamiento de técnicas de aprendizaje automático (ML) se vuelve imperativo en este esfuerzo, dada la abundancia de datos en lenguaje natural repleto de patrones discernibles. Las metodologías de ML muestran una notable capacidad para adaptarse a las sutilezas específicas del dominio, garantizando así resultados de alta precisión dependiendo del conjunto de características utilizado. Mientras que las técnicas basadas en léxico, que se basan en diccionarios externos, ofrecen robustez al mitigar el sobreajuste, pueden fallar al capturar las complejidades contextuales dentro de dominios especializados. Por el contrario, las técnicas estadísticas basadas en corpus requieren volúmenes de datos sustanciales para lograr la estabilización. Las metodologías basadas en redes complejas, por otro lado, surgen como altamente útiles, hábiles para preservar el orden, la proximidad, el contexto y las relaciones dentro del texto analizado. Conceptualmente, el análisis de sentimientos, sinónimo de minería de opiniones, profundiza en el ámbito de las afirmaciones subjetivas, aprovechando el procesamiento del lenguaje natural (NLP) para recopilar y examinar expresiones cargadas de sentimiento (Khan et al., 2016). El discernimiento de actitudes subjetivas dentro de vastos repositorios de datos sociales representa un punto focal dentro de los dominios de la minería de datos y el NLP, reflejando la creciente importancia otorgada a la extracción e interpretación del sentimiento en el discurso contemporáneo.

Para el presente proyecto, el análisis de sentimientos es de gran utilidad para poder determinar la proporción de comentarios que pueden apoyar o refutar a los candidatos de la contienda electoral.

Experimentos y resultados

En el preámbulo de nuestros experimentos, llevamos a cabo una selección meticulosa de una muestra representativa de los comentarios recopilados, seguida por una clasificación manual exhaustiva de los mismos. Este proceso implica la evaluación por parte de expertos humanos, quienes, basándose en su profundo conocimiento del tema, su comprensión semántica y su juicio informado, categorizan cada comentario en diversas polaridades en función de las emociones reflejadas en el discurso. En el marco de este estudio, la tarea de clasificación recayó en una cohorte de estudiantes de la destacada carrera de Ingeniería en Datos e Inteligencia Artificial, específicamente dentro del curso de Probabilidad. Las clases de interés se definieron con precisión: los comentarios se consideraron positivos si manifestaban una actitud favorable hacia el proceso o algún candidato en particular, mientras que se catalogaron como negativos aquellos que reflejaban una postura desfavorable en relación con el proceso o los candidatos en cuestión. Tras un riguroso escrutinio, la muestra inicial de 1600 comentarios se vio reducida a aproximadamente 1300 comentarios viables, listos para el análisis subsiguiente.

De entre los comentarios obtenidos, se categorizaron 582 como positivos y 718 como negativos, lo que sugiere una probable disparidad en el conjunto de datos. Con el fin de abordar esta disparidad y, al mismo tiempo, preservar la información esencial, se recurrió a técnicas de reducción de dimensionalidad. En particular, se aplicó el método de información mutua, según lo propuesto por Xu et al. (2007), para identificar el vocabulario con mayor capacidad discriminativa en cada clase. En la Tabla 1 se presentan las palabras de mayor relevancia dentro de cada categoría.

Tabla 1 Palabras más destacadas en base a estadísticas ordenadas por polaridad

Polaridad	Palabra
Positivo	felicidades, aplauso, continuidad, conciencias, gob, noroña, experiencia, haciendo, vamos, claroooo, continuara, apoya, excelente, proyectos, adelante, conocemos, arriba, marcha, cien, bravo, mx, oaxaca, claudia presidenta, obviamente, contigo, dudarlo, cuarta.
Negativo	quitar, ridícula, viejo, hermana, metro, cdmx, cara, crees, misma, botarga, discurso, robar, idiota, precampaña, junto, saben, federal, escribir, pegar, gana, cierre, gobierna, cuenta, perfil, seria, hará, nacional, dinero, vez, pasa, inseguridad, robot, basura, vende, falto, jamás, perder, bots, según, corrupta.

Aunque los conjuntos de palabras resultantes proporcionan información esclarecedora, su interpretación puede resultar desafiante para aquellos lectores no familiarizados con el tema. Con el propósito de abordar esta dificultad y facilitar la comprensión de los hallazgos, recurrimos a la herramienta Chat GPT para generar frases utilizando las palabras obtenidas en la fase previa. En la Tabla 2, se presentan las frases elaboradas mediante Chat GPT, proporcionando así una explicación más accesible y clara de los resultados.

Tabla 2 Interpretación Chat Generative Pre-Trained Transformer(Chat GPT) por polaridad

Polaridad	Interpretación Chat GPT
Positivo	¡Bravo! ¡Felicidades por esa continuidad! Con tu experiencia y el apoyo de todos, vamos ¡arriba! Marcha con cien aplausos, México y Oaxaca saben que ClaudiaPresidenta es la opción. Obviamente, contigo todo es claro, ¡haciendo historia! ¡Excelente! ¡Continuará el impulso de proyectos adelante, porque conocemos tus conciencias.
Negativo	La hermana del viejo que gobierna la CDMX se puso una botarga ridícula en el metro durante la precampaña, y en su discurso idiota, creyó que pegar carteles y robar dinero federal la haría ganar. Pero no cuenta con que la inseguridad jamás se va a quitar. Al cierre de su campaña, junto a sus bots y el robot basura, pasa lo mismo de siempre: un perfil corrupto que vende promesas falsas y serias, según los que nunca saben. Falta de seriedad, parece que esta vez va a perder.

A partir de las Tablas 1 y 2, se observa que en el apartado positivo, un conjunto de palabras hace referencia clara a la candidata de la coalición Morena-PT-Verde, utilizando términos como: "continuidad", "norroña" (Noroña), "cuarta", "claudia presidenta". Es importante destacar que a los alumnos que contribuyeron al etiquetado de los datos se les instruyó mantener la imparcialidad y evaluar el sentimiento general de los comentarios, sin importar el candidato mencionado. Aunque estos experimentos proporcionan información sobre el vocabulario utilizado para caracterizar los comentarios positivos y negativos, la muestra utilizada es demasiado pequeña para realizar generalizaciones en esta etapa.

Como se indicó previamente, se recopilaron 50,000 comentarios durante el periodo de enero a mayo de 2024. Estos comentarios fueron preprocesados para eliminar elementos irrelevantes y ajustarlos adecuadamente para los algoritmos de aprendizaje automático. En la Tabla 3 se muestran los comentarios recolectados cada mes y la cantidad de comentarios útiles tras el proceso de limpieza.

Tabla 3 Comentarios recolectados antes y después de limpiarlos

Mes	Comentarios sin procesar	Comentarios limpios
Enero	8,907	6,540
Febrero	15,977	11,946
Marzo	15,231	11,201
Abril	12,879	12,399
Mayo	19,102	17,217

Subsecuentemente, utilizando métodos para medir la similitud entre textos, dos integrantes del equipo (diferentes de los analistas) realizaron particiones en la base de datos para identificar los comentarios más similares a las consultas realizadas. De esta manera, fue posible clasificar el corpus en tres categorías: comentarios de partidarios de Xóchitl Gálvez, comentarios de partidarios de Claudia Sheinbaum y comentarios de partidarios de Jorge Maynez. Aunque estas particiones no son perfectas y pueden incluir comentarios de partidarios, opositores y personas desinteresadas, los algoritmos empleados pueden detectar patrones subyacentes en la información. Esto permitirá al analista determinar la pertinencia de los hallazgos.

Fig 1. Nube de palabras obtenida de los comentarios identificados como de partidarios de Claudia Sheinbaum.



Fig 3. Nube de palabras obtenida de los comentarios identificados como de partidarios de Jorge Maynez.



Este primer experimento nos ha permitido confirmar que las particiones realizadas son adecuadas, ya que el nombre de cada candidato es el elemento predominante en cada figura, junto con términos relacionados, como la conexión Maynez-MC y Claudia-MORENA. En el caso de la candidata Xóchitl Gálvez, al ser candidata de una coalición, esta relación es más compleja; sin embargo, aparece la conexión Xóchitl-PRI.

Con el objetivo de comprender los intereses y preocupaciones de los votantes, se realizó un análisis de tópicos en la partición correspondiente a cada candidato. Debido a la gran cantidad de temas identificados, la Tabla 4 presenta solo los tópicos más relevantes obtenidos del análisis del corpus utilizando ChatGPT.

T

Tabla 4 Tópicos de los candidatos por cada trimestre.

1er trimestre	2do trimestre
Claudia Sheinbaum Pardo - Desconfianza hacia las instituciones gubernamentales. - Seguridad: Se menciona la inseguridad en el país y se hace referencia a casos específicos, como la caída del metro y la violencia. - Propuestas y promesas de campaña: Se mencionan diversas propuestas y promesas de campaña de los candidatos. - Opiniones sobre la gestión de la pandemia de COVID-19. - Descontento y decepción con el gobierno actual.	- Crisis económica. - Estrategias de campaña y percepción pública: análisis sobre las estrategias de campaña, la percepción pública de los candidatos y la influencia de los medios de comunicación. - Propuestas relacionadas con la educación, salud y seguridad. - Cuestionamientos sobre la veracidad de la información y los discursos de los candidatos. - Corrupción y narcotráfico.
Xóchitl Gálvez - Debate sobre la privatización y extranjerización de recursos nacionales (Pemex). - Opiniones sobre la participación de la ciudadanía en las elecciones. - Análisis sobre la historia y comportamiento de los partidos políticos tradicionales. - Crítica a la falta de preparación y experiencia de algunos candidatos. - Críticas y burlas hacia Xóchitl Gálvez.	- Transparencia gubernamental. - Reducción de programas sociales. - Partidos políticos (PRI, PAN, PRD). - Debate político - Seguridad pública y narcotráfico.
Jorge Álvarez Máynez - Campañas políticas y estrategias de marketing. - Popularidad y percepción de Jorge Álvarez Máynez. - Experiencia política y preparación de los candidatos. - Jóvenes como votantes y su influencia en las elecciones. - Análisis de propuestas y viabilidad de las mismas.	- Influencia de las redes sociales en la opinión pública y el voto. - Incertidumbre sobre la efectividad de las propuestas y promesas de campaña. - Análisis de encuestas y resultados. - Comparación entre candidatos (ponentes en debates, propuestas, preparación académica, etc.) - Desarrollo de infraestructura educativa y social.

De acuerdo a los tópicos encontrados en el análisis, podemos darnos cuenta que a los votantes de los diferentes candidatos les preocupa la inseguridad, la corrupción, la desconfianza hacia las instituciones gubernamentales, la gestión de la pandemia de COVID-19, la crisis económica y la transparencia gubernamental. Además, les interesa las propuestas relacionadas con la educación, la salud, la seguridad, evitar la reducción de programas sociales, la privatización de recursos nacionales como Pemex, la participación ciudadana en las elecciones y el desarrollo de infraestructura educativa y social.

Finalmente, se realizó un análisis de sentimientos en las particiones correspondientes a cada candidato para determinar los porcentajes proporcionales de su base votante. Es importante aclarar que, debido a la dificultad de identificar con precisión los comentarios de los votantes declarados de un candidato específico, es posible que algunos comentarios aparezcan en los conjuntos de datos de más de un candidato. Por esta razón, aunque los porcentajes obtenidos no son exactos, proporcionan una estimación valiosa sobre las posibilidades de éxito de cada candidato. Además, es necesario tener en cuenta que muchos de los comentarios recolectados consisten únicamente en cadenas de emojis y signos de admiración, lo que afecta la cantidad de comentarios útiles para el análisis. En las figuras 4 a 6 se muestran los gráficos de pastel de los porcentajes de comentarios a favor y en contra de cada candidato.

Figura 4. Graficas de pastel de comentarios positivos y negativos del 1er y 2do trimestre Claudia Sheinbaum. Se utilizaron un total de 4930 comentarios para el 1er trimestre y 3557 para el segundo.

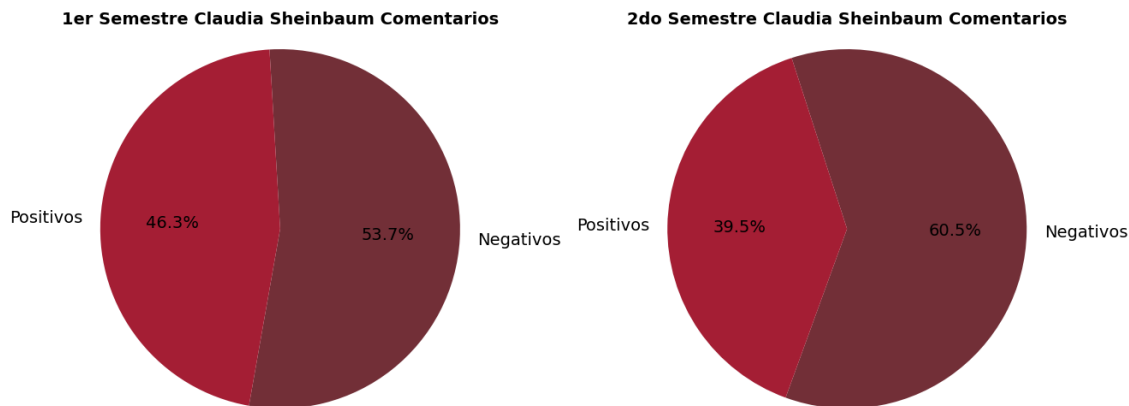


Figura 5. Graficas de pastel de comentarios positivos y negativos del 1er y 2do trimestre Xochitl Gálvez. Se utilizaron un total de 2746 comentarios para el 1er trimestre y 1879 para el segundo.

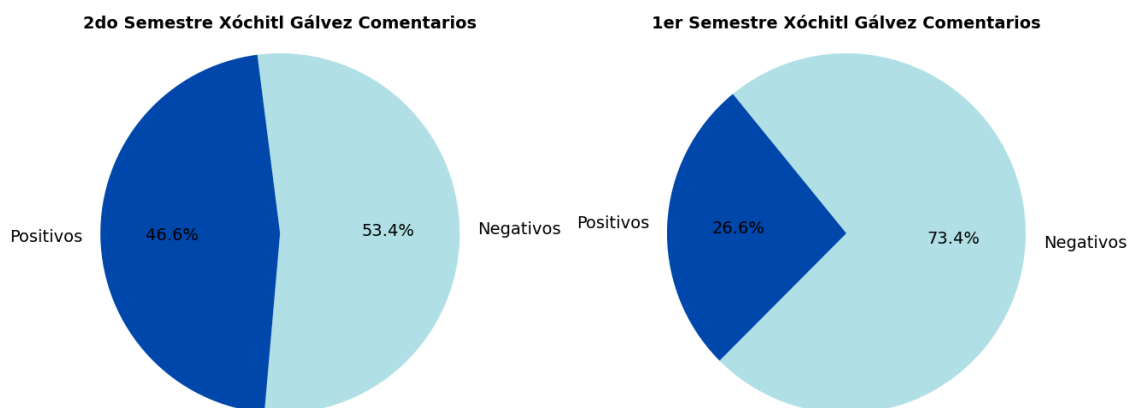
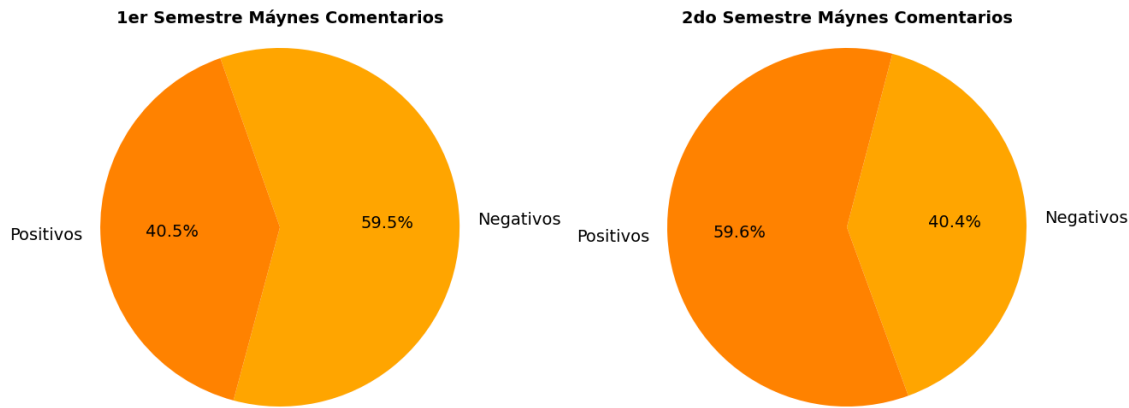


Figura 6. Graficas de pastel de comentarios positivos y negativos del 1er y 2do trimestre Máynez. Se utilizaron un total de 1022 comentarios para el 1er trimestre y 2097 para el segundo.



A partir de los gráficos anteriores, se observa que todos los candidatos presentan una mayor proporción de comentarios negativos en comparación con los positivos. Es crucial tener en cuenta que no todos los comentarios son utilizables debido a las razones previamente mencionadas. Esto resulta en una variabilidad en la cantidad de comentarios disponibles, lo que impide una comparación directa entre los gráficos de pastel. Para facilitar una mejor comprensión, la Figura 7 muestra un gráfico de tendencias que ilustra la cantidad de comentarios positivos recibidos por los candidatos y su evolución a lo largo de los periodos evaluados. Se observa un marcado descenso en las opiniones que respaldan los proyectos políticos de ambas candidatas. En contraste, el candidato Jorge Maynez muestra un comportamiento creciente durante el segundo trimestre. Al comparar estos resultados con los finales (ver Fig. 8), nuestro método demostró ser capaz de identificar las tendencias predominantes en el proceso, así como los pensamientos, preocupaciones e intereses de los votantes respecto a sus candidatos. Sin embargo, existe una desviación significativa en el caso de Jorge Maynez, quien terminó en tercer lugar según los resultados finales de la contienda electoral.

Figura 7. Gráfica de tendencia de los comentarios positivos por cada candidato para ambos trimestres.

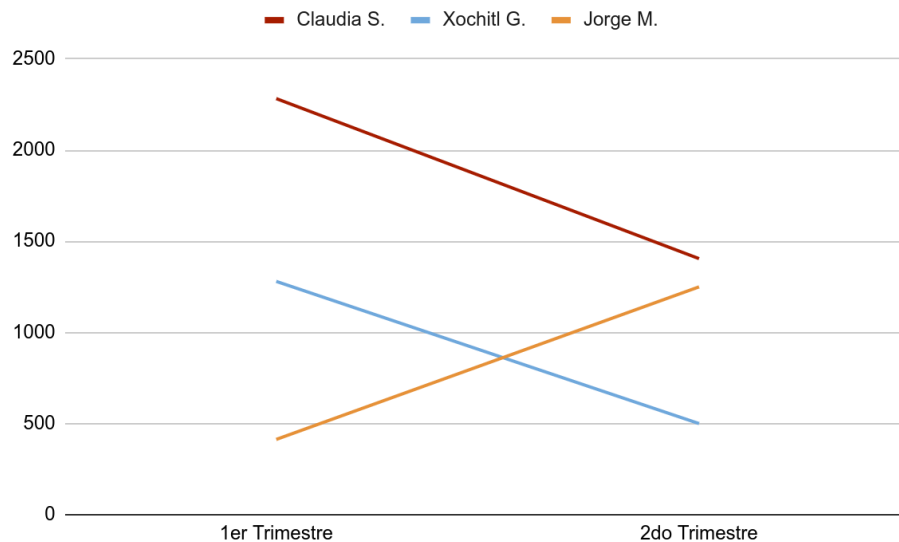
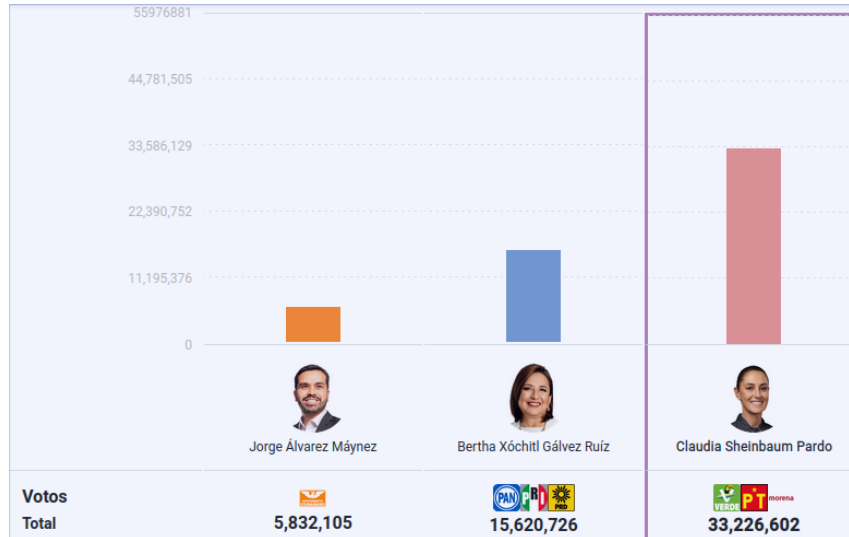


Figura 8. Resultados finales para la contienda electoral de 2024. Fuente: <https://prep2024.ine.mx/publicacion/nacional/presidencia/nacional/candidatura>



Discusión

El análisis llevado a cabo en este estudio ha demostrado la eficacia del uso de técnicas de procesamiento del lenguaje natural (NLP) para identificar tendencias y preferencias electorales en el contexto de las elecciones presidenciales de 2024 en México. Los métodos propuestos permitieron acertar con precisión la victoria de Claudia Sheinbaum Pardo, destacando la capacidad de nuestras herramientas para captar el sentimiento predominante en las redes sociales.

Sin embargo, el método falló al predecir los resultados del segundo y tercer lugar del proceso electoral, lo que sugiere áreas de mejora en nuestra aproximación. Una razón importante detrás de este fallo es el sesgo implícito en el medio elegido (TikTok), que está orientado a un público joven y, por lo tanto, carece de representatividad en otros sectores de la población. Además, el léxico utilizado en la plataforma es limitado y saturado de emojis, lo que reduce la cantidad de información utilizable. A pesar de estos desafíos, la identificación precisa del ganador demuestra que nuestras técnicas son prometedoras. No obstante, es crucial abordar estos problemas para mejorar la precisión de las predicciones en todos los aspectos del proceso electoral.

Los métodos basados en NLP no solo permitieron identificar al ganador, sino que también revelaron el vocabulario y los temas de interés de los votantes. Este análisis detallado ofreció una visión profunda de las preocupaciones y prioridades de la ciudadanía, proporcionando un mapa detallado del respaldo porcentual a cada candidato. Estos hallazgos son de gran valor, ya que permiten una comprensión más matizada de las dinámicas electorales y los factores que influyen en la decisión del voto.

Un aspecto clave a destacar es la capacidad del método propuesto para adaptarse al entorno digital contemporáneo, aprovechando la vasta cantidad de datos generados en redes sociales. Este enfoque presenta una alternativa viable y moderna a las encuestas telefónicas tradicionales, que aunque validadas, no reflejan adecuadamente la realidad actual debido a su sesgo hacia los usuarios de líneas fijas y su alta demanda logística y económica.

En resumen, los hallazgos de este estudio refuerzan la viabilidad del uso de técnicas de inteligencia artificial y análisis de redes sociales para evaluar las preferencias electorales. A medida que continuamos mejorando nuestros métodos, se espera que estos enfoques proporcionen una herramienta cada vez más precisa y útil para capturar el pulso de la nación, contribuyendo a un proceso electoral más transparente y representativo.

Conclusiones y trabajo futuro

Este estudio ha demostrado la efectividad de las técnicas de procesamiento del lenguaje natural (NLP) y el análisis de datos provenientes de redes sociales para evaluar las preferencias electorales en el contexto de las elecciones presidenciales de 2024 en México. A través de estas metodologías, fue posible identificar correctamente a la ganadora de la contienda, Claudia Sheinbaum Pardo, lo que subraya la capacidad de nuestras herramientas para captar el sentimiento predominante en plataformas digitales.

No obstante, se evidenciaron limitaciones en la precisión de las predicciones, particularmente en la identificación del segundo y tercer lugar. Este fallo se atribuye en gran parte al sesgo implícito en el medio elegido, TikTok, que está orientado a un público joven y, por lo tanto, carece de representatividad en otros sectores de la población. Además, el léxico limitado y saturado de emojis en esta plataforma redujo la cantidad de información utilizable. Estos desafíos resaltan la necesidad de ajustar y mejorar nuestros modelos para abarcar una representación más amplia y precisa del electorado.

A pesar de estas limitaciones, los métodos basados en NLP ofrecieron una visión profunda de los temas de interés y las preocupaciones de los votantes, proporcionando un mapa detallado del respaldo porcentual a cada candidato. Este enfoque ha mostrado ser una alternativa viable y moderna a las encuestas telefónicas tradicionales, que aunque validadas, no reflejan adecuadamente la realidad actual debido a su sesgo hacia los usuarios de líneas fijas y su alta demanda logística y económica.

En conclusión, nuestros hallazgos refuerzan la viabilidad del uso de técnicas de inteligencia artificial y análisis de redes sociales para evaluar las preferencias electorales. A medida que continuamos mejorando nuestros métodos, se espera que estos enfoques proporcionen herramientas cada vez más precisas y útiles para capturar el pulso de la nación, contribuyendo a un proceso electoral más transparente y representativo. Esto permitirá a los responsables de políticas públicas, investigadores y partes interesadas tomar decisiones informadas y fomentar un discurso democrático más inclusivo y receptivo.

Como se mencionó anteriormente en el documento, esta investigación es aún trabajo en curso, de lo cual se desprenden las siguientes actividades a realizar en subsecuentes ediciones del verano científico.

- Ampliación y revisión de los conjuntos de datos de texto.

- Construcción de modelos de aprendizaje automático para la clasificación de temáticas y polaridad de texto en español de México.
- Pruebas y validación en casos de estudio subsecuentes.

Bibliografía/Referencias

- Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., ...Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part B), 10125–10144. doi: 10.1016/j.jksuci.2022.10.010
- Arts, K., & Thomassen, J. (2008). Satisfaction with democracy: Do institutions matter? In *Electoral Studies* (Vol. 27, Issue 1, pp. 5–18). Elsevier BV. <https://doi.org/10.1016/j.electstud.2007.11.005>
- Bhadauria, A. S., Shukla, M., Kumar, P., & Dwivedi, N. (2024). Forecasting Election Sentiments: Deep Learning vs. Traditional Models. 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECs), 1–6. <https://doi.org/10.1109/SCEECs61402.2024.10482071>
- Brito, K. D. S., Filho, R. L. C. S., & Adeodato, P. J. L. (2021). A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions. In *IEEE Transactions on Computational Social Systems* (Vol. 8, Issue 4, pp. 819–843). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/tcss.2021.3063660>
- Crow, D. (2010). The Party's Over: Citizen Conceptions of Democracy and Political Dissatisfaction in Mexico. In *Comparative Politics* (Vol. 43, Issue 1, pp. 41–61). Comparative Politics CUNY. <https://doi.org/10.5129/001041510x12911363510358>
- Dharani, D. G., Hemalatha, R., Pradeep, R., & Jagan, G. (2023). Political Opinion Mining for Popularity Prediction using Machine Learning Techniques. 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 1–5. <https://doi.org/10.1109/RMKMATE59243.2023.10369941>
- Deng, N., Liu, J., Dai, Y., & Li, H. (2019). Different cultures, different photos: A comparison of Shanghai's pictorial destination image between East and West. In *Tourism Management Perspectives* (Vol. 30, pp. 182–192). Elsevier BV. <https://doi.org/10.1016/j.tmp.2019.02.016>
- Díaz-Pacheco, Á., Guerrero-Rodríguez, R., Álvarez-Carmona, M. Á., Rodríguez-González, A. Y., & Aranda, R. (2024). Quantifying differences between UGC and DMO's image content on Instagram using deep learning. In *Information Technology & Tourism*. Springer Science and Business Media LLC. <https://doi.org/10.1007/s40558-023-00282-9>
- Díaz-Pacheco, A., Álvarez-Carmona, M. A., Rodríguez-González, A. Y., Carlos, H., & Aranda, R. (2023). Measuring the Difference Between Pictures From Controlled and Uncontrolled Sources to Promote a Destination. A Deep Learning Approach. In *International Journal of Interactive Multimedia and Artificial Intelligence: Vol. In Press* (Issue In Press, pp. 1–14). Universidad Internacional de La Rioja. <https://doi.org/10.9781/ijimai.2023.10.003>
- Guerrero-Rodríguez, R., Álvarez-Carmona, M. Á., Aranda, R., & Díaz-Pacheco, Á. (2024). Big data analytics of online news to explore destination image using a comprehensive deep-learning approach: a case from Mexico. *Inf. Technol. Tourism*, 26(1), 147–182. doi: 10.1007/s40558-023-00278-5
- Jorge, J. E., & Miró, E. M. (2021). La falla de las encuestas en las elecciones argentinas de 2019.
- Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). What low response rates mean for telephone surveys. *Pew Research Center*, 15(1), 1-39
- Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. In *Complex Adaptive Systems Modeling* (Vol. 4, Issue 1). Springer Science and Business Media LLC. <https://doi.org/10.1186/s40294-016-0016-9>
- Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. In *ICST Transactions on Scalable Information Systems* (Vol. 0, Issue 0, p. 159623). European Alliance for Innovation n.o. <https://doi.org/10.4108/eai.13-7-2018.159623>
- Pasek, J. (2015). Predicting Elections: Considering Tools to Pool the Polls. In *Public Opinion Quarterly* (Vol. 79, Issue 2, pp. 594–619). Oxford University Press (OUP). <https://doi.org/10.1093/poq/nfu060>
- Sirasapalli, J. J., & Malla, R. M. (2023). A deep learning approach to text-based personality prediction using multiple data sources mapping. In *Neural Computing and Applications*. Springer Science and Business Media LLC. <https://doi.org/10.1007/s00521-023-08846-w>
- Wang, J., & Dong, Y. (2020). Measurement of Text Similarity: A Survey. In *Information* (Vol. 11, Issue 9, p. 421). MDPI AG. <https://doi.org/10.3390/info11090421>
- Xu, Y., Jones, G. J., Li, J., Wang, B., & Sun, C. (2007). A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, 3(3), 1007-1012.