Transformación Digital
Agencia de Transformación Digital y Telecomunicaciones

INFOTEC

Biblioteca INFOTEC

Ciudad de México, a 24 de enero de 2025

VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Predicción de la aprobación presidencial usando aprendizaje computacional y datos de X (antes Twitter)"

Desarrollado por el alumno: **Pablo Xavier Noriega Robles**, bajo la asesoría del **Dr. Mario Graff Guerrero**, cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que el alumno no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Dr. Juan Antonio Vega Garfias
Subgerente de Innovación Gubernamental

JAVG/jah

C.c.p. Mtra. Analy Mendoza Rosales. – Encargada de la Gerencia de Capital Humano. - Para su conocimiento.
Pablo Xavier Noriega Robles. – Alumno de la Maestría en Ciencia de Datos e Información. – Para su conocimiento.

2025 Año de La Mujer Indígena

MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

"PREDICCIÓN DE LA APROBACIÓN PRESIDENCIAL USANDO APRENDIZAJE COMPUTACIONAL Y DATOS DE X (ANTES TWITTER)"

TESIS
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

PABLO XAVIER NORIEGA ROBLES

Asesor:

DR. MARIO GRAFF GUERRERO

Ciudad de México, enero, 2025

INFOTEC POSGRADOS

Agradecimientos

En primer lugar, quiero expresar mi más profundo agradecimiento a mis padres, cuyo esfuerzo incansable me permitió recibir una educación de calidad y alcanzar este importante logro en mi vida. Su ejemplo y apoyo han sido pilares fundamentales en este camino.

A mis abuelos, por todas las historias que me contaron cuando era niño, las cuales no solo despertaron mi imaginación, sino que también me enseñaron a cuestionar, analizar y comprender el mundo que me rodea desde una edad muy temprana. Su sabiduría y experiencia marcaron profundamente mi desarrollo y forjaron en mí una visión crítica y empática hacia la vida.

A mi esposa, quien ha sido la piedra angular de mis días, le agradezco por su amor, paciencia y por estar siempre a mi lado, incluso en los momentos más desafiantes. Eres mi fortaleza y mi mayor apoyo. Tu presencia ha llenado de significado este recorrido, y no tengo palabras suficientes para expresar cuánto valoro tu compañía y dedicación.

A mi hija, mi luz y mi fuerza, le agradezco por ser la fuente de inspiración que me motiva cada día a ser una mejor persona y a luchar por el mejor futuro posible, para ella y mi familia.

A mis suegros, quiero expresar mi gratitud por su apoyo incondicional y constante hacia nuestra familia día con día. Su generosidad y respaldo han sido fundamentales en este camino, además de agradecerles todas las increíbles experiencias que hemos vivido juntos en estos años de conocernos.

A mis amigos y amigas, gracias por compartir experiencias y momentos que han dejado una huella imborrable en mi vida. Las vivencias junto a ustedes han moldeado mi carácter y mi visión del mundo, aportando alegría y aprendizajes invaluables.

A mi asesor Mario, quiero agradecerle por brindarme las herramientas, el conocimiento y la guía para llevar a cabo este trabajo. Su experiencia y acompañamiento han sido claves para superar los retos de esta etapa académica.

A INFOTEC, le agradezco por la oportunidad de haberme aceptado en la maestría, abriéndome las puertas para crecer académica y profesionalmente.

Finalmente, agradezco a la vida por guiarme hacia el camino correcto en cada decisión importante que he tomado, incluso cuando el destino final aún sea incierto. Cada paso dado ha sido una oportunidad para crecer y aprender, y por ello me siento profundamente agradecido.

Índice general

Índice de figuras

Índice de tablas

Abreviaturas y acrónimos

ML	Aprendizaje automático (del inglés "Machine Learning").
AI	Inteligencia artificial (del inglés "Artificial Intelligence").
PLN	Procesamiento del Lenguaje Natural
SVR	Support Vector Regressor
XGBOOST	(del inglés “Extreme Gradient Boosting”)

Glosario

“A”

Algoritmo: Un conjunto de instrucciones o reglas lógicas diseñadas para resolver un problema o realizar una tarea específica.

“B”

Big Data: Conjunto de datos tan grandes y complejos que resulta difícil procesarlos con herramientas convencionales de análisis de datos.

“C”

Conjunto de entrenamiento: Subconjunto de datos utilizado para entrenar un modelo de aprendizaje automático.

“M”

Machine learning: En español “aprendizaje de máquina”, es un subcampo de las ciencias de la computación y una rama de la inteligencia artificial, que tiene como objetivo desarrollar técnicas que posibiliten que las computadoras aprendan (Murphy, 2012).

“N”

Normalización: Segmentación de las palabras, formateo de las palabras y segmentación de las oraciones en el texto.

“T”

Tokenización: Una instancia de un tipo en un texto dado, por ejemplo una sola palabra.

“X”

X: Red social que anteriormente se conocía con el nombre de Twitter. Una de las plataformas digitales con más actividad en cuanto a mensajes e interacciones políticas.

Resumen

El estudio de la aprobación presidencial tradicionalmente se ha ligado al ámbito de los análisis económicos; en México no se ha explorado a profundidad la relación entre la dinámica de las conversaciones en X (anteriormente Twitter) y su potencial para estimar la aprobación presidencial a través del análisis de sus interacciones.

La presente investigación introduce un modelo de aprendizaje automático que se vale de diversas técnicas de Procesamiento del Lenguaje Natural para analizar la conversación en X, utilizando esta información para calcular la aprobación presidencial durante períodos específicos. Con este enfoque, se busca abrir nuevas vías para comprender cómo las expresiones públicas en plataformas digitales pueden reflejar, y posiblemente predecir, la percepción pública hacia la figura presidencial.

Introducción

En la era digital actual, las redes sociales se han convertido en un medio cada vez más importante para la participación ciudadana y la comunicación social, constituidas como el principal canal de la opinión pública. Las redes sociales también se han convertido en una plataforma para la discusión política, donde los usuarios expresan sus puntos de vista sobre los eventos sociales y el actuar gubernamental. Tres premisas claves fundamentan la relevancia y propósito de este trabajo:

1) La era digital ha transformado la manera en que la sociedad se comunica e interactúa, siendo las redes sociales un canal fundamental para la expresión de la opinión pública (Lloyd, 2022). En particular, X se ha consolidado como una plataforma clave para la discusión política.

2) Es una práctica histórica de los gobiernos y figuras políticas querer conocer la opinión de los ciudadanos respecto a su actuar político (Blumenthal et al., 2018). Razón por la cual se han desarrollado y refinado los métodos de medición de aprobación política a lo largo del tiempo.

3) El entrelazamiento de la ciencia de datos y el Procesamiento del Lenguaje Natural (PLN) ha desarrollado una herramienta analítica de la opinión pública en redes sociales, posibilitando el medir la influencia de los actores más relevantes y evaluar el impacto de sus mensajes (Khurana et al., 2020).

Capítulo 1
Generalidades

Capítulo 1. Generalidades

Desde el siglo pasado, con la instauración de los análisis estadísticos en los estudios políticos, la medición del nivel de aprobación de un presidente ha sido una herramienta crucial para la comprensión y la evaluación de la efectividad de su gobierno. La opinión pública es una fuerza poderosa en la política, y el nivel de aprobación de un presidente es un indicador clave de cómo está siendo percibido su liderazgo (Apablaza & Jiménez, 2009).

Medir el nivel de aprobación del presidente es importante porque proporciona una instantánea del estado de ánimo del público en un momento dado. La popularidad de un presidente puede variar en función de los acontecimientos políticos, económicos y sociales, por lo que es muy útil medir con regularidad el nivel de aprobación del presidente para detectar cualquier cambio en el sentimiento público. Si la aprobación del presidente está disminuyendo, puede ser una señal de que el público está descontento con su liderazgo o con las políticas que está promoviendo.

Conocer este nivel de aprobación también es algo que le interesa mucho al presidente en turno, ya que si cuenta con una alta aprobación, es más probable que se le recuerde como un líder exitoso y efectivo, y por lo contrario, una baja aprobación le puede afectar en su legado político. En pocas palabras, las mediciones regulares del nivel de aprobación del presidente son útiles para entender cómo está siendo percibido su liderazgo y para hacer ajustes estratégicos en consecuencia.

Este ímpetu por cuantificar el nivel de aprobación presidencial ha evolucionado significativamente a lo largo de los años, desde la utilización de encuestas por correo hasta la implementación de técnicas más avanzadas, como el seguimiento en tiempo real de las redes sociales.

1.1 Planteamiento del problema

Con la presente investigación se quiere conocer si la adopción de técnicas de PLN y aprendizaje computacional aplicadas al procesamiento de la conversación en X, posibilita una evaluación detallada y dinámica del sentir popular, particularmente en lo que respecta a la figura presidencial, con el objetivo de calcular la aprobación presidencial mensual de forma automatizada.

Se realiza este planteamiento debido a que la ciencia de datos permite abarcar una gama más amplia de casos y opiniones en comparación a las encuestas tradicionales, ofreciendo así una representación más integral, amplia y diversa de las percepciones públicas.

1.2 Protocolo de investigación

A continuación se presenta de manera esquemática tanto el objetivo general como los objetivos específicos que guiaron el curso de esta investigación, así como las preguntas de investigación planteadas para la misma. De igual forma se presentan las hipótesis que dieron respuesta preliminar a las preguntas planteadas.

PREGUNTAS	OBJETIVOS	HIPÓTESIS
¿Es posible calcular el grado de aprobación presidencial en méxico, a partir del procesamiento de la conversación en X usando aprendizaje computacional?	Calcular el grado de aprobación presidencial a partir de analizar y clasificar la conversación en X entorno al presidente de méxico usando un modelo de aprendizaje computacional.	Sí es posible calcular la aprobación presidencial al implementar un modelo de ciencia de datos que tenga como entrada parámetros extraídos de la conversación en X entorno al presidente de México.
¿Cómo procesar la conversación en X en torno al presidente de México para realizar un cálculo de la aprobación presidencial con los datos recabados?	Construir una base de datos con una desagregación mensual (diciembre 2019 - diciembre 2022) que incluya los tuits con mención al presidente de méxico y su nivel de aprobación, a la par de identificar los algoritmos necesarios para el preprocesamiento de los	Existe un conjunto de variables y algoritmos específicos dentro de las bibliotecas enfocadas al procesamiento del lenguaje natural para realizar el preprocesamiento de la base de datos.

	elementos del conjunto de datos.	
¿Cuáles son los mejores parámetros para realizar el cálculo del grado de la aprobación presidencial a partir del procesamiento del lenguaje natural de los tuits recopilados?	Extraer y ajustar los parámetros que permitan realizar predicciones sobre el nivel de aprobación presidencial dada una entrada de tuits.	Existen diferentes opciones de pesado y modelado dentro de los algoritmos de aprendizaje computacional para encontrar un adecuado ajuste en el modelaje del cálculo del grado de aprobación presidencial.
¿Cuáles son los mejores métodos estadísticos para medir el desempeño del modelo ajustado enfocado al cálculo del grado de aprobación presidencial?	Realizar un análisis estadístico para medir el desempeño y la precisión del modelo ajustado con el fin de conocer los parámetros óptimos del mismo.	Existen diversas herramientas estadísticas que nos ayudarán a evaluar el desempeño de los modelos ajustados en el estudio.

1.3 Justificación

La presente investigación es conveniente en términos científicos porque proporciona una metodología innovadora para cuantificar la opinión pública expresada en las redes sociales y su relación con la aprobación presidencial en México —a diferencia de los métodos tradicionales de encuesta directa. Esto es de suma importancia en un mundo cada vez más digitalizado, donde la opinión pública en las redes sociales puede tener un impacto significativo en la percepción de la actuación de los líderes políticos.

A su vez cuenta con una relevancia social, si consideramos la aprobación como un indicador clave de la percepción pública sobre la gestión gubernamental. Estudiar cómo se refleja esta aprobación en las redes sociales (en particular en X), aporta a la comprensión de la dinámica entre la opinión pública y el actuar político. Además, este estudio podría beneficiar a actores políticos y tomadores de decisiones al proporcionar una herramienta útil y viable para monitorear a corto plazo la opinión pública y la percepción social de sus acciones políticas.

Las implicaciones prácticas de esta investigación están encaminadas con el objetivo de la misma; al establecer una correlación significativa entre la conversación en X y los niveles de aprobación presidencial, se podría predecir tendencias de aprobación basadas en análisis en tiempo real de las redes sociales. Esto podría ser valioso para actores políticos, instituciones gubernamentales e incluso organizaciones de la sociedad civil que buscan posicionar mensajes estratégicos de comunicación, a la par de responder a las dinámicas cambiantes de la opinión pública.

Paralelamente, el valor teórico de este estudio radica en la contribución al cuerpo académico y práctico sobre cómo las redes sociales están reconfigurando la forma en que la sociedad se involucra con la política (Castells, 2009). Al explorar la relación entre la actividad en X y la aprobación presidencial, se podría establecer un acercamiento válido desde la ciencia de datos para extraer las opiniones vertidas en las plataformas digitales que reflejan el sentir social respecto al actuar presidencial.

La relevancia metodológica de esta investigación radica en la propuesta de un ejercicio práctico y replicable que introduce una innovación en la medición histórica de la aprobación presidencial. Esto se logra mediante la aplicación de métodos de aprendizaje computacional al análisis de la conversación en X. En esencia, la investigación contribuye al desarrollo de un instrumento en español que emplea técnicas de Procesamiento del Lenguaje Natural (PLN), impulsando el avance de modelos de ciencia de datos aplicados a la generación de conocimiento sociopolítico, un campo escasamente explorado en español y aún menos desarrollado en el contexto mexicano.

Al encontrarse en un área de estudio emergente, los resultados de esta investigación tienen la posibilidad de relacionarse con estudios posteriores, siendo de utilidad para otros científicos de datos o investigadores interesados en el tema, ya sea de forma local o internacional. Además, los resultados pueden aportar metodológicamente a la optimización de los procesos de medición de la aprobación presidencial tradicional e incluso a la disminución de costos.

Sumando los elementos anteriormente planteados, esta investigación se presenta como pionera en el abordaje de la aprobación presidencial en México mediante la ciencia de datos, teniendo como campo de estudio la conversación en X, con el fin de modelar las mejores variables que se relacionan de forma transversal con la aprobación presidencial y su cálculo.

1.4 Límites y alcances

1.4.1 Los límites encontrados

La primera limitante versa en torno a la *calidad y representatividad de los datos,* dado que la población activa en X no necesariamente representa a la totalidad de la opinión pública en México. Por lo tanto, existe el riesgo de un sesgo de segmentación de los tuits, en el cual las opiniones capturadas podrían no reflejar fielmente el sentimiento general de la población. Además las cuentas falsas podrían influenciar un enfoque o tema de conversación que no se relacione con la realidad de la opinión pública que se midió en su momento por la casa encuestadora.

La segunda limitante se relaciona con la *ambigüedad en la interpretación del español.* Aunque los métodos y módulos de PLN han avanzado mucho en el idioma inglés, la interpretación computacional del español podría ser un gran desafío, especialmente en contextos irónicos, sarcásticos o ambivalentes (como los albures mexicanos). Las connotaciones en los tuits podrían no ser completamente capturadas o podrían ser malinterpretadas, lo que podría afectar la precisión del análisis de sentimiento aplicado para la clasificación de la polaridad de cada tuit.

La tercera limitante se encuentra en la *poca viabilidad para aplicación del modelo en el futuro por cambios en X.* Actualmente se logró conseguir una base de datos de tuits de años pasados (previos a 2023), pero ante los cambios recientes en el manejo de la plataforma X (incluso de nombre, al ser anteriormente conocida como Twitter), será más difícil y costoso poder extraer tuits para la aplicación del modelo a futuro (por ejemplo en los próximos sexenios), lo que deja

en duda la viabilidad de obtener a futuro bases de datos tan extensas como la trabajada en el presente estudio.

1.4.2 Los alcances más relevantes

El principal alcance identificado es de carácter *exploratorio*, ya que la investigación aborda un tema novedoso y poco estudiado en español. Existen pocos trabajos que integren el uso de técnicas de Procesamiento del Lenguaje Natural (PLN) en nuestro idioma, y son aún más escasos aquellos enfocados en analizar la sociedad mexicana y su interacción con las plataformas digitales.

El segundo alcance identificado es de naturaleza *correlacional*. Si la investigación logra establecer una correlación significativa entre la conversación en X y las variaciones en los niveles de aprobación presidencial, podría abrir la posibilidad de predecir tendencias de aprobación mediante análisis en tiempo real de las redes sociales. Al identificar patrones o cambios en el sentimiento y las conversaciones en línea, se podría anticipar un incremento o descenso en la aprobación presidencial en períodos subsecuentes. Esto tendría un valor significativo no solo en el ámbito académico, sino también para actores políticos y tomadores de decisiones, al ofrecer herramientas para responder de manera ágil y estratégica a las dinámicas cambiantes de la opinión pública.

Por último se propone un *alcance descriptivo,* debido a que esta investigación tiene el potencial de contribuir significativamente al cuerpo académico y práctico sobre cómo las redes sociales están reconfigurando la forma en que la sociedad se involucra con la política. Al explorar la relación entre la discusión en X y la aprobación presidencial, este estudio abona al tema de cómo las opiniones y discursos en plataformas digitales reflejan, influencian o distorsionan el sentir social.

Capítulo 2
Base de datos

Capítulo 2. Base de datos

A continuación se describen los pasos implementados para la construcción de la base de datos de tuits requerida para este proyecto, los cuales se utilizan como técnicas de preprocesamiento enfocadas a la limpieza y transformación de datos, así como su integración y normalización de los mismos.

2.1 Construcción de la base de datos

2.1.1 Fuentes de información

Para la realización del presente estudio se obtuvo una base de datos que contiene la recopilación de millones de tuits delimitados por palabras claves de los nombres de diversos actores políticos mexicanos. Por lo cual la base contiene una gran cantidad de tuits referentes al presidente de México en turno durante la recopilación de 2018 a 2023. La fuente que proporcionó estos tuits históricos fue el INFOTEC, al haberlos recopilados anteriormente para diversos estudios de Análisis de Sentimientos y PLN como el que se encuentra en (Graff et al., 2022). Afortunadamente la base de datos es bastante extensa para el entrenamiento e implementación del modelo de ciencia de datos planteado en el capítulo anterior.

2.1.2 Creación de la base de datos

La integración de datos fue necesaria debido a la existencia de múltiples archivos en formato JSON, cada uno correspondiente a un día de recopilación de tuits. Estos archivos abarcan todos los meses de los primeros cuatro años del gobierno del presidente López Obrador (1 de diciembre de 2018 - 31 de diciembre de 2022), sumando un total de 1,430 archivos. Para facilitar esta integración, se desarrolló un algoritmo que procesa los archivos de cada mes, combinándolos mediante el método *merge* de la biblioteca Pandas. Este enfoque permitió consolidar los datos de manera eficiente y organizada para su posterior análisis.

La estructura por defecto de la base de datos originaria por día, cuenta con las columnas de “id”, “text”, “user_id”, “user_nickname”, “tiemestamp_ms”, “lables”, “full_name”, “location” y “place”.

Las variables de mayor relevancia son el texto de los tuits y la fecha de creación (“tiemestamp_ms”), puesto que es el texto lo que se analiza y clasifica mediante algoritmos de PLN, mientras que la fecha de creación sirve para poder relacionar cada tuit con el nivel de aprobación presidencial desagregado mensualmente.

En la fase de limpieza de datos se excluyen las columnas que no son relevantes para esta investigación, como es el caso de “user_id” y “user_nickname”. De igual forma no se trabaja con las variables de geolocalización “full_name” y “place”.

2.1.3 Homogeneización de variables

Se homologó el nombre de la columna con la fecha de creación de cada tuit, puesto que no es el mismo para todos los archivos (en unos aparece como “tiemestamp_ms” y en otros como “created_at”). Esto es importante porque se necesita homogeneizar la fecha de creación con el mismo nombre de columna, por lo cual se transformó el primer nombre mencionado por el segundo, a la par de transformar esta columna a una de tipo *date.time* en orden para su manipulación.

2.1.4 Preprocesamiento de la base de datos

En la etapa final, se llevó a cabo un proceso de normalización de datos, enfocado en el preprocesamiento de la columna de texto de cada tuit. Para optimizar el Procesamiento del Lenguaje Natural (PLN), se aplicaron diversas técnicas, como la conversión del texto a minúsculas, la eliminación de acentos, puntuación y palabras vacías (stopwords). Estas acciones permiten preparar los datos de manera óptima para entrenar el modelo propuesto, el cual analiza la conversación y calcula el grado de aprobación presidencial correspondiente.

2.2 Análisis exploratorio de los datos

A continuación se describen las principales técnicas exploratorias que se utilizan en la base de datos, el objetivo de emplearlas y las herramientas estadística que se usan para su estudio. En la tabla 1 se presenta de forma visual las gráficas y la numeralia extraídas del Análisis Exploratorio de Datos (AED).

7,590,216 Tuits totales de la base de datos	**85,547,985** Total de palabras en el corpus (sin *stopwords*)	**2,991,049** Palabras únicas en el corpus (sin *stopwords*)
86 Caracteres por tuit en promedio	**178,172** Hashtags únicos utilizados en el corpus de tuits.	**61- 81** Mínimo y máximo de aprobación presidencial (2019-2022)

Tabla 1. *Numeralia extraída del AED.* Fuente: Elaboración propia.

En la figura 1 se expone una nube de palabras para proporcionar una representación visual de aquellas palabras que aparecen con mayor frecuencia. Para ello, se utilizó la biblioteca *WordCloud*, empleada regularmente para este propósito.

Figura 1. *Nube de palabras con mayor frecuencia en el corpus.* Fuente: Elaboración propia.

La figura 2 se enfoca en la frecuencia de palabras, con el propósito de identificar las más comunes dentro del conjunto de datos examinado. Se prestó especial atención a cada palabra individual presente en el corpus de texto, empleando tablas de frecuencia como la herramienta estadística principal. Para lograr esto, se implementó el proceso de tokenización, que permite dividir el texto en palabras individuales, seguido de un conteo de éstas.

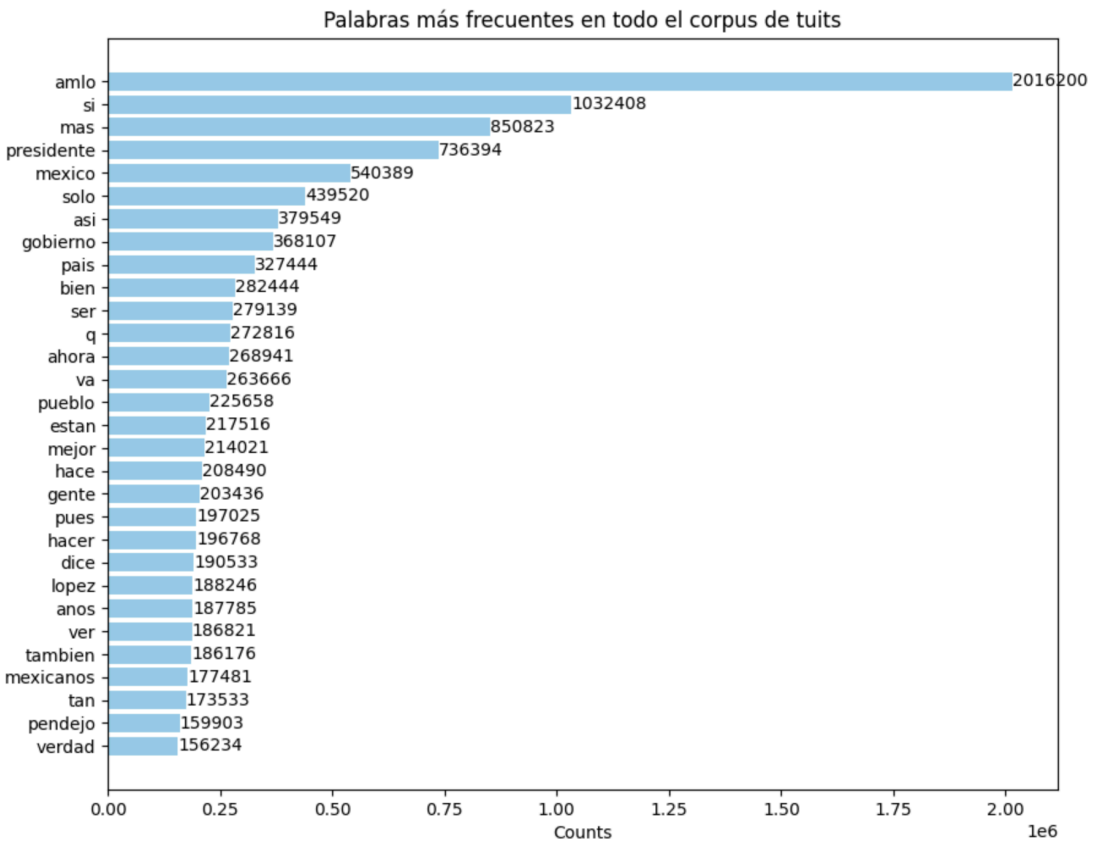

Figura 2. *Histograma de frecuencia de palabras.* Fuente: Elaboración propia.

En la figura 3 se encuentra un estudio sobre la frecuencia de las etiquetas (*hashtags*), con el objetivo específico de determinar las etiquetas más utilizadas en los primeros cuatro años del mandato presidencial, entre 2019 y 2022. Las características de interés en este caso han sido las palabras precedidas por el símbolo '#', analizadas mediante el uso de tablas de frecuencia. Para la extracción y conteo de hashtags, se recurrió a expresiones regulares y al método *Counter* de Python, respectivamente.

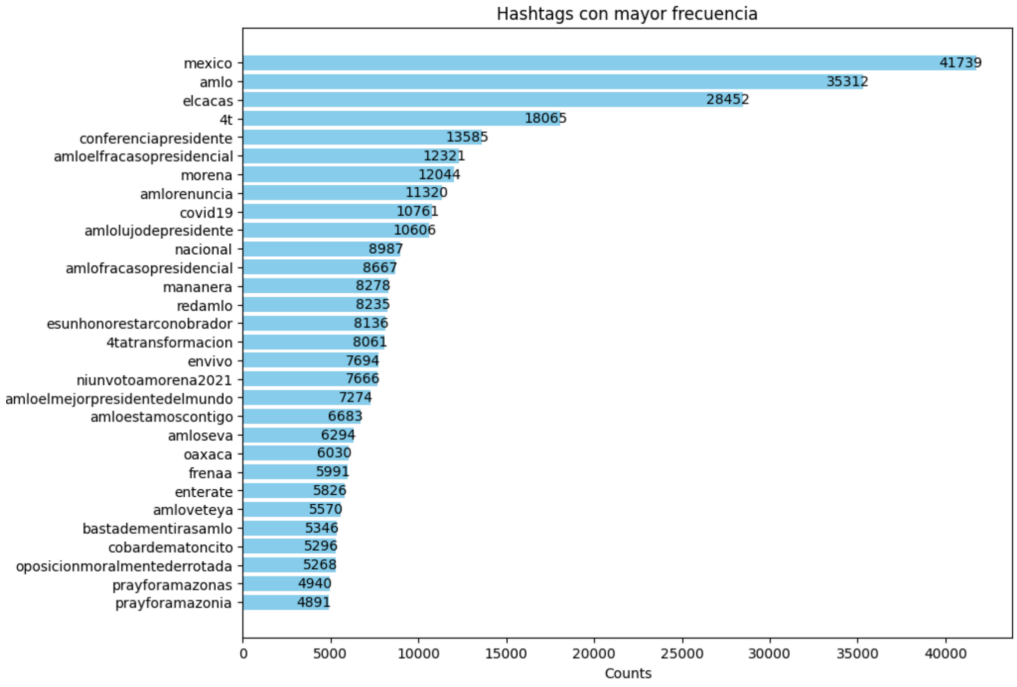

Figura 3. *Histograma de frecuencia de hashtags.* Fuente: Elaboración propia.

Finalmente, en las figuras 4, 5 y 6, se explora la distribución de la cantidad de tuits por año y la longitud promedio de los tuits mediante el uso de histogramas. Estos gráficos se generaron con la ayuda de la biblioteca *Matplotlib* y ofrecen una vista más clara de las distribuciones por año, pudiendo reconocer que la mayor cantidad de tuits presentes en la base de datos son del 2019.

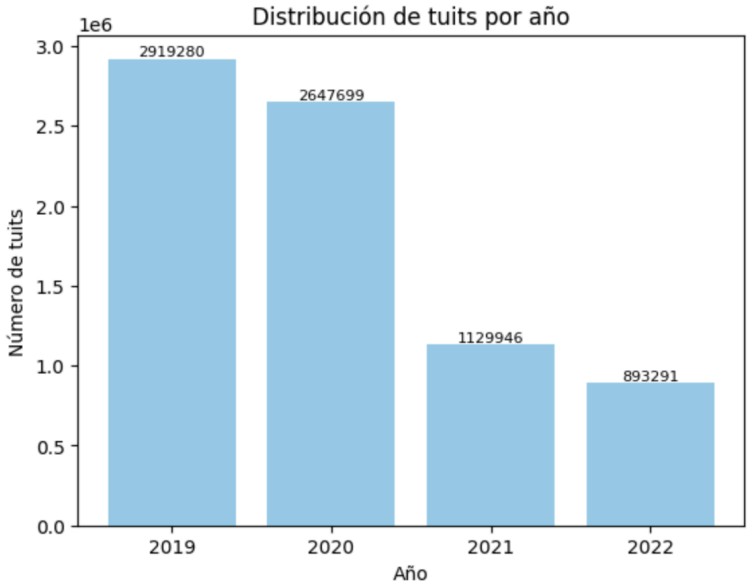

Figura 4. *Histograma de número de tuits por año.* Fuente: Elaboración propia.

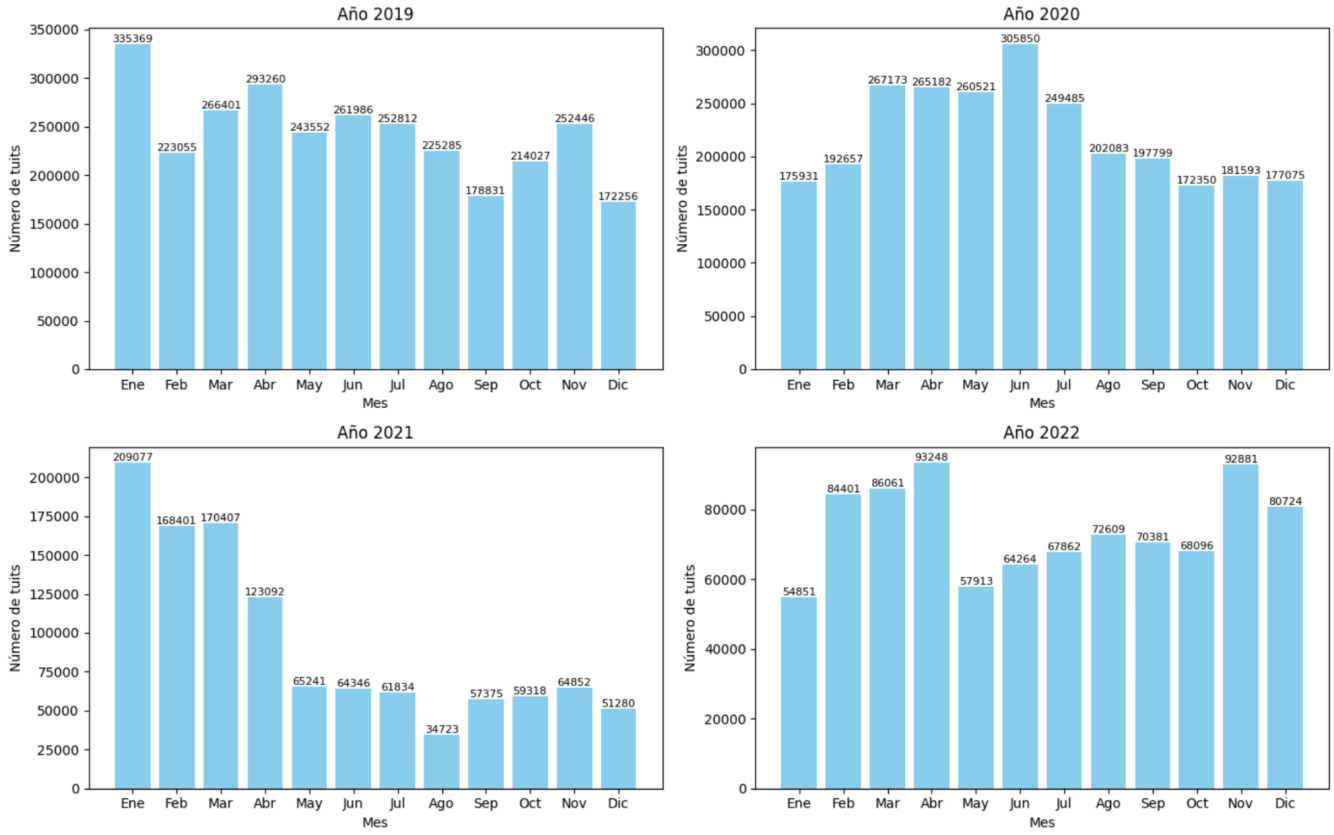

Figura 5. *Histogramas de número de tuits por mes en cada año.* Fuente: Elaboración propia

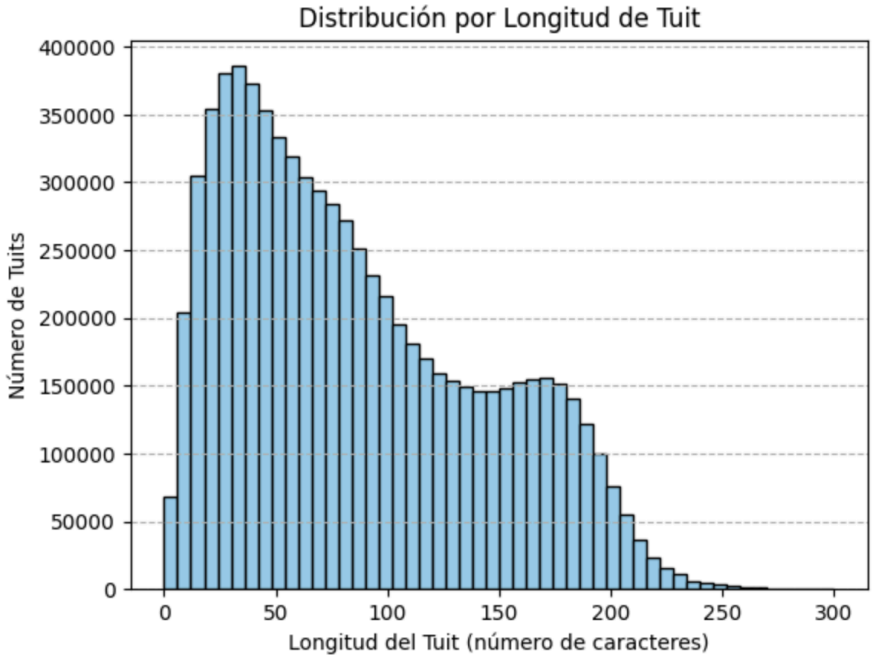

Figura 6. *Histograma de la distribución de la longitud de los tuits.* Fuente: Elaboración propia.

Capítulo 3

Marco teórico, diseño del estudio y ajuste de modelo

Capítulo 3. Marco teórico, diseño del estudio y ajuste de modelo

A continuación se presenta el marco teórico que sirve como base conceptual para la implementación práctica del modelo propuesto de predicción de la aprobación presidencial. A su vez se incluye una revisión de la literatura en la cual se describe el estado del arte y los trabajos de investigación que se relacionan con el objeto de estudio. Posteriormente se describen la metodología utilizada con el fin de lograr la implementación y ajuste del modelo.

3.1 Marco teórico

Como base teórica y campo de estudio de esta investigación, tenemos que la ciencia de datos es un campo interdisciplinario que combina técnicas estadísticas, matemáticas y de programación para extraer conocimiento y comprensión de los datos, la cual se utiliza en una amplia gama de aplicaciones, desde el análisis de datos empresariales hasta la investigación científica (Provost, 2013).

La ciencia de datos también es un componente clave de la Inteligencia Artificial (IA) y del aprendizaje automático (ML, por sus siglas en inglés) (Provost, 2013). Los científicos de datos pueden utilizar técnicas de ML para crear modelos predictivos y sistemas de recomendación que se utilizan en una vasta gama de aplicaciones, como puede ser la publicidad que vemos en línea, o incluso la conducción autónoma de vehículos. La ciencia de datos es una herramienta esencial en diversos campos al permitir a los investigadores y a los tomadores de decisiones extraer conocimiento y comprensión de grandes conjuntos de datos.

Por su parte, el Procesamiento del Lenguaje Natural (PLN) es el uso de la inteligencia artificial para interactuar con sistemas computacionales mediante el lenguaje humano (Gandhi, 2021). Una de las aplicaciones más comunes del PLN es el análisis de sentimientos en plataformas como X o Facebook. Además, el PLN permite el desarrollo de aplicaciones más avanzadas como chatbots, reconocimiento de voz, traducción automática, corrección ortográfica, búsqueda

de palabras clave, extracción de información y personalización de anuncios. Con estas herramientas, se puede crear una aplicación de X capaz de comprender el lenguaje humano (Gandhi, 2021). El PLN es una parte esencial de la ciencia de datos que facilita la comprensión automática y el análisis de textos de manera eficaz. Gracias a ello, es posible manejar grandes volúmenes de datos textuales, realizar tareas complejas y ofrecer soluciones para una amplia gama de problemáticas relacionadas.

El Aprendizaje Profundo (Deep Learning) es una parte de las técnicas de aprendizaje automático que utiliza redes neuronales artificiales. Estas redes tienen varios niveles de capas que ayudan a sacar características importantes de los datos brutos (Gandhi, 2021). Se usa en varios campos, como el procesamiento de señales e imágenes. Esto se debe a que las Redes Neuronales Profundas tienen varias redes neuronales conectadas, donde el resultado de una red se convierte en la entrada de la siguiente, y así sucesivamente. El aprendizaje profundo puede analizar y aprender de los datos textuales por sí mismo, examinando varias capas de características para hacer predicciones. Por ejemplo, en el análisis de sentimientos se ha utilizado el aprendizaje profundo para clasificar la polaridad de tuits entre positivo, negativo o neutral (Ali et al., 2021; Kumar et al., 2019; Budiharto & Meiliana, 2018).

3.1.1 Modelado y vectorizado del texto

Para entender la ruta de trabajo en la aplicación del modelo de aprendizaje automático que se quiere implementar, así como el vectorizado/pesado de las palabras mediante modelados de texto, primero debemos definir lo que es una representación vectorial y un word embedding.

La representación de objetos en forma de vectores ha sido muy favorecida y usada en el área de recuperación de información. Cada documento analizado (por ejemplo un tuit) es representado como un vector. En forma vectorial, generalmente, se ordenan las palabras considerando algún orden (por ejemplo un orden léxico) como cada una de las dimensiones del vector, donde cada

dimensión representa una característica, es decir, una palabra del vocabulario (Pennington et al., 2014).

Recientemente ha habido un auge en aplicar enfoques distribucionales del significado de las palabras en los textos, donde se asume que las palabras que co-ocurren en el mismo contexto tienden a tener el mismo significado (Firth, 1957). Lo anterior ha llevado al desarrollo de modelos que implementan la semántica distribucional, como es el caso de Latent Semantic Analysis (LSA), Hyperspace Analogue to Language (HAL), y enfoques de word embeddings como Global Vectors y Word2Vec (Pennington, 2014).

La idea básica con los word embeddings es recolectar grandes cantidades de información y codificar en vectores de alta dimensión cómo se distribuyen las palabras de acuerdo a su contexto. Aquí se asume que los vectores construidos contienen información semántica y sintáctica de la palabra codificada por medio de dichos vectores. Dichas representaciones de vectores de palabras se han usado en diferentes tareas como son la identificación de analogías, clasificación de textos, agrupamiento de datos, análisis de sentimientos, entre otras.

A diferencia de la matriz de término-documento en la representación vectorial, las dimensiones en esta representación de word embeddings no están asociadas a ningún término, sino que codifica la “semántica” de las palabras con las que se relaciona la palabra a la que representa al vector. Por lo que palabras cercanas semánticamente, tendrán vectores similares (Mikolov, 2013).

Con el paso del tiempo, distintos modelos de embeddings han surgido, como FastText, BERT, OpenAI GPT, InferSent, XLNET, y muchas otras variantes que intentan codificar las oraciones de manera diferente y que capture mejor la semántica. En general, estos enfoques se basan en construir el modelo basándose en redes neuronales, con grandes cantidades de datos de internet como Wikipedia, Common Crawl, Google News, y otros corpus; por lo que cada enfoque se ha centrado en encontrar las dimensiones óptimas para su modelo.

3.1.2 Aprendizaje Supervisado

El Aprendizaje Supervisado (AS) es un paradigma fundamental en el ámbito del aprendizaje computacional que se basa en la construcción de un modelo a partir de un conjunto de datos de entrenamiento previamente etiquetado. Los datos de entrenamiento consisten en pares de entrada-salida, donde la entrada representa las características o variables observadas y la salida representa la variable objetivo o el valor que se desea predecir. El objetivo del AS es entrenar un modelo que pueda aprender la relación entre las características de entrada y la variable objetivo, de manera que pueda realizar predicciones precisas para nuevos datos no vistos (Géron, 2019).

Existen dos tipos principales de AS, ya sea un aprendizaje de regresión o uno de clasificación. El aprendizaje por regresión se utiliza cuando la variable objetivo es continua, es decir, puede tomar cualquier valor dentro de un rango. Por ejemplo, predecir el precio de una vivienda en función de su tamaño, ubicación, características y otros factores relevantes. En cambio, el aprendizaje para clasificación se utiliza cuando la variable objetivo es categórica, es decir, solo puede tomar un número finito de valores discretos (Murphy, 2012). Por ejemplo, se puede clasificar correos electrónicos como spam o no spam, identificar la especie de una flor a partir de una imagen o sus características, o también clasificar la polaridad de un texto entre positiva, negativa o neutra.

Existen diversos algoritmos de AS que se pueden emplear para entrenar modelos predictivos. Algunos de los más comunes son la Regresión Lineal, la Regresión Logística, los Árboles de Decisión o el modelo K-Nearest Neighbors (KNN). En la parte del Aprendizaje Profundo (Deep Learning), existen las Redes Neuronales Artificiales (RNA). La elección del algoritmo adecuado depende del tipo de problema, la naturaleza de los datos y los objetivos específicos del análisis (Friedman, 2001; Breiman, 2001 & Bengio, 2006).

3.1.3 Análisis de sentimientos

El análisis de sentimientos es una técnica del PLN que se centra en la identificación y clasificación de las opiniones, emociones y actitudes expresadas en textos. El objetivo del análisis de sentimientos es extraer información de una variedad de fuentes, como redes sociales, reseñas de productos, foros en línea y artículos de noticias con el fin de comprender las tendencias de opinión pública sobre eventos sociales, políticos y económicos (Cambria & Hussain, 2016).

En general, un método de aprendizaje computacional enfocado al PLN realiza los siguientes pasos: el preprocesamiento de los textos; calcular las probabilidades o transformar los datos de texto a una forma vectorial; aprender el modelo de datos asociado a las categorías definidas en el conjunto de entrenamiento; y, por último, predecir nuevos datos que sirvan de validación del método, es decir, qué tan bien se desempeña al utilizar los datos de prueba.

Por ejemplo, en el caso de clasificación, x puede ser el vector que representa a un texto (tuit) con su respectivo pesado (TF, TFI-DF, word embedding) mientras que y las posibles clases en las que se puede clasificar el texto (positivo, negativo, neutro). En este caso, y sería un vector en el que cada dimensión representa una clase y contendría la probabilidad de pertenencia a la clase.

Hay que tomar en cuenta que los métodos del PLN aplicados a los textos de las redes sociales se enfrentan a dificultades debido a la ortografía no estándar, el ruido y los conjuntos limitados de funciones para la agrupación y clasificación automáticas (Tellez et. al., 2017). Sin embargo, el análisis de sentimientos mediante la ciencia de datos es una técnica útil para analizar grandes cantidades de datos no estructurados y determinar la polaridad de las opiniones expresadas en ellos.

3.2 Revisión de la literatura y antecedentes

3.2.1 Antecedentes históricos

En la década de 1940 comenzaron a medirse los niveles de aprobación presidencial de manera regular, por medio de encuestas telefónicas y por correo postal (Blumenthal, 2018). En la década de 1950, la compañía Gallup comenzó a medir la aprobación presidencial de manera semanal, lo que permitió una mayor precisión en las mediciones y un seguimiento más cercano de la opinión pública. A medida que la tecnología avanzó, se fueron implementando nuevas técnicas para medir la opinión pública.

En la década de 1980, se comenzaron a utilizar encuestas por muestreo telefónico aleatorio, lo que permitió una mayor representatividad de la muestra. A principios del nuevo milenio, se comenzaron a utilizar encuestas por internet y por teléfono móvil, facilitando una mayor accesibilidad para los encuestados. El mayor cambio en la historia moderna se presentó en la década pasada, cuando las redes sociales cambiaron la forma y el medio en que la opinión pública se distribuye, al convertirse en su canal principal (Blumenthal, 2018).

Paralelamente, en los últimos años las redes sociales se han convertido en una herramienta fundamental en la comunicación, el entretenimiento y la información para millones de personas en todo el mundo. Su influencia en la sociedad es cada vez mayor, y aunque existen tanto ventajas como desventajas, “no se puede negar que las redes sociales han cambiado la forma en que las personas interactúan, se comunican y se relacionan entre sí” (Lloyd, 2022, p.56).

En lo referente al campo político, gracias a que las redes sociales han democratizado la información y han permitido que cualquier persona tenga la posibilidad de expresar sus ideas y opiniones en un espacio público, éstas se han convertido en un medio cada vez más importante para la participación ciudadana y la formación de opiniones, constituyéndose como el principal canal de la opinión pública en la actualidad (Espinoza, 2022).

Para el sociólogo Manuel Castells (2009), la opinión pública es el resultado de un proceso de comunicación simbólica entre individuos y grupos que comparten valores, intereses y percepciones sobre una determinada cuestión. De acuerdo con esta perspectiva, la opinión pública no es algo dado o estático, sino que se construye y se transforma a lo largo del tiempo a través de la interacción y diálogo entre los distintos actores sociales. Este proceso se ha visto influenciado por las tecnologías de la comunicación y ha generado una estructura de la comunicación en red que posibilita la creación de una opinión pública global y conectada. Castells considera que la opinión pública no es un concepto homogéneo, sino que "está formada por una multiplicidad de opiniones y perspectivas que pueden ser tanto complementarias como contradictorias" (p. 58).

Con las plataformas digitales actuales, los usuarios pueden compartir noticias, pensamientos, imágenes y vídeos, teniendo la capacidad de llegar a un público masivo en cuestión de segundos. Esto ha generado un impacto significativo en la forma en que se produce y se consume información, ya que las redes sociales han logrado derribar las barreras de entrada a los medios de comunicación tradicionales y han permitido que la gente tenga acceso a una gran variedad de fuentes de información (Martínez, 2017).

La ciencia de datos aplicada a las redes sociales posibilita el análisis de amplios conjuntos de datos en tiempo real, lo que permite a los analistas identificar patrones y tendencias en las interacciones políticas en las redes sociales (Khurana et al., 2020). Además, la ciencia de datos puede ayudar a identificar a los actores políticos más influyentes en las redes sociales y a medir el impacto de sus mensajes en la opinión pública.

3.2.2 Trabajos relacionados que miden la aprobación presidencial

El trabajo con mayor acercamiento contextual y metodológico a la problemática planteada se encuentra en Solis (2022), el cual tiene como objetivo analizar los tuits con contenido político redactados por ciudadanos peruanos en X con la finalidad de descubrir su valoración y aceptación respecto al presidente del Perú.

Para dicha encomienda, Solis desarrolla un clasificador basado en redes neuronales artificiales, el cual fue entrenado con 3400 tuits recolectados durante agosto y diciembre del 2021. El autor concluye que al comparar los resultados obtenidos de sus modelos con los informes reportados por encuestadoras nacionales, sus estimaciones están fuertemente correlacionadas, estipulando así que “este método ofrece una alternativa rápida, eficiente y de bajo costo para monitorear y obtener la tasa de aprobación del presidente peruano” (p. 11).

El trabajo muestra una detallada explicación de los modelos de redes neuronales que se implementaron en el estudio, pero al ser un artículo editorial, se comparte con poca profundización el código y bibliotecas para la elaboración del modelo. Aún así, el estudio es muy valioso por el acercamiento al trabajo aquí planteado y sus conclusiones sirven de base para realizar el modelo que aquí se propone.

Paralelamente, Solis formó parte de un equipo más grande de investigación junto a Vidalón et al. (2022), quienes implementaron un modelo de red neuronal convolucional (CNN) para predecir la aprobación de políticos peruanos basándose en datos de X, enfocándose en predecir el respaldo popular del presidente de Perú. El modelo fue capaz de aprender y clasificar la diversidad de mensajes en los tuits según su polaridad sentimental como positivos/negativos o de aprobación/desaprobación, ponderando los resultados positivos y negativos según el número de mensajes procesados para proporcionar una tasa final de aprobación o desaprobación. Los resultados obtenidos se compararon con los reportados por encuestadoras especializadas peruanas, encontrando correlación entre las predicciones modeladas y los resultados de las encuestas. Los investigadores concluyen que con su modelo lograron demostrar la eficacia de ese enfoque para analizar la aprobación política mediante el análisis de sentimientos en redes sociales.

También existe otro estudio relacionado con la medición de la aprobación presidencial en Fer et al. (2021), quienes examinan la viabilidad de usar datos de redes sociales, específicamente de X, como complemento o sustituto de

encuestas tradicionales para realizar la medición. Señalan que los tuits políticos generalmente muestran claramente el apoyo u oposición a figuras o políticas gubernamentales, y los consideran una fuente potencialmente fiable para su utilización. El estudio desarrolla un marco para evaluar la correlación entre el sentimiento expresado en los tuits y los resultados de las encuestas de opinión pública, usando pruebas múltiples para interpretar la fuerza de estas correlaciones. Específicamente, calcularon la correlación entre la aprobación presidencial medida por encuestas y el sentimiento de tuits que mencionaban la palabra "Trump", utilizando el método VADER para asignar una puntuación de sentimiento continua a cada tuit.

Un último estudio referente a la aprobación presidencial se encuentra en Strong & Kohli (2019) quienes desarrollaron un modelo de clasificación para predecir los resultados de las elecciones presidenciales estadounidenses, utilizando una amplia gama de datos multifactoriales. Aunque se inspiraron en modelos previos, se diferencian en incorporar variables como indicadores económicos, sondeos de aprobación y política exterior, para lograr un entendimiento más completo de la situación reputacional del presidente. Utilizando un conjunto de datos desde 1948 hasta 2016, el modelo empleó Árboles de Decisión impulsados por la metodología de bootstrap para superar la limitación de tener solo 18 observaciones. Se destacó la tasa promedio de aprobación presidencial como variable significativa, mostrando un sólido desempeño del modelo en la predicción de resultados electorales y enfatizando la importancia de la aprobación presidencial en modelos predictivos electorales.

3.2.3 Trabajos relacionados que utilizan el análisis de sentimiento

Con el propósito de responder si es posible emplear X para observar cómo la manifestación popular afecta a un personaje político, en Shaghaghi et al. (2021) llevaron a cabo un análisis de sentimientos sobre el nivel de afectividad de la población hacia el presidente Donald Trump. Se recopilaron tuits mediante la librería Tweepy y la API de X utilizando la palabra clave "Donald Trump". Los

autores etiquetaron manualmente la información recolectada y utilizaron dos tipos de algoritmos clasificadores: Naive Bayes y la red neuronal de tipo Long-Short Term Memory (LSTM). Los resultados mostraron una tasa de aceptación del 63% para Naive Bayes y del 69% para LSTM.

En otro estudio relacionado, Ansari et al. (2020) se centraron en los sentimientos de los usuarios de X hacia los principales partidos políticos nacionales que participaron en las elecciones de la India en 2019. El objetivo fue utilizar arquitecturas de aprendizaje profundo y compararlas con modelos clásicos de aprendizaje automático para inferir los resultados de las elecciones. En primer lugar, se recopilaron tuits relacionados con los dos partidos más populares en ese momento. Para clasificar los datos, se utilizaron diferentes modelos, incluyendo LSTM, SVM, un Árbol de Decisiones y Regresión Logística, siendo este último modelo el que obtuvo mayor precisión.

Por su parte, Yavaria et al. (2022) realizaron un estudio basado en el análisis de datos de X y el análisis de sentimientos, para utilizar la proporción de la tasa de mensajes positivos entre la tasa de mensajes negativos como un indicador efectivo para predecir elecciones.

Lovera y Cardinale (2023) realizan un estudio comparativo sobre diferentes técnicas de análisis de sentimientos en X, comparando métodos de Machine Learning y Deep Learning. Los autores proponen una estrategia metodológica que incluye preprocesamiento de datos, construcción y evaluación de los modelos predictivos. Para su estudio utilizaron el conjunto de datos Sentiment140, aplicando técnicas de PLN para la limpieza y preparación de los datos. Los resultados mostraron que el modelo de Deep Learning basado en LSTM superó a los modelos clásicos y otros modelos de Deep Learning en métricas de precisión y F1, demostrando la efectividad del LSTM para entender relaciones en textos cortos como los tuits.

En cuanto a la aplicación práctica del análisis de sentimientos, en la investigación de Tellez et al. (2017), un equipo de científicos de datos se dio a la tarea de realizar un extenso análisis sobre las posibles configuraciones de

parámetros que existen en la parte transformacional de un texto (modelaje del texto), la cual es antecesora a la decisión del algoritmo clasificador a utilizar. Estos afirman que es posible mejorar la parte clasificatoria al mejorar el modelaje, tras experimentar con dos grandes conjuntos de datos sobre miles de tuits provenientes de México, con los cuales cotejaron los resultados de los diferentes parámetros utilizados para su procesamiento y posterior puntuación de su precisión clasificatoria.

3.3 Marco metodológico

El presente apartado describe el proceso metodológico adoptado para transformar una amplia base de datos de tuits –la cual abarca de enero de 2019 a diciembre de 2022- en un conjunto de datos procesado apto para alimentar un modelo predictivo de la aprobación presidencial. Ante la ausencia de variables predefinidas y la naturaleza no estructurada de los datos iniciales, se implementó un proceso de ingeniería de características, a partir de la aplicación de modelos preentrenados, para definir de manera práctica el problema en términos de aprendizaje supervisado.

El marco propuesto para la valoración de aprobación basada en el análisis conversacional de X se muestra en la Figura 7. El modelo se entrenó para aprender a predecir la aprobación presidencial dadas las distintas características de entrada de nuevos datos desconocidos. Este enfoque está diseñado para procesar la conversación en X en torno al presidente de México.

Figura 7. *Proceso metodológico de la investigación.* Fuente: Elaboración Propia.

3.3.1 Unificación de los datos

La primera etapa del proceso metodológico se centró en la unificación de una vasta colección de tuits por día, acumulados desde enero de 2019 hasta diciembre de 2022. La recopilación diaria de los tuits la realizaron académicos de INFOTEC dentro del periodo mencionado anteriormente, a partir de la utilización de la API del entonces Twitter. Este paso fue crucial para garantizar una base de datos coherente y estandarizada, donde cada entrada de tuit se normalizó para homogeneizar formatos, como el tipo de fecha de creación. En total se trabajó con 48 conjuntos de datos, uno por cada mes dentro del periodo de tiempo anteriormente descrito.

3.3.2 Preprocesamiento y filtrado de los datos

Tras la unificación, los datos pasaron por un proceso de preprocesamiento y filtrado. Este proceso implicó la limpieza del texto mediante la utilización de expresiones regulares, eliminando elementos no esenciales como URLs, emojis, y cuentas de usuarios.

Además, se normalizó el texto a minúsculas y se aplicaron técnicas de tokenización para descomponer el contenido en unidades básicas de análisis, esto con el apoyo de la librería *unicodeta* en Python para el PLN. También se eliminaron las *stopwords* de cada texto al aplicar la librería NLTK.

Dado que los archivos diarios estaban previamente categorizados por palabras claves referentes a políticos mexicanos en la columna de “layers”, un filtro adicional (aplicando la función *str.contains*) recopiló únicamente los tuits con etiquetas referentes al presidente de México mediante la búsqueda de términos específicos como “amlo”, “@_lopezobrador”, “andrés manuel”, “peje”, entre otras palabras, a la vez que se excluían los términos referentes a otros políticos mexicanos.

Una vez realizada la normalización de los textos, se procedió a la creación de una nueva columna con el nombre de ‘tuitsProcesados’ para cada uno de los

datasets mensuales, con la cual se realizó gran parte de la ingeniería de características posterior.

3.3.3 Ingeniería de características para la construcción del problema de aprendizaje supervisado

La ingeniería de características fue el núcleo del proceso metodológico, transformando el texto no estructurado en un conjunto estructurado de variables analíticas mediante el uso de modelos de PLN preentrenados, los cuales forman parte de la biblioteca EvoMSA[1] (Graff et al., 2020).

Esta biblioteca de modelos de lenguaje preentrenados facilita transformar los textos en características significativas para modelos de aprendizaje automático, facilitando la identificación de patrones relacionados con sentimientos positivos, negativos o neutrales; y a la vez permite el análisis de probabilidades para representar los textos vectorizados en palabras clave o emojis.

El uso de EvoMSA en esta investigación permite evaluar la polaridad de sentimientos de los tuits (positiva, negativa, neutral), mediante dos tipos de modelos preentrenados diferentes: INEGI y TASS2016, el primero entrenado en español con tuits provenientes de usuarios mexicanos, y el segundo también entrenado en español pero con tuits provenientes de usuarios de España.

La biblioteca EvoMSA de igual forma permite detectar emociones específicas que tengan mayor o menor probabilidad de ser representadas por un emoji, en este caso fueron seleccionados deliberadamente los emojis de alegría (😄), tristeza (😔), enojo (😡) y risa (😂).

Por otra parte, se emplean funciones específicas en Python para calcular la cantidad total de tuits por cada mes dentro del periodo de tiempo de la investigación. Además, se emplea una métrica de diversidad léxica, definida como la proporción de palabras únicas sobre el total de palabras en un tuit. Esta medida proporciona una indicación de la riqueza y variación del lenguaje utilizado.

[1] Para conocer la documentación de esta librería, consultar: https://evomsa.readthedocs.io/en/docs/

De igual forma se utiliza la biblioteca *textstat* para calcular el índice de legibilidad *Flesch Reading Ease* de cada tuit, ofreciendo un métrica valiosa sobre la facilidad con la que se pueden leer y entender los mensajes.

Por último, se ingresó de forma manual el promedio mensual de diferentes casas encuestadores en su estimación de la aprobación presidencial para un mes determinado, el cual se extrajo de la plataforma Oraculus[2], con el fin de dotar de mayor objetividad este estudio.

En la Tabla 2 se presenta un ejemplo visual de las características extraídas de cada uno de los 48 meses procesados, correspondientes al periodo de enero de 2018 a diciembre de 2022. La tabla muestra la primera entrada del conjunto de datos general, que consta de 48 registros, uno por mes. El conjunto de datos incluye 16 columnas en total, de las cuales las columnas 2 a 15 se utilizaron como variables independientes, mientras que la columna 16 se empleó como variable objetivo del modelo, representando el grado de aprobación.

1	2	3	4	5	6	7	8	9
fecha	cantidad_tuits	longitud_media	diversidad_lexica	legibilidad	inegi_postivo	inegi_neutro	inegi_negativo	tass2016_negativo
2019-01-01	335369	22.352331	0.918241	37.727828	0.256437	0.179080	0.398782	0.422374

10	11	12	13	14	15	16
tass2016_neutro	tass2016_postivo	emocion_alegría 😁	emocion_tristeza 😔	emocion_enojo 😡	emocion_risa 😂	grado_aprobacion
0.012896	0.229857	0.022541	0.018540	0.006232	0.017428	80

Tabla 2. *Primera entrada del conjunto de datos de características del modelo.*
Fuente: Elaboración propia.

La variable “longitud_media” expresa el promedio de caracteres que tiene cada tuit correspondiente al mes de su procesamiento, en este caso los tuits de enero 2018 tuvieron 22 caracteres en promedio.

La variable “diversiad_lexica” expresa el cociente entre el número de palabras únicas utilizadas en el mes (vocabulario) y el número total de palabras de

[2] Para conocer más sobre este proyecto y obtener los registros del promedio de aprobación presidencial refierase a https://oraculus.mx/aprobacion-presidencial/

ese mes, donde 0 indica poca diversidad al haber pocas palabras únicas en relación al total y 1 indica una alta diversidad léxica, al tener todas las palabras únicas.

La variable “legibilidad” se obtuvo del índice de *Flesch Reading Ease*, una medida de legibilidad que indica qué tan fácil es comprender un texto. El índice tiene un rango numérico entre 0 y 100 donde los valores altos representan textos más claros y sencillos de interpretar. En este caso se extrajo la legibilidad promedio de cada uno de los meses procesados.

Las variables “inegi_positivo”, “inegi_neutro” e “inegi_negativo” representan la polaridad promedio de los tuits procesados durante cada mes, según los modelos preentrenados de EvoMSA basados en INEGI. Estas variables reflejan la probabilidad promedio de que el contenido de los tuits sea positivo, neutro o negativo, respectivamente, en un rango de 0 a 1. Valores más altos en cada variable indican una mayor probabilidad de que los tuits correspondan a la polaridad señalada, permitiendo identificar tendencias generales de la conversación en X durante el periodo analizado.

De manera similar, las variables “tass2016_positivo”, “tass2016_neutro” y “tass2016_negativo” representan la polaridad promedio de los tuits procesados según los modelos preentrenados de EvoMSA basados en TASS2016. Este modelo, entrenado en español con datos provenientes de usuarios de España, también mide la probabilidad de que el contenido de los tuits sea positivo, neutro o negativo, utilizando el mismo rango de 0 a 1. Estas variables complementan el análisis al incorporar una perspectiva diferente en la evaluación de los sentimientos, basada en un conjunto de datos de origen distinto, enriqueciendo la comprensión de las polaridades mensuales observadas.

Las variables “emocion_alegría 😁”, “emocion_tristeza 😔”, “emocion_enojo 😡” y “emocion_risa 😂” representan la probabilidad promedio mensual de que el conjunto de palabras procesadas en los tuits de cada mes exprese la emoción asociada al emoji correspondiente. Cada variable utiliza un rango de 0 a 1, donde valores más altos indican una mayor probabilidad de que las palabras del mes reflejen dicha emoción. Por ejemplo, un valor elevado en “emocion_alegría 😁”

sugiere que los tuits tienen una mayor tendencia a transmitir sentimientos felices, mientras que un valor alto en “emocion_tristeza 😔” indica una inclinación hacia sentimientos de aflicción.

La correlación de estas características extraídas de los datos iniciales no estructurados con los niveles históricos de aprobación presidencial, permitió formular un problema de aprendizaje supervisado, definiendo de manera clara la variable objetivo, así como las variables de entrada para el modelo predictivo.

3.3.4 Entrenamiento de los modelos de aprendizaje automático

Con el objetivo de desarrollar un modelo predictivo capaz de estimar la aprobación presidencial, se evaluaron diversos modelos de aprendizaje automático, seleccionados por su capacidad para manejar variables continuas y por sus diferencias fundamentales en cuanto a enfoques de modelado. Cada modelo se eligió para explorar cómo diferentes técnicas podrían capturar y reflejar las complejidades de los datos disponibles.

Se inició con la regresión lineal debido a su simplicidad y transparencia en el modelado de relaciones lineales entre variables independientes y la variable objetivo. Este modelo es fundamental en estadísticas y ofrece una base sólida para comparar la eficacia de enfoques más complejos, además de proporcionar una comprensión clara de la influencia de cada característica en la variable de aprobación presidencial.

También se puso a prueba el modelo basado en Árboles de Decisión (Decision Tree Regressor), por su habilidad para modelar relaciones no lineales y su capacidad para manejar interacciones complejas entre características sin necesidad de transformaciones explícitas de los datos. Su estructura jerárquica facilita la interpretación de cómo las decisiones son tomadas por el modelo, lo cual es útil para entender qué factores influyen más en la aprobación presidencial.

Como una extensión de los Árboles de Decisión, el modelo Random Forest utiliza un enfoque de ensemble que promedia múltiples árboles para mejorar la precisión y controlar el sobreajuste. Este modelo también se puso a prueba debido

a su robustez y mejor rendimiento general en comparación con un solo Árbol de Decisión, lo que lo hace adecuado para datos que pueden contener numerosas anomalías y variaciones como los tratados en este estudio.

De igual forma, se utilizó el modelo Support Vector Regression (SVR) para evaluar su eficacia en encontrar un hiperplano que mejor se ajuste a los datos en un espacio de mayor dimensión. Aunque típicamente más utilizado en clasificación, en su forma de regresión, SVR puede ser muy potente para capturar relaciones complejas y no lineales, pero con la desventaja de requerir un ajuste cuidadoso de sus parámetros.

Asimismo, el modelo Gradient Boosting Machine (GBM) es un método que construye secuencialmente árboles de decisión, cada uno intentando corregir los errores del anterior. Se seleccionó GBM por su eficiencia en reducir el sesgo y la varianza, ofreciendo una estrategia sólida para mejorar progresivamente la precisión a través del aprendizaje de gradientes.

Como una implementación optimizada de GBM, el modelo XGBoost fue probado por su popularidad en competencias de modelado predictivo y su capacidad para manejar grandes volúmenes de datos de manera eficiente.

Cada modelo fue evaluado en términos del Error Cuadrático Medio (MSE por sus siglas en inglés) y del coeficiente de determinación (R^2), para determinar su eficacia en predecir la aprobación presidencial. La diversidad de estos modelos permite una exploración exhaustiva de diferentes aspectos de los datos, desde relaciones lineales simples hasta interacciones complejas y no lineales, proporcionando así un panorama amplio de cómo diferentes enfoques pueden ser utilizados para abordar la tarea de predicción en un contexto político real.

Como se mencionó anteriormente, el modelo de regresión lineal se eligió como primer modelo de aprendizaje automático para este estudio. El primer paso consistió en definir una variable $\mathcal{X}$ con las distintas características extraídas de la conversación presente en los tuits, al mismo tiempo de establecer la variable $\boldsymbol{y}$ con el promedio mensual de la aprobación presidencial. Posteriormente se dividió el

conjunto de datos en el de entrenamiento y prueba mediante el módulo de *train_test_split* de scikit learn.

El siguiente paso fue ajustar el modelo *LinearRegression* a las variables de entrenamiento, lo cual implementa el aprendizaje computacional del modelo para poder posteriormente predecir los datos del conjunto de prueba. Una vez realizado lo anterior mediante las métricas de rendimiento de *mean_squared_error* y *r2_score* que también proporciona la biblioteca *scikit learn*, se procedió a comparar las predicciones del modelo con la variable y del conjunto de prueba.

En la tabla 3 se encuentran de forma agrupada los diferentes enfoques y herramientas utilizados en cada parte del proceso metodológico.

Proceso	Técnica/Enfoque Utilizado	Bibliotecas/Herramientas Empleadas
Unificación de base de datos	Manipulación de datos. Funciones iterativas para unir los archivos por día.	Pandas Funciones básicas de Python
Preprocesamiento y filtrado	Normalización de texto. Eliminación de usuarios y URL's. Eliminación de stopwords. Filtrado de palabras y unión de términos.	Unicodeta.normalize Expresiones regulares nltk.corpus str.contains; str.join

Ingeniería de características	Cantidad de tuits	Funciones aritméticas
	Diversidad léxica media	Función propia
	Longitud media	Funciones propia
	Legibilidad media	textstat.flesch_reading_ease
	Polaridad (modelo INEGI)	EvoMSA.DenseBoW (Atributo DataSet)
	Polaridad (modelo Tass2016)	EvoMSA.DenseBoW (Atributo DataSet)
	Representación de emociones con emojis	EvoMSA.DenseBoW (Atributo Emoji)
Entrenamiento del modelo	División del conjunto de entrenamiento y prueba	sklearn.model_selection (train_test_split)
	Entrenamiento y ajuste de los modelos	sklearn.linear_model (LinearRegression) sklearn.tree (DecisionTreeRegressor) sklearn.ensemble (RandomForestRegressor) (GradientBoostingRegressor) sklearn.svm (SVR) (xgboost)
	Medición del rendimiento del modelo ajustado	sklearn.metrics (*mean_squared_error; r2_score*)
	Eliminación Recursiva de características	sklearn.feature_selection (RFE)

Tabla 3. *Técnicas utilizadas en el proceso metodológico.* Fuente: Elaboración propia.

3.4 Ajuste de los modelos

El primer resultado con el modelo de regresión lineal no mostró valores prometedores para utilizarlo como modelo predictivo de datos nuevos, puesto que obtuvo un MSE de 15.88 y un valor negativo de -0.53 para la R^2. Por tal motivo en esta sección se muestra el ajuste de los hiperparámetros y las características de entrada empleadas para mejorar este modelo y servir como base para la estandarización de parámetros en los demás tipos de modelos entrenados.

Para lograr la optimización del modelo de regresión lineal, el primer ajuste del modelo se realizó con el método de Eliminación Recursiva de Características (RFE), el cual extrae las variables que más se correlacionan con la variable objetivo. Tras realizar este análisis se obtuvo que las mejores características para entrenar el modelo fueron:

- ‘diversidad_léxica’
- ‘inegi_neutro'
- 'tass2016_neutro'
- 'emocion_alegría 😁'
- 'emocion_tristeza 😔',
- 'emocion_enojo 😡'
- 'emocion_risa 😂'

Con este enfoque discriminatorio de las mejores características, se entrenó un nuevo modelo de regresión lineal, el cual resultó con un MSE de 9.70 y una R^2 de 0.49, consiguiendo de esta forma un valor positivo.

Para mejorar aún más el rendimiento, se aplicó otro método de selección de mejores características: la selección univariante (*SelectKBest).* Este método permite experimentar con la cantidad de características que se quieren obtener. Al probar con k=6 se obtuvo el mejor rendimiento del modelo de regresión lineal. Las mejores características seleccionadas mediante este método fueron:

- 'cantidad_tuits'
- 'inegi_neutro'

- 'longitud_media'
- 'emocion_risa 😂'
- 'tass2016_neutro'
- ‘inegi_negativo'

Al entrenar el modelo de regresión lineal con estas características y un tamaño del conjunto de prueba de 30% (test_size=0.3), se obtuvo un MSE de 5.89 y una R^2 de 0.71, mejorando bastante la explicación de la variabilidad para este modelo.

El modelo más óptimo que se probó fue el de árboles de decisión, el cual fue entrenado con todas las características y un conjunto de prueba también de 30%, obteniendo el mejor rendimiento de todos los modelos, con un MSE de 4.2 y una R^2 de 0.80.

3.5 Análisis de los resultados

Como se mencionó en la introducción, el valor deductivo de esta investigación fue la creación de su propio problema de estudio, al pasar de datos no estructurados a un problema de aprendizaje supervisado. La metodología descrita en la sección anterior, posibilita la estandarización de los parámetros de los modelos de aprendizaje automático implementados para la obtención de mejores valores y rendimientos a la hora de poner a prueba la generalización de los modelos.

A continuación se analizan los resultados estadísticos más significativos del conjunto de datos de características; posteriormente se muestran los valores de rendimiento de cada modelo predictivo implementado para complementar este análisis.

3.5.1 Análisis estadísticos de las mejores características

Una vez obtenido el conjunto de datos de características, a partir del procesamiento de información de los conjuntos mensuales de tuits, se realizó un análisis exploratorio de las características en cuanto a su correlación con la variable objetivo de aprobación presidencial (fig. 8).

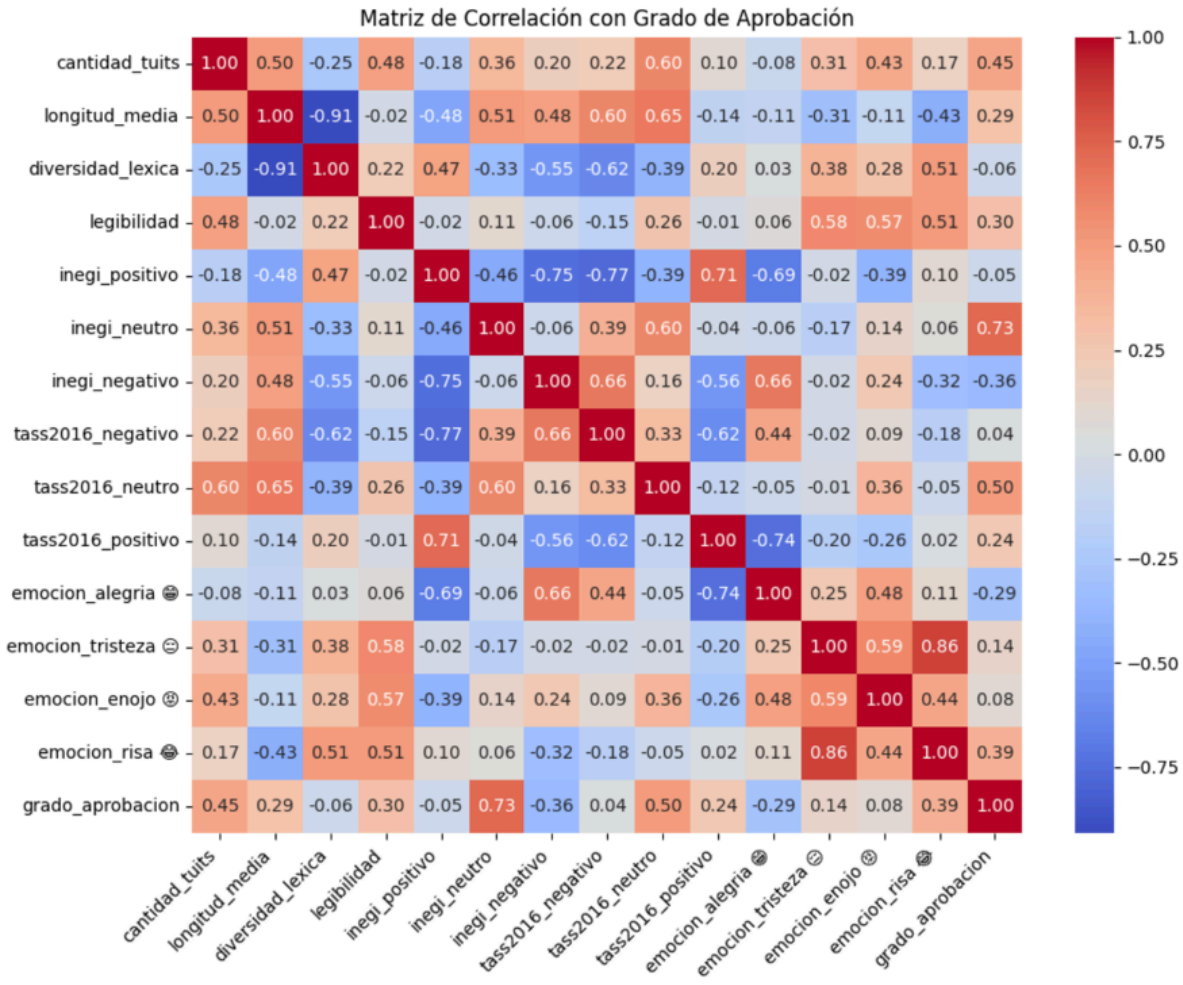

Figura 8. *Matriz de correlaciones con el grado de aprobación presidencial.*

Fuente: Elaboración propia.

El primer hallazgo importante es que existen 3 variables con una correlación positiva con la aprobación presidencial bien marcada, siendo la más predominante 'inegi_neutro' con un 0.73 de correlación, seguida de 'tass2016_neutro', con 0.50 de correlación y la cantidad de tuits, con un valor correlacional de 0.45.

Otro hallazgo refiere a las cuatro variables con correlación negativa, y cercanas a cero, las cuales son:

- 'emocion_alegría 😁'
- "inegi_negativo"

- “inegi_positivo”
- “diversidad léxica”

Para un análisis más detallado, en las figuras 9, 10 y 11 se presentan los diagramas de dispersión de las variables con mayor correlación positiva con la aprobación presidencial: “inegi_neutro”, “tass2016_neutro” y “cantidad de tuits”. Estas gráficas permiten observar la positividad de la correlación gracias a la visualización de la línea de regresión, que muestra una inclinación más pronunciada en estas tres variables independientes. En los diagramas 9 y 10, las unidades del eje x oscilan entre 0 y 1, representando la probabilidad promedio de que los datos mensuales se alineen más o menos con la polaridad correspondiente. Incluso una pequeña variación de tan solo 0.01 en estas probabilidades puede influir significativamente en la correlación con la aprobación presidencial, resaltando la sensibilidad de estas variables en el análisis.

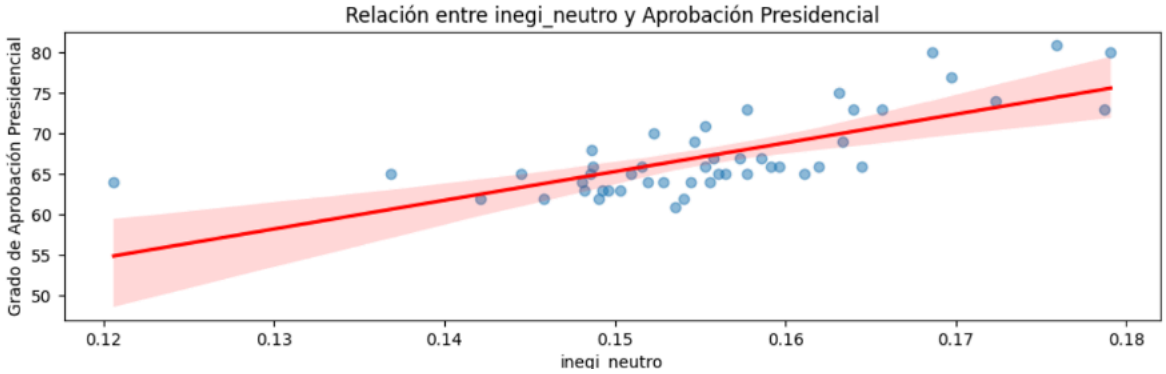

Figura 9. *Diagrama de dispersión de la variable “inegi_neutro” v.s aprobación presidencial.*

Fuente: Elaboración propia.

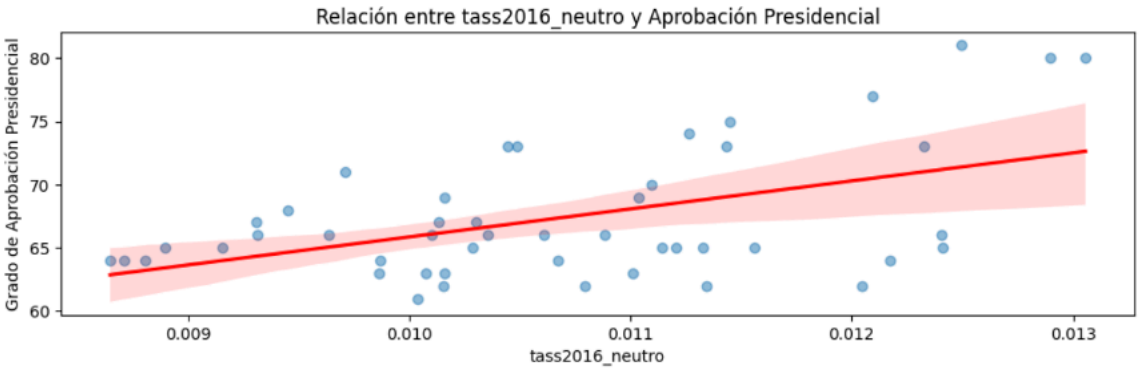

Figura 10. *Diagrama de dispersión de variable "tass2016_neutro" v.s aprobación presidencial.*

Fuente: Elaboración propia.

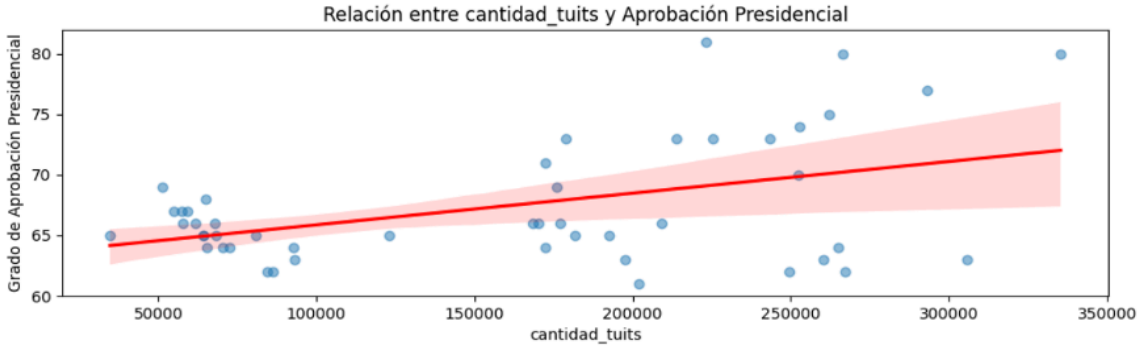

Figura 11. *Diagrama de dispersión de la cantidad de tuits v.s aprobación presidencial.*

Fuente: Elaboración propia.

Además del diagrama de dispersión con la línea de regresión, en la figura 12 se compara la serie de tiempo de las variable con mayor correlación a la aprobación presidencial ("inegi_neutro"), lo cual es posible gracias al formato *date.time* de la columna "fecha" en el conjunto de datos de características. El eje "y" de esta figura, de igual forma oscila entre 0 y 1 como la probabilidad de que el conjunto de tuits de cada mes tienda a la neutralidad en su estilo de lenguaje.

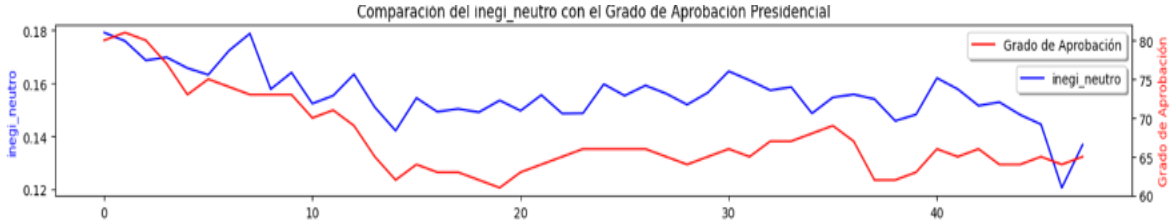

Figura 12. *Serie de tiempo entre "inegi_neutro " y aprobación presidencial.*

Fuente: Elaboración propia.

3.5.2 Análisis de rendimiento por modelo

A continuación, en la Tabla 4 se presentan los mejores resultados obtenidos por tipo de modelo, evaluados a través de las dos métricas de rendimiento seleccionadas: el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2). Los mejores rendimientos fueron recopilados tras realizar un total de 24 entrenamientos, considerando cuatro variaciones metodológicas aplicadas a cada uno de los seis modelos analizados.

La primera variante utilizó las 14 características descritas previamente en la Tabla 2. En la segunda variante, se emplearon todas las características, pero los datos fueron estandarizados mediante la herramienta *StandardScaler* de Python. La tercera variante aplicó un proceso de selección univariante de características mediante la herramienta *SelectKBest*, seleccionando las 6 características más significativas en relación con la aprobación presidencial. Finalmente, la cuarta variante consistió en entrenar el modelo utilizando las 6 características seleccionadas, pero con los datos estandarizados.

Adicionalmente, las figuras 13 y 14 proporcionan una comparación gráfica de los resultados para ambas métricas, permitiendo un análisis visual del desempeño de cada modelo.

Modelo	Mejor MSE	Mejor R^2
Regresión lineal	5.89	0.71
Árbol de decisión	4.2	0.80
Random Forest	5.57	0.73
SVR	14.61	0.30
GBM	8.43	0.59
XGBoost	6.865	0.67

Tabla 4. *Registro estadísticos de métricas por cada tipo de modelo.*

Fuente: Elaboración propia.

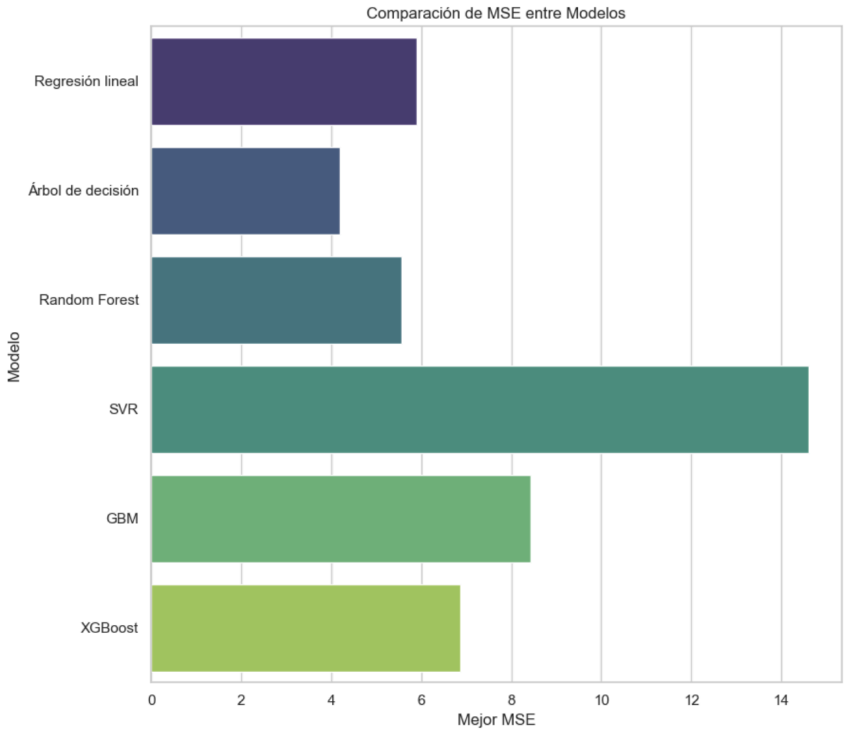

Figura 13. *Comparación visual del mejor MSE obtenido por cada modelo.* Fuente: Elaboración propia.

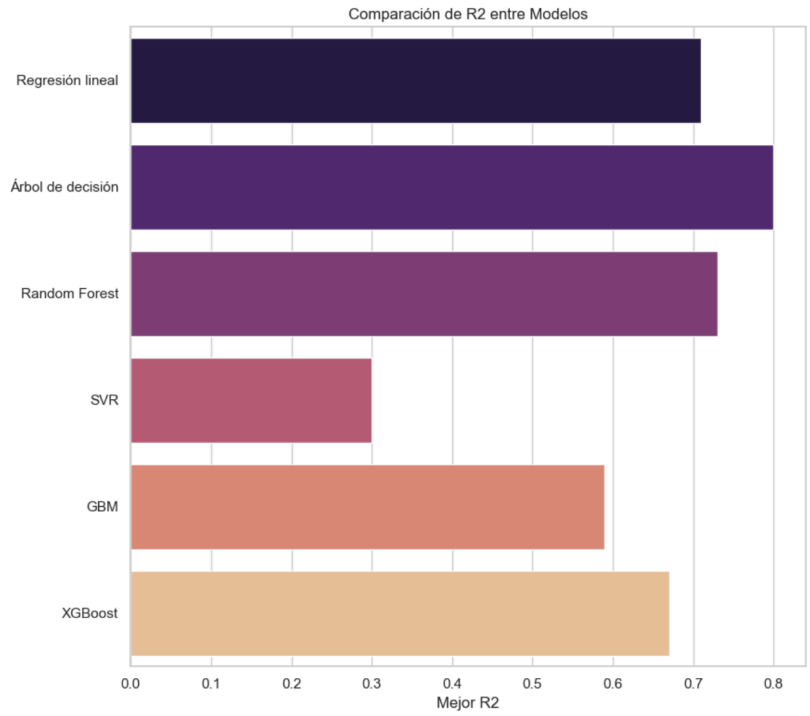

Figura 14. *Comparación visual de la mejor métrica de R^2 obtenida por cada modelo.* Fuente: Elaboración propia.

La regresión lineal mostró un desempeño notablemente sólido, con una R^2 de 0.71, indicando que puede explicar el 71% de la variabilidad en la aprobación presidencial. Este modelo es apreciado por su simplicidad y la interpretabilidad directa de sus resultados, lo que lo hace particularmente útil para inferencias estadísticas en contextos donde la relación lineal es adecuada, como en este caso.

El árbol de decisión superó a la regresión lineal con el MSE más bajo y la mejor R^2 de todos los modelos evaluados (80%), destacando su capacidad para manejar complejidades no lineales y su habilidad para capturar interacciones entre características sin necesidad de transformaciones específicas de los datos.

Como un modelo de ensamble basado en árboles de decisión, Random Forest mejoró la generalización sobre el árbol de decisión único, ofreciendo un rendimiento robusto y evitando el sobreajuste, característico de modelos más simples. Su R^2 sugiere que es capaz de explicar el 73% de la variabilidad, una mejora respecto a muchos modelos básicos.

El SVR no logró rendimientos comparables a los modelos basados en árboles, reflejando limitaciones posiblemente relacionadas con la selección de parámetros o la necesidad de kernels más efectivos para capturar la dinámica subyacente de los datos. Su bajo valor de R^2 indica una capacidad limitada para predecir con precisión la aprobación presidencial.

El GBM, a pesar de ser un método potente para enfrentar diversas problemáticas predictivas, se quedó corto en este análisis, con un rendimiento medio en comparación con los otros modelos. Esto puede atribuirse a la necesidad de ajustes más finos en sus parámetros y posiblemente una mejor adaptación de las características utilizadas.

El modelo XGBoost, reconocido por su destacada eficacia en competiciones de modelado predictivo, obtuvo buenos resultados, aunque no superó el rendimiento del árbol de decisión. No obstante, su coeficiente de determinación R^2 de 0.67 evidencia su capacidad para manejar grandes conjuntos

de datos y abordar complejidades inherentes a ellos, especialmente cuando se aplica una afinación adecuada de hiper-parámetros.

Cada modelo presenta fortalezas y debilidades que se reflejan en su desempeño para este conjunto de datos de características extraídas de los tuits. El Árbol de Decisión sobresalió al lograr el coeficiente de determinación R^2 más alto, lo que indica que su capacidad para capturar relaciones complejas resulta especialmente adecuada en este contexto. No obstante, técnicas como Random Forest y XGBoost también arrojaron resultados prometedores, sugiriendo que los métodos de ensemble pueden ser particularmente eficaces para esta tarea. Por otro lado, modelos como SVR y GBM requieren un ajuste más minucioso y, posiblemente, una revisión en la selección de características para optimizar su rendimiento.

3.5.3 Aplicación del modelo y predicción de la aprobación presidencial

En esta sección se realiza el análisis de resultados de cada modelo entrenado a la hora de predecir los valores de aprobación presidencial, con entradas nuevas de datos de los meses de enero 2023 y febrero 2023, últimos conjuntos de tuits que se obtuvieron para esta investigación.

Para analizar los resultados de las predicciones de los diferentes modelos en comparación con los números reales proporcionados por Oraculus para enero y febrero de 2023, podemos evaluar cada modelo en términos de su precisión y consistencia en el acercamiento a los valores reales. Estos análisis pueden ayudar a identificar qué modelo se comporta mejor en el contexto de predecir la aprobación presidencial.

Se observan resultados mixtos en cuanto a la precisión y consistencia entre los modelos para enero y febrero de 2023. El valor real de aprobación presidencial reportado por Oraculus fue de 64% para enero y 66% para febrero, proporcionando una base sólida para comparar las predicciones de cada modelo.

La regresión lineal, con predicciones de 68.91% en enero y 64.87% en febrero, mostró una tendencia a sobreestimar en enero y a subestimar ligeramente

en febrero. A pesar de no ser el más preciso, el modelo demostró una variabilidad relativamente baja en sus errores entre los dos meses, sugiriendo una cierta consistencia en su rendimiento.

El Árbol de Decisión produjo resultados más cercanos al valor real, con una predicción de 64.0% en enero y 64.78% en febrero, manteniéndose consistentemente cerca del valor real y mostrando una ligera subestimación. Este comportamiento indica una habilidad del modelo para manejar bien las interacciones no lineales entre las características.

Por su parte, Random Forest, con 65.54% en enero y 64.73% en febrero, también mantuvo una consistencia en sus predicciones, similar al árbol de decisión pero con una ligera tendencia general a la subestimación, especialmente notable en febrero donde se alineó casi perfectamente con el valor real de Oraculus.

El modelo SVR tuvo un rendimiento más variado, con una predicción de 65.92% en enero y 63.89% en febrero, mostrando una sobreestimación en el primer mes seguida de una subestimación más significativa en el segundo. Esta fluctuación destaca una posible sensibilidad a variaciones mensuales en los datos que podrían afectar su estabilidad y precisión. El modelo de GBM, por otro lado, registró predicciones de 65.30% en enero y 64.05% en febrero, acercándose al valor real con una leve subestimación en febrero, indicando un rendimiento razonable aunque no excepcional.

Finalmente, XGBoost mostró mejor rendimiento en enero con un 67.86%, pero experimentó una caída significativa en febrero a 63.86%, lo que podría reflejar una sensibilidad a cambios en el conjunto de datos entre los meses. Este comportamiento sugiere que, aunque XGBoost puede ser capaz de capturar bien la dinámica en ciertos periodos, podría requerir ajustes o reevaluación de sus parámetros para mantener la consistencia.

En la tabla 5 se observa la recopilación de los valores predichos por cada tipo de modelo para cada mes de datos nuevos, en este caso enero y febrero del

2023. Posteriormente, en las figuras 15 y 16 se muestra de forma visual las diferencias de cada modelo con el valor real medido por la plataforma *Oraculus*.

Modelo	Enero 2023 (Oraculus= 64%)	Febrero 2023 (Oraculus = 66%)
Regresión lineal	68.91	64.87
Árbol de decisión	64.0	64.78
Random Forest	65.54	64.73
SVR	65.92	63.89
GBM	65.30	64.05
XGBoost	67.86	63.86

Tabla 5. *Comparación de las predicciones por modelo con datos nuevos.*
Fuente: Elaboración propia.

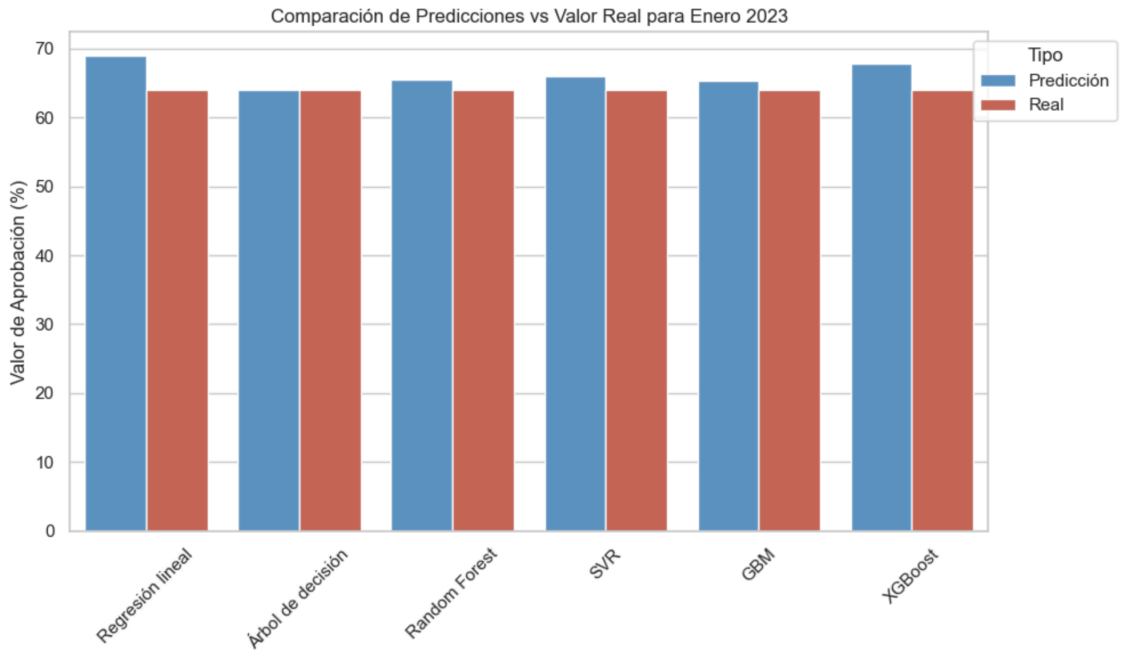

Figura 15. *Comparación visual de la predicción por modelo v.s. el valor real en enero de 2023.*
Fuente: Elaboración propia.

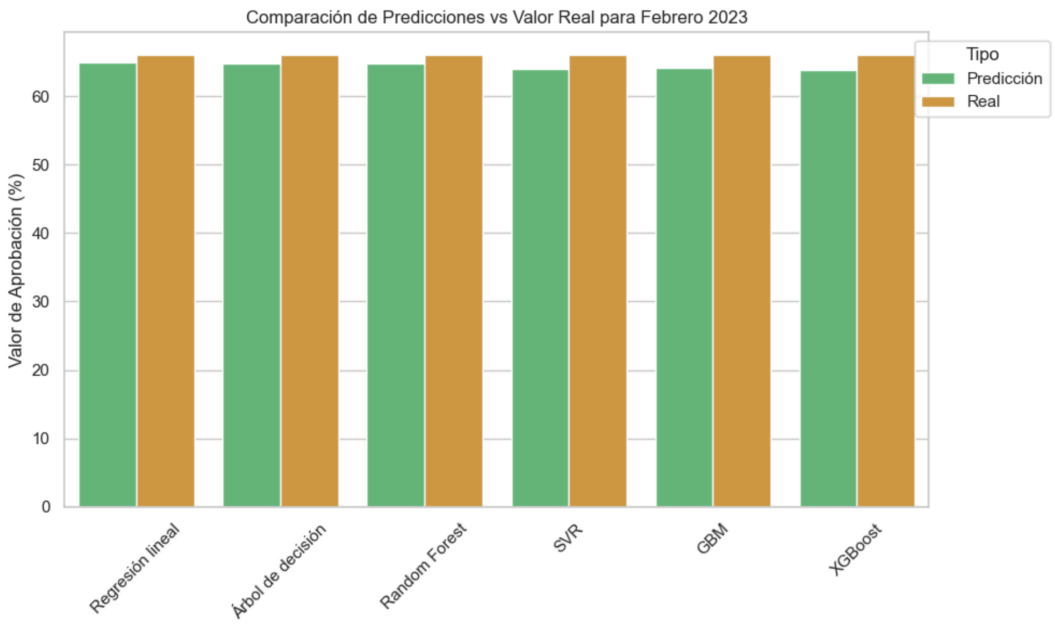

Figura 16. *Comparación visual de la predicción por modelo v.s. el valor real en febrero de 2023.*
Fuente: Elaboración propia.

A la hora de la aplicación con nuevos datos desconocidos, los modelos de Árbol de Decisión y *Random Forest* demostraron ser los más consistentes y precisos en general, con predicciones cercanas a los valores reales y una baja variabilidad entre meses. Estos modelos se observan como los preferibles por su estabilidad y capacidad para manejar la variabilidad en los datos de aprobación presidencial.

Por otro lado, modelos como SVR y XGBoost, aunque mostraron buenos resultados en algunos casos, presentaron mayor variabilidad en sus errores, lo que puede requerir un examen más profundo de sus configuraciones o una adaptación más detallada a las características específicas del conjunto de datos.

Esta evaluación sugiere la importancia de no solo considerar la precisión media de las predicciones de un modelo, sino también su consistencia y fiabilidad a lo largo del tiempo, especialmente en aplicaciones prácticas como la predicción de tendencias políticas donde la precisión y la estabilidad son cruciales.

3.6 Discusión de los resultados

En el análisis de la sección previa sobre el desempeño de los modelos de aprendizaje automático utilizados para predecir la aprobación presidencial, se identificaron diferencias significativas en su capacidad para procesar los datos de los tuits y generar predicciones efectivas. La comparación mostró que algunos modelos lograron manejar con éxito la complejidad y no linealidad de los datos, mientras que otros enfrentaron limitaciones que impactaron tanto su precisión como su consistencia.

El modelo de Árbol de Decisión y el modelo Random Forest se destacaron por su capacidad para integrar y procesar la variedad y complejidad de los datos, reflejando no solo una alta precisión sino también una notable estabilidad en sus predicciones a través del tiempo. Esto se debe posiblemente a la habilidad de estos modelos para capturar relaciones no lineales sin la necesidad de transformaciones previas del conjunto de datos, permitiendo una interpretación directa de cómo las características influyen en las predicciones.

Por otro lado, modelos como SVR y XGBoost, aunque capaces de manejar grandes volúmenes de datos y complejidades inherentes a estos, mostraron variabilidad en su desempeño, lo que sugiere que podrían beneficiarse de un ajuste más fino en la selección y ponderación de características, así como en la configuración de sus parámetros.

En el contexto de este estudio, un aspecto crucial para la evaluación de los modelos de aprendizaje automático es el volumen de datos utilizados para el entrenamiento y la prueba. Aunque el conjunto de datos original consistió en millones de tuits recolectados a lo largo de varios años, la transformación de estos datos en un formato estructurado para el análisis resultó en solo 48 entradas mensuales. Esta limitación en el número de entradas de datos para el aprendizaje y la prueba, tiene implicaciones significativas en la capacidad de aprendizaje y generalización de los modelos.

Con un conjunto limitado de datos, existe un riesgo significativo de sobreajuste, donde los modelos aprenden a predecir perfectamente según los datos de entrenamiento, pero fallan al generalizar a nuevos datos. Esto es particularmente problemático para modelos complejos como XGBoost y SVR, que son capaces de aprender de grandes conjuntos de datos pero pueden ajustarse excesivamente a las peculiaridades de un conjunto de datos pequeño, afectando su estabilidad y confiabilidad en las predicciones.

Continuando con la discusión de resultados, un hallazgo interesante en el análisis fue el papel predominante de la variable "inegi_neutro" en la correlación con la aprobación presidencial. Esta variable, que mide la neutralidad en los tuits, puede interpretarse como un indicador del tono general de la discusión pública sobre el presidente. Una posible explicación de su alta correlación con la aprobación presidencial podría ser que un tono más neutral en las discusiones dentro de X refleja una polarización menos aguda entre los partidarios y los detractores del presidente.

En contextos políticos, donde la polarización puede ser alta, una mayor neutralidad podría indicar una aceptación más generalizada del liderazgo del presidente, lo que a su vez podría correlacionarse con índices de aprobación más altos.

Los resultados del estudio destacan la importancia de elegir y ponderar cuidadosamente las características utilizadas en los modelos de predicción. Además, subrayan la necesidad de comprender el contexto social y político en el que se generan los datos, ya que este contexto puede influir significativamente en las dinámicas que los modelos buscan capturar. Por ejemplo, en períodos de elecciones, manifestaciones importantes o eventos políticos significativos, el volumen y el tono de las conversaciones en plataformas sociales pueden cambiar drásticamente. Estos cambios pueden afectar la polaridad de los comentarios, haciendo que los modelos que no ajustan por estos factores contextuales puedan interpretar erróneamente las emociones o sentimientos generalizados.

En términos prácticos, ignorar el contexto y la dinámica de los datos puede llevar a interpretaciones erróneas y a tomas de decisión basadas en información inexacta. Por ejemplo, un modelo que no ajusta adecuadamente por cambios contextuales, puede sobreestimar o subestimar la aprobación presidencial, lo que podría influir negativamente en las decisiones políticas o las estrategias de campaña.

Tras esta discusión de resultados, se llega a la conclusión que comprender el contexto y la dinámica de los datos no solo mejora la precisión de los modelos de predicción, sino que también enriquece la interpretación de los resultados, asegurando que las conclusiones y acciones basadas en estos análisis sean fundadas y relevantes, abonando a que cuando se aplica la ciencia de datos a contextos complejos como la política, la interacción entre la técnica y la teoría es esencial para el desarrollo de soluciones efectivas y confiables.

Conclusiones y recomendaciones

Conclusiones y recomendaciones

Esta investigación ha proporcionado una visión detallada del uso de diversos modelos de aprendizaje automático para predecir la aprobación presidencial, utilizando un conjunto de datos derivado de tuits relacionados con la política mexicana. A través de técnicas de PLN y una ingeniería de características, se ha logrado transformar datos no estructurados de redes sociales, en un formato estructurado que sirva como características de entrada de los modelos predictivos. A continuación, se resumen los hallazgos más importantes y las conclusiones derivadas de este trabajo, así como las recomendaciones a futuros estudios en el campo.

A nivel de los resultados obtenidos en la implementación de los diferentes modelos predictivos, los modelos de Árbol de Decisión y Random Forest demostraron ser los más efectivos en predecir la aprobación presidencial. El Árbol de Decisión, en particular, logró el mejor desempeño con el menor MSE y el mayor valor de R^2, explicando hasta el 80% de la variabilidad de los datos, y destacando su capacidad para manejar la complejidad y las interacciones no lineales entre características.

De esta manera, el enfoque metodológico adoptado en este estudio subraya la importancia fundamental de una selección cuidadosa de las características para optimizar el desempeño de los modelos predictivos. La elección de las variables correctas para incluir en el análisis no solo influye en la precisión de las predicciones, sino que también determina la eficacia general del modelo en capturar y reflejar las dinámicas subyacentes de los datos.

Variables como "inegi_neutro" o “tass2016_neutro” demostraron tener una fuerte correlación con la aprobación presidencial, posiblemente reflejando la neutralidad en los discursos públicos que puede indicar una menor polarización. Este hallazgo resalta la necesidad de entender profundamente el contexto y la semántica de los datos utilizados para la formación de modelos predictivos.

A nivel teórico, esta investigación demuestra el potencial de la ciencia de datos aplicada al análisis político para proporcionar insights valiosos y oportunos sobre la opinión pública. Las técnicas y modelos explorados aquí no solo avanzan en nuestra comprensión de cómo los datos de redes sociales pueden ser utilizados para la predicción política, sino que también establecen un marco para futuros trabajos en este campo emergente y dinámico. Comprender los elementos contextuales de lo social permite a los científicos de datos ajustar sus modelos de manera más precisa, logrando reflejar las dinámicas subyacentes de los datos en lugar de limitarse a reaccionar ante fluctuaciones superficiales.

A nivel práctico, este estudio enfatiza la eficacia del marco de trabajo propuesto para realizar análisis comprensivos y viables del contenido de tuits, particularmente en contextos políticos. Este estudio abona a lo demostrado en las investigaciones previas revisadas, las cuales indican que mediante el uso de herramientas avanzadas de análisis de texto y de sentimientos, es posible extraer y sintetizar información valiosa de grandes volúmenes de datos no estructurados, ofreciendo una base sólida para la toma de decisiones informadas y la formulación de estrategias.

Una ventaja significativa de la metodología implementada en esta investigación es su adaptabilidad y escalabilidad, al poder aplicarse a diferentes conjuntos de datos y ajustarse para responder a diversos objetivos o necesidades prácticas, como es el seguimiento de cambios en la opinión pública. Así, los modelos desarrollados en este estudio ofrecen una herramienta prometedora para monitorizar la opinión pública en tiempo real, proporcionando una alternativa más rápida y económicamente viable a las encuestas tradicionales. Esta capacidad es especialmente valiosa en entornos dinámicos y rápidamente cambiantes como el político, donde comprender la opinión pública de manera oportuna es crucial para la toma de decisiones.

Futuras investigaciones podrían explorar la integración de modelos más avanzados de PLN y técnicas de Aprendizaje Profundo para mejorar la comprensión y el análisis de los datos. Además, expandir el estudio para incluir

otras regiones geográficas o diferentes contextos políticos podría proporcionar *insights* adicionales sobre la universalidad y adaptabilidad de los modelos desarrollados.

Dada la limitación observada en este estudio respecto al volumen de datos estructurados derivados de millones de tuits para análisis, se recomienda para futuros trabajos incrementar significativamente el número de entradas de datos para entrenar y probar los modelos de aprendizaje automático, mejorando así su capacidad de aprendizaje y generalización. Una estrategia efectiva sería aumentar la granularidad de los datos, considerando una segmentación temporal más fina como semanal o diaria en lugar de mensual, lo que permitiría capturar variaciones más sutiles en la opinión pública y aumentar el número de puntos de datos disponibles.

También se recomienda enriquecer la ingeniería de características para incluir metadatos del tuit, como ubicación, hora de publicación y datos demográficos del usuario, junto con otras características textuales avanzadas, lo cual podría ofrecer una matriz de información más completa. Asimismo, combinar datos de múltiples plataformas sociales o fuentes de noticias podría proporcionar un conjunto de datos más robusto y diversificado.

De igual forma, extender el período durante el cual se recopilan los datos puede proporcionar una muestra más representativa y diversa, especialmente importante para capturar tendencias a largo plazo y efectos estacionales en la opinión pública.

Incorporar un modelo que pueda manejar múltiples idiomas también es recomendable, especialmente dado el contexto global y la diversidad lingüística de las poblaciones. Esto permitirá una comprensión más profunda y una aplicación más amplia de los modelos en diferentes contextos geográficos y culturales. Al implementar estas mejoras y estrategias descritas, futuros estudios podrían superar las limitaciones encontradas en este estudio y desarrollar modelos de aprendizaje automático más precisos y capaces de generalizar a partir de un conjunto de datos más representativo.

Finalmente, para que las herramientas predictivas sean útiles y confiables en el análisis político, es fundamental impulsar la educación y la transparencia en su uso. Esto implica informar al público sobre el funcionamiento de estos modelos, sus posibles aplicaciones y sus limitaciones. Al hacerlo, se puede generar mayor comprensión y confianza en estas herramientas, lo que a su vez fomentará debates más informados y constructivos sobre su impacto en la sociedad. La combinación de mejoras técnicas en el desarrollo de estos modelos y un enfoque centrado en las necesidades de los usuarios proporciona una base sólida para su uso futuro en el ámbito político. De esta manera, se puede garantizar que sean tanto científicamente rigurosos como socialmente responsables.

Fuentes de consulta

Ali, H., Farman, H., Yar, H., Khan, Z., Habib, S., & Ammar, A. (2021). Deep learning-based election results prediction using Twitter activity. *Soft Computing, 26*(16), 7535-7543. https://doi.org/gpz3s8

Ansari, Z. (2020). Analysis of Political Sentiment Orientations on Twitter. *Procedia Computer Science, 167*, 1821–1828.

Apablaza, C. & Jiménez, F. (2009). Factores Explicativos de la Aprobación Presidencial. *Revista Libertad y Desarrollo*

Ascolano Ruiz, F., Cazorla Quevedo, M. A., Alfonso, M. I., Colomina Pardo, O. & Lozano Ortega, M. A. (2003). Inteligencia artificial: modelos, técnicas y áreas de aplicación. Editorial Paraninfo.

Atefeh, A., & Inkpen, D. (2022). *Natural Language Processing for Social Media.* Morgan & Claypool Publishers

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications.* John Wiley & Sons

Bengio, Y., LeCun, Y., & Hinton, G. (2006). Deep Learning. *Nature*, *437*(7065), 1072-1074.

Blumenthal, M., & Klarner, C. (2018). *Polling on the presidency: A handbook of presidential election polls and presidential approval ratings.* Routledge

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Cambria, E., & Hussain, A. (2016). *Opinion mining and sentiment analysis*. Cambridge University Press.

Chaves-Montero, A. (ed.) (2017). *Comunicación Política y Redes Sociales.* España: Ediciones Egregius

Douglas, B. (1995). CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33–38.

Edward, A. (2021). An Extensive Guide to collecting tweets from Twitter API v2 for academic research using Python 3. *Medium*. *Towards Data Science* [En línea]

Ferg, R., Conrad, F., & Gagnon-Bartsch, J. (2021). A Critical Evaluation of Tracking Public Opinion with Social Media: A Case Study in Presidential Approval. *Methods, Data, Analyses, 15*(2), 215-240. https://doi.org/gpz568

Firth, J R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society).*

Friedman, J. H. (2001). Gradient Boosting for Regression and Classification. *Journal of Statistical Software*, *38*(1), 1-24.

Gandhi, U. D., Kumar, P. M., Babu, G. C., & Karthick, G. S. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). *Wireless Personal Communications.* https://doi.org/gj3g55

García, S., Ramírez-Gallego, S., Luengo, J., & Herrera, F. C. (2016). Big Data: Preprocesamiento y calidad de datos. Novática, 237, 17-23.

Géron, A. (2019). *Introducción al aprendizaje automático con scikit-learn, Python y TensorFlow* (2ª Ed.). Vicens Vives.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Graff, M., Miranda-Jiménez, S., Téllez, E., & Moctezuma, D. (2020). EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis [Application Notes]. *IEEE Computational Intelligence Magazine, 15*(1), 76-88. https://doi.org/mmnm

Graff, M., Moctezuma, D., Miranda-Jiménez, S., & Tellez, E. S. (2022). A Python library for exploratory data analysis on twitter data based on tokens and aggregated origin–destination information. *Computers & Geosciences, 159, 105012.* https://doi.org/m9nz

Gualda, E. (2022). Social big data, sociología y ciencias sociales computacionales. *Empiria: Revista de metodología de ciencias sociales (53)*, 147-177.

Ketkar, N. (2017). *Deep Learning with Python: A Hands-on Introduction.* Apress.

Khurana Batra, P. (2020). Election result prediction using twitter sentiments analysis. *6th International Conference on Parallel, Distributed and Grid Computing,* 182–185

Kumar, S., & Soman, K. (2019). Deep Learning Based Part-of-Speech Tagging for Malayalam Twitter Data (Special Issue: Deep Learning Techniques for Natural Language Processing). *Journal Of Intelligent Systems, 28*(3), 423-435. https://doi.org/mhq2

Lloyd, S. (2022). *Social Media as Social Science Data.* Cambridge University Press.

Lovera, F., & Cardinale, Y. (2023). Análisis de sentimientos en Twitter: Un estudio comparativo. *Revista Científica de Sistemas e Informática, 3*(1). https://doi.org/mg67

Martínez, J. (2017). *Minería de opiniones mediante análisis de sentimientos y extracción de conceptos en Twitter.* [Tesis doctoral, Universidad Complutense de Madrid]

McCallum, Q. E. (2012). *Bad data handbook: cleaning up the data so you can get back to work*. O'Reilly Media, Inc

Meruane, O. S., & Balin, D. L. (2012). Descripción de las formas de justificación de los objetivos en artículos de investigación en español de seis áreas científicas. *Onomázein, (25)*, 315-344.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (1ª Ed.). MIT Press.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data, 1*(1), 51-59

Quintana, A. (2008). *Planteamiento del problema: errores de la lectura superficial de libros de texto de metodología. Revista de investigación en psicología, 11*(1), 239-253.

Scott, Z. (s.f.). *Working with RNNs*. Recuperado de:

https://www.tensorflow.org/guide/keras/working_with_rnns

Solis, L. (2022). Aprobación del presidente de Perú basado en análisis de sentimientos en Twitter. *TECHNO Review, 11*(1), 1-13. https://doi.org/mg63

Shaghaghi, N. (2021). Twitter Sentiment Analysis and Political Approval Ratings for Situational Awareness. *Proceedings - 2021 IEEE International Conference on Cognitive and Computational Aspects of Situation Management*, 59–65.

Strong, S., & Kohli, I. S. (2019). Approval Ratings and Predicting United States Presidential Elections. *Social Science Research Network*. https://doi.org/mg64

Tellez, E., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O., &

TensorFlow documentation. (s.f.). Understanding masking & padding. Recuperado de: https://www.tensorflow.org/guide/keras/understanding_masking_and_padding

Vidalón, J., Solis, L., Porras, E., Lagos, M., & Tinoco, E. C. (2022). Approval Rating of Peruvian Politicians and Policies using Sentiment Analysis on Twitter. *International Journal Of Advanced Computer Science And Applications*, *13*(6). https://doi.org/mhbm

Villaseñor, E. (2017). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems With Applications, 81,* 457-471. https://bit.ly/3A4a7w5

Yavaria, A., Hassanpour, H., Rahimpour Camib, B., & Mahdavi, M. (2022). Election Prediction Based on Sentiment Analysis using Twitter Data. *International Journal Of Engineering. Transactions B: Applications, 35*(2), 372-379. https://doi.org/mg62

Zumárraga, M. (2022). Uso político de redes sociales y su efecto sobre la participación política offline: un análisis de mecanismos mediadores. *Revista Aposta (92)*, 64 - 86. https://bit.ly/3ZaIyLc