

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“ANÁLISIS VISUAL DE SENTIMIENTOS EN TWITTER Y SU CORRELACIÓN CON LOS RESULTADOS ELECTORALES EN LOS ÚLTIMOS 10 AÑOS”

JAIME ANTONIO SCHWEIR MENESES

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: José Luis Martí Lara
Profesor Correferente: Mauricio Olivares Faúndez

Junio - 2024

AGRADECIMIENTOS

Quisiera expresar mi agradecimiento a todas las personas que me ayudaron durante mi etapa universitaria, quienes fueron un pilar fundamental a lo largo de todo este tiempo.

En primer lugar, quiero agradecer a mi familia. A mis padres, quienes me han apoyado incondicionalmente y confiaron en mí desde el comienzo hasta el día de hoy. Sin su apoyo y sacrificio para que nunca me falte algo, no hubiera sido posible llegar hasta aquí. A mis hermanas pequeñas, que siempre se preocuparon por mí, llevándome comida mientras estudiaba y ayudándome en todo lo que podían. Y a mi perrita Luna, quien me acompañó durante mis largas noches de estudio.

También quiero agradecer a mis amigos, con quienes formé un grupo humano increíble durante todo este tiempo en la universidad. Ellos me ayudaron a sobrellevar esta etapa, en donde pasamos tardes de estudio, noches de juegos y nos apoyabamos mutuamente en todas las complicaciones que se nos ponían en frente.

Finalmente, quiero expresar mi gratitud a todos los profesores que contribuyeron a mi formación. En especial, agradezco a mi profesor guía, José Luis Martí, por su labor durante toda mi estadía en la universidad, su apoyo constante, su paciencia y su disposición para guiarme durante en este último proceso.

A todos ellos, y a todos los que de alguna manera participaron en mi formación como Ingeniero, les extiendo mi más profundo agradecimiento. Muchas gracias.

RESUMEN

Resumen— Este trabajo de memoria se enfoca en el análisis de tweets relacionados con ciertas elecciones para identificar una posible correlación con los resultados oficiales. Para ello, se emplearon algoritmos de clasificación, modelos de procesamiento del lenguaje natural y técnicas de visualización de datos. Los hallazgos indican que, aunque no se obtienen resultados exactos, los márgenes de error son aceptables y las tendencias de los tweets reflejan adecuadamente los resultados oficiales, sugiriendo una correlación. La relevancia de este estudio radica en su potencial para desarrollar herramientas que puedan mejorar las estrategias de comunicación política y ofrecer predicciones más precisas en futuros escenarios electorales.

Palabras Clave— Análisis de tweets, elecciones, procesamiento del lenguaje natural, algoritmos de clasificación, visualización de datos.

ABSTRACT

Abstract— This thesis work focuses on the analysis of tweets related to certain elections to identify a possible correlation with the official results. For this purpose, classification algorithms, natural language processing models, and data visualization techniques were employed. The findings indicate that, although exact results are not obtained, the error margins are acceptable, and the trends in the tweets adequately reflect the official results, suggesting a correlation. The relevance of this study lies in its potential to develop tools that can improve political communication strategies and provide more accurate predictions in future electoral scenarios.

Keywords— Analysis of tweets, elections, natural language processing, classification algorithms, data visualization.

GLOSARIO

API: *Application Programming Interface*.

BERT: *Bidirectional Encoder Representations from Transformers*.

CRISP-DM: *Cross-Industry Standard Process for Data Mining*.

CSV: *comma-separated values*.

GPT: *Generative Pretrained Transformer*.

IDF: *Inverse Document Frequency*.

KDD: *Knowledge Discovery in Databases*.

KPI: *Key Performance Indicators*.

ML: *Machine Learning*.

NLP: *Natural Language Processing*.

SEMMA: *Sample, Explore, Modify, Model, Assess*.

SW: *Sentiment Weighting*.

TF: *Term Frequency*.

UTFSM: *Universidad Técnica Federico Santa María*.

WW: *Word Weighting*.

ÍNDICE DE CONTENIDOS

RESUMEN	II
ABSTRACT	II
GLOSARIO	III
ÍNDICE DE FIGURAS	VI
ÍNDICE DE TABLAS	VI
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	2
1.1 Contexto	2
1.2 Situación Actual	3
1.3 Encuestas tradicionales	3
1.4 Objetivos	5
1.4.1 Objetivo General	5
1.4.2 Objetivos Específicos	5
CAPÍTULO 2: MARCO CONCEPTUAL	6
2.1 Las redes sociales	6
2.1.1 Historia de las redes sociales	6
2.1.2 Influencia de las Redes sociales	7
2.1.3 Consecuencias de las redes sociales	8
2.1.4 Twitter	9
2.2 Analisis de sentimientos	11
2.2.1 Definición de analisis de sentimientos	11
2.2.2 Descripción de las técnicas y tecnologías	11
2.2.3 Niveles de Análisis de Sentimientos	12
2.2.4 Análisis de sentimiento en Python	13
2.2.5 KPI's en el analisis de sentimiento	15
2.2.6 Estudios previos	16
2.3 Metodología para Análisis de Datos	17
2.3.1 KDD	17
2.3.2 SEMMA	19
2.3.3 CRISP-DM	20
2.4 Visualización de datos	22
2.4.1 Tableau	22
2.4.2 Flourish	23
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN	25
3.1 Metodología seleccionada	25

3.2	Desarrollo de la Solución Propuesta	26
3.2.1	Selección de datos	26
3.2.2	Preprocesamiento de datos	28
3.2.3	Transformación de Datos	29
3.2.4	Minería de Datos	30
CAPÍTULO 4: VISUALIZACION Y ANÁLISIS DE RESULTADOS		34
4.1	Minería de datos	34
4.1.1	Ánálisis de los modelos clasificadores de sentimiento	34
4.1.2	Análisis de resultados por algoritmo de análisis de sentimiento	39
4.1.3	Elecciones presidenciales 2017	42
4.1.4	Proceso constitucional 2020	49
4.1.5	Elecciones presidenciales 2021	54
4.1.6	Elecciones constitucionales 2022	61
4.2	Evaluación e interpretación	67
4.2.1	Resultados	67
4.3	Consolidación del Conocimiento	68
4.3.1	Caso de estudio: Elecciones constitucionales 2023	68
4.4	Conclusiones generales	73
CONCLUSIÓN		76
REFERENCIAS BIBLIOGRÁFICAS		78

ÍNDICE DE FIGURAS

1	Cantidad de cuentas de twitter por el mundo en el 2012	9
2	Proceso de Knowledge Discovery in Databases (KDD).	18
3	Proceso de Sample, Explore, Modify, Model and Assess (SEMMA).	19
4	Proceso de Cross-Industry Standard Process for Data Mining (CRISP-DM). . .	21
5	Ejemplo de archivo obtenido tras la extracción de datos.	28
6	Estructura de archivos JSON obtenidos tras la transformación de datos. . . .	30
7	Comparación de modelos clasificadores de sentimiento	35
8	Grafico de los rendimiento de modelos luego de las adaptaciones	38
9	Grafico de rendimiento por porcentaje de confianza	39
10	Resultados de la primera vuelta de las elecciones presidenciales de 2017. . .	45
11	Palabras más repetidas sobre el candidato Sebastián Piñera en 2017.	46
12	Palabras más repetidas sobre el candidato Alejandro Guillier en 2017.	47
13	Resultados sobre el analisis de connotación de tweets.	47
14	Comparación entre resultados oficiales y porcentajes de comentarios en Twitter.	48
15	Comparación final de resultados de las elecciones presidenciales año 2017 .	49
16	Palabras más repetidas sobre la opción Apruebo en 2020.	51
17	Palabras más repetidas sobre la opción Rechazo en 2020.	52
18	Comparación entre resultados oficiales y porcentajes de comentarios en Twitter del año 2020.	54
19	Comparación final de resultados del proceso constitucional año 2020.	55
20	Candidatos de las primarias presidenciales del año 2021.	56
21	Resultados de primarias presidenciales del año 2021	57
22	Palabras más repetidas referentes a Gabriel Boric en las elecciones de 2021. .	58
23	Palabras más repetidas referentes a Jose Antonio Kast en las elecciones de 2021.	59
24	Comparación entre resultados oficiales y connotaciones de comentarios en Twitter en las elecciones presidenciales del año 2021.	60
25	Comparación final de resultados de las elecciones del 2021.	61
26	Consituyentes electos en el año 2021.	62
27	Palabras más repetidas referentes a el Apruebo en las elecciones de 2022. . .	63
28	Palabras más repetidas referentes a el Rechazo en las elecciones de 2022. . .	65
29	Resultados sobre el analisis de tweets en las elecciones del año 2022.	66
30	Comparación entre resultados oficiales y connotaciones de comentarios en Twitter en las elecciones del 2021.	66
31	Comparación final de resultados de las elecciones del 2021.	67
32	Palabras mas repetidas con relación a la opción ``En contra" del plebiscito 2023	70
33	Palabras mas repetidas con relación a la opción ``A favor" del plebiscito 2023	71
34	Connotaciones de los tweets en el plebiscito 2023	73
35	Resultado final previsto del plebisicito 2023	74

ÍNDICE DE TABLAS

1	Comparativa de errores en convención constitucional en plebiscito 2020. . .	4
2	Comparativa de errores en convención constitucional en plebiscito 2020. . .	4
3	Resumen de datos referentes a las votaciones más significativas.	27
4	Parámetros para la búsqueda de datos.	27
5	Cantidades de datos posteriores a las iteraciones.	29
6	Resultados de testeo de los modelos de análisis de sentimiento.	31
7	Mejora de los resultados tras la implementación de emoticones y palabras clave.	32
8	Candidatos y coaliciones en las elecciones presidenciales de 2017.	44
9	Resultados sobre el análisis de connotación de <i>tweets</i> 2020.	53
10	Resultados oficiales de elección de primarias 2021	56
11	Resultados sobre el análisis de <i>tweets</i> en las elecciones del 2021.	60
12	Resumen de resultados obtenidos	67
13	Recolección de <i>tweets</i> para el plebiscito 2023	69

INTRODUCCIÓN

El presente informe se realiza en un contexto de frecuentes elecciones en el país y un considerable uso de las redes sociales para manifestar opiniones respecto a diferentes medidas o candidatos. Este estudio tiene como objetivo determinar una posible correlación entre la cantidad de *tweets* publicados por los votantes y los resultados finales de las elecciones. La creciente influencia de las redes sociales en la opinión pública y su capacidad para reflejar tendencias sociales y políticas hace pertinente investigar esta relación.

La propuesta de solución se basa en la adaptación de algoritmos de clasificación para analizar los *tweets* de las elecciones más recientes e importantes. Se emplean modelos de clasificación de sentimientos para identificar si existen tendencias claras que se reflejen en los resultados oficiales. La metodología incluye la recolección y preprocesamiento de datos de *tweets*, la aplicación de técnicas de procesamiento de lenguaje natural (NLP) y la visualización de los datos para interpretar los resultados.

Para una estructura clara y comprensible del trabajo, este se ha dividido en diferentes capítulos. El capítulo 1, **Definición del Problema**, establece el contexto del estudio, describiendo el entorno electoral y el uso de las redes sociales, además de los objetivos específicos que se desean alcanzar. El capítulo 2, **Marco Conceptual**, proporciona las bases teóricas necesarias para comprender el trabajo realizado, incluyendo el procesamiento de lenguaje natural, el análisis de sentimientos y las metodologías de minería de datos.

El capítulo 3, **Propuesta de Solución**, describe en detalle la metodología empleada, desde la selección y preprocesamiento de datos hasta la minería de datos. El capítulo 4, **Visualización y Análisis de Resultados**, presenta el desarrollo completo de la metodología, continuando minería de datos hasta la consolidación del conocimiento, evaluando los algoritmos y obteniendo los resultados.

Finalmente, se presentan las **Conclusiones**, que abarca detalladamente el análisis de cada resultado obtenido, las posibles mejoras al trabajo realizado y las implicaciones futuras de este estudio. Estas conclusiones permiten entender cómo los hallazgos pueden utilizarse en futuras investigaciones y aplicaciones prácticas en el campo de la predicción electoral.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

En este capítulo, se demarcó el contexto que dio origen al problema investigado y que motivó el análisis. Además, se presentaron tanto los objetivos específicos como el objetivo general que guiaron a la resolución del problema identificado.

1.1. Contexto

Chile ha atravesado una serie de cambios durante la última década, marcadas por eventos de envergadura en contextos políticos, sociales y ambientales. La secuencia de acontecimientos ha sido tanto diversa como profunda: terremotos devastadores, escándalos de corrupción que han sacudido la confianza en las instituciones, movilizaciones sociales de gran escala y elecciones que han alterado el curso político del país.

Cada uno de dichos acontecimientos ha llevado a la ciudadanía a adoptar sus posturas frente a cada suceso, generando una amplia gama de opiniones que han sido expresadas a través de distintos medios de comunicación, fomentando así intensos debates en la sociedad. Un claro ejemplo de esto fue la convención y el plebiscito constitucional, eventos que aumentaron considerablemente la tensión entre todos aquellos que tenían distintas opiniones, dando lugar a una polarización política y social en el país [Gaspar, 2022], lo cual se vio confirmado en las elecciones presidenciales del año 2021.

Las manifestaciones de estas posturas han encontrado como lugar seguro los medios de comunicación y, de manera más notoria, las redes sociales. Twitter, en particular, ha servido como una plataforma dinámica para la discusión y el intercambio de ideas, actuando como un barómetro de la opinión pública. A través de esta red, se han articulado y confrontado perspectivas, especialmente en torno a hitos como la convención constitucional y el plebiscito constitucional, los cuales fueron decisivos en la escalada de tensiones entre grupos con visiones políticas contrapuestas.

A pesar de que en términos de usuarios activos Twitter no encabeza las estadísticas en Chile, su influencia en el discurso público es indiscutible. Se estima que para el año 2022, aproximadamente 2,9 millones de chilenos usaban esta plataforma [Carrasco, 2022], lo que representa el 15,1 % aproximadamente de la población del país, los cuales se involucran en debates políticos y sociales con una intensidad que no siempre se ve reflejada en otras redes sociales.

Sin embargo, la relevancia de Twitter en la ciudadanía chilena va más allá de los números. Su estructura favorece la formación de comunidades de discusión y la rápida propagación de información, aspectos que han sido cruciales en la formación de opinión y en la organi-

zación de movimientos sociales. Esta interacción ha tenido un rol central durante periodos electorales, donde la plataforma ha servido como un espacio para la campaña y el debate.

No obstante, a pesar del auge de Twitter en los últimos años y el uso los usuarios chilenos para discutir asuntos políticos, informarse y organizarse, aún persiste una brecha importante en la comprensión de cómo las opiniones vertidas en estas plataformas reflejan la opinión pública general en Chile.

Este panorama resalta la necesidad de desarrollar métodos más modernos para medir las percepciones políticas de los ciudadanos, especialmente a través del análisis de las opiniones expresadas en las redes sociales. De esta manera, este trabajo tuvo como objetivo abordar aquel desafío, contribuyendo así a una interpretación más exacta y representativa de la opinión pública.

1.2. Situación Actual

Actualmente Chile se encuentra en una fase decisiva de su historia política, con la mirada puesta en las elecciones constitucionales de diciembre, las cuales tienen el potencial de redirigir el futuro del país. En medio de esta atmósfera cargada de cambio y expectativa, las redes sociales emergen de nuevo como un vital espacio de debate público. Estos foros digitales se han alzado como plataformas donde la ciudadanía ejerce su libertad de expresión, compartiendo reflexiones y expectativas sobre los sucesos políticos que se están desarrollando.

El incremento en la participación digital ha diversificado el espectro de opiniones y, consecuentemente, ha generado un volumen creciente de datos que reflejan el pulso político y social de los chilenos frente a estas nuevas elecciones. Sin embargo, actualmente el análisis de estos datos no está a la altura de las circunstancias. Los estudios son escasos y tienden a quedarse en la superficie, limitándose a métricas visibles pero limitadas como el número de seguidores, resultados de encuestas, *likes* y comentarios. Esto deja un vacío significativo en el aprovechamiento de esta información que es publicada en estas redes sociales. Una inmersión más profunda en estos datos podría revelar patrones y perspectivas que serían de gran valor para entender el clima político actual del país.

1.3. Encuestas tradicionales

Las instituciones de encuestas tradicionales han enfrentado retos considerables en la recopilación de datos, fundamentalmente a causa de la creciente polarización detectada en el país [Gaspar, 2022]. Esta polarización ha impactado de manera directa en las predicciones llevadas a cabo por los encuestadores, que han tenido que lidiar con errores de gran magnitud en sus proyecciones [Solís y Lee, 2020]. Estas imprecisiones han provocado que se registren

márgenes de error bastante notables, siendo un hecho que pone en una situación complicada a la reputación, credibilidad y confianza hacia las encuestadoras.

Como caso ilustrativo, en las votaciones referentes al porcentaje de la opción Apruebo en el plebiscito del año 2020, se identificó un margen de error del 6 % en las encuestas tradicionales tal como se puede ver en la Tabla 1, una cifra ya de por sí preocupante. Sin embargo, la encuesta realizada por Cadem, una de las encuestadoras más prestigiosas a nivel nacional, reveló un resultado preocupante con un error del 18 %, siendo este el más acertado en comparación con todas las otras instituciones encuestadoras, que registraron errores que fluctuaban entre el 18 % y el 28 % en las encuestas tradicionales previas al plebiscito (Tabla 2). Estos márgenes de error tan amplios han desembocado en una pérdida de la credibilidad sobre las encuestas y destacan la urgencia de revisar y mejorar las metodologías de sondeo.

ERROR DEL ÚLTIMO SONDEO DE CADA ENCUESTADORA RESPECTO DEL 'APRUEBO' EN EL PLEBISCITO CONSTITUCIONAL DE 2020				
Encuestadora	Fecha de la última encuesta sobre el Plebiscito 2020	Apruebo según el sondeo (%)	Resultado del Apruebo en Plebiscito Constitucional de 2020 (%)	Error del sondeo respecto al Apruebo (%)
Cadem	07/2020	71	78	-7
Criteria	09/2020	72	78	-6
TuInfluyes	10/2020	69	78	-9
Mori-Fiel	09/2020	66	78	-12
Activa	10/2020	65,4	78	-12,6

Tabla 1: Comparativa de errores en convención constitucional en plebiscito 2020.

Fuente: [Solís y Lee, 2020]

ERROR DEL ÚLTIMO SONDEO DE CADA ENCUESTADORA RESPECTO DE LA OPCIÓN 'CONVENCIÓN CONSTITUCIONAL' EN EL PLEBISCITO CONSTITUCIONAL DE 2020					
Encuestadora	Fecha de la última encuesta sobre el Plebiscito 2020	A favor de Convención Constitucional según sondeo (%)	Convención Mixta (con parlamentarios) (%)	Resultado a favor de la opción 'Convención Constitucional' en el Plebiscito 2020	Error respecto del resultado a favor de la CC (79%)
Cadem	07/2020	55	35	79	-24
Criteria	09/2020	59	28	79	-20
TuInfluyes	10/2020	61	21	79	-18
Mori-Fiel	09/2020	51	27	79	-28
Activa	10/2020	56,2	20,1	79	-22,8

Tabla 2: Comparativa de errores en convención constitucional en plebiscito 2020.

Fuente: [Solís y Lee, 2020]

Es evidente que la polarización y la volatilidad política actual presenta desafíos significativos para las encuestadoras, ya que capturar con precisión la opinión pública se ha vuelto una tarea cada vez más ardua. Siendo fundamental que se realicen ajustes metodológicos tanto en la recolección como en el análisis de los datos y en la interpretación de los resultados.

La persistencia de métodos anticuados en la captura y análisis de la opinión pública se ha demostrado ineficaz, y si no se abordan estos problemas, los actores políticos pueden continuar malinterpretando el sentir de la población. Esto tiene el potencial de desencadenar una serie de respuestas mal calibradas: desde políticas que no reflejan la diversidad de la sociedad

hasta campañas electorales que no capturan las verdaderas preocupaciones de los votantes. Tal desconexión concluye en un desgaste en la confianza que los ciudadanos depositan en las encuestas y, por extensión, en el proceso político en general.

Por lo tanto, el desafío principal trasciende la mera discrepancia entre los resultados de las encuestas tradicionales y la realidad de la opinión pública. Radica en la incapacidad para adaptarse a la rápida evolución tecnológica y en la inercia al innovar en métodos de análisis de datos. En un mundo donde la tecnología avanza a pasos agigantados, persistir en enfoques obsoletos conduce en este caso a una representación distorsionada de las opiniones políticas, lo que acarrea implicancias sustanciales para la integridad del proceso democrático y la formulación de políticas en Chile.

1.4. Objetivos

1.4.1. Objetivo General

El objetivo general del trabajo consiste en desarrollar un análisis sobre las opiniones políticas de los ciudadanos en redes sociales a través de visualizaciones y determinar su correlación con los resultados electorales.

1.4.2. Objetivos Específicos

Se plantean los siguientes objetivos específicos para cumplir y complementar el objetivo general previamente presentado:

- Establecer y aplicar un método eficiente y efectivo, para recopilar comentarios de Twitter que reflejen las opiniones políticas de los ciudadanos chilenos.
- Analizar y evaluar los rendimientos de los modelos de sentimientos con una o más entidades, para determinar cuales son los efectivos procesando textos de Twitter.
- Analizar la correlación entre las opiniones políticas expresadas en Twitter y los resultados electorales reales, para determinar la precisión y el valor predictivo de los análisis realizados.

CAPÍTULO 2

MARCO CONCEPTUAL

Este capítulo está dedicado a establecer el marco conceptual necesario para una comprensión del estudio. Se presenta un recorrido histórico y analítico de las redes sociales, proporcionando el contexto y los conocimientos fundamentales para abordar el caso de estudio.

2.1. Las redes sociales

En este segmento del estudio, se abordó el contexto histórico de las redes sociales y exploramos las implicancias significativas que ha implicado su avance.

En este punto se explica el contexto histórico de las redes junto las implicancias de esta.

2.1.1. Historia de las redes sociales

Las redes sociales, tal como se conocen en la actualidad, son un fenómeno relativamente reciente, con sus orígenes en el amanecer del siglo XXI. Sin embargo, la idea de conectar personas a través de la tecnología es mucho más antigua, remontándose a las primeras formas de comunicación en línea como los servicios de correo electrónico y foros de discusión, que comenzaron a popularizarse en las décadas de 1970 y 1980 ¹.

Fue en la década de 1990 cuando se comenzó a ver el surgimiento de plataformas que se asemejan más a las redes sociales que se conocen hoy en día. La evolución de estos espacios digitales llevó a la creación de sitios que no solo permitían la interacción básica entre usuarios, sino también la creación de perfiles personales y la formación de redes de contactos. *SixDegrees.com*, lanzado en 1997, se destaca como uno de los primeros ejemplos de una red social moderna, ofreciendo a los usuarios la posibilidad de crear perfiles y listas de amigos.

A inicios del 2000, plataformas como Friendster, MySpace y LinkedIn marcaron hitos significativos en el desarrollo de las redes sociales. Estos sitios introdujeron características innovadoras que permitían a los usuarios compartir fotos y música, y en el caso de LinkedIn, establecer conexiones en el ámbito profesional. Sin embargo, el verdadero punto de inflexión en la historia de las redes sociales se produjo con el lanzamiento de Facebook en 2004 y Twitter en 2006. Estas plataformas no solo atrajeron a millones de usuarios en todo el mundo, sino que también revolucionaron la forma en que las personas interactúan, comparten información y se organizan en línea.

¹Esta sección es un resumen de lo que se encuentra en <https://www.rdstation.com/es/redes-sociales/>

Desde entonces, las redes sociales han revolucionado la manera en que las personas interactúan, comparten información y establecen conexiones. Estas plataformas en línea han evolucionado para convertirse en herramientas indispensables, no solo para la interacción social sino también para la creación y mantenimiento de una presencia digital personal y profesional. Permiten a los usuarios crear perfiles detallados de sus personas y también conectar con otros, compartiendo una diversidad de contenido como lo son textos, fotografías, videos, enlaces, etc.

El impacto de las redes sociales en la sociedad y la cultura contemporáneas es indiscutible, transformando los métodos de comunicación tradicionales, fomentando formas de interacción más instantáneas y directas. Esta capacidad de compartir información y experiencias en tiempo real ha generado una nueva dimensión en la comunicación humana, facilitando conexiones que trascienden las barreras geográficas y culturales.

En un mundo cada vez más interconectado, las redes sociales han democratizado la información, otorgando a cada usuario la capacidad de ser tanto consumidor como creador de contenido. Este fenómeno ha cambiado no solo cómo las personas se comunican entre sí, sino también cómo interactúan con el mundo que les rodea, influyendo en una variedad de ámbitos que van desde la política hasta el entretenimiento.

2.1.2. Influencia de las Redes sociales

Además de transformar la comunicación interpersonal y cultural, las redes sociales han ejercido una influencia significativa en diferentes áreas como el comercio y la economía. En el ámbito comercial las empresas comenzaron a adoptar rápidamente estas plataformas para interactuar de manera más directa y personal con sus clientes, y siempre promocionando sus productos y servicios. Asimismo, las redes sociales han sido catalizadores en el surgimiento de nuevas formas de comercio y economía, tales como el comercio social y la economía colaborativa, donde la interacción entre los usuarios y la confianza mutua juegan un papel crucial.

En el ámbito político, las redes sociales se han consolidado como un espacio fundamental para la comunicación política, la movilización ciudadana y el debate público. Los políticos utilizan estas plataformas para comunicarse directamente con los ciudadanos, exponer sus ideas y campañas, y movilizar a sus seguidores. Paralelamente, los ciudadanos recurren a las redes sociales para debatir sobre temas políticos, participar activamente en discusiones políticas y organizar o unirse a movimientos sociales [Enli, 2017].

El impacto de las redes sociales también se extiende a iniciativas benéficas y campañas de concienciación. Un ejemplo destacado es el *Ice Bucket Challenge*, que se convirtió en un fenómeno viral en el verano de 2014. Este desafío involucró a personas de todas las edades y estatus, desde niños en sus hogares hasta celebridades como Mark Zuckerberg y Robert Downey Jr. La campaña, que consistía en grabarse vertiendo un balde de agua helada sobre sí

mismo, buscaba promover la conciencia sobre la Esclerosis Lateral Amiotrófica (ELA) y alentar las donaciones para su investigación [Sifferlin, 2014].

Con la ayuda de las redes sociales, el *Ice Bucket Challenge* alcanzó una audiencia global, con más de 17 millones de personas subiendo videos del desafío, llegando a aproximadamente 440 millones de espectadores [Facebook, 2014]. Esta campaña no solo incrementó la concienciación sobre la ELA, sino que también fue un éxito rotundo en términos de recaudación de fondos, con la Asociación de ELA reportando ingresos superiores a los 115 millones de dólares en el verano de 2014, en comparación con los 2.8 millones recaudados en 2013. Además, los fondos recaudados contribuyeron significativamente a la investigación de la enfermedad, incluyendo el descubrimiento de un nuevo gen asociado con la ELA [NEJM, 2016].

2.1.3. Consecuencias de las redes sociales

El uso generalizado de las redes sociales, a pesar de sus numerosos beneficios, no está exento de desafíos y preocupaciones. El avance de estas tecnologías ha traído consigo problemas en varios ámbitos que han resultado difíciles de controlar.

Uno de los problemas más prominentes es la privacidad personal. Muchos usuarios comparten información personal sin considerar plenamente en las consecuencias de esto, incluyendo fotos, videos y direcciones. Esta práctica incrementa el riesgo de que dicha información caiga en manos equivocadas, poniendo en peligro su seguridad e integridad.

Otra consecuencia del uso constante de las redes sociales es la despersonalización de las relaciones interpersonales. Cada vez más, las personas se sienten cómodas expresándose a través de una pantalla en lugar de en interacciones cara a cara, llevando a relaciones que son menos significativas y expresivas. Esta tendencia puede llevar a un sentimiento de aislamiento, con individuos sintiéndose solos a pesar de estar constantemente conectados, sumergidos en un flujo incesante de noticias y actualizaciones.

Además, las redes sociales pueden ser una herramienta muy peligrosa, debido a que las personas suelen sentirse seguros detras de un usuario, dando opiniones que en su día a día no darían e insultando a personas desconocidas por cualquier motivo lo cual es el punto clave para el hostigamiento y el ciberacoso que está ocurriendo hoy en día. Este tipo de acoso en línea afecta a una proporción alarmante de usuarios. Según Panda Security (2023), alrededor del 38 % de las personas han experimentado alguna forma de acoso en línea, y hasta un 25 % de los estudiantes que sufren ciberacoso han llegado a autolesionarse, lo que evidencia la gravedad del problema [Web, 2023].

Por último, está el fenómeno de las *fake news*. Estas noticias falsas o engañosas se han propagado ampliamente con el auge de las redes sociales. Su rápida difusión puede alterar significativamente la opinión pública y derrumbar la credibilidad de los medios de comunicación. La propagación de noticias falsas no solo presenta desafíos para la verificación de la infor-

mación, sino que también tiene implicaciones profundas en la política, la salud pública y la confianza en las instituciones.

2.1.4. Twitter

Twitter, desde su lanzamiento en 2006, se ha consolidado como una de las redes sociales más influyentes y revolucionarias en el ámbito de la comunicación. A pesar de haber surgido como una plataforma de microblogging destinada al intercambio de actualizaciones personales, su evolución ha sido notable, transformándose en una herramienta indispensable para la comunicación en tiempo real, la propagación de noticias y el debate de asuntos de actualidad [Enli, 2017].

A lo largo de su trayectoria, Twitter ha proporcionado a sus usuarios un espacio para expresar sus pensamientos y experiencias de manera instantánea mediante la publicación de mensajes concisos de hasta 280 caracteres, denominados tweets. Su creciente popularidad y relevancia atrajo a millones de usuarios alrededor del mundo que utilizan la plataforma para emitir y recibir información tal como se aprecia en la figura 1, estableciéndose así como un pilar fundamental de las redes sociales con un impacto global significativo.

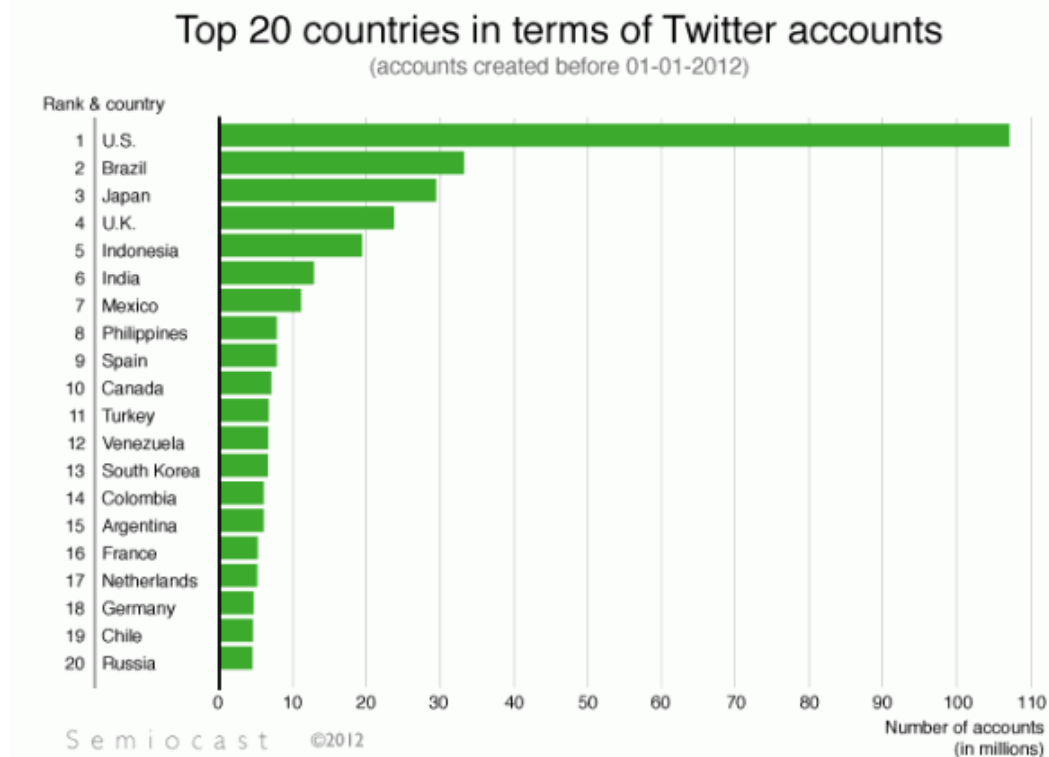


Figura 1: Cantidad de cuentas de twitter por el mundo en el 2012

Fuente: [Maturana, 2012]

La sinteticidad de los *tweets* ha fomentado un estilo de comunicación conciso y directo, favoreciendo el uso de *hashtags* y menciones para estructurar conversaciones y promover la interacción entre numerosos usuarios en torno a temáticas concretas.

En este contexto, Twitter se ha establecido como un medio ideal para el debate, especialmente en asuntos políticos. Su posición como red social predominante para la discusión de temas políticos es innegable, y su influencia en este campo es considerable. Los políticos utilizan Twitter para comunicarse directamente con los ciudadanos, exponiendo sus ideas, políticas y logros sin la intervención de intermediarios. Esta interacción directa ha transformado la forma en que los políticos interactúan con la población, permitiéndoles gestionar su mensaje e imagen de una manera que era impensable hasta hace poco [Enli, 2017].

Twitter, además, ha jugado un papel crucial en la facilitación de la participación ciudadana en el escenario político, ofreciendo a los usuarios una plataforma para expresar sus opiniones, participar en debates y movilizarse en torno a causas políticas y sociales. Las discusiones políticas en Twitter tienen el potencial de alcanzar a un vasto público y causar un impacto relevante en la opinión pública, amplificando voces y generando debates que trascienden fronteras geográficas y sociales.

Políticos y grupos de interés han sabido aprovechar Twitter como una herramienta para movilizar a sus seguidores, organizar eventos y protestas, y promover la participación ciudadana. Esto ha quedado particularmente patente en movimientos como *la Primavera Árabe* y *Occupy Wall Street*, donde Twitter desempeñó un papel crucial en la organización y difusión de las protestas [Tufekci, 2017].

No obstante, no todo es positivo. En línea con el constante aumento de usuarios en Twitter a nivel mundial [Galeano, 2023], también ha habido un crecimiento exponencial en la difusión de noticias falsas que han tenido un impacto directo en diversas campañas políticas en todo el mundo. Además, se ha observado una proliferación de discursos de odio infundados dirigidos hacia los usuarios de la plataforma [Allcott y Gentzkow, 2017].

La propagación de *fake news* en Twitter ha planteado serias preocupaciones en el ámbito político. Estas noticias falsas pueden tener un efecto perjudicial en la percepción pública, distorsionando la realidad y afectando la toma de decisiones informadas. Las campañas políticas se han visto afectadas por la propagación de información engañosa, lo que puede comprometer la integridad del proceso democrático.

2.2. Análisis de sentimientos

2.2.1. Definición de análisis de sentimientos

El análisis de sentimientos, también conocido como minería de opiniones, es un campo de estudio interdisciplinario que se centra en examinar, interpretar y entender las opiniones, sentimientos, valoraciones, percepciones y emociones de las personas hacia diferentes entidades. Estas entidades pueden incluir productos, servicios, organizaciones, individuos, problemas, eventos, temas, y sus correspondientes atributos [Liu, 2012].

Este campo de estudio se apoya en técnicas y métodos provenientes del procesamiento del lenguaje natural (NLP), *machine learning* (ML) y la minería de texto. Estos métodos permiten identificar, extraer y estudiar información subjetiva a partir de diversas fuentes de datos. El análisis de sentimientos puede clasificar las opiniones en diversas categorías tales como positivas, negativas o neutrales. Adicionalmente, puede identificar la intensidad del sentimiento expresado y hacerlo a varios niveles, incluyendo el nivel del documento, el nivel de la oración, y el nivel de la entidad y el aspecto [Pang y Lee, 2008].

Un aspecto importante a considerar en el análisis de sentimientos es su capacidad para manejar y procesar grandes volúmenes de datos, y extraer información útil de ellos. Con la explosión de la cantidad de datos de texto generados en línea a través de las redes sociales, blogs, foros de discusión y otros medios digitales, la capacidad para analizar y comprender estos datos se ha vuelto cada vez más crucial.

El análisis de sentimientos encuentra su aplicación en una amplia gama de áreas. En el ámbito empresarial, es utilizado para la monitorización de las redes sociales, la gestión de la reputación de las empresas, el análisis de la voz del cliente, y el análisis de mercado. En el campo de la política, el análisis de sentimientos puede ayudar a entender la opinión pública y monitorizar el sentimiento hacia políticas, partidos y candidatos.

2.2.2. Descripción de las técnicas y tecnologías

El análisis de sentimientos se fundamenta en una variedad de técnicas y tecnologías nombradas anteriormente; un mayor detalle de las mismas se da a continuación:

Procesamiento del Lenguaje Natural (NLP): Es una rama de la inteligencia artificial que se ocupa de la interacción entre los computadores y el lenguaje humano. Su objetivo es capacitar a las máquinas para leer, comprender y extraer significado de los datos en lenguaje humano de una manera valiosa y útil [Goldberg, 2017]. En el análisis de sentimientos, el NLP se emplea para extraer características de los textos que luego se pueden utilizar para determinar los sentimientos expresados.

Machine learning (ML): Es un enfoque de análisis de datos que automatiza la construcción de modelos analíticos. Se trata de una disciplina de la inteligencia artificial que se basa en la idea de que los sistemas pueden aprender de los datos, identificar patrones y tomar decisiones con una intervención humana mínima [Bishop, 2006]. En el análisis de sentimientos, se entrenan modelos de ML para clasificar los textos en diferentes categorías de sentimientos.

Minería de Texto: Es el proceso de extraer información valiosa de los textos. Comprende técnicas de NLP, ML y análisis de datos para descubrir y extraer patrones y tendencias significativas en los datos de texto [Feldman y Sanger, 2006]. En el análisis de sentimientos, la minería de texto se utiliza para extraer y cuantificar los sentimientos y opiniones expresadas en los textos.

Estas tecnologías trabajan en conjunto para permitir el análisis de sentimientos. El NLP ayuda a comprender el texto, el ML ayuda a clasificar los sentimientos y la minería de texto ayuda a extraer información relevante de los datos de texto. Con el avance de estas áreas, el análisis de sentimientos se ha vuelto cada vez más sofisticado y preciso, permitiendo una comprensión más profunda de las emociones y opiniones expresadas por las personas en los textos.

2.2.3. Niveles de Análisis de Sentimientos

El análisis de sentimientos puede llevarse a cabo en diferentes niveles de detalles, dependiendo del objetivo del análisis y de la naturaleza de los datos de entrada. Los tres niveles más comunes son los presentados a continuación:

Análisis a nivel de documento: Considera cada documento como una unidad básica de análisis. Este enfoque asume que cada documento expresa un solo sentimiento hacia un objeto o tema específico. Por ejemplo, en el análisis de sentimientos de las revisiones de productos, cada revisión puede considerarse como un documento individual y el objetivo es determinar si la revisión es positiva, negativa o neutral hacia el producto en cuestión [Pang *et al.*, 2002]. El análisis a nivel de documento es útil cuando se desea obtener una visión general del sentimiento general expresado en un conjunto de documentos, como la opinión pública sobre un tema específico.

Análisis a nivel de oración: Descompone cada documento en oraciones individuales y realiza el análisis de sentimientos en cada oración por separado. Este nivel de análisis permite capturar los sentimientos y opiniones expresados en cada oración específica. Puede ser útil cuando un documento contiene sentimientos relacionados con múltiples objetos o temas diferentes [Wiebe *et al.*, 2005]. Al analizar las oraciones individualmente, se puede obtener una comprensión más detallada de los sentimientos expresados en diferentes partes del texto.

Análisis a nivel de entidad y aspecto: Busca identificar los sentimientos expresados hacia diferentes entidades y aspectos dentro de una oración o documento. En este nivel de análisis, se consideran las entidades (por ejemplo, productos, personas, empresas) y los aspectos (por ejemplo, precio, calidad, servicio) mencionados en el texto. Se analizan los sentimientos expresados hacia cada entidad y aspecto específico. Por ejemplo, en una revisión de un producto, un cliente puede expresar un sentimiento positivo hacia el precio del producto (un aspecto) pero un sentimiento negativo hacia la calidad del producto (otro aspecto) [Liu, 2015]. Este nivel de análisis permite una comprensión más detallada de los sentimientos hacia elementos específicos mencionados en el texto.

La elección del nivel de análisis depende del objetivo de la investigación y de la naturaleza de los datos de entrada. El análisis a nivel de documento proporciona una visión general del sentimiento general, mientras que el análisis a nivel de oración y a nivel de entidad y aspecto permiten un análisis más específico.

2.2.4. Análisis de sentimiento en Python

El análisis de sentimientos se ha convertido en una herramienta esencial en el campo del *Natural language processing* (NLP), y Python se destaca como uno de los lenguajes de programación más populares y accesibles para implementar esta técnica. Su popularidad se debe en gran medida a la disponibilidad de una serie de bibliotecas especializadas que simplifican enormemente el proceso de procesamiento y análisis de datos textuales, facilitando así una amplia gama de investigaciones en el campo del NLP.

Estas bibliotecas de Python no solo ofrecen una variedad de herramientas para el análisis de sentimientos, sino que cada una ha sido diseñada con un propósito específico, abordando diferentes aspectos y niveles de complejidad en el análisis. A continuación, se presentan algunas de las bibliotecas más prominentes en este campo, destacando sus enfoques y características particulares.

El ***Natural Language Toolkit* (NLTK)** en Python es una herramienta indispensable en el campo del NLP. Esta biblioteca extensiva no solo proporciona funcionalidades esenciales como la tokenización, la eliminación de palabras irrelevantes (*stop words*) y el *stemming*, sino que también es fundamental en la preparación y procesamiento de datos para el análisis de sentimientos [Python, 2023].

NLTK es reconocida por su amplia gama de recursos, que incluyen acceso a corpora textuales y léxicos, herramientas de clasificación y etiquetado, así como algoritmos de análisis sintáctico. Esta riqueza de funcionalidades hace de NLTK una biblioteca versátil y poderosa en el NLP.

Uno de los aspectos más destacados de NLTK es su conjunto de clasificadores preentrenados, entre los que se encuentra VADER (*Valence Aware Dictionary for Sentiment Reasoning*). VA-

DER es especialmente eficaz en la identificación y categorización de la intensidad de los sentimientos en textos, clasificándolos como positivos, negativos o neutros. Su eficacia es particularmente notable en el análisis de textos provenientes de redes sociales, donde el lenguaje utilizado tiende a ser informal y cargado de emotividad. El puntaje compuesto de VADER, que oscila entre -1 y +1, proporciona una medida cuantitativa de la polaridad general del texto, facilitando así la interpretación objetiva de los sentimientos expresados [DataCamp, 2023].

Además, NLTK permite a los usuarios personalizar y afinar los modelos de análisis de sentimientos para adaptarlos a contextos específicos, lo que aumenta significativamente su aplicabilidad en diversos ámbitos de investigación y desarrollo.

TextBlob es una biblioteca versátil en Python que simplifica el procesamiento de datos textuales para tareas comunes de NLP. Su API ofrece facilidades para una amplia gama de tareas de NLP, como la extracción de frases nominales, análisis de sentimientos, clasificación y traducción, lo que la convierte en una herramienta ideal tanto para principiantes como para usuarios avanzados en el campo del NLP [Loria, 2020].

Una de las funcionalidades más prácticas de TextBlob es la corrección ortográfica, que permite a los usuarios identificar y corregir errores en los textos, mejorando así la calidad y la legibilidad de los datos analizados. Además, la capacidad de TextBlob para traducir textos de un idioma a otro, apoyándose en el textitAPI de Google Translate en su backend, amplía significativamente su aplicabilidad en entornos multilingües [Loria, 2020].

El cuanto a su análisis de sentimientos, TextBlob emplea un enfoque basado en un léxico y reglas predefinidas, que se centra en la identificación y clasificación de palabras según su polaridad y subjetividad inherentes. Este método implica el uso de un léxico que contiene palabras comunes asociadas con sentimientos positivos, negativos y neutros. Al analizar un texto, TextBlob compara las palabras del texto con este léxico para asignar un puntaje de polaridad y subjetividad. La polaridad se mide en una escala de -1 a +1, donde -1 representa sentimientos completamente negativos y +1 representa sentimientos completamente positivos. La subjetividad, por otro lado, varía de 0 a 1, indicando en qué medida el texto expresa opiniones personales en lugar de hechos objetivos [Tran, 2020].

Este enfoque léxico hace que TextBlob sea eficaz para analizar textos donde el tono emocional es claro y directo. Sin embargo, puede enfrentar limitaciones en textos con lenguaje más sutil, donde el contexto y las connotaciones pueden influir significativamente en el sentimiento expresado. A pesar de estas limitaciones, TextBlob sigue siendo una herramienta valiosa por su simplicidad y accesibilidad, proporcionando una excelente plataforma para una implementación básica y rápida de análisis de sentimientos en proyectos de NLP.

La biblioteca **Transformers de Hugging Face** representa un avance significativo en el campo del NLP, proporcionando acceso fácil a modelos de lenguaje preentrenados como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pretrained Transformer). Estos modelos, basados en la arquitectura de transformer, han revolucionado el tratamiento de tareas complejas de procesamiento de texto [Wolf et al., 2020].

En el análisis de sentimientos, la biblioteca Transformers se destaca por su capacidad para realizar clasificaciones de emociones en textos con una eficiencia y precisión notables. A diferencia de otras bibliotecas que pueden depender de enfoques léxicos o basados en reglas, Transformers utiliza modelos de aprendizaje profundo que han sido entrenados en enormes conjuntos de datos, lo que les permite captar matices y contextos más complejos en el lenguaje. Esto es particularmente valioso en escenarios donde el texto es ambiguo o contiene sutilezas lingüísticas que modelos más simples podrían no interpretar correctamente.

Una de las fortalezas clave de esta biblioteca es su habilidad para manejar grandes volúmenes de datos sin sacrificar la precisión. Esto la hace ideal para aplicaciones en entornos donde se generan grandes cantidades de texto, como en redes sociales, reseñas de productos o foros en línea. Además, los modelos como BERT y GPT han demostrado ser excepcionalmente eficaces en una variedad de tareas de NLP, no solo en el análisis de sentimientos, sino también en la comprensión del lenguaje y la generación de texto [Devlin *et al.*, 2018].

El uso de Transformers en el análisis de sentimientos va más allá de la mera clasificación de textos como positivos o negativos. Su capacidad para comprender el contexto y la relación entre las palabras en una oración permite una interpretación más matizada del sentimiento, capturando aspectos como la ironía, el sarcasmo y las emociones complejas. Esto lo convierte en una herramienta poderosa para investigadores y practicantes del NLP que buscan una comprensión más profunda y detallada del sentimiento y la tonalidad expresada en el lenguaje natural.

2.2.5. KPI's en el analisis de sentimiento

El análisis de sentimientos se ha convertido en una herramienta esencial en el procesamiento del lenguaje natural, aprovechando técnicas de la lingüística computacional, el aprendizaje automático y la minería de datos para interpretar las opiniones expresadas en textos. Este proceso implica evaluar textos para determinar las emociones, ya sean positivas, negativas o neutrales. Un factor crucial para la precisión y relevancia de los resultados obtenidos en el análisis de sentimientos es la metodología de ponderación aplicada. Esta metodología asigna importancia relativa a diferentes elementos del texto, como palabras o frases, basándose en su capacidad para influir en la orientación sentimental de un documento.

Métodos tradicionales como la **frecuencia de términos** (TF), que cuenta cuántas veces aparece una palabra en un documento, y la **frecuencia de documentos inversa** (IDF), que reduce el peso de las palabras que aparecen frecuentemente en muchos documentos y por lo tanto son menos informativas, han sentado las bases para este proceso. Sin embargo, la evolución hacia técnicas de ponderación más sofisticadas es esencial para capturar la complejidad del lenguaje humano.

En este contexto, los **Indicadores Clave de Rendimiento** (KPIs) juegan un papel vital en la evaluación y mejora continua de las técnicas de análisis de sentimientos. KPIs específicos,

como la precisión, la cual se centra en la proporción de identificaciones correctas de sentimiento entre todas las realizadas, indicando cuán fiable es el modelo al etiquetar datos como positivos o negativos, el *recall*, el cuál evalúa la proporción de identificaciones correctas entre todos los ejemplos relevantes en el conjunto de datos, reflejando la capacidad del modelo para detectar todas las instancias relevantes y la medida F1 que ofrece un balance entre precisión y *recall*, proporcionando una visión integral de la performance del modelo al combinar ambos en una sola métrica, permiten evaluar la efectividad de sus métodos, incluidas las técnicas de ponderación.

Todo esto concluye en un avance en la comprensión y rendimiento de los análisis de sentimiento y también con ello en las metodologías de ponderación tal como se ilustra en el estudio “Statistically Weighted Reviews to Enhance Sentiment Classification” [Pang y Lee, 2015], que propone un enfoque combinando la **Ponderación de Palabras (WW)** con la **Ponderación de Sentimientos (SW)**. Este enfoque demuestra cómo la aplicación de funciones estadísticas específicas a la ponderación de términos mejora la clasificación de sentimientos, superando a los métodos tradicionales en términos de precisión y estableciendo un precedente para la evaluación de técnicas de ponderación mediante KPIs específicos.

La integración de KPIs en el análisis de sentimientos, especialmente para evaluar métodos de ponderación, es esencial para el progreso de este campo. El estudio mencionado refuerza la importancia de enfoques estadísticamente fundamentados por las ponderaciones y destaca la importancia de los KPIs en la mejora continua de las técnicas de análisis de sentimientos. A medida que el campo evoluciona, la identificación y aplicación de KPIs seguirán siendo cruciales para el desarrollo de metodologías capaces de capturar con precisión la complejidad del sentimiento humano.

2.2.6. Estudios previos

A pesar de los grandes avances logrados por la inteligencia artificial en tiempos recientes, el análisis de sentimientos ha progresado más lento debido a los desafíos que se enfrenta. Estos desafíos incluyen la ambigüedad, subjetividad, ironía y sarcasmo inherentes al lenguaje humano, elementos que complican considerablemente la interpretación precisa de los textos [Lu *et al.*, 2023]. Como resultado, los estudios relacionados con el análisis de sentimientos han encontrado grandes limitaciones en varios aspectos de su aplicación y desarrollo.

En particular, estudios referentes a la predicción de resultados electorales se han llevado a cabo en distintas partes del mundo. Un caso es el estudio “Sentiment Analysis and Its Application in Predicting Electoral Outcomes: A Survey” [Maharana *et al.*, 2023], en donde se intentó predecir resultados electorales a partir de la recopilación de información de redes sociales, aplicando el análisis a algunas elecciones en Estados Unidos. Aunque el estudio concluyó con resultados prometedores, enfrentó problemas significativos relacionados con la calidad de los datos, especialmente debido a la dificultad de interpretar información proveniente de diferentes redes sociales.

Otro estudio relevante es “Análisis de sentimiento político en redes sociales, como instrumento para la toma de decisiones estratégicas pre electorales” [Urgilés *et al.*, 2022], que se centró exclusivamente en Twitter para evitar los problemas derivados de la diversidad de fuentes de datos. Este análisis, aplicado a una elección seccional en Ecuador, identificó algunas tendencias, aunque con márgenes de error considerables.

En Chile, también se han realizado estudios similares, como la tesis “Análisis de Sentimientos y Predicción de Eventos en Twitter” [García, 2014], que se enfocó en crear un modelo de análisis de sentimientos para las elecciones presidenciales primarias de 2013. Aunque el modelo obtuvo buenos resultados, su validez se limitó a esa elección y los datos específicos utilizados, lo que dificulta afirmar su eficacia en elecciones futuras.

Por lo que, este trabajo busca abordar y solucionar estos problemas, diferenciándose mediante el uso de cuatro elecciones de prueba, tanto constitucionales como presidenciales, para confirmar que el modelo puede aplicarse a distintos tipos de elecciones sin perder eficacia. Además, se con un énfasis considerable en la extracción constante de datos en diferentes fechas y en la limpieza de estos datos para maximizar su calidad y minimizar el ruido. Finalmente, el modelo se probó en una elección real para confirmar su viabilidad y efectividad.

2.3. Metodología para Análisis de Datos

El análisis de sentimientos es una especialización dentro de la minería de datos y el NLP, y estos proyectos requieren un enfoque metodológico estructurado para asegurar un proceso claro y confiable. La adopción de una metodología adecuada proporciona un marco sólido para las distintas etapas del proyecto, abarcando desde la recolección inicial de datos hasta su análisis detallado y la evaluación final. Y en la presente sección se presentan las metodologías estructuradas mas comunes.

2.3.1. KDD

El proceso de Knowledge Discovery in Databases (KDD) es una metodología compleja y multidisciplinaria que se centra en la extracción de conocimientos valiosos y accionables de grandes volúmenes de datos. Este proceso integra técnicas provenientes de disciplinas como la inteligencia artificial, la estadística y la gestión de bases de datos, reflejando su naturaleza interdisciplinaria y su aplicabilidad en un amplio espectro de campos [Conexión ESAN, 2018].

En la era del *big data*, donde la cantidad y variedad de información disponible crece exponencialmente, el KDD se ha convertido en una herramienta esencial para transformar grandes conjuntos de datos en información significativa y conocimiento práctico.

El proceso de KDD se desglosa en varias etapas clave, que se detallan en la Figura 2:

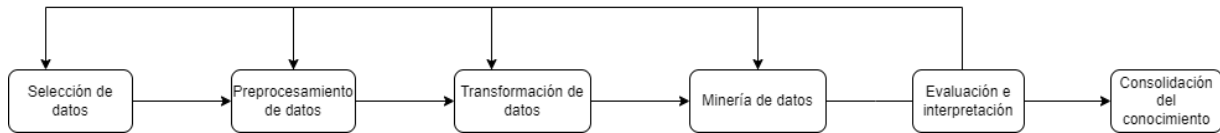


Figura 2: Proceso de Knowledge Discovery in Databases (KDD).

1. **Selección de Datos:** Aquí se seleccionan los datos específicos necesarios para el análisis. Esta fase implica determinar qué datos son relevantes para los objetivos del proyecto y pueden incluir ciertos criterios de búsqueda o la selección de subconjuntos de datos para el análisis.
2. **Preprocesamiento de Datos:** Se realiza la mejora de la calidad de los datos mediante la corrección o eliminación de información incorrecta, duplicada o incompleta, y el manejo de valores atípicos para asegurar la precisión y fiabilidad de los datos.
3. **Transformación de Datos:** Los datos se transforman en un formato adecuado para el análisis. Esto puede incluir la normalización, la creación de variables derivadas, la codificación de categorías, y la preparación de estructuras de datos adecuadas.
4. **Minería de Datos:** Se aplican técnicas avanzadas de minería de datos para identificar patrones, tendencias y correlaciones significativas. Es una etapa clave donde se descubre la información valiosa.
5. **Evaluación e Interpretación:** Los resultados obtenidos de la minería de datos se evalúan y se interpretan en relación con los objetivos iniciales del proyecto, validando los patrones encontrados y entendiendo su significado.
6. **Consolidación del Conocimiento:** El conocimiento descubierto se integra de manera efectiva para mejorar procedimientos, productos o decisiones estratégicas. Esta fase implica la validación final de los hallazgos y su preparación para una implementación práctica y efectiva. Los resultados son formalizados en informes detallados, presentaciones y recomendaciones claras para asegurar que el conocimiento pueda ser utilizado de manera óptima por todas las partes interesadas.

Una de las principales ventajas del KDD es su capacidad para adaptarse a cambios y nuevas informaciones durante el proceso de análisis. Esta flexibilidad es esencial en proyectos dinámicos donde las condiciones y los requerimientos pueden cambiar con el tiempo. KDD permite a los investigadores revisar y ajustar las fases anteriores según sea necesario, asegurando que el proceso sea iterativo y adaptable. Además, el enfoque en el análisis continuo desde la etapa de preprocesamiento garantiza la calidad y relevancia de los datos desde el inicio, lo que es crucial para la extracción de conocimientos significativos y accionables.

El enfoque interdisciplinario del KDD permite un análisis de datos más completo y profundo, siendo particularmente eficaz en descubrir patrones ocultos y relaciones complejas en gran-

des volúmenes de datos. Su adaptabilidad a diversos campos y su capacidad para soportar decisiones informadas basadas en datos lo convierten en una herramienta indispensable en el análisis moderno de datos [rajput, 2023].

2.3.2. SEMMA

Desarrollado por el SAS Institute, SEMMA es una metodología diseñada para facilitar el proceso de selección, exploración y modelado de grandes volúmenes de datos, con el objetivo de descubrir patrones de negocio desconocidos. SEMMA, un acrónimo de *Sample*, *Explore*, *Modify*, *Model* y *Assess*, describe las cinco fases esenciales de esta metodología. A continuación, se detalla cada fase y se amplía su aplicación y relevancia 3:

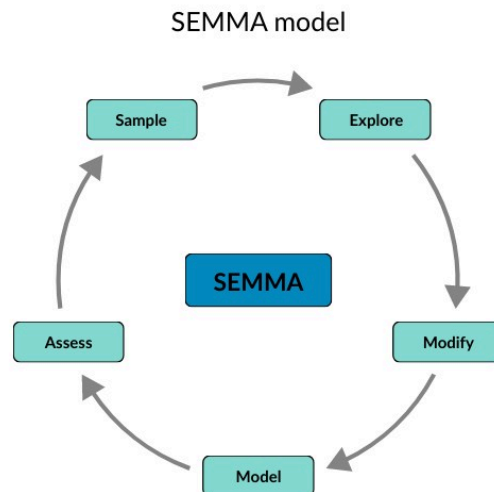


Figura 3: Proceso de Sample, Explore, Modify, Model and Assess (SEMMA).

1. **Sample (Muestreo):** Esta fase inicial se centra en la extracción de una muestra representativa de datos. La representatividad es crucial para garantizar que los resultados del análisis sean generalizables. El muestreo aleatorio simple es una técnica común, pero también pueden emplearse métodos más complejos dependiendo de la naturaleza de los datos y el problema específico.
2. **Explore (Exploración):** En esta etapa, se utilizan herramientas de visualización y técnicas estadísticas para simplificar el problema y determinar las variables explicativas clave. Esta fase implica un examen detallado de las características de los datos, incluyendo la distribución, la correlación entre variables y posibles anomalías o patrones inesperados.
3. **Modify (Modificación):** Basándose en los hallazgos de la fase de exploración, los datos se manipulan y preparan para el análisis. Esto puede incluir la transformación de varia-

bles, la creación de nuevas variables derivadas, la normalización de datos y la solución de problemas como valores atípicos.

4. **Model (Modelado):** Aquí se realiza el análisis y modelado de los datos para establecer relaciones entre las variables explicativas y los objetivos del proyecto. Se emplean métodos estadísticos y algoritmos basados en datos para construir modelos predictivos o descriptivos, dependiendo de los objetivos de la minería de datos.
5. **Assess (Evaluación):** La última fase implica la valoración de los modelos construidos. Esto incluye la evaluación de la bondad del modelo, la interpretación de los resultados y su validación, ya sea mediante métodos estadísticos o aplicando el modelo a nuevas poblaciones muestrales para verificar su robustez y precisión.

Las ventajas de SEMMA incluyen su estructura clara y sistemática, que ofrece una guía explícita para cada fase del proceso de minería de datos, facilitando así la comprensión y ejecución de proyectos complejos. Esta metodología se destaca por su enfoque en la eficiencia del modelado, proporcionando herramientas específicas para optimizar la selección de modelos y asegurar la precisión en la evaluación. Además, SEMMA es particularmente útil para proyectos que requieren un método riguroso y detallado, ya que promueve un análisis exhaustivo y una validación rigurosa de los modelos antes de su implementación final.

A diferencia de otras metodologías, SEMMA se centra más específicamente en las tareas de modelado de la minería de datos, lo que la hace particularmente adecuada para situaciones donde el objetivo principal es el análisis estadístico y la modelización de datos [Álvarez J.; Mesa J.; González A., 2013].

2.3.3. CRISP-DM

La metodología CRISP-DM, concebida por Chapman en 1999, representa un estándar en la minería de datos, estructurada en cuatro niveles de abstracción jerárquicos, desde generalidades hasta detalles específicos. Este proceso integral se organiza en seis fases principales, con cada fase estructurada en tareas generales que a su vez se desglosan en tareas específicas para casos concretos.

En el nivel más general, CRISP-DM se organiza en seis fases principales (Figura 4):

1. **Comprensión del Negocio:** Donde se definen los objetivos y requerimientos del proyecto desde una perspectiva empresarial, traduciéndolos en objetivos técnicos y un plan de proyecto.
2. **Comprensión de los Datos:** Incluye la recolección inicial de datos, evaluación de su calidad y exploración para identificar relaciones clave.

2.4. Visualización de datos

La visualización de datos es un componente crucial en la interpretación y comunicación de información compleja de manera comprensible. Los principios de visualización de datos enfatizan la importancia de representar la información de manera que se alinee con la percepción y el procesamiento cognitivo humano. Un principio fundamental es el uso efectivo de canales visuales como la posición, longitud, ángulos, área, brillo, y color, lo cual se basa en la investigación sobre cómo se detectan los patrones y las comparaciones visuales [Reynolds, 2024]. Este enfoque ayuda a asegurar que la representación de datos sea intuitiva y eficaz.

Conocer a la audiencia es vital para el diseño de visualizaciones de datos, ya que diferentes audiencias pueden tener distintas necesidades y niveles de comprensión. Esto implica considerar quiénes serán los usuarios finales de los datos y qué necesitan entender a partir de ellos. La simplicidad también juega un papel crucial en la efectividad de las visualizaciones. Mantener los diseños simples y libres de elementos superfluos permite a los usuarios enfocarse en la información más importante sin distracciones.

La elección del tipo de gráfico correcto es otra consideración clave. Cada tipo de gráfico tiene sus fortalezas y es adecuado para representar diferentes tipos de relaciones y datos [Reynolds, 2024]. Por ejemplo, los gráficos de barras son efectivos para comparaciones precisas entre categorías, mientras que los gráficos de dispersión son ideales para mostrar correlaciones y agrupaciones. La sabiduría en el uso del color no solo mejora la estética de la visualización sino que también puede guiar la atención del usuario hacia la información más relevante, ayudando a comunicar efectivamente los puntos clave.

Hoy en día, la tecnología ha avanzado hasta el punto de proporcionar una amplia gama de herramientas diseñadas para abordar estos aspectos fundamentales de la visualización de datos. Estas herramientas están diseñadas para facilitar el proceso de diseño y implementación de visualizaciones de datos efectivas, permitiendo a los usuarios concentrarse en el mensaje y los datos que se desean comunicar, sin necesidad de preocuparse excesivamente por los detalles técnicos. Este enfoque democratiza el acceso a técnicas de visualización de datos avanzadas, permitiendo a un público más amplio, no solo a especialistas, crear representaciones de datos que son al mismo tiempo informativas y visualmente atractivas.

2.4.1. Tableau

Tableau es una potente herramienta de visualización de datos diseñada para facilitar la comprensión y análisis de grandes conjuntos de datos mediante la creación de visualizaciones interactivas. Desde su lanzamiento en 2003, Tableau ganó popularidad en diversas industrias debido a su interfaz intuitiva y su capacidad para manejar datos de varias fuentes ².

²Esta sección es un resumen de lo que se encuentra en <https://www.tableau.com/why-tableau/what-is-tableau>

La herramienta se distingue por su capacidad para permitir a los usuarios arrastrar y soltar diferentes campos de datos para crear visualizaciones complejas sin necesidad de programación. Esto democratiza el acceso a técnicas avanzadas de visualización y análisis de datos, haciendo posible que personas con poca o ninguna habilidad en programación puedan generar información significativa a partir de sus datos.

Tableau soporta una amplia variedad de tipos de visualizaciones, desde gráficos de barras y líneas básicas hasta mapas geográficos interactivos y gráficos de dispersión complejos. Esta flexibilidad hace de Tableau una herramienta versátil, adecuada tanto para tareas simples de visualización de datos como para aplicaciones empresariales avanzadas. Además, Tableau ofrece funcionalidades de análisis avanzado, como el modelado predictivo y el análisis de tendencias, lo que permite a los usuarios no solo visualizar sus datos, sino también descubrir patrones y proyecciones futuras.

Una gran virtud de Tableau es su enfoque en la interactividad y la compartibilidad. Las visualizaciones creadas con Tableau pueden ser fácilmente compartidas y accedidas a través de la web, permitiendo a los equipos colaborar y tomar decisiones basadas en datos obtenidos rápidamente. Esta capacidad de compartir y interactuar con las visualizaciones en tiempo real amplía el alcance del análisis de datos, facilitando el acceso a información valiosa.

Finalmente, Tableau se integra con una variedad de fuentes de datos, desde hojas de cálculo simples hasta grandes bases de datos y servicios en la nube. Esta capacidad de integración facilita a los usuarios el acceso y la combinación de datos de diferentes fuentes, lo que es crucial para el análisis de datos en entornos empresariales complejos. La constante innovación y las actualizaciones regulares aseguran que Tableau permanezca a la vanguardia de la tecnología de visualización de datos, ofreciendo a los usuarios nuevas y mejoradas formas de ver y entender sus datos.

2.4.2. Flourish

Flourish es una plataforma de visualización de datos basada en la web, lanzada en 2016, que ha ganado popularidad rápidamente entre periodistas, investigadores y analistas de datos por su enfoque en la creación de visualizaciones de datos interactivas y animadas sin necesidad de programación. La plataforma se destaca por su facilidad de uso, permitiendo a los usuarios convertir datos en gráficos atractivos, mapas, y visualizaciones complejas a través de una interfaz intuitiva de arrastrar y soltar [UDGVirtual, 2024].

Flourish ofrece una amplia gama de plantillas de visualización personalizables, lo que facilita la creación de visualizaciones de datos sofisticadas y profesionalmente diseñadas. Estas plantillas abarcan desde gráficos de barras, líneas y áreas hasta visualizaciones más avanzadas como mapas de calor, gráficos de red y diagramas de Sankey. Esta variedad asegura que los usuarios puedan encontrar la mejor manera de representar sus datos, independientemente del contexto o el mensaje que deseen comunicar.

Una característica distintiva de Flourish es su capacidad para crear visualizaciones interactivas y animaciones que pueden mejorar significativamente la participación del espectador. Las visualizaciones pueden ser incrustadas fácilmente en sitios web y plataformas de medios sociales, lo que permite a los usuarios compartir sus datos de manera efectiva y atractiva con una audiencia amplia. Además, Flourish soporta la colaboración en equipo y la gestión de proyectos, lo que lo hace una herramienta valiosa para grupos de trabajo que requieren crear y gestionar visualizaciones de datos de manera conjunta.

La plataforma se integra sin problemas con diversas fuentes de datos, incluyendo hojas de cálculo en línea y bases de datos, permitiendo a los usuarios cargar y actualizar datos de manera eficiente. Esta integración facilita el mantenimiento de visualizaciones actualizadas sin esfuerzo manual adicional, una característica crucial para el seguimiento de datos en tiempo real o el análisis de tendencias a lo largo del tiempo.

Además, Flourish se compromete con la accesibilidad y la compatibilidad en diferentes dispositivos, asegurando que las visualizaciones creadas sean accesibles para todos los usuarios, independientemente del dispositivo que estén utilizando. Esto es especialmente importante en un entorno digital cada vez más móvil, donde la capacidad para ver y interactuar con datos en smartphones y tablets es esencial [Flourish, 2024].

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

En esta sección, se presenta y detalla el enfoque metodológico diseñado para abordar de manera clara y estructurada los objetivos propuestos en la sección 1.4.

3.1. Metodología seleccionada

Para seleccionar la metodología más adecuada para este proyecto, se revisaron las tres metodologías presentadas en la sección 2.3. Tras evaluarlas, se eligió la metodología KDD por su flexibilidad y su enfoque en el análisis continuo. Estas características son cruciales para este trabajo, ya que permiten una constante búsqueda de patrones y la capacidad de revisar fases anteriores, asegurando así unas mejores conclusiones.

En la Figura 2, se presenta el diagrama que ilustra la metodología KDD empleada en el desarrollo de este trabajo. Inicialmente, para la fase de **Selección de Datos**, se seleccionó la red social Twitter como fuente principal, logrando obtener un amplio conjunto de datos mediante técnicas de *Web Scraping*. Tras acumular una cantidad significativa de información, se procedió al **Preprocesamiento de Datos**. En este paso hubo un enfoque directo en la limpieza de datos, por lo que se realizó una cuidadosa selección para retener únicamente aquellos datos que aportaban información relevante para alcanzar los objetivos previamente establecidos.

Es importante destacar que, durante este procedimiento, se descartó una considerable cantidad de datos. Esta situación, en ocasiones, hizo necesario regresar a la etapa de **Selección de Datos** y repetir el ciclo, con el fin de asegurar que se contara con un volumen adecuado de información relevante para el análisis y los resultados posteriores. Este ciclo se mantuvo hasta conseguir un conjunto de datos óptimo y representativo que permitiera avanzar de manera efectiva en las siguientes fases del trabajo.

Concluida la fase de **Preprocesamiento de Datos**, se dio inicio a la fase de **Transformación de datos**, etapa en donde se transformaron los datos limpios en datos de menor tamaño los cuales permitieran una mayor velocidad de lectura.

Posterior a esta fase, se comenzó con la **Minería de Datos** etapa crucial en la cual los datos fueron categorizados en términos de su tono positivo o negativo hacia una entidad específica. Esta clasificación se llevó a cabo mediante técnicas de *Análisis de Sentimiento*, procesando cada uno de los elementos de datos para asignarles la categoría correspondiente.

Durante este paso, se optó por utilizar mas de una biblioteca de Análisis de Sentimiento, buscando seleccionar aquellas con la mayor precisión y rendimiento para cada caso especí-

fico, con el objetivo de lograr conclusiones más precisas. Esta estrategia aseguró una mayor exactitud y la fiabilidad de los resultados del análisis, permitiendo interpretaciones más acertadas. Adicionalmente, se realizó una visualización y análisis detallado de los datos, con el fin de comparar los resultados clasificados con los obtenidos realmente en las elecciones.

Con las visualizaciones iniciales en mano, se llevó a cabo el análisis correspondiente de los resultados. Este análisis profundizó en la interpretación de los datos, buscando patrones, tendencias y correlaciones que pudieran aportar al estudio. Durante esta fase de análisis, se generaron nuevas visualizaciones con el fin de complementar y enriquecer la comprensión de los resultados, proporcionando así una base sólida para la toma de decisiones y la elaboración de conclusiones.

Al finalizar la fase de **Minería de datos**, comienza la fase de **Evaluación e Interpretación**. En este punto, se llevó a cabo una minuciosa síntesis de los descubrimientos más importantes, extrayendo valiosas conclusiones de todo el proceso investigativo. Este esfuerzo de síntesis fue esencial para dar respuesta a cada uno de los objetivos planteados en las etapas iniciales del estudio. Se analizaron cuidadosamente las correlaciones, patrones y tendencias durante la investigación, con el propósito de construir una interpretación clara, precisa y bien fundamentada de los resultados obtenidos.

3.2. Desarrollo de la Solución Propuesta

En esta parte, se brinda una descripción detallada y contextualizada del estudio para cada uno de los pasos previamente mencionados, con el objetivo de analizar las elecciones más significativas de la última década.

3.2.1. Selección de datos

La fase inicial de esta investigación está marcada por la tarea de **Selección de datos**. Para este propósito, se eligió *Twitter* como la principal fuente de información, pero ¿Por qué optar por *Twitter* y no otra plataforma social? La respuesta radica en las características de *Twitter*. A diferencia de otras redes sociales, que tienden a enfatizar el contenido multimedia, *Twitter* se centra en el contenido textual. Esta particularidad permite captar de manera más precisa y directa las opiniones y percepciones de los usuarios sobre diversos temas, brindando un panorama directo de la discusión pública.

Con esta premisa en mente, el objetivo de esta etapa fue recolectar una variedad de tweets relacionados con las elecciones que han tenido mayor relevancia en la última década. Para identificar cuáles eran las elecciones relevantes, se adoptaron criterios basados en datos concretos, específicamente, aquellos que mostraban el mayor porcentaje de participación y la mayor cantidad de votantes en cada proceso electoral de los últimos diez años, tal y como

se puede apreciar en la Tabla 3. A partir de este análisis, se determinó que cuatro eventos electorales sobresalían del conjunto:

- Elecciones constitucionales de 2022
- Elecciones presidenciales de 2017
- Elecciones presidenciales de 2021
- Proceso constitucional de 2020

Elección	Año	Total de Electores Registrados	Participación Electoral (%)	Total de Votos
Elecciones constitucionales	2022	15173905	85,86%	13025739
Elecciones presidenciales	2021	15030974	55,03%	8364481
Proceso constitucional	2020	14855719	50,98%	7573914
Elecciones presidenciales	2017	14347288	48%	7032878
Eleccion Senadores	2021	15033562	34,05%	5119452
Elecciones presidenciales	2013	13373108	22,51%	3010890

Tabla 3: Resumen de datos referentes a las votaciones más significativas.
Fuente: SERVEL.

Con estas selecciones basadas en datos objetivos se busca garantizar que el estudio se centre en los eventos que generaron la mayor movilización y participación ciudadana, y así con ello impacto en las redes sociales, asegurando así que la investigación tenga una cantidad de datos representativa.

Tras determinar cuáles serían las votaciones a analizar, el siguiente paso consistió en iniciar la Extracción de datos. Para llevar a cabo esta tarea, fue necesario seleccionar con precisión las palabras clave que serían utilizadas en la búsqueda, así como definir el rango de fechas en el que se realizaría. Todos estos parámetros, cruciales para la obtención de resultados acertados y útiles, están detalladamente especificados en la Tabla 4.

Elección	Año	Fecha mínima	Fecha máxima	Palabras Clave	Palabras Opcionales
Elecciones constitucionales	2022	30-07-2022	30-08-2022	Apruebo, Rechazo	constitucional, plebiscito
Elecciones presidenciales	2021	10-11-2021	10-12-2021	Gabriel Boric, Jose Kast	elecciones, chile
Proceso constitucional	2020	15-09-2020	15-10-2020	Apruebo, Rechazo	constitucional, plebiscito
Elecciones presidenciales	2017	12-12-2017	12-11-2017	Sebastian Piñera, Alejandro Guillier	elecciones, chile

Tabla 4: Parámetros para la búsqueda de datos.
Fuente: Elaboración propia.

Una vez establecidos todos estos parámetros, se dio inicio a la tarea de *Web Scraping*, en donde se analizó la estructura HTML de Twitter y se buscaron aquellos contenedores que obtenían la información deseada para extraerlas mediante *Web Scraping*. Este método permitió llevar a cabo una búsqueda personalizada, con la posterior recolección y almacenamiento de la información necesaria en un archivo CSV, cuyo contenido se puede apreciar en la imagen 5.

	A	B	C	D	E	F	G	H	I	J	K
1	tweet_avatar,"images[0]"	"username","text","in_reply_to","replies","retweets","quotes","likes","url","timestamp","images","tweet_links									
2	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
3											
4	Cómo no va a ganar el Apruebo loco, todo puede pasar",,"0","1","0","3","https://twitter.com/cmconsalvez/status/1561865726679404544","20										
5	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
6	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
7	PUEBLO CALLEJERO CON SU PUEBLO.										
8	REALIDAD.",,"0","0","0","0","https://twitter.com/Manuel99826503/status/1561865638099894272","2022-08-22 23:59:00+00:00",,,,,,,,,,,,,,										
9	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
10	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
11	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn										
12	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
13	#Apruebo",,"0","0","0","0","https://twitter.com/CLAUDIA74100865/status/1561865276097851392","2022-08-22 23:57:00+00:00",,"#Apruebo",										
14	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
15	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
16	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
17	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
18	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										
19	https://pbs.twimg.com/profile_images/enc/cHJvZmlsZV9pbWFnZXNvMTU1OTM2MDE3OTc3NDE2OTA4OC96UzZvZWVhZm9iaWdnZXIuanBn_400x										

Figura 5: Ejemplo de archivo obtenido tras la extracción de datos.

Fuente: Elaboración propia.

Este procedimiento garantiza no solo la recopilación de una gran cantidad de datos pertinentes y valiosos, sino también su organización y almacenamiento de manera accesible, concluyendo de esta manera la etapa de **Selección de datos** y con ello también sentando las bases sólidas para las etapas siguientes del estudio.

3.2.2. Preprocesamiento de datos

Una vez culminada la etapa de **Selección de datos**, se da paso a la fase de **Preprocesamiento de datos**. Este procedimiento implica someter el conjunto de datos recabado a un meticuloso trabajo de análisis y selección, con el objetivo de depurar y descartar aquellos elementos que no aporten valor significativo a la investigación, ya que, existen datos que lejos de ser útiles, resultan ser redundantes e irrelevantes.

La importancia de esta etapa radica en la necesidad de optimizar los recursos y tiempo dedicados a las fases subsiguientes del estudio. Además, la depuración de datos juega un papel fundamental en la preservación de la integridad de los resultados finales, minimizando el riesgo de conclusiones erróneas derivadas de la inclusión de datos de baja calidad.

En este contexto, se han definido los siguientes parámetros para llevar a cabo la Filtración de datos.

1. **Imágenes:** Se decidió excluir todos aquellos tweets que incluyen imágenes. La razón detrás de esta decisión es que, en muchos casos, el significado o el contexto de un tweet puede estar intrínsecamente ligado a su imagen adjunta. Además, en ciertas situaciones, el tweet podría consistir únicamente en una imagen sin texto acompañante. Considerando el elevado volumen de datos recogidos, no resulta factible la revisión individual de cada imagen.

2. **Retweets:** Se optó por eliminar aquellos *tweets* que contenían *retweets*. Esta decisión se basa en que el contenido de un *retweet* a menudo depende del *tweet* original. Además, en los casos de *retweets* de usuarios con alta popularidad, estos pueden representar datos redundantes y no aportar valor adicional al análisis.
3. **Bots:** Se estableció una estrategia para intentar filtrar los mensajes generados por bots. La estrategia consistió en incluir como máximo 3 *tweets* con un contenido textual idéntico. De esta manera, se busca limitar la influencia de los bots, manteniendo un equilibrio para no excluir datos genuinos que puedan tener similitudes.
4. **Noticias:** Se decidió excluir todo tipo de noticias relacionadas con los candidatos o las opciones de voto. Esto se debe a que, tras una evaluación preliminar, se determinó que muchas de estas noticias no aportaban valor al tipo de análisis propuesto y podrían sesgar o desviar los resultados.

La fase de **Preprocesamiento de datos** se inició inmediatamente después de concluir la etapa de **Selección de datos**, momento en el cual se dedicó el tiempo necesario a examinar y procesar la información acumulada. Se utilizó un *script* programado en Python especialmente diseñado para limpiar y depurar los datos, seleccionando exclusivamente aquellos elementos que resultaban de utilidad para los fines investigativos.

No obstante, es pertinente destacar que la cantidad de datos útiles se vio considerablemente reducida tras este paso, lo que condujo a la decisión de reiterar la etapa de **Selección de Datos**. Este procedimiento se efectuó en varias ocasiones, hasta lograr acumular un volumen de información suficientemente amplio y de la calidad requerida para el estudio. Tal situación se evidencia en la Tabla 5, donde se observa la cantidad de datos después de cada iteración.

Elección	Año	Iteración 1	Iteración 2	Iteración 3	Iteración 4
Elecciones constitucionales	2022	8207	14576	18307	-
Elecciones presidenciales	2021	8115	15031	18429	-
Proceso constitucional	2020	8215	14813	18284	-
Elecciones presidenciales	2017	6936	10525	14328	18055

Tabla 5: Cantidades de datos posteriores a las iteraciones.

Fuente: Elaboración propia.

Este método iterativo aseguró que, a pesar de los criterios de filtrado aplicados, se contara con suficiente información de calidad para proceder con las etapas subsiguientes del estudio.

3.2.3. Transformación de Datos

Tras completar la etapa de **Preprocesamiento de datos**, se procedió a la etapa de **Transformación de Datos**. Esta fase fue fundamental para optimizar el tiempo de procesamiento de los datos, una tarea importante debido a la cantidad de información recopilada.

El enfoque principal fue conservar exclusivamente aquellos elementos de datos que eran de importancia central para el análisis, descartando otros aspectos que ya habían cumplido su propósito en etapas previas o que no aportaban valor significativo. Esto incluyó la eliminación de datos como *likes*, *quotes*, avatares, entre otros, centrándose en preservar primordialmente el texto escrito por el usuario.

Una vez realizada esta filtración y transformación, se obtuvo un conjunto de datos más limpio, conciso y directamente relevante para los objetivos del estudio. Este conjunto de datos refinado fue entonces almacenado en un formato estructurado, específicamente en un archivo JSON. La estructura final de estos datos transformados se muestra en la Figura 6, proporcionando una visualización clara de cómo se organizó la información para las siguientes fases del análisis.

```
[{
  "text": "No había visto la franja del apruebo en que participa Gustavo Gatica. La vi hoy. Se me llenaron los ojos de lágrimas. No puedo con tanta miseria. Tenemos que refundar este país. No hay otra opción. Por Gustavo, por Fabiola, por Anthony. Por todos.",
},
{
  "text": "la franja del rech4zo me da asco, más ganas de votar apruebo me da",
},
{
  "text": "La sola presencia de Jorge González en la franja del apruebo tiene más peso que todos los argumentos del Rechazo",
},
]
```

Figura 6: Estructura de archivos JSON obtenidos tras la transformación de datos.

Fuente: Elaboración propia.

Este enfoque metódico en la transformación de datos asegura que el análisis siguiente sea eficiente y se concentre en información de alta relevancia, maximizando así la utilidad y aplicabilidad de los hallazgos obtenidos.

3.2.4. Minería de Datos

Una vez completada la **Transformación de Datos**, se inició la etapa de **Minería de Datos**. El enfoque principal en esta fase fue procesar e interpretar todos los conjuntos de datos obtenidos, transformándolos de meros registros a conclusiones significativas basados en estadísticas y análisis.

Para esto se inició con el procesamiento de datos realizando clasificación de cada uno de los *tweets*; el primer paso fue emplear la librería TextBlob, que se introdujo previamente en la sección 2.2.3. Esta herramienta fue utilizada para traducir los *tweets* del español al inglés. La razón detrás de este procedimiento radica en que la tecnología de análisis de sentimiento es más avanzada y eficaz en inglés que en español, por lo cual era un paso importante para unos resultados mas exactos.

Una vez que los *tweets* estuvieron en inglés, se procedió con el *testing* de cada uno de los modelos de análisis de sentimiento seleccionados para este estudio. Los modelos que fueron considerados para el *testing* son los siguientes:

- **TextBlob**: Utilizado inicialmente para la traducción, esta biblioteca también ofrece la funcionalidad `blob.sentiment.polarity`, que se emplea para determinar la polaridad de los *tweets*. Debido a su doble utilidad, fue una elección lógica para este análisis.
- **NLTK**: Se incorporó el uso de la biblioteca NLTK, en particular su módulo `nltk.sentiment.vader`, que está específicamente dirigido para analizar texto de redes sociales, gracias a su capacidad de interpretar la polaridad de los mensajes.
- **Transformers (de Hugging Face)**: Por último se optó por la biblioteca Transformers que, mediante su función *pipeline*, se pueden utilizar diversos modelos que se enfocan en diferentes tipos de texto, en este caso se seleccionó **“cardiffnlp/twitter-roberta-base-sentimen”** el cual es un modelo que destaca por su habilidad para entender los textos considerando el contexto analizado en su totalidad y además fue entrenado en base a textos de Twitter lo cual facilita la comprensión de estos.

La implementación de estos tres modelos se sometió a evaluación mediante pruebas con una muestra de aproximadamente 1300 *tweets*. Estos *tweets* fueron recopilados de diversas fuentes y posteriormente se les asignaron las connotaciones adecuadas a cada uno, con el objetivo de evaluar la eficacia de cada modelo en términos de su alineación con las connotaciones correspondientes. Los resultados de estas pruebas preliminares se ilustran en la Tabla 6, donde se detalla el porcentaje de acierto en cada caso.

Biblioteca	% Tweets positivos	% Tweets negativos	% Tweets Neutros	% General
TextBlob	68,59%	39,18%	48,21%	52,13%
NLTK	87,13%	61,22%	59,37%	70,59%
Transformers	94,93%	95,01%	57,17%	90,03%

Tabla 6: Resultados de testeo de los modelos de análisis de sentimiento.
Fuente: Elaboración propia.

Analizando los resultados obtenidos en la fase de prueba, se observó que la función *pipeline* de la biblioteca Transformers mostró un rendimiento sobresaliente con un porcentaje general de éxito del 85,07 %. En segundo lugar, NLTK obtuvo un 77,09 %, mientras que TextBlob quedó en una posición más distante con un 58,13 %. A pesar de ser resultados prometedores, se identificó la posibilidad de mejorarlos aún más mediante la implementación de estrategias adicionales:

- **Emoticones**: Se construyó un diccionario compuesto por 23 emoticones, con el propósito de ajustar la polaridad obtenida por los métodos de análisis de sentimientos. Cada emoticón puede incrementar, disminuir o neutralizar la polaridad, según corresponda.

- **Palabras Clave:** A través de un *script* en Python, se examinaron las palabras más repetidas en todos los *tweets* recopilados. Se identificaron las 100 palabras más comunes asociadas con la jerga chilena y de difícil comprensión en inglés. A estas palabras se las sustituyó por sus equivalentes en español para mantener su significado tras la traducción.

La integración de estas mejoras tuvo un impacto positivo en los resultados de todos los modelos, tal como se muestra en la Figura 7. Sin embargo, se observó que algunos modelos se beneficiaron más que otros de estos ajustes. TextBlob, por ejemplo, experimentó un aumento considerable en su rendimiento, lo cual era esperable dado su bajo rendimiento inicial y su orientación hacia textos más formales. NLTK también mostró una mejora, aunque menos pronunciada, ya que su especialización en texto de redes sociales ya incluye un análisis integrado de emoticones, por lo tanto, su avance se debió en gran medida a la adición de palabras clave. Por su parte, la biblioteca Transformers experimentó solo una leve mejora, ya que sus resultados iniciales eran altamente efectivos. Su diseño para analizar textos informales y su habilidad para comprender múltiples idiomas, incluyendo algunos emoticones, de por sí le otorgaban una lectura muy completa al analizar *tweets*.

Biblioteca	% Tweets positivos	% Tweets negativos	% Tweets Neutros	% General
TextBlob	75,00%	44,18%	52,15%	57,92%
NLTK	90,55%	69,21%	71,55%	78,92%
Transformers	98,20%	95,32%	61,81%	91,55%

Tabla 7: Mejora de los resultados tras la implementación de emoticones y palabras clave.

Fuente: Elaboración propia.

La evaluación detallada de los resultados ha permitido discernir cuáles son los modelos de análisis de sentimientos más eficaces para este estudio en particular, destacándose la función *pipeline* de la biblioteca Transformers por su alta efectividad.

En preparación para los siguientes pasos, los cuales se explican con mayor detalle en el capítulo siguiente, se han seleccionado cuidadosamente los modelos más adecuados basándose en su capacidad para detectar connotaciones positivas, negativas y neutrales en los *tweets*. Una vez elegidos, se diseñaron algoritmos específicos para analizar exhaustivamente todos los *tweets* recopilados, asignándoles una connotación definida por los modelos seleccionados. Esto permitió obtener una visión integral de las cantidades y porcentajes de *tweets* con connotaciones positivas y negativas. Estos datos, luego, se compararon con los resultados reales de las elecciones, utilizando algunas visualizaciones para enriquecer y clarificar el análisis.

Adicionalmente, se realizó un análisis general que incluyó la creación de una visualización de las palabras más utilizadas en relación con cada opción candidata. Este análisis se hace con el objetivo de discernir de manera más clara cuáles eran los argumentos y temas más comunes a favor o en contra de cada candidato, proporcionando una comprensión más profunda de las dinámicas de la opinión pública.

Por otro lado, se implementó una metodología cuantitativa para asignar un “puntaje” a cada opción de candidato, basado en las connotaciones derivadas de los *tweets*. Este enfoque permite calcular un porcentaje comparativo entre las opciones, reflejando una estimación final la cual se puede contrastar directamente con los resultados electorales reales.

Este paso se aplicó a todos los procesos constitucionales y elecciones estudiadas. Además, se realizó un caso de estudio específico: las elecciones constitucionales de 2023. En este escenario, se utilizaron datos recopilados antes del anuncio de los resultados electorales para poner a prueba el modelo desarrollado en un contexto real, proporcionando una gran oportunidad para evaluar la capacidad predictiva de la metodologías en un entorno práctico y dinámico.

CAPÍTULO 4

VISUALIZACION Y ANÁLISIS DE RESULTADOS

En este capítulo, se prosigue con la metodología planteada, enfocándose en el análisis de los resultados obtenidos y la derivación de conclusiones a partir de estos.

4.1. Minería de datos

Continuando con la etapa de **Minería de Datos**, es fundamental analizar los modelos clasificadores de sentimientos y sus resultados para determinar como utilizarlos y ponerlos a prueba en las elecciones seleccionadas.

4.1.1. Análisis de los modelos clasificadores de sentimiento

Como se mencionó anteriormente, la elección de los modelos clasificadores de sentimientos adecuado es fundamental para asegurar la precisión en el análisis de los datos. En el capítulo previo, se dió a conocer que los tres modelos clasificadores de sentimiento mostraron resultados muy positivos tras ser sometidos a pruebas de rendimiento. Esto se puede visualizar en la Figura 7, donde se compararon los resultados obtenidos por los tres modelos mediante un gráfico de barras, utilizando la herramienta *Flourish*.

El análisis de la figura revela que el rendimiento de la función de Transformers sobresale consistentemente sobre los otros modelos en cada categoría evaluada, demostrando ser la herramienta más confiable para la clasificación de sentimientos.

Aunque los resultados son prometedores, se identificó la posibilidad de mejorar aún más su precisión. Dada la importancia de obtener los resultados más fiables posibles para este proyecto, se implementaron dos nuevas estrategias para mejorar la efectividad de los modelos de clasificación:

- **Emoticones:** se construyó un diccionario compuesto por 23 emoticones, con el propósito de ajustar la polaridad obtenida por los métodos de análisis de sentimientos.
- **Palabras Clave:** a través de un script en Python, se examinaron las palabras más repetidas en todos los tweets recopilados. Se identificaron las 100 palabras más comunes asociadas con la jerga chilena, y de difícil comprensión en inglés, y se sustituyeron.

La presencia de emoticones en un texto juega un papel importante en su interpretación, pudiendo hacer que un mensaje sea más difícil de leer y, en consecuencia, más complica-

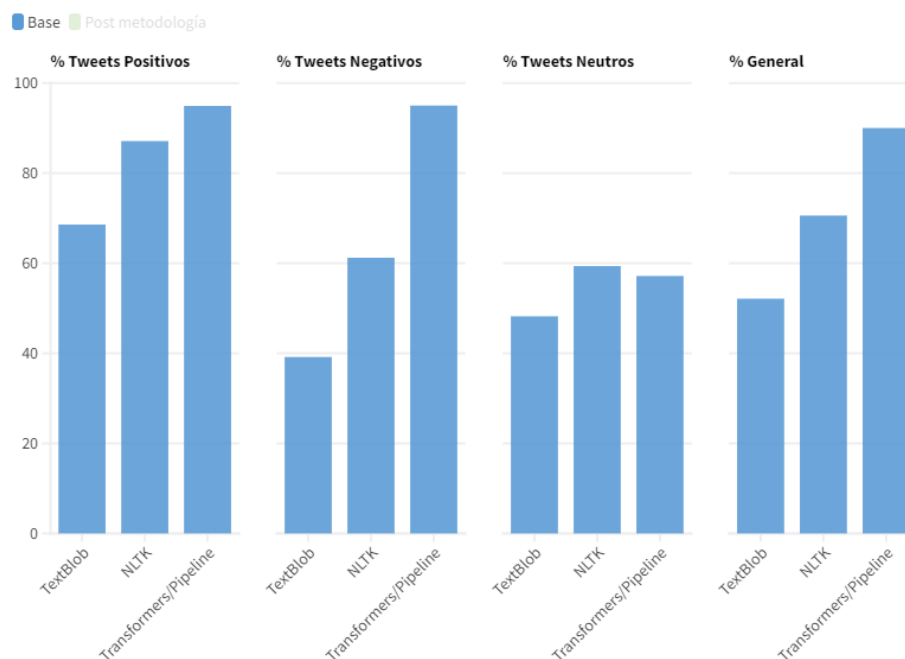


Figura 7: Comparación de modelos clasificadores de sentimiento

Fuente: Elaboración propia.

do de analizar adecuadamente. Reconociendo esta influencia, se desarrolló un diccionario compuesto por 23 emoticones diseñado para ajustar el puntaje de polaridad otorgado por los modelos de análisis de sentimientos.

Un ejemplo ilustrativo de esta influencia se observa en la biblioteca TextBlob. Al procesar la frase "Me encanta este candidato :D", TextBlob asigna una polaridad de 0,50, indicando correctamente que se trata de una expresión positiva con un 50 % de certeza. Sin embargo, al analizar la frase "Me encanta este candidato" sin el emoticón, el modelo entrega el mismo nivel de polaridad. Esto sugiere que, en su estado original, los emoticones no incrementan ni disminuyen el nivel de confianza asignado a una connotación específica.

Para abordar esta limitación, se implementó un enfoque que extrae el emoticón del texto antes de su procesamiento. De esta forma, se obtiene el nivel de polaridad y connotación del modelo y también se ajusta la puntuación según corresponda basándose en el emoticón extraído. Este ajuste no solo altera la magnitud de la polaridad, sino que también puede cambiar completamente la connotación del texto.

Un claro ejemplo de este fenómeno es la frase "Boric es uno de los presidentes más jóvenes del mundo", la cual obtiene una polaridad de 0,29 con TextBlob, indicando una tonalidad mayormente neutra con una leve inclinación positiva. Sin embargo, al modificar la frase a "Boric es uno de los presidentes más jóvenes del mundo >c:", transforma radicalmente el mensaje a uno de connotación negativa, reflejando descontento hacia la juventud e inexperiencia

del presidente. En este escenario, la polaridad se ajusta a -0,29, efectivamente invirtiendo su valor para capturar el cambio en la emoción transmitida por el mensaje, demostrando la capacidad del diccionario de emoticones para modificar de manera efectiva la interpretación del modelo.

Al igual que con los emoticones, el uso de palabras clave y jerga local presenta un desafío similar en el análisis de sentimientos, particularmente en contextos específicos como el chileno. En Chile, se emplean numerosas expresiones y palabras en el día a día que son poco comunes en otras partes de Latinoamérica y, por supuesto, en el resto del mundo. Este uso particular del lenguaje puede dificultar la comprensión no solo entre personas, sino también para los modelos de análisis de sentimientos, especialmente cuando se trabaja con textos traducidos al inglés. Por esta razón, se decidió implementar una estrategia de reemplazo de palabras para adaptar la jerga local al contexto de estos modelos.

Esta estrategia de adaptación funciona de la siguiente manera: se identifican las palabras más utilizadas en los *tweets* recopilados que pueden ser incomprensibles o perder su significado en la traducción al inglés, y se reemplazan por términos equivalentes en español más estándar. Por ejemplo, el insulto chileno “weon” o “wn” se sustituye por “tonto”. Así, una frase como “el candidato es muy wn” se transformaría en “el candidato es muy tonto”. Este cambio no solo hace que el texto sea más comprensible para un público más amplio, sino que también mejora la capacidad del modelo para analizar correctamente la polaridad del mensaje.

En términos de análisis de sentimientos con TextBlob, este cambio tiene un impacto significativo. Ya que la frase original “el candidato es muy weon” recibe una polaridad de 0.0, indicando una falta de comprensión y, por lo tanto, una incapacidad para asignar una connotación, la frase modificada “el candidato es muy tonto” recibe una polaridad de -0.53. Este puntaje refleja una clara connotación negativa y una confianza del 53 %, lo cual demuestra la efectividad de esta estrategia para mejorar la precisión del modelo en el análisis de sentimientos.

Con este nuevo enfoque en mente, se procedió a procesar los 1300 datos de prueba disponibles. Se reemplazaron todas las palabras difíciles de entender y se realizó la modificación en los emoticones. Los resultados obtenidos tras estas adaptaciones se muestran en la ya presentada Tabla 7, proporcionando una visión clara del impacto de estas estrategias en la efectividad del análisis de sentimientos.

Tras la implementación de las estrategias de ajustes y modificación, todos los modelos experimentaron cambios en su rendimiento, algunos leves y otros más drásticos. Sin embargo, cada ajuste influyó positivamente en el aumento de los porcentajes generales de rendimiento de cada biblioteca utilizada.

En el caso de TextBlob, se observó un incremento del 5,79 % en su rendimiento general, logrando resultados significativamente mejores en cada aspecto evaluado. En *tweets* positivos, el rendimiento aumentó en un 6,41 %, alcanzando el 75 % de rendimiento, el cual no es un

mal porcentaje pero sigue estando por debajo del resto de bibliotecas, en *tweets* neutros se registró un aumento del 3,94 %, mientras que en *tweets* negativos sigue con un mal porcentaje. Estas mejoras pueden atribuirse a varios factores, pero principalmente se debe a que TextBlob, como se mencionó en la sección 2.2.3, es una biblioteca con múltiples enfoques y cuyo análisis de sentimientos está más orientado a textos formales, generando así algunas dificultades para interpretar los mensajes informales de redes sociales. Por lo tanto, la estrategia implementada fue crucial para reorientar estos *tweets* informales hacia un estilo más formal, mejorando así la precisión de los resultados obtenidos con TextBlob.

Por otro lado, NLTK fue la biblioteca que más se benefició de las estrategias aplicadas, con un aumento del 8,33 % en su rendimiento general, una mejora notable considerando que sus resultados iniciales ya eran bastante buenos. En *tweets* positivos, su precisión mejoró un 3,42 %, superando el 90 %; en *tweets* negativos aumentó un 7,99 %, lo cual fue un gran aumento pero dado el porcentaje base aun queda bastante margen de mejora; y en *tweets* neutros se elevó un 12,18 %, el mayor incremento entre todas las categorías y bibliotecas. Este notable progreso se debe a que NLTK, especializada en el análisis de textos informales, intentaba interpretar palabras incomprensibles de manera errónea, lo cual se evitó en este segundo *testing* en donde las últimas modificaciones clarificaron los textos, facilitando la tarea del análisis de sentimientos.

Finalmente, la función *pipeline* de la biblioteca Transformers con su modelo *twitter-roberta-base-sentiment*. Este modelo obtuvo unos muy buenos resultados desde un comienzo, siendo muy difícil mejorarlos en la teoría y se confirmó en la práctica con un incremento del 1.52 % en su rendimiento general. Aún así los *tweets* positivos mejoraron un 3,27 %, alcanzando un impresionante 98 % de rendimiento, los negativos tuvieron un cambio mínimo, y los neutros mejoraron un 4,51 %, beneficiándose principalmente del reemplazo de palabras incomprensibles. Este modelo tuvo muy buenos resultados desde un comienzo gracias a su gran capacidad para interpretar los contextos de los textos analizados y entender cada uno de los emoticones debido a su entrenamiento basado en *tweets*. Por esta razón, la ligera mejora en resultados se debe netamente a la estrategia del reemplazo de palabras, la cual ayudó al modelo a entender algunos *tweets* que en un comienzo no podía descifrar.

Con estos resultados y la Figura 8, se puede concluir que la biblioteca Transformers con su modelo *twitter-roberta-base-sentiment* supera ampliamente a las demás en la mayoría de los aspectos, excepto en el análisis de *tweets* neutros, donde NLTK muestra una ligera ventaja. Por lo tanto, estos fueron los modelos principales utilizados para los análisis posteriores.

Con dichos resultados es sencillo determinar cual es la mejor estrategia para analizar *tweets* positivos y negativos, pero para los *tweets* neutros es algo más complejo ya que el mejor rendimiento alcanzado es un 71,55 %, un porcentaje que aún no inspira suficiente confianza para los análisis subsiguientes.

Para abordar este problema, se decidió emplear una combinación de dos modelos analizados, en particular modelo de NLTK y el modelo *twitter-roberta-base-sentiment* de la función *pipeline* de Transformers. La biblioteca TextBlob fue excluida debido a su menor aporte en

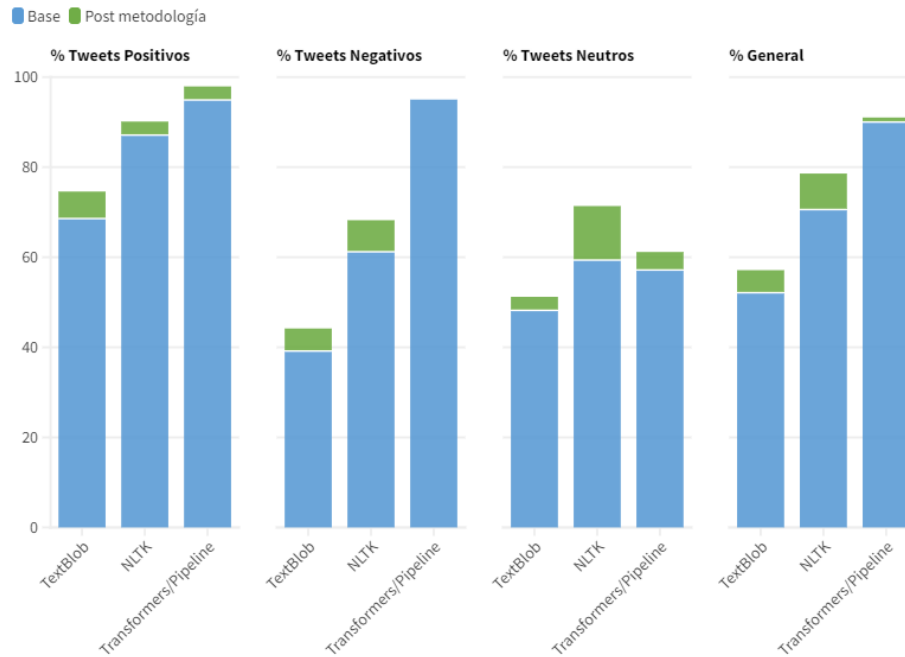


Figura 8: Gráfico de los rendimiento de modelos luego de las adaptaciones
Fuente: Elaboración propia.

términos de mejora del rendimiento y su mayor tiempo de compilación. Los criterios establecidos para determinar la neutralidad en la connotación de un *tweet* fueron los siguientes:

1. Si NLTK clasifica un *tweet* como neutral con más de un 23 % de confianza, se considera neutral.
2. Si NLTK clasifica un *tweet* como neutral con menos de un 23 % de confianza, se consulta el resultado de la función *pipeline*. Si la conclusión de *pipeline* tiene más de un 23 % de confianza, se toma como definitiva.
3. Si NLTK clasifica un *tweet* como neutral con menos de un 23 % de confianza, y la conclusión de *pipeline* también tiene menos de un 23 % de confianza, se opta por el resultado con mayor porcentaje de confianza.
4. Si NLTK clasifica un *tweet* como positivo o negativo, se utiliza la conclusión de *pipeline* independiente de cual sea.

El umbral del 23 % de confianza se seleccionó tras realizar varios *tests* para maximizar el rendimiento en la clasificación de *tweets* neutros, sin sacrificar significativamente el rendimiento en los *tweets* positivos y negativos. Se experimentó con un margen de confianza desde el 15 % hasta el 35 %, encontrando que el 23 % ofrecía el mejor equilibrio tal como se ve en la figura 9.

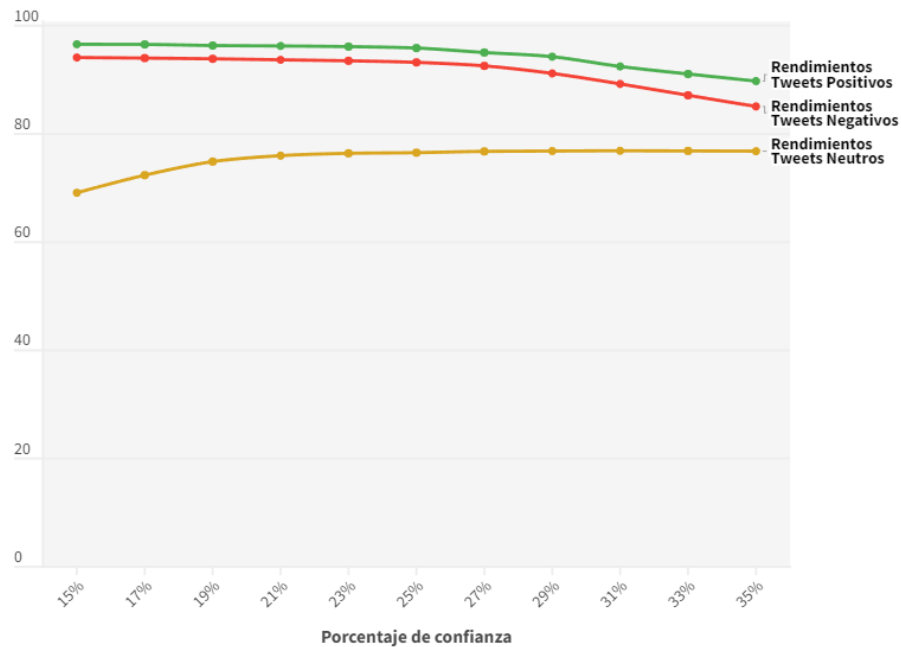


Figura 9: Grafico de rendimiento por porcentaje de confianza
Fuente: Elaboración propia.

Con estas condiciones, se logró un rendimiento final del 96,13 % en tweets positivos, 93,51 % en negativos y 76,41 % en neutros, alcanzando un rendimiento general del 92,13 %. Este algoritmo se convierte así en la base para el análisis de sentimientos de los tweets recopilados de Twitter, como se detalla en el Algoritmo 1.

4.1.2. Análisis de resultados por algoritmo de análisis de sentimiento

El principal objetivo de este trabajo es examinar detalladamente los *tweets* publicados durante las elecciones y determinar si existe alguna correlación significativa entre el contenido de estos *tweets* y los resultados electorales. Para esto es necesario determinar de alguna manera los porcentajes, por ello el algoritmo de análisis de sentimiento es un punto clave en este trabajo, al permitir la clasificación de cada *tweet* según su connotación.

Para realizar un análisis interpretable, es necesario procesar cada línea del archivo JSON, donde cada línea representa un *tweet* recopilado individualmente. Estos *tweets* deben ser analizados por el Algoritmo 1, que determinará su connotación como positiva, negativa o neutral. A continuación, se contabilizarán los *tweets* según su connotación para calcular los porcentajes de *tweets* positivos y negativos. En este estudio, se decidió no considerar los *tweets* neutros para la interpretación de resultados, ya que estos no contribuyen a la comprensión de las tendencias o sentimientos generales y, además, guardan paralelismo con los votos en blanco y nulos en las elecciones, que no se cuentan en el resultado final de votos.

Algoritmo 1 Algoritmo de análisis de sentimiento con manejo de emoticones

Require: Un tweet como entrada

Ensure: La connotación del tweet como salida

#Quitar emoticones de los tweets

$(tweet, conn_emoticon, conf_emoticon) \leftarrow eliminar_emoticon(tweet)$

#Proceso de analisis de sentimientos

$array_NLTK \leftarrow NLTK(tweet)$

$array_pipeline \leftarrow Pipeline(tweet)$

#Ajuste de connotación y confianza considerando los emoticones

$(conn_NLTK, conf_NLTK) \leftarrow rec_emoticon(array_NLTK, conn_emoticon, conf_emoticon)$

$(conn_pipeline, conf_pipeline) \leftarrow rec_emoticon(array_pipeline, conn_emoticon, conf_emoticon)$

#Asignación de variables finales

$(connotacion_NLTK, confianza_NLTK) \leftarrow (conn_NLTK, conf_NLTK)$

$(connotacion_pipeline, confianza_pipeline) \leftarrow (conn_pipeline, conf_pipeline)$

if $connotacion_NLTK = "Neutral"$ & $confianza_NLTK \geq 0,23$ **then**

return $connotacion_NLTK$

else if $connotacion_NLTK = "Neutral"$ & $confianza_NLTK < 0,23$ **then**

if $confianza_pipeline \geq 0,23$ **then**

return $connotacion_pipeline$

else

if $confianza_pipeline \geq confianza_NLTK$ **then**

return $connotacion_pipeline$

else

return $connotacion_NLTK$

end if

end if

else

return $connotacion_pipeline$

end if

Para determinar los porcentajes correspondientes, se empleó el Algoritmo 2, que procede a leer cada uno de los tweets, aplica el Algoritmo 1 para establecer la connotación de estos y, finalmente, calcula el porcentaje final.

Algoritmo 2 Cálculo de porcentajes de tweets negativos y positivos

Require: Nombre de archivo como entrada

Ensure: Porcentaje de tweets negativos y positivos como salida

```
archivo ← lectura_csv(nombre_de_archivo)
contador_positivo ← 0
contador_negativo ← 0

while tweet ∈ archivo do
    tweet ← reemplazar_palabras(tweet)
    connotacion ← algorithm1(tweet)
    if connotacion = "Positivo" then
        contador_positivo ← contador_positivo + 1
    else if connotacion = "Negativo" then
        contador_negativo ← contador_negativo + 1
    end if
end while

porcentaje_positivo ←  $\frac{\text{contador\_positivo}}{\text{contador\_positivo} + \text{contador\_negativo}}$ 
porcentaje_negativo ←  $\frac{\text{contador\_negativo}}{\text{contador\_positivo} + \text{contador\_negativo}}$ 

return porcentaje_positivo, porcentaje_negativo
```

Utilizando el Algoritmo 2, se calculan los porcentajes de tweets positivos y negativos para cada conjunto de datos de elección. Este enfoque proporciona una medida cuantitativa clara de los sentimientos generales expresados en Twitter durante los periodos electorales, permitiendo un análisis más profundo de las tendencias y patrones subyacentes.

Finalmente, se añade un sistema de clasificación diseñado para ofrecer un resultado preciso que pueda actuar como predictor de los resultados electorales. Este sistema emplea un método sencillo pero eficaz, basado en el estudio “*Statistically Weighted Reviews to Enhance Sentiment Classification*”, que propone que un enfoque combinando la ponderación de Palabras con la ponderación de Sentimientos es una de las mejores manera de obtener un rendimiento general, se va a utilizar en este caso, asignando como resultado final la suma de las ponderaciones asignadas a cada tweet analizado. Dichos puntajes se calculan mediante una fórmula directa, que consiste en la sumatoria de las multiplicaciones de la confianza obtenida por cada biblioteca de analisis de sentimiento con su respectivo porcentaje de efectividad.

$$\sum_{i=1}^n (confianza_i \times efectividad_i) \quad (1)$$

donde n es el número total de librerías.

El algoritmo 3 aplica las ponderaciones antes mencionadas, teniendo en cuenta dos factores clave, la eficacia demostrada de cada librería según los resultados de pruebas anteriores y la confianza que cada una tiene en sus propias predicciones. De esta manera, se logra un equilibrio óptimo entre la precisión y la fiabilidad de las predicciones generadas a partir de los datos de Twitter.

Con todos los algoritmos necesarios ya desarrollados, ha llegado el momento de la etapa final de la fase de **Minería de Datos**. Este último paso implica la aplicación de estos algoritmos a cada elección presidencial y constitucional, con el objetivo de observar y analizar sus resultados finales.

4.1.3. Elecciones presidenciales 2017

El año 2017 marcó un momento políticamente complejo y trascendental en la historia reciente de Chile. Durante este período, el país experimentó intensos escándalos políticos y una creciente división social, que se reflejó en una tasa de desaprobación del 70 % para la entonces presidenta, Michelle Bachelet.

Esta época estuvo caracterizada por una notable fragmentación y polarización en el panorama político chileno. Las divisiones internas dentro de los partidos y alianzas tradicionales, así como el creciente descontento estado actual del país, llevaron a una amplia diversidad de candidaturas presidenciales. Este escenario reflejaba un amplio espectro de opiniones políticas y una clara señal de la demanda de cambio por parte de la ciudadanía.

La fragmentación política se tradujo en una elección presidencial con un número significativo de candidatos, concluyendo con un total de ocho postulantes que representaban a distintas coaliciones y espectros políticos. Estos candidatos provenían de una variedad de orígenes, incluyendo cuatro independientes, representantes del Partido Demócrata Cristiano, el Partido Progresista, la Unión Patriótica y el Partido Político País. Cada uno de ellos, apoyado por sus respectivas coaliciones, se embarcó en la contienda electoral con el objetivo de llegar a la presidencia (Tabla 8).

Bajo este contexto, Chile se dirigió a las urnas el 19 de noviembre de 2017. Según las encuestas previas, Sebastián Piñera aparecía como el claro favorito, seguido a distancia por Alejandro Guillier, mientras que Beatriz Sánchez se proyectaba en un tercer lugar con una intención de voto que fluctuaba entre el 10 % y el 15 %.

Algoritmo 3 Algoritmo de análisis de sentimiento con puntuación

Require: tweets como entrada

#Quitar emoticones de los tweets

$(tweet, conn_emoticon, conf_emoticon) \leftarrow eliminar_emoticon(tweet)$

#Análisis de sentimientos

$array_NLTK \leftarrow NLTK(tweet)$

$array_pipeline \leftarrow Pipeline(tweet)$

$array_TBlob \leftarrow TextBlob(tweet)$

#Ajuste de connotación y confianza considerando los emoticones

$(conn_NLTK, conf_NLTK) \leftarrow rec_emoticon(array_NLTK, conn_emoticon, conf_emoticon)$

$(conn_pipeline, conf_pipeline) \leftarrow rec_emoticon(array_pipeline, conn_emoticon, conf_emoticon)$

$(conn_TBlob, conf_TBlob) \leftarrow rec_emoticon(array_TBlob, conn_emoticon, conf_emoticon)$

#Asignación de variables finales

$(connotacion_NLTK, confianza_NLTK) \leftarrow (conn_NLTK, conf_NLTK)$

$(connotacion_pipeline, confianza_pipeline) \leftarrow (conn_pipeline, conf_pipeline)$

$(connotacion_TBlob, confianza_TBlob) \leftarrow (conn_TBlob, conf_TBlob)$

$(porcentaje_TBlob_pos, porcentaje_TBlob_neg) \leftarrow (0,75, 0,52)$

$(porcentaje_NLTK_pos, porcentaje_NLTK_neg) \leftarrow (0,91, 0,69)$

$(porcentaje_pipeline_pos, porcentaje_pipeline_neg) \leftarrow (0,98, 0,95)$

if $connotacion_NLTK = "Positiva"$ **then**

$(puntuacion) + (confianza_NLTK * porcentaje_NLTK_pos) \leftarrow (puntuacion)$

else

$(puntuacion) - (confianza_NLTK * porcentaje_NLTK_neg) \leftarrow (puntuacion)$

end if

if $connotacion_pipeline = "Positiva"$ **then**

$(puntuacion) + (confianza_pipeline * porcentaje_pipeline_pos) \leftarrow (puntuacion)$

else

$(puntuacion) - (confianza_pipeline * porcentaje_pipeline_neg) \leftarrow (puntuacion)$

end if

if $connotacion_TBlob = "Positiva"$ **then**

$(puntuacion) + (confianza_TBlob * porcentaje_TBlob_pos) \leftarrow (puntuacion)$

else

$(puntuacion) - (confianza_TBlob * porcentaje_TBlob_neg) \leftarrow (puntuacion)$

end if

Candidato(a)		Partido		Coalición/Partido	
	Sebastián Piñera Echenique		Ind.		Chile Vamos
	Alejandro Guillier Álvarez		Ind.		La Fuerza de la Mayoría
	Beatriz Sánchez Muñoz		Ind.		Frente Amplio
	José Antonio Kast Rist		Ind.		Independiente
	Carolina Goic Borojevic		PDC		Partido Demócrata Cristiano
	Marco Enríquez-Ominami Gumucio		PRO		Partido Progresista
	Eduardo Artés Brichetti		UPA		Unión Patriótica
	Alejandro Navarro Brain		País		País

Tabla 8: Candidatos y coaliciones en las elecciones presidenciales de 2017.

Fuente: Elaboración propia.

Finalmente, los resultados de la elección mostraron a Sebastián Piñera liderando, aunque con un margen menos amplio de lo esperado, seguido por Alejandro Guillier, cuyo porcentaje de votos estuvo bastante alineado con las predicciones de las encuestas. Sin embargo, la verdadera sorpresa de la elección fue Beatriz Sánchez, quien alcanzó un 20,27 % de los votos, casi desplazando a Guillier de la segunda vuelta por una diferencia de solo un 2,5 %. Este resultado no fue anticipado por ninguna de las encuestadoras. El resto de los candidatos obtuvo porcentajes menores, siendo José Antonio Kast el más destacado con un 7 %, un porcentaje muy relevante para sus futuras aspiraciones políticas (Figura 10).

Como resultado, Alejandro Guillier y Sebastián Piñera avanzaron a la segunda vuelta electoral. Este escenario proporcionó una gran oportunidad para aplicar y evaluar el algoritmo de análisis de sentimientos desarrollado, analizando las conversaciones y opiniones expresadas en Twitter durante el periodo electoral. Este análisis permitió dar los primeros acercamientos hacia los objetivos propuestos, y determinar si efectivamente existe una correlación entre los sentimientos expresados en la red social y los resultados electorales finales, poniendo a prueba la eficacia de los algoritmos desarrollados en un contexto político.

Para iniciar este análisis, se analizaron los tweets recopilados con el fin de realizar un estudio visual sobre las palabras más utilizadas y comprender mejor el contexto y las temáticas asociadas a cada candidato. En la imagen 11 se destacan las palabras más repetidas en los tweets referentes al candidato Sebastián Piñera. A primera vista, los tweets recopilados ofrecen un gran resumen de los temas principales que se comentaban en relación con la candidatura de

Votos y porcentaje por candidato

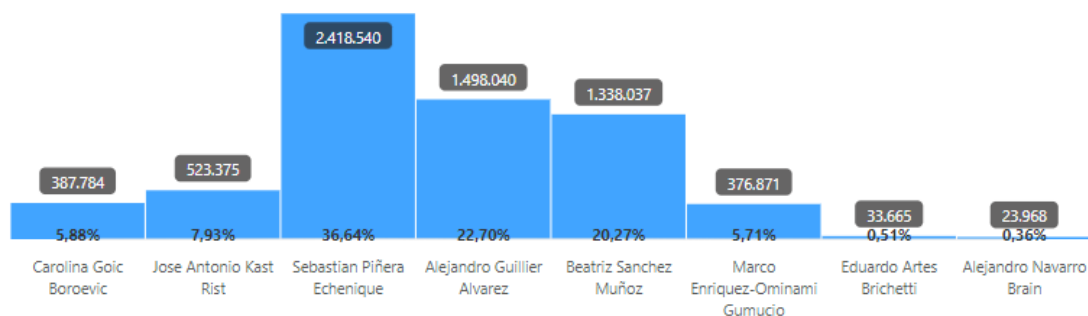


Figura 10: Resultados de la primera vuelta de las elecciones presidenciales de 2017.
Fuente: Elaboración propia.

Sebastián Piñera. En el análisis visual, los temas negativos asociados a Piñera están resaltados en rojo, incluyendo críticas que lo califican como un retroceso para el país o acusaciones de corrupción. Por otro lado, también se destacan temas utilizados como argumentos a favor de su candidatura, las cuales se resaltan en verde, como las referencias hacia el progreso que representaría su elección, comparaciones negativas de Guillier con Venezuela (o “Chilezuela”), y un enfoque en políticas de libre mercado y reducción de impuestos. Finalmente, las palabras en naranja representan los temas principales de debate, incluyendo comentarios tanto positivos como negativos. Estos abarcan desde la eterna discusión entre posturas de izquierda y derecha hasta comparaciones entre candidatos.

Este análisis sugiere que los *tweets* reflejan de manera general, pero acertada, las discusiones y argumentos predominantes durante la campaña de Sebastián Piñera. Esto demuestra que las opiniones en Twitter estaban alineadas con los temas y acontecimientos políticos reales de ese período.

Continuando con el análisis de palabras sobre las elecciones presidenciales de 2017, se examinó la conversación pública en torno al candidato Alejandro Guillier. A diferencia del análisis realizado para Sebastián Piñera, los *tweets* relacionados con Guillier presentan un enfoque distinto, como se observa en la Figura 12.

En los *tweets* referentes a Guillier, se destaca la mención frecuente a terceras personas y entidades políticas, tal como se evidencia en las palabras resaltadas en naranja. Figuras como Bachelet, el Frente Amplio, el alcalde Sharp y la periodista Ximena Rincón aparecen con notable recurrencia. Este patrón sugiere que la conversación en torno a Guillier estaba fuertemente influenciada por su relación con otras figuras políticas y temas de actualidad.

En cuanto a los aspectos positivos asociados a Guillier, se mencionan temas como la igualdad, las reformas impulsadas por la presidenta Bachelet, la reestructuración de las AFP, la educación y el rechazo a la campaña del terror. Sin embargo, estos temas positivos tuvieron una presencia relativamente baja en comparación con otros temas.

Los resultados obtenidos del análisis de los *tweets* durante las elecciones presidenciales de 2017 revelan una tendencia interesante. Según los datos, el candidato Sebastián Piñera tuvo un desempeño favorable en Twitter, con un porcentaje más alto de comentarios positivos y un porcentaje más bajo de comentarios negativos en comparación con su contrincante, Alejandro Guillier. Esta tendencia en las redes sociales sugiere una mayor inclinación hacia un resultado electoral positivo para Piñera, lo que efectivamente coincidió con el resultado final de las elecciones.

Esta correlación entre la opinión pública en Twitter y los resultados electorales se evidencia aún más en la figura 14. En ésta, se compara la proporción de comentarios positivos y negativos de cada candidato con sus respectivos porcentajes de votos oficiales. Se observa que el porcentaje de comentarios positivos hacia Piñera actúa como una cota superior aproximada a su resultado electoral, mientras que el porcentaje de comentarios negativos se presenta como una cota inferior.

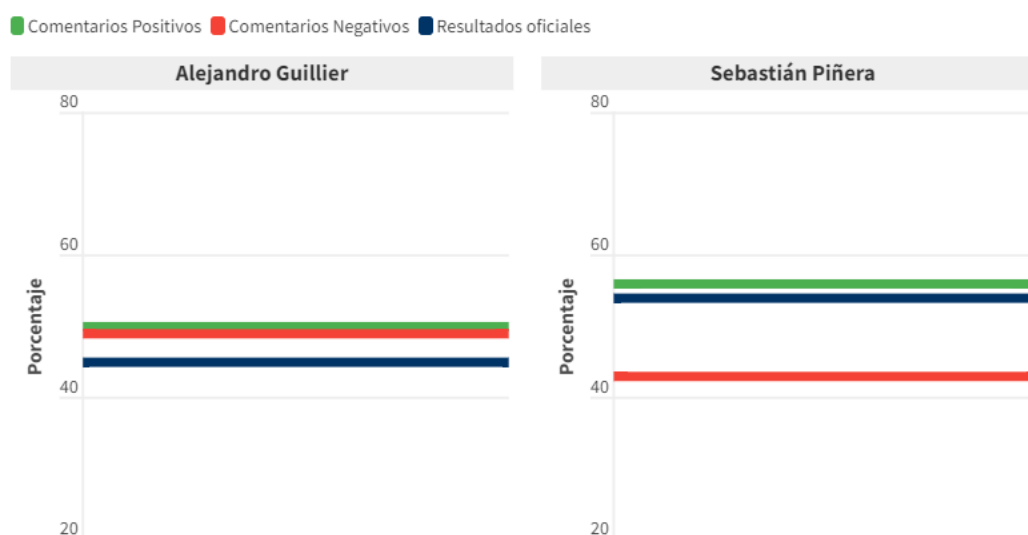


Figura 14: Comparación entre resultados oficiales y porcentajes de comentarios en Twitter.
Fuente: Elaboración propia.

En el caso de Alejandro Guillier, los porcentajes derivados de Twitter no reflejan de manera tan exacta los resultados electorales. Con un 50,2 % de comentarios positivos y un 49,8 % de comentarios negativos, los datos están alejados del 45,53 % obtenido en los resultados oficiales. A pesar de esta discrepancia, la tendencia general en Twitter hacia Guillier fue más baja que la observada para Piñera, lo que de alguna manera también señala una tendencia hacia un resultado menos favorable para Guillier.

Por lo que en base a este análisis, la tendencia observada en los datos de Twitter durante las elecciones presidenciales de 2017 apunta hacia un claro favoritismo por Sebastián Piñera. Sin embargo, para obtener una evaluación más precisa y cuantitativa, se decidió utilizar la última fórmula y con ello el Algoritmo 3. Esta fórmula, detallada en la ecuación 1, se diseñó

para calcular un puntaje final que refleje un porcentaje más exacto de la comparación entre los candidatos.

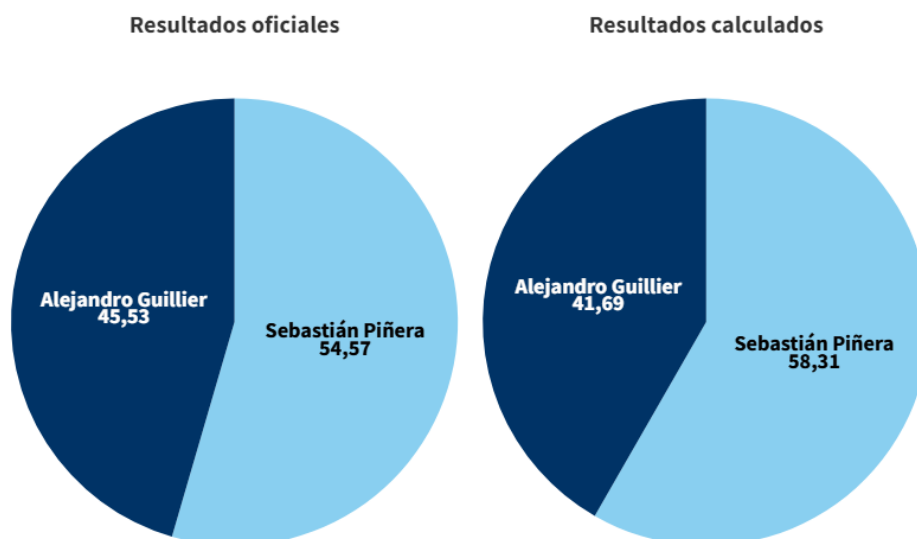


Figura 15: Comparación final de resultados de las elecciones presidenciales año 2017
Fuente: Elaboración propia.

Tras la implementación del último algoritmo de análisis, se lograron obtener los resultados finales, los cuales se presentan en la Figura 15. Estos resultados confirmaron la hipótesis inicial, es decir, existía una preferencia notable por Sebastián Piñera en comparación con Alejandro Guillier en las conversaciones de Twitter. Interesantemente, el porcentaje obtenido mediante este cálculo mostró una gran precisión, con un margen de un 3,74 %. Este resultado es particularmente significativo, ya que indica un principio de correlación que servirá para el análisis de las siguientes elecciones.

4.1.4. Proceso constitucional 2020

El año 2020 fue uno de los años más importantes en la historia de Chile, marcado por la realización del Plebiscito Nacional el día 25 de octubre, un evento trascendental en la democracia chilena. Este plebiscito, resultado directo del estallido social de 2019, representó una respuesta institucional a las masivas demandas de cambios estructurales en el país, en especial en lo que respecta a su marco constitucional.

En el centro de este proceso plebiscitario se encontraban dos preguntas fundamentales. La primera interrogaba a la ciudadanía sobre la aprobación o rechazo de la redacción de una

nueva Constitución, marcando así un potencial fin para la Constitución del 1980. La segunda pregunta se centraba en el mecanismo para la redacción de la nueva carta magna: una Convención Mixta Constitucional, compuesta por miembros electos y parlamentarios en ejercicio, o una Convención Constitucional, integrada exclusivamente por ciudadanos electos para este fin.

El plebiscito se presentó como un momento de alta participación ciudadana y gran expectativa nacional e internacional. Las opciones “Apruebo” y “Rechazo” polarizaron a la sociedad chilena, reflejando no sólo diferentes visiones sobre la Constitución, sino también sobre la identidad y el futuro del país. El “Apruebo” se identificaba con la demanda de una sociedad más justa, igualitaria y con mayores garantías de derechos sociales, mientras que el “Rechazo” se asociaba a la defensa del orden institucional existente y el temor a los riesgos de una reescritura constitucional.

Este plebiscito no solo implicaba una decisión sobre el texto constitucional, sino que también reflejaba los valores que Chile quería proyectar hacia el resto del mundo. La campaña, tanto para el “Apruebo” como para el “Rechazo”, se caracterizó por un intenso debate público, manifestaciones y una significativa movilización en redes sociales y medios de comunicación.

La particularidad del Plebiscito Nacional de 2020 en Chile, centrado en la decisión entre “Apruebo” y “Rechazo” para la redacción de una nueva Constitución, era una nueva oportunidad para aplicar el algoritmo de análisis de sentimientos en un escenario diferente al de las elecciones presidenciales. En este caso, el foco no estaba en los candidatos, sino en las opciones propuestas, siendo este un reto distinto para el algoritmo.

Al igual que en la subsección anterior 4.1.3, se procedió a analizar los *tweets* recopilados en torno a la opción “Apruebo” (Figura 16) y “Rechazo” (Figura 17). Este análisis tuvo como objetivo identificar las palabras y temas más recurrentes entre los partidarios de una nueva Constitución, brindando así una visión del discurso predominante y las motivaciones detrás del apoyo a cada opción.

La visualización de las palabras más repetidas en torno a la opción “Apruebo” durante el proceso constitucional ofrece una clara visión de los argumentos y las emociones que permeaban el debate público en ese momento. Es interesante el análisis de las palabras neutras (en color dorado) que, si bien podrían considerarse ambivalentes, jugaron un papel crucial en la construcción de argumentos tanto a favor como en contra del “Apruebo”. Este fenómeno subraya términos como la franja electoral y la figura del presidente Sebastián Piñera, independiente de las opiniones negativas asociadas a su mandato, se convirtieron en elementos centrales del debate, utilizados como argumentos positivos o negativos respecto a la necesidad de una nueva Constitución.

Entre los argumentos positivos que inclinaban a los votantes hacia el “Apruebo”, resalta la figura de Jorge González, el reconocido cantante chileno que no solo apoyó esta opción, sino que también participó activamente en la franja electoral, generando reacciones positivas entre los partidarios. Además, las grandes caravanas y marchas a favor del “Apruebo” destacan

Este análisis visual sugiere, una correlación entre los temas discutidos en las redes sociales y las conversaciones cotidianas, reflejando diversas inquietudes de la población ante distintas opciones. A pesar de tratarse de una elección constitucional y no presidencial, en una primera instancia parece ser efectivo para captar el pensamiento de los ciudadanos, aunque algunos términos pueden ser irrelevantes para los análisis. Sin embargo, es importante subrayar que esta observación inicial funciona como un indicativo preliminar y no como una confirmación absoluta de la tendencia.

Con base en este entendimiento, se hace necesario proceder al análisis de cada *tweet* mediante técnicas de análisis de sentimientos, específicamente utilizando el Algoritmo 2. Este algoritmo, siguiendo la metodología empleada en la sección anterior, se dedicará a procesar y clasificar cada *tweet* de acuerdo con su connotación. Los resultados de esta clasificación se exponen en la Tabla 9, proporcionando una visión clara de las tendencias y preferencias manifestadas por los usuarios en relación con la elección constitucional en cuestión.

Elecciones	Comentarios Positivos	Comentarios Negativos
Apruebo	59,81%	40,19%
Rechazo	23,40%	76,60%

Tabla 9: Resultados sobre el análisis de connotación de *tweets* 2020.

Fuente: Elaboración propia.

El análisis realizado a los *tweets* ha revelado una tendencia pronunciada hacia la opción del “Apruebo”, destacándose por un elevado porcentaje de comentarios positivos y una mínima cantidad de comentarios negativos. Por otro lado, la opción del “Rechazo” se distingue por un porcentaje significativamente alto de comentarios negativos en comparación con el resto, lo que sugiere la presencia de una intensa campaña en contra de esta opción. Este patrón sugiere un resultado favorable para el “Apruebo” por un margen considerable, lo cual se alinea estrechamente con los resultados oficiales.

La correlación mencionada se puede visualizar gráficamente en el Figura 18, donde se observa una clara inclinación de los comentarios hacia la opción “Apruebo”. Sin embargo, cabe destacar que, aunque la tendencia es correcta, la magnitud de la inclinación es muy diferente en comparación con los resultados oficiales en el caso de los comentarios hacia el “Apruebo”, presentando una diferencia de más del 20 %.

Por otro lado, los comentarios relacionados con el “Rechazo” se reflejan con una mayor precisión, con un 23,40 % de comentarios positivos y un 76,6 % de comentarios negativos, cifras que se acercan al 78,31 % de votos para el “Apruebo” y el 21,69 % para el “Rechazo” que se obtuvieron en los resultados oficiales.

A pesar de la precisión observada en la tendencia y la exactitud de los comentarios hacia el “Rechazo”, la discrepancia notable en la cantidad de comentarios hacia el “Apruebo” indica que este es solo un resultado interpretativo y que resulta difícil considerarlo como un indicador confiable. Por esta razón, se optó por recurrir nuevamente al Algoritmo 3 en busca de un resultado más claro y que ofrezca una base para una comparación directa.

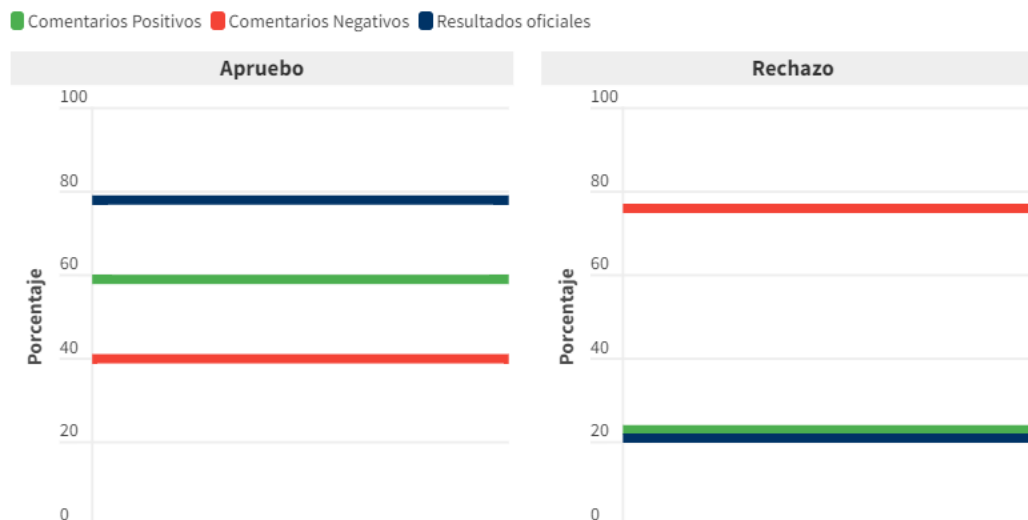


Figura 18: Comparación entre resultados oficiales y porcentajes de comentarios en Twitter del año 2020.

Fuente: Elaboración propia.

Los resultados finales confirman que la opción “Apruebo” se sitúa con una mayoría significativa, alcanzando un porcentaje del 74,57 %, tal como se ilustra en la figura 19. Este análisis, que toma en cuenta tanto la cantidad como la claridad de los comentarios hacia cada opción, revela un margen de error de solo el 3,74 %, demostrando una alta precisión en los porcentajes obtenidos.

4.1.5. Elecciones presidenciales 2021

Enmarcadas en un periodo de significativo cambio social y siguiendo el histórico Plebiscito Constitucional de 2020, las elecciones presidenciales de 2021 en Chile representaron un momento crucial en el proceso de reevaluación nacional y en la búsqueda de un nuevo contrato social, iniciado con el estallido social de octubre de 2019.

Este ciclo electoral destacó por la notable diversidad y polarización en las propuestas de los candidatos, abarcando un espectro político que se extendía desde la derecha más conservadora hasta la izquierda más radical. Los ciudadanos chilenos enfrentaban una decisión fundamental sobre la dirección futura del país, con opciones que variaban entre profundizar las reformas sociales y políticas, y aquellas enfocadas en reestablecer el orden y la estabilidad.

Las elecciones primarias jugaron un papel fundamental en definir el escenario político, seleccionando a los candidatos que liderarían los principales pactos electorales. Por el lado de Chile Vamos, Sebastián Sichel, ex presidente del Banco Estado, obtuvo un notable 49,08 %, mientras que por Apruebo Dignidad, Gabriel Boric, destacado líder de la izquierda, emergió

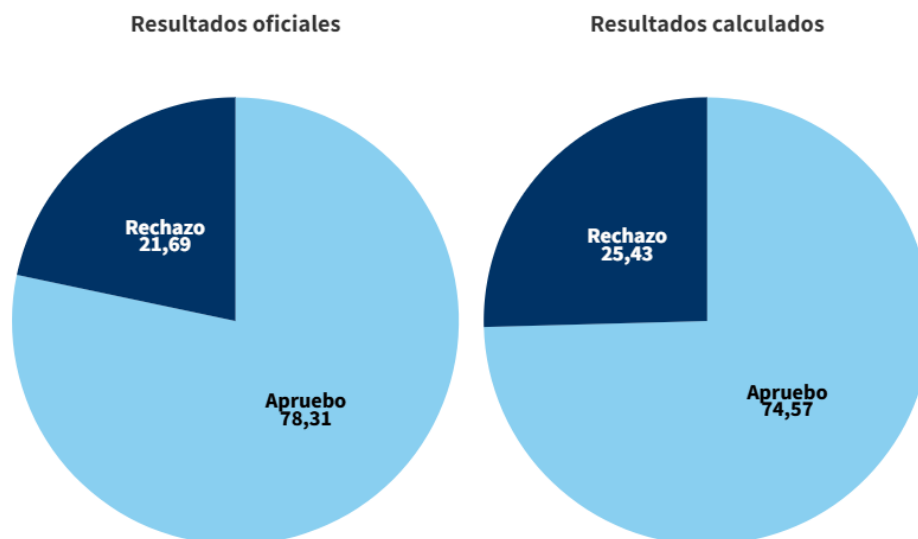


Figura 19: Comparación final de resultados del proceso constitucional año 2020.

Fuente: Elaboración propia.

victorioso con un impresionante 60,43 %, dejando atrás a Daniel Jadue tal como se puede ver en la Tabla 10 . Estos resultados, reflejados en las primarias, no solo delinearon las contiendas internas dentro de cada bloque, sino que también anticiparon el tenor de la campaña electoral a seguir.

Con la definición de los candidatos principales, el panorama electoral se conformó por una lista de siete aspirantes a la presidencia presentados en la Tabla 10, cada uno de ellos presentando un conjunto único de políticas y visiones para el futuro de Chile. Esta diversidad propició un debate público intenso y multifacético sobre temas críticos como la reforma constitucional, el modelo económico del país, el sistema de pensiones, la gestión de la crisis sanitaria y las políticas de bienestar social.

El 18 de noviembre de 2021, la ciudadanía acudió a las urnas, marcando con su voto el rumbo que deseaban para Chile. Los resultados confirmaron las tendencias previas de las encuestas, posicionando a José Antonio Kast y Gabriel Boric como los más votados con un 27,91 % y un 25,83 % de los votos, respectivamente, tal como se aprecia en la tabla 21. La sorpresiva tercera mayoría de Franco Parisi, por un margen estrecho respecto a Sebastián Sichel, reflejó la volatilidad y la imprevisibilidad del escenario político.

La definición de Gabriel Boric y José Antonio Kast como contendientes de la segunda vuelta presidencial situó a Chile frente a dos visiones radicalmente distintas respecto al futuro del país. La elección entre estos dos candidatos no solo determinaría la dirección inmediata de

	Candidato	Partido	Apoyo político	Votos	% Pacto	% Total
A1	 Joaquín Lavín Infante	UDI	Unión Demócrata Independiente	420 691	 31,30 %	 13,59 %
A2	 Ignacio Briones Rojas	EVOP	Evolución Política	131 957	 9,82 %	 4,26 %
A3	 Sebastián Sichel Ramírez	Ind	Independiente	659 570	 49,08 %	 21,31 %
A4	 Mario Desbordes Jiménez	RN	RN-PRI	131 674	 9,80 %	 4,25 %
Total de votos pacto Chile Vamos				1 343 892	100 %	43,42 %
B5	 Gabriel Boric Font	CS	Frente Amplio	1 058 027	 60,43 %	 34,19 %
B6	 Daniel Jadue Jadue	PCCh	Chile Digno	692 862	 39,57 %	 22,39 %
Total de votos pacto Apruebo Dignidad				1 750 889	100 %	56,58 %

Tabla 10: Resultados oficiales de elección de primarias 2021

Fuente: Elaboración propia.

						Partido de la Gente
Gabriel Boric	José Antonio Kast	Yasna Provoste	Sebastián Sichel	Eduardo Artés	Marco Enríquez-Ominami	Franco Parisi
Para vivir mejor	Atrévete	Ahora Yasna	Se puede	¡A refundar Chile!	El cambio es hoy	La gente al poder
						

Figura 20: Candidatos de las primarias presidenciales del año 2021.

Fuente: Elaboración propia.

la política chilena, sino que también tendría profundas implicaciones en el proceso de la nueva Constitución y en cómo Chile enfrentaría sus desafíos socioeconómicos y políticos en los años venideros.

Este escenario proporcionó una gran oportunidad para evaluar los algoritmos de análisis de sentimientos en redes sociales, buscando establecer correlaciones entre las opiniones expresadas en Twitter y los resultados electorales. Un análisis preliminar, realizado en la sección 4.1.3, había sugerido la existencia de una correlación. Sin embargo, para confirmar la fiabilidad de estos algoritmos y determinar si los resultados anteriores fueron un acierto metodológico o un caso aislado, se consideró imprescindible su aplicación en este nuevo contexto electoral.

El estudio inicial sobre el candidato Gabriel Boric reveló un espectro de palabras claves que capturaban la esencia del debate en torno a su candidatura, tal como se muestra en la figura 22. Este análisis de palabras más repetidas refleja no solo los temas centrales de su campaña, sino también los argumentos empleados tanto por sus partidarios como por sus detractores.

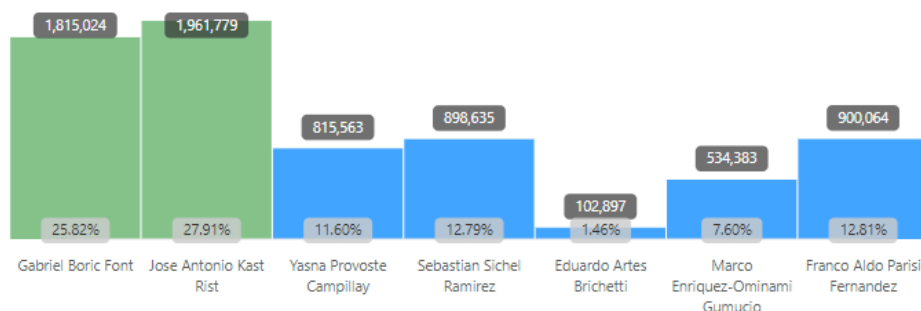


Figura 21: Resultados de primarias presidenciales del año 2021

Fuente: Elaboración propia.

Tal como se mencionó, las elecciones presidenciales de 2021 en Chile se desarrollaron en un contexto singular, marcado por el proceso constitucional en curso, lo que les otorgó una relevancia y cobertura mediática excepcionales. En este escenario, las discusiones en torno a la candidatura de Gabriel Boric captaron particularmente la atención, reflejando una sociedad profundamente dividida en sus percepciones sobre el futuro del país. Las palabras más mencionadas en relación con Boric, especialmente aquellas resaltadas en color dorado, revelan los argumentos clave que movilizaban tanto el apoyo como el rechazo hacia su figura. Por un lado, sus programas y propuestas generaban opiniones polarizadas; mientras algunos los veían como la promesa de un cambio social profundamente necesario, otros los criticaban por considerarlos inviables o demasiado radicales.

La comparación constante con José Antonio Kast, su principal adversario, subrayó la polarización del debate político. Estas comparaciones no se limitaban a políticas específicas, sino que también tocaban temas más amplios como la lucha de clases, el comunismo y la izquierda. Para algunos votantes, Boric representaba una lucha legítima contra las desigualdades estructurales de la sociedad chilena, mientras que para otros, simbolizaba una amenaza al orden establecido y al modelo económico del país.

Entre los argumentos más negativos mencionados en contra de Boric, destacaban las críticas a su supuesta tolerancia o apoyo indirecto a los actos violentos del 19 de octubre de 2019, un evento que marcó profundamente el estallido social. Además, se cuestionaba su capacidad para manejar la economía, un punto que los debates parecían confirmar como uno de sus aspectos más débiles. Esta preocupación económica, sumada a las acusaciones de inconsistencia en sus posiciones, las llamadas “volteretas”, planteaba dudas sobre su fiabilidad como líder.

No obstante, los argumentos a favor de Boric subrayaban su compromiso con la igualdad de género, contrastándolo con la figura de Kast y posicionándolo como el “candidato del pueblo”. Este enfoque en la dignidad social y la justicia, junto con su propuesta de reestructurar las AFP y reformar el sistema de salud junto a la constitución, resonó con un sector significativo de la población. Para sus seguidores, Boric encarnaba la esperanza de un Chile más justo y equitativo, respondiendo a las demandas surgidas durante el estallido social.

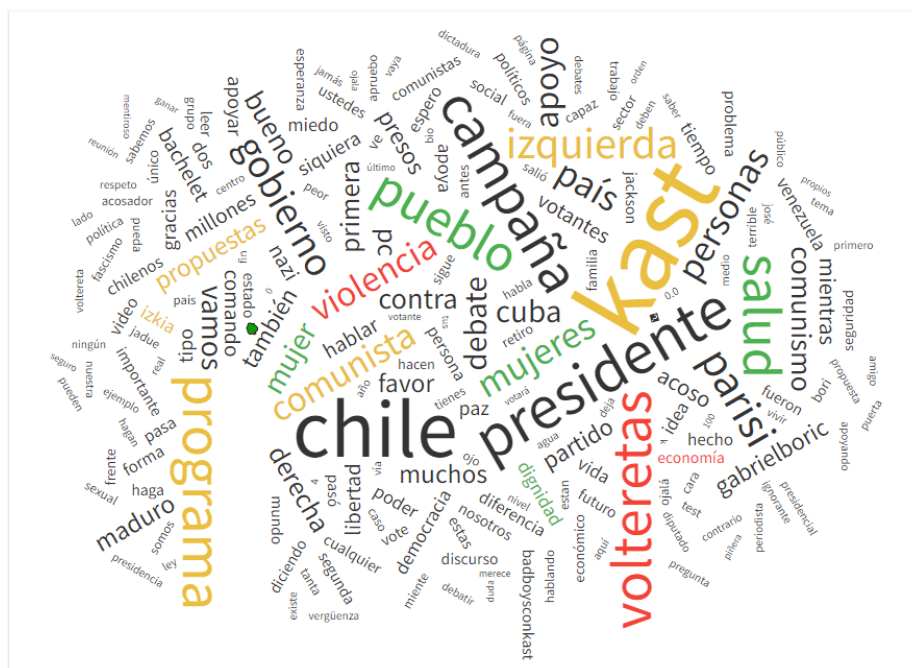


Figura 22: Palabras más repetidas referentes a Gabriel Boric en las elecciones de 2021.

Fuente: Elaboración propia.

Tras el análisis de los comentarios en redes sociales en torno a Gabriel Boric, el foco se desplaza ahora hacia José Antonio Kast, revelando un escenario de debate y polarización similar. La figura que despliega las palabras más repetidas relacionadas con Kast durante las elecciones de 2021, observadas en la figura 23, ofrece un reflejo de las dinámicas políticas que influyeron en su campaña y la percepción pública de su candidatura.

El análisis de las interacciones y discusiones en redes sociales relativas a José Antonio Kast durante la campaña para las elecciones presidenciales de 2021 destaca la influencia significativa de la polarización política en la percepción pública de los candidatos. La prevalencia de términos como “izquierda” y “derecha” en las menciones asociadas a Kast subraya cómo las identidades políticas tradicionales siguen siendo ejes centrales de división y debate entre los electores chilenos. La mención frecuente de la “campaña” refleja un interés particular en las estrategias adoptadas por Kast para atraer a los votantes, así como las consecuencias de estas estrategias en términos de ganancia o pérdida de apoyo entre el electorado.

Los comentarios negativos hacia a Kast superaron en número a los contabilizados para Boric, aunque no se concentraron en una única crítica dominante. En su lugar, se dispersaron entre varias líneas de críticas, incluyendo acusaciones de machismo, que ponen de relieve la preocupación sobre su postura respecto a los derechos y el trato hacia las mujeres. Propuestas controvertidas como la construcción de una zanja para detener la inmigración ilegal provocaron grandes debates, reflejando las divisiones sobre cómo abordar aquellos problemas [Jara, 2020]. Además, las etiquetas de “nazi” y “mentiroso” sugieren una percepción negativa de su ideología y la confiabilidad de sus declaraciones.

tan pronunciada sobre José Antonio Kast, lo cual se asemeja a la tendencia observada en los resultados oficiales.

Candidatos	Comentarios Positivos	Comentarios Negativos
Boric	56,97%	43,03%
Kast	52,67%	47,33%

Tabla 11: Resultados sobre el análisis de tweets en las elecciones del 2021.

Fuente: Elaboración propia.

Estos resultados evidencian una inclinación hacia Gabriel Boric con un 56,97 % de comentarios positivos frente a un 43,03 % de connotación negativa. José Antonio Kast, por su parte, recibió un 52,67 % de comentarios positivos contra un 47,33 % negativos, indicando igualmente una tendencia favorable, pero menor en comparación con Boric por un margen de 4,3 %. Esta diferencia se visualiza en la Figura 24, corroborando la capacidad del algoritmo para captar la tendencia electoral.

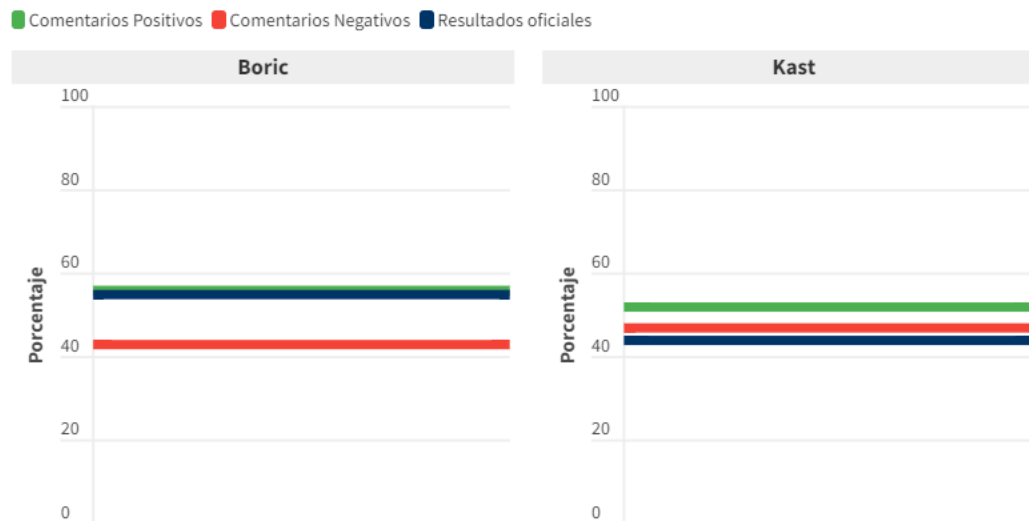


Figura 24: Comparación entre resultados oficiales y connotaciones de comentarios en Twitter en las elecciones presidenciales del año 2021.

Fuente: Elaboración propia.

La confirmación de la tendencia a favor de Boric, captada por el algoritmo, alinea los resultados de la red social con los oficiales, reafirmando la viabilidad de esta metodología para futuras elecciones presidenciales. Sin embargo, la interpretación directa de los porcentajes requiere cautela. Por ello, se aplicó el Algoritmo 3, el cual determinará las puntuaciones y obtendrá un cálculo más preciso. Este resultado se ve culminando en la Figura 25, que exhibe el apoyo consolidado hacia Boric con un margen de error de 3,61 %.

Este enfoque confirma que la metodología aplicada es efectiva para el análisis de elecciones presidenciales con un leve margen de error, además de esto los pasos anteriores fueron

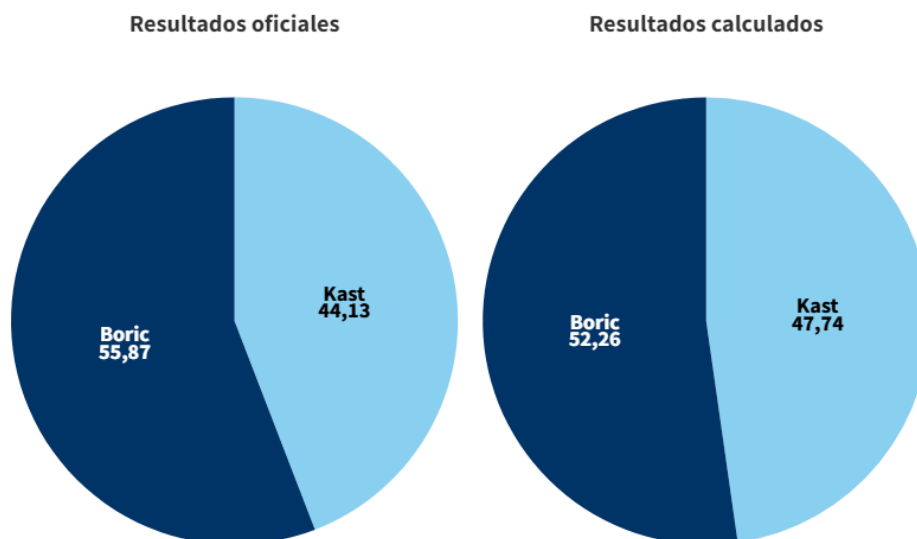


Figura 25: Comparación final de resultados de las elecciones del 2021.

Fuente: Elaboración propia.

capaces de discernir las preocupaciones ciudadanas hacia los candidatos y establecer tendencias a través del análisis de sentimientos en Twitter.

4.1.6. Elecciones constitucionales 2022

La elección presidencial de Gabriel Boric en 2021 marcó el inicio de una nueva etapa en el proceso constitucional chileno, un proceso que no solo buscaba reemplazar la Constitución de 1980, heredada de la dictadura de Pinochet, sino también incorporar demandas sociales largamente postergadas. Este proceso fue significativo no solo por su objetivo de redactar una nueva Constitución, sino también por su enfoque inclusivo y democrático, eligiendo para ello un órgano constituyente compuesto exclusivamente por ciudadanos.

En el 2021, los chilenos fueron convocados a las urnas para elegir a los miembros de la Convención Constitucional encargada de redactar la nueva Constitución. Esto fue una primicia mundial debido a su composición equitativa en términos de género, y por incluir representantes de los pueblos indígenas, asegurando una diversidad de voces y perspectivas en la redacción del texto constitucional, como se puede apreciar en la Figura 26.

La campaña para la elección de los constituyentes se destacó por debates intensos sobre el futuro del modelo económico chileno, la ampliación de derechos sociales y ambientales, el reconocimiento constitucional de los pueblos originarios, y la estructura y forma del Estado.

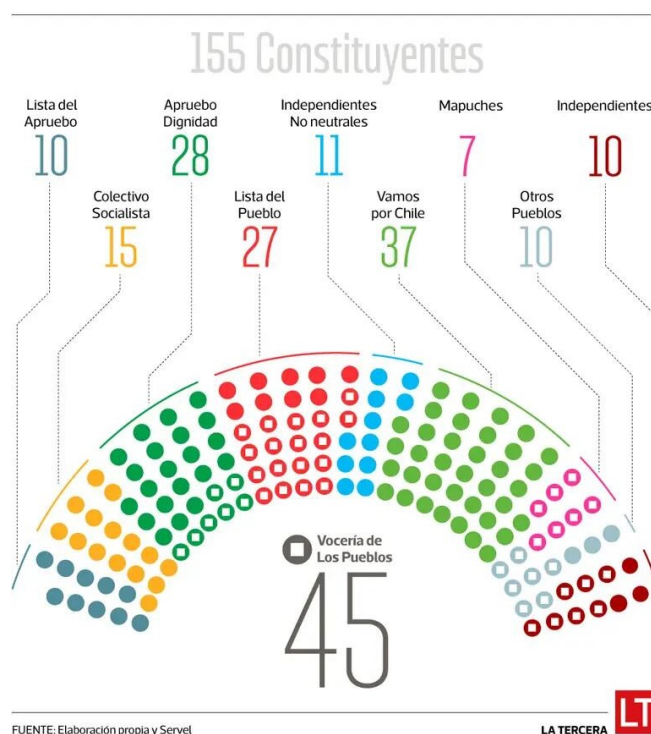


Figura 26: Consituyentes electos en el año 2021.
Fuente: La Tercera.

Este período reflejó la pluralidad de visiones dentro de la sociedad chilena y la esperanza de muchos por construir un país más justo e igualitario a través de una nueva Constitución.

El proceso de redacción de la nueva Constitución, liderado por una mayoría de izquierda y apoyado por el gobierno de Boric, enfrentó el desafío de mantener el apoyo popular en medio de encuestas que mostraban una disminución en la aprobación del texto propuesto. Este escenario puso a prueba la capacidad del gobierno para movilizar el apoyo necesario y lograr la aprobación de la nueva carta magna.

Ante este contexto, se optó por emplear la elección para aceptar o rechazar la nueva Constitución como un caso de estudio destinado a evaluar la eficacia de la metodología que se había experimentado en subsecciones previas. Sin embargo, la validez completa de esta metodología aún no había sido comprobada en el contexto específico de referendos constitucionales. A pesar de haber obtenido resultados prometedores en la subsección 4.1.4, resulta esencial someter la metodología a un proceso de evaluación adicional.

Para abordar este análisis, se inició nuevamente examinando las palabras más repetidas en los tweets relacionados con las opciones del “Apruebo” y el “Rechazo” durante el referéndum para aceptar o rechazar la nueva Constitución propuesta en 2022. Este enfoque buscó comprender cómo las palabras utilizadas en las discusiones en línea podrían reflejar las argumentaciones y preocupaciones de los ciudadanos en aquel momento.

puestas y la esperanza de contar con una Constitución que, según sus partidarios, el país necesita y demanda. Este análisis ofrece una visión general de las dinámicas sociales y políticas que influenciaron la decisión de los votantes.

En contraste, los comentarios negativos se centraron, principalmente, en temas como la violencia observada durante el estallido social y su continuación en algunas manifestaciones, la preocupación por un posible aumento en los impuestos como resultado de la implementación de la nueva Constitución, el temor a la expropiación de propiedades privadas, y la figura del presidente Gabriel Boric, cuya popularidad estaba en descenso en ese momento. Estos puntos negativos reflejan las principales críticas y miedos expresados por los opositores al “Apruebo”.

Para complementar el análisis previo enfocado en el “Apruebo”, se prosigió con el estudio de los tweets relacionados con la opción del “Rechazo”. Esta evaluación buscaba entender las dinámicas discursivas y las preocupaciones de los ciudadanos que se inclinaban por esta opción, analizando las palabras y hashtags más utilizados en este contexto. La Figura 28 muestra un conjunto diverso de términos que reflejan las discusiones predominantes entre los defensores del “Rechazo”.

Al igual que en el análisis del “Apruebo”, los *hashtags* y palabras claves desempeñaron un papel importante en las discusiones *online*. Sin embargo, en el caso del “Rechazo”, se observó una variedad más amplia de temas en debate. Entre los puntos destacados estaban las franjas electorales que, a pesar de algunos aciertos, también enfrentaron críticas por decisiones consideradas inapropiadas por parte de la audiencia. Otro tema relevante fue el apoyo de actores y personalidades públicas al “Apruebo”, lo cual fue objeto de críticas frecuentes por parte de los partidarios del “Rechazo”. Además, los políticos de ambos bandos fueron criticados, evidenciando una desconfianza generalizada hacia la clase política.

Los comentarios negativos hacia el “Rechazo” incluyeron críticas a ciertos segmentos de su franja política, acusándola de promover mensajes controvertidos sobre no denunciar actos de acoso, el desespero y la creación de propuestas tardías para evitar una nueva constitución, y la acusación de que muchas de las afirmaciones del “Rechazo” se basaban en falsedades, promoviendo una campaña de desinformación y animadversión hacia el “Apruebo”.

Por otro lado, los comentarios positivos hacia el “Rechazo” resaltaron la paz y tranquilidad que, según sus partidarios, rechazar la nueva constitución traería al país, evitando así futuros problemas. La consigna de votar sin miedo reflejaba un llamado a enfrentar lo que percibían como intimidaciones por parte del “Apruebo”. En cuanto a la figura del presidente Gabriel Boric, mientras que en el contexto del “Apruebo” su asociación fue mayormente negativa, afectando potencialmente la inclinación a votar por esta opción debido a su decreciente popularidad, en el marco del “Rechazo” esta percepción negativa hacia su gestión se convirtió en un argumento a favor del “Rechazo”. La disminución en el apoyo a Boric fue utilizado por los partidarios del “Rechazo” como una razón importante para oponerse a la propuesta constitucional que él respaldaba.



Figura 28: Palabras más repetidas referentes a el Rechazo en las elecciones de 2022.

Fuente: Elaboración propia.

El análisis de las palabras más frecuentes en los tweets asociados a las opciones del “Apruebo” y el “Rechazo” durante el referéndum de 2022 para la aceptación o rechazo de la nueva Constitución arroja información muy valiosa. No obstante, la calidad de las palabras y los contextos en los que se utilizan estas palabras no resultan ser las más apropiadas para este tipo de análisis, lo que requiere una búsqueda más exhaustiva para extraer aquella información.

Concluyendo el análisis de las palabras más repetidas, se avanza hacia el examen detallado de cada tweet, enfocándose en su connotación y tendencias, replicando la metodología empleada en análisis anteriores. Para ello, se recurre nuevamente al Algoritmo 2, con el propósito de verificar los resultados obtenidos de Twitter y efectuar una comparación de los resultados obtenidos con las subsecciones anteriores

El procesamiento de todos los tweets a través del algoritmo 2 reveló una marcada inclinación hacia la opción de “Rechazo”, evidenciado en la tabla 29, donde se muestra que un 68,54 % de los tweets apoyaban esta opción, contrastando con solo un 31,46 % de comentarios negativos. Esta proporción es significativamente distinta al 53,87 % de comentarios negativos asociados con la opción del “Apruebo”, teniendo menos de la mitad comentarios positivos, específicamente un 46,13 %.

Estas cifras, de alguna manera, reflejan el resultado final, donde el “Rechazo” prevaleció con un 61,89 %, mientras que el “Apruebo” alcanzó un 38,11 %. Aunque existe una correlación entre los datos de Twitter y los resultados oficiales, la precisión no es tan clara debido a

Candidatos	Comentarios Positivos	Comentarios Negativos
Apruebo	46,13%	53,87%
Rechazo	68,54%	31,46%

Figura 29: Resultados sobre el análisis de tweets en las elecciones del año 2022.

Fuente: Elaboración propia.

la diferencia en magnitudes. En algunas subsecciones anteriores era posible interpretar los datos de manera que se acercaran más a los resultados oficiales. En cambio, en este análisis, se observa un margen de error aproximado de casi un 7 % utilizando los resultados obtenidos para la opción del “Rechazo”, como se ilustra en el gráfico 30.

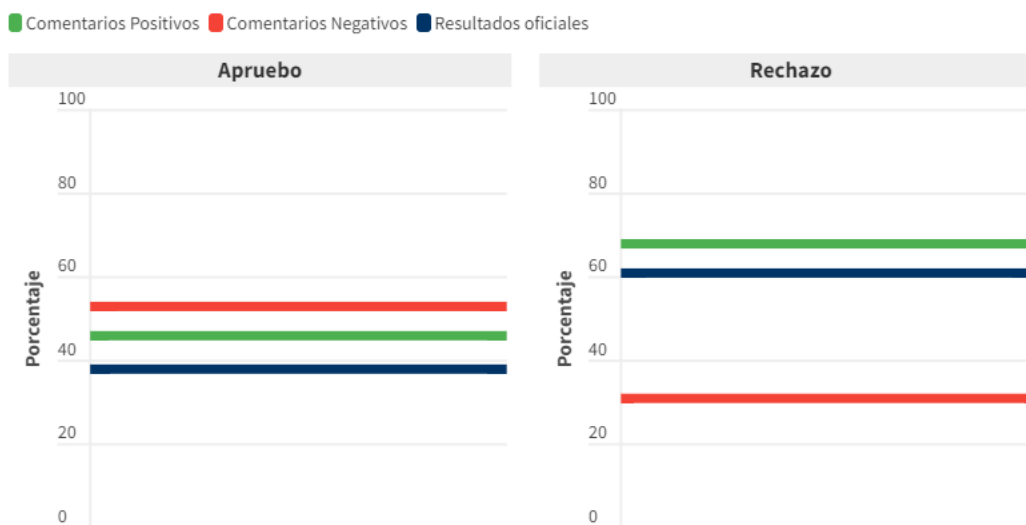


Figura 30: Comparación entre resultados oficiales y connotaciones de comentarios en Twitter en las elecciones del 2021.

Fuente: Elaboración propia.

Ante esta situación, era indispensable la aplicación del Algoritmo 3 para tratar de obtener un resultado más preciso o determinar si este también se vio afectado por la naturaleza de los tweets recopilados. Después de aplicar el mencionado algoritmo, los resultados mostraron una mayor fiabilidad y un margen de error reducido. Tal como se ilustra en la Figura 31, este algoritmo permitió la estimación de un resultado esperado del 42,16 % para el “Apruebo” y del 57,84 % para el “Rechazo”. Esta proyección se ajusta de manera coherente con los resultados oficiales, presentando un margen de error del 4,05 %, lo que indica un desempeño satisfactorio.

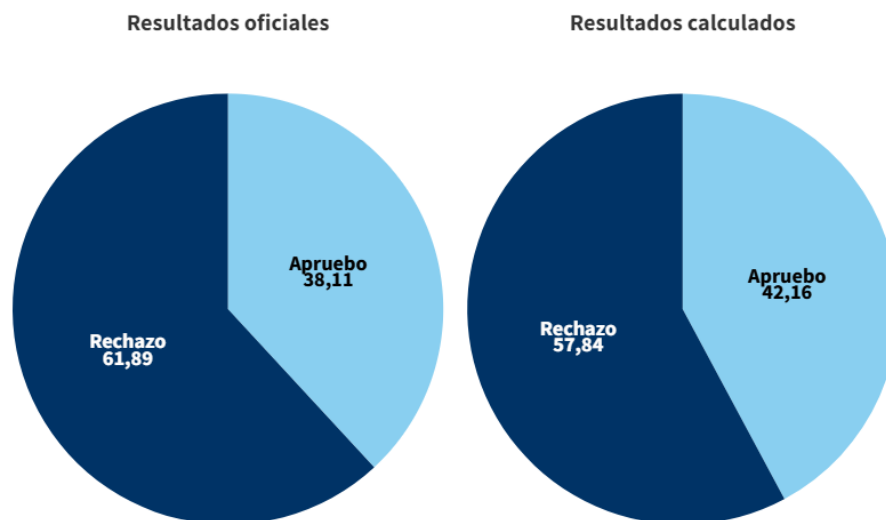


Figura 31: Comparación final de resultados de las elecciones del 2021.
Fuente: Elaboración propia.

4.2. Evaluación e interpretación

Tras finalizar la etapa de **Minería de Datos**, en la cual se probaron los algoritmos junto con los modelos clasificadores de sentimientos, se dio paso a la etapa de **Evaluación e Interpretación**. En esta fase, se recopilaron y analizaron los resultados obtenidos anteriormente, interpretando los datos para extraer las primeras conclusiones.

4.2.1. Resultados

Los resultados obtenidos fueron muy positivos, mostrando márgenes de error bastante bajos, tal como se aprecia en la Tabla 12. En esta tabla se muestra un resumen de los resultados proyectados y los oficiales, además de los márgenes de error de cada una de las elecciones. El mayor margen de error fue para las elecciones constitucionales del año 2023 con un 4,05 %, y el menor para las elecciones presidenciales del año 2021 con un 3,61 %.

Elecciones	Porcentaje de opción 1 proyectada	Porcentaje de opción 1 real	Porcentaje de opción 2 proyectada	Porcentaje de opción 2 real	Margen de error
Presidenciales 2017	54,57%	58,31%	45,53%	41,69%	3,74%
Constitucionales 2020	78,31%	74,57%	54,57%	58,31%	3,74%
Presidenciales 2021	55,87%	52,26%	44,13%	47,74%	3,61%
Constitucionales 2023	57,84%	61,89%	42,16%	38,11%	4,05%

Tabla 12: Resumen de resultados obtenidos

Estos márgenes de error no solo son razonables, sino que también demuestran la capacidad del modelo para capturar tendencias significativas. A pesar de los márgenes de error, las opciones electas fueron consistentemente las proyectadas por los datos, incluso en elecciones con resultados ajustados.

Además, los márgenes de error observados son bastante consistentes entre las pruebas realizadas, lo cual permite establecer un margen de error estimado para el Algoritmo 3 utilizado. Para elecciones presidenciales, el margen de error se puede aproximar a un 3,7 %, mientras que para plebiscitos u otros tipos de elecciones, el margen de error podría ser ligeramente mayor, alcanzando hasta un 4,1 %.

En resumen, los resultados indican que el análisis de sentimientos aplicado a los tweets puede ser una herramienta efectiva para prever los resultados electorales con un margen de error aceptable, lo que subraya la relevancia y utilidad del enfoque propuesto en este estudio.

4.3. Consolidación del Conocimiento

Al concluir la etapa de **Evaluación e Interpretación**, se inició la última fase de **Consolidación del Conocimiento**. Durante esta etapa, los algoritmos y el conocimiento previamente desarrollados se aplicaron en tiempo real para verificar su usabilidad y practicidad en este contexto. Se realizó una demostración detallada del procedimiento, lo que permitió identificar y visualizar las posibles mejoras. Esta fase es necesaria para asegurar que la solución propuesta no solo sea efectiva, sino también práctica y ajustable según las necesidades en un entorno real.

4.3.1. Caso de estudio: Elecciones constitucionales 2023

Como se destacó en la sección anterior, el enfoque propuesto demostró ser exitoso, arrojando resultados prometedores. No obstante, como último paso, es adecuado probar el enfoque en un contexto real y dinámico, por ello se seleccionó como último análisis el plebiscito constitucional programado el día 17 de diciembre del año 2023.

Para proporcionar un marco de referencia, es relevante recordar las elecciones constitucionales de 2022, en donde la propuesta presentada por sectores de izquierda, en su mayoría, fue rechazada por la población chilena. Este acontecimiento marcó un giro en la orientación política del país, contrastando con los triunfos previos de la izquierda en las elecciones constitucionales de 2020 y las presidenciales de 2021. La derrota en el plebiscito de 2022 representó un revés para estos sectores, mientras que para la derecha chilena emergió como una gran oportunidad para proponer y someter a votación su versión de una nueva constitución el 17 de diciembre.

Este contexto, cargado de intriga e incertidumbre [País, 2023], proporcionó el escenario perfecto para aplicar la metodología en un entorno real. El objetivo no es solo analizar el impacto y la percepción pública a través de las redes sociales, sino también validar la precisión y relevancia de esta herramienta para predecir y entender las dinámicas electorales futuras.

Para el análisis preliminar del plebiscito constitucional de 2023, se inició la recopilación de datos a través de Twitter, abarcando el periodo del 13 de noviembre al 12 de diciembre del 2023. Este intervalo de tiempo fue seleccionado para capturar una amplia gama de discusiones y perspectivas relacionadas con el plebiscito, garantizando así que los datos reflejaran de manera equilibrada los temas predominantes a lo largo de el periodo previo a la votación.

La estrategia de recolección se centró en asegurar una representación balanceada de las opiniones “A favor” y “En contra” de la nueva propuesta constitucional, con el fin de evitar sesgos que pudieran influir en el análisis debido a la concentración en temas de un día específico o de un periodo corto. Como resultado de este enfoque, se logró recopilar un total de 36,319 *tweets*, distribuidos equitativamente entre las dos opciones, con 18,356 *tweets* referentes a la opción “A favor” y 17,963 *tweets* relacionados con la opción “En contra”.

La estrategia de recolección empleó técnicas de *web scraping*, similares a las utilizadas en secciones previas de este trabajo. La búsqueda avanzada de *tweets* se realizó utilizando frases clave como “A favor” y “En contra”, complementadas con términos relacionados como “constitución”, “plebiscito” y “elecciones”. Este enfoque aseguró la captura de una amplia gama de discursos relevantes al tema en cuestión.

La recolección de datos se detalla en la Tabla 13, que ilustra los resultados obtenidos tras aplicar técnicas de extracción y realizar varias iteraciones para asegurar la diversidad y el equilibrio en el conjunto de datos. Esta base de *tweets* proporcionó el fundamento para los análisis subsiguientes, los cuales se enfocaron en determinar la connotación y las tendencias predominantes en el discurso público relacionado con el plebiscito.

	Semana 1	Semana 1 Filtrada	Semana 2	Semana 2 Filtrada	Semana 3	Semana 3 filtrada	Total
A favor/Apruebo	12037	7391	11011	5835	10025	5130	18356
En contra/Rechazo	12059	6512	12086	6257	11512	5194	17963
Total	24096	13903	23097	12092	21537	10324	36319

Tabla 13: Recoleccion de *tweets* para el plebiscito 2023

Fuente: Elaboración propia.

El análisis de los *tweets* recolectados para el plebiscito de 2023 comienza con una exploración de las palabras más repetidas, buscando identificar los principales motivos y debates en torno a cada opción de voto. Este enfoque proporciona una comprensión inicial de las preocupaciones y argumentaciones predominantes en el discurso público.

El análisis de las palabras más repetidas en los *tweets* referentes a la opción “En contra” del plebiscito de 2023, como se muestra en la figura 32, revela temas claves y las preocupaciones centrales entre los participantes del debate. La mención de “franja” indica que las estrategias de comunicación de campaña son un tema de discusión caliente, reflejando tanto críticas

perdiciar los recursos invertidos en el proceso constituyente. “Seguridad” emerge como un valor central, argumentando que la nueva constitución fortalecería la seguridad nacional y ciudadana. La inclusión de “expertos” en la redacción busca enfatizar la competencia y la base sólida de la propuesta, en respuesta a críticas anteriores sobre la preparación de la constitución anterior.

Realizando un análisis global de los argumentos y debates surgidos en torno a las opciones del plebiscito de 2023, es posible identificar diversas áreas de preocupación y oportunidad que cada facción podría abordar para mejorar su propuesta y responder de manera efectiva a las inquietudes del electorado. Los temas destacados en las discusiones ofrecen una ventana a los aspectos que más resuenan con la población, ya sea en forma de críticas o como puntos de apoyo. En este sentido, el análisis cumple un rol esencial al momento actual, proveyendo datos valiosos para que las campañas ajusten sus estrategias y mensajes de cara a las preocupaciones y expectativas de la ciudadanía, buscando no solo mitigar los aspectos negativos sino también potenciar aquellos que generan mayor afinidad y apoyo hacia su propuesta.

Tras concluir el análisis de las palabras más mencionadas, se avanzó hacia el estudio de los *tweets* para identificar sus connotaciones y tendencias, replicando el mismo trabajo aplicado en secciones previas. Este paso implica un examen detallado de los *tweets* asociados a cada opción del plebiscito, con el fin de clasificarlos según su connotación positiva o negativa hacia las propuestas en discusión. Este análisis se lleva a cabo mediante el algoritmo 2 que, utilizando bibliotecas especializadas en análisis de sentimiento, determina la connotación de cada *tweet*.

Aplicando el algoritmo mencionado se reveló una preferencia en Twitter hacia la opción “En contra”, con 10,492 *tweets* manifestando una connotación positiva hacia esta elección, frente a 7,425 *tweets* con connotaciones negativas. Esto se traduce en un 58,55 % de *tweets* positivos en contraste con un 41,46 % de comentarios negativos para la opción “En contra”. En cambio, la opción “A favor” registró 8,721 comentarios positivos frente a 9,042 negativos, lo que resulta en un 49,09 % de *tweets* positivos y un 50,90 % de negativos, como se ilustra en la figura 34. Estos datos sugieren una menor inclinación de apoyo en la red social hacia la opción “A favor”.

Como se ha demostrado, existe una tendencia clara hacia la opción “A favor”, y con ello obteniendo la primera conclusión. Por otro lado, basado en análisis previos, es observable que, en ciertas ocasiones, los resultados oficiales tienden a alinearse con la tendencia predominante, lo que sugiere que el resultado oficial podría haberse aproximado a un 58,55 % para la opción “En contra” y un 41,56 % para la opción “A favor”. Sin embargo, a pesar de que en los porcentajes finales se acercaron a los resultados aproximados, esta correlación no es una constante, por lo que es necesario llevar a cabo un análisis más detallado.

Para profundizar en este análisis, se recurrió nuevamente al Algoritmo 3. Este método se basa en evaluar tanto la connotación como la confianza expresada en cada *tweet* mediante bibliotecas de análisis de sentimiento, asignando un puntaje específico a cada uno para obtener una valoración global de cada opción. Aplicado a los datos recopilados del plebiscito



Figura 34: Connotaciones de los tweets en el plebiscito 2023

Fuente: Elaboración propia.

2023, este enfoque ha generado resultados consistentes con las tendencias observadas anteriormente. Como se muestra en la figura 35, se estimó que la opción “A favor” iba a alcanzar aproximadamente un 41,33 %, mientras que la opción “En contra” obtendría un 58,67 %. Experiencias previas con este método indican un margen de error aproximado de 3,5 %, lo que, incluso en un escenario pesimista, señalaba a la opción “En contra” como la ganadora. Por lo tanto, los algoritmos utilizados indicaron sin duda que la opción “En contra” debía ser la ganadora a partir de los comentarios en Twitter.

4.4. Conclusiones generales

Este informe ha explorado la correlación entre las tendencias de Twitter y los resultados electorales con las elecciones mas importantes de los últimos 10 años, revelando cómo las expresiones en redes sociales no solo reflejan las opiniones públicas sino que también tienen el potencial de influir en las dinámicas políticas. Durante este periodo, se analizaron cinco casos electorales, cada uno proporcionando una excelente oportunidad para obtener información sobre el estado de cada opción en diferentes momentos del proceso electoral.

Los resultados de estos análisis han sido muy positivos, cumpliendo con el objetivo general de investigar las precisiones y evaluar un posible valor predictivo. Este estudio logró destacar a Twitter como un indicador eficaz del comportamiento electoral y subrayó la importancia de la plataforma como una importante herramienta para el análisis de tendencias políticas.

La metodología KDD empleada, incluyó varias fases, siendo la recopilación, filtración y análisis de datos las mas desafiantes, en particular en el caso de estudio, ya que era necesario

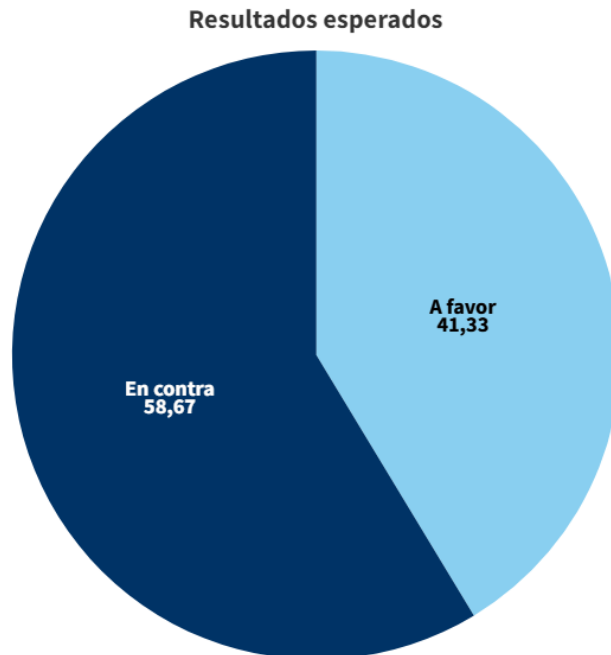


Figura 35: Resultado final previsto del plebiscito 2023
Fuente: Elaboración propia.

obtener la información dentro de fechas específicas para permitir su análisis posterior. Esta tarea se convirtió en un trabajo exigente al realizarse en tiempo real. Aún así se pudo realizar con diferentes herramientas de *Web Scrapping* cumpliendo el primer objetivo específico, desarrollar y aplicar un método efectivo para extraer datos relevantes de Twitter. Esta base de datos bien fundamentada permitió realizar análisis subsiguientes con una mayor calidad.

Progresando hacia un análisis más detallado, cada *tweet* relacionado con entidades específicas fue examinado para determinar su tendencia y estimar el apoyo en redes sociales, utilizando el Algoritmo 2. Para lograr esto, se pusieron a prueba diversos modelos de sentimiento seleccionados en todos los ámbitos posibles. Esto permitió extraer todas las conclusiones pertinentes de cada modelo, para posteriormente utilizar su mejor versión en el algoritmo mencionado. Este paso cumplió con otro de los objetivos específicos, aplicar y evaluar los modelos de sentimiento para clasificar y cuantificar las tendencias en los datos. Los resultados obtenidos fueron muy satisfactorios, mostrando una alineación clara con los resultados electorales oficiales. Esto demostró la efectividad del algoritmo y de los modelos seleccionados para identificar patrones de apoyo o rechazo hacia candidatos específicos, confirmando su utilidad como herramientas de predicción en contextos políticos.

Sin embargo, no todos los aspectos del estudio resultaron ser exitosos. En términos de magnitudes, para las elecciones presidenciales fue posible interpretar los datos como cotas superiores e inferiores, utilizando los resultados del candidato con mayor tendencia positiva. Incluso en las elecciones presidenciales del año 2021, se alcanzó un margen de error de casi

un 1 %. No obstante, este enfoque no resultó tan efectivo al analizar las elecciones constitucionales.

La falta de efectividad en el análisis de las elecciones constitucionales sugiere que se requería un nuevo enfoque para ese obtener ese tipo de conclusiones. Por lo tanto, en este punto, lo único que se puede asegurar con certeza es que el algoritmo demostró ser altamente eficaz para determinar las tendencias de acuerdo con los resultados oficiales en el contexto de elecciones presidenciales.

Esto llevó a la necesidad de realizar un análisis final que permitiera al menos una comparación directa entre los datos obtenidos de Twitter y los resultados oficiales. Por esta razón, se procedió al último análisis, en el cual se empleó un sistema de puntuación basado en la ecuación 1, fundamentado en los principios del Algoritmo 2. Este enfoque dio origen al Algoritmo 3, que arrojó resultados bastante satisfactorios.

En las elecciones presidenciales, se observó un margen de error promedio de 3,6 %, mientras que en las elecciones constitucionales hubo una pequeña diferencia, en particular, en el estudio de las elecciones constitucionales del año 2022, en donde el error alcanzó un 4,05 %, incrementando ligeramente este margen. Por otro lado, en el otro caso de estudio de elecciones constitucionales, el margen estuvo más alineado con lo observado en las elecciones presidenciales, con un 3,74 %. Esto demuestra que, efectivamente, existe una correlación significativa entre los resultados electorales y la información recopilada de Twitter.

El último paso del estudio involucró la implementación de un sistema de puntuación para realizar una comparación directa entre los datos de Twitter y los resultados electorales. Este enfoque, basado en el Algoritmo de análisis de puntuación, no solo proporcionó resultados satisfactorios sino que también confirmó una correlación significativa entre los resultados electorales y la información recopilada de Twitter, cumpliendo así con el último de los objetivos específicos planteados en un comienzo.

En conclusión, este trabajo ha validado con éxito tanto el objetivo general como cada uno de los objetivos específicos mediante la aplicación de la metodología KDD y un análisis detallado. Los resultados obtenidos no solo enriquecen la comprensión de la dinámica electoral a través de las redes sociales, sino que también proporcionan una base sólida para futuras investigaciones en este campo en constante evolución.

CONCLUSIÓN

Este informe ha demostrado cómo las plataformas de redes sociales, especialmente Twitter, reflejan y pueden influir en las dinámicas políticas. A través del análisis de las tendencias de Twitter y su relación con los resultados electorales de los últimos diez años, se ha corroborado la idea inicial de que Twitter actúa como índice de la opinión pública. Los cinco casos electorales analizados han proporcionado información valiosa sobre el estado del proceso electoral en cada momento, confirmando la utilidad predictiva de las redes sociales en contextos electorales variados.

Este éxito fue posible gracias a una meticulosa recopilación de datos de alta calidad, facilitada por la metodología KDD. La versatilidad de esta metodología permite una mejora continua y habilitó una extracción constante de datos relevantes después de cada proceso de filtrado, transformando las fases de recopilación en una constante iteración entre filtrado y recopilación. Sin embargo, esta flexibilidad también tiene puntos negativos, la metodología KDD es tan adaptable que puede resultar costosa en términos de tiempo y recursos. Este aspecto destaca un desafío crítico en el empleo de técnicas avanzadas de análisis de datos, mientras que permiten una gran profundización en los fenómenos estudiados, también exigen una inversión considerable en la gestión y análisis de los datos recopilados.

Gracias a estos desafíos, el estudio ha identificado áreas clave para futuras mejoras, con el potencial de reducir los márgenes de error y, eventualmente, desarrollar un algoritmo capaz de predecir resultados electorales con gran certeza. Mejoras en la automatización de tareas, una filtración más eficiente de los tweets recopilados y optimizaciones en los algoritmos empleados son algunas de las vías factibles para perfeccionar estos resultados. Estas mejoras mencionadas hacen viable continuar avanzando en este enfoque, el cual inevitablemente presentará nuevos desafíos, como la influencia que podría tener la reimplementación del voto obligatorio, cuyo impacto en el comportamiento electoral aún es incierto.

Por lo cual, este estudio sienta una base sólida para los análisis políticos en futuras elecciones en Chile, donde se puede obtener mucho provecho desde el punto de vista político, entendiendo cuáles son las preocupaciones de la gente y las razones por las que apoyan las opciones que eligen. Más allá de esto, los pilares principales del estudio permiten su extensión a otras áreas, continuando con la constante evolución del análisis de sentimientos y el análisis de redes sociales. Estas herramientas tienen un potencial amplio y versátil, capaz de adaptarse y evolucionar en respuesta a los cambios en el comportamiento y las preferencias de los usuarios de redes sociales.

A partir de esto, se concluye que el estudio no solo enriquece el entendimiento del panorama político actual, sino que también equipa a los analistas, políticos y estrategas con las herramientas necesarias para anticiparse y responder más efectivamente a las tendencias emergentes. Al integrar de manera proactiva el análisis de redes sociales en la estrategia política, se pueden diseñar campañas más resonantes y eficaces que realmente aborden las

preocupaciones y expectativas de los votantes.

Por otro lado, a nivel técnico, es necesario tener bases informáticas para llevar a cabo este estudio, en este caso en particular, el desarrollo fue apoyado por conocimientos adquiridos en diversas ramas informáticas. En específico, los cursos de Bases de Datos y Bases de Datos Avanzadas los cuales fueron esenciales para la recopilación y organización de los datos, mientras que el curso de Visualización de Datos facilitó la representación gráfica de los resultados y el análisis posterior de estos datos. Además, el curso de Gestión de Proyectos de Informática jugó un papel muy importante en la gestión de tareas internas durante todo el transcurso del trabajo, permitiendo una excelente organización de ideas y distribución de trabajo.

Finalizando este trabajo, se puede apreciar el inmenso mundo de las redes sociales y la infinidad de información que se puede obtener y trabajar. Lo explorado hasta ahora es tan solo la punta del iceberg, representando información que puede ser vital para decisiones muy importantes y, con ello, el futuro de un país. Este estudio no solo subraya la relevancia de las competencias técnicas en el análisis moderno de datos, sino que también destaca cómo la integración de estas habilidades puede influir significativamente en decisiones políticas de gran importancia.

REFERENCIAS BIBLIOGRÁFICAS

- [Allcott y Gentzkow, 2017] Allcott, H. y Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.
- [Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- [Carrasco, 2022] Carrasco, D. (2022). Redes sociales en Chile: perfiles y número de usuarios, consultado el 08 de julio del 2023. *recuperado desde* <https://marketing4ecommerce.cl/redes-sociales-en-chile-numero-usuarios/>.
- [Conexión ESAN, 2018] Conexión ESAN (2018). Minería de datos: ¿en qué consiste el knowledge discovery in databases?
- [DataCamp, 2023] DataCamp (2023). Nltk sentiment analysis tutorial: Text mining & analysis in python. Consultado el 16-11-2023.
- [Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Consultado el 16-11-2023.
- [Enli, 2017] Enli, G. (2017). Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European Journal of Communication*, 32(1):50–61.
- [Facebook, 2014] Facebook (2014). Ice Bucket Challenge: 17 Million Videos Shared on Facebook. Facebook Newsroom.
- [Feldman y Sanger, 2006] Feldman, R. y Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- [Flourish, 2024] Flourish (2024). Flourish | data visualization & storytelling. Último acceso: 16 de marzo de 2024.
- [Galeano, 2023] Galeano, S. (2023). Cuáles son las redes sociales con más usuarios del mundo (2023), consultado el 08 de julio del 2023. *Recuperado de* <https://marketing4ecommerce.cl/cuales-redes-sociales-con-mas-usuarios-mundo-ranking/>.
- [García, 2014] García, L. M. (2014). Análisis de sentimientos y predicción de eventos en twitter. Memoria para optar al título de Ingeniero Civil Eléctrico, Universidad de Chile.
- [Gaspar, 2022] Gaspar, G. (2022). Chile polarizado: aprendiendo de los errores del pasado, consultado el 08 de julio del 2023. *Recuperado desde* <https://www.elmostrador.cl/destacado/2022/05/31/chile-polarizado-aprendiendo-de-los-errores-del-pasado/>.

- [Goldberg, 2017] Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.
- [Jara, 2020] Jara, M. (2020). Kast: el candidato chileno que quiere construir zanjas en las fronteras con Perú y Bolivia.
- [Liu, 2012] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan Claypool Publishers.
- [Liu, 2015] Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions.
- [Loria, 2020] Loria, S. (2020). textblob documentation.
- [Lu et al., 2023] Lu, X., Zhu, Y., Li, W., y Huang, C. (2023). Recent advancements and challenges of nlp-based sentiment analysis: A systematic survey. *ScienceDirect*.
- [Maharana et al., 2023] Maharana, A., Priyadarshi, N., y Nayak, D. (2023). Sentiment analysis and its application in predicting electoral outcomes: A survey. *Journal of Computational Social Science*.
- [Maturana, 2012] Maturana, J. (2012). España, el tercer país más activo en twitter, consultado el 08 de julio del 2023. *Recuperado desde <https://www.muycomputer.com/2012/02/01/espana-el-tercer-pais-mas-activo-en-twitter/>*.
- [NEJM, 2016] NEJM (2016). NEK1 Variants Confer Susceptibility to Amyotrophic Lateral Sclerosis. *New England Journal of Medicine*, 375:2016–2026.
- [Pang y Lee, 2008] Pang, B. y Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1–2):1–135.
- [Pang y Lee, 2015] Pang, B. y Lee, L. (2015). Statistically weighted reviews to enhance sentiment classification. *Karbala International Journal of Modern Science*, 1(1).
- [Pang et al., 2002] Pang, B., Lee, L., y Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics.
- [País, 2023] País, E. (2023). Chile entra en la recta final de la campaña constitucional tras cuatro años de incertidumbre. *El País*.
- [Python, 2023] Python, R. (2023). Sentiment analysis: First steps with python's nltk library. Consultado el 16-11-2023.
- [rajput, 2023] rajput, A. (2023). Kdd process in data mining.
- [Rehbein, 2019] Rehbein, C. (2019). Cntv informa distribución del tiempo de la franja televisiva del plebiscito.

- [Reynolds, 2024] Reynolds, C. (2024). The 12 basic principles of data visualization. <https://datafloq.com>.
- [Sifferlin, 2014] Sifferlin, A. (2014). What the Ice Bucket Challenge Actually Did for ALS Research. *Time*.
- [Solís y Lee, 2020] Solís, C. y Lee, L. (2020). Cuando las encuestadoras erraron en un 19 constitucional en el plebiscito de 2020, consultado el 08 de julio del 2023. *Recuperado desde* <https://interferencia.cl/articulos/cuando-las-encuestadoras-erraron-en-un-19-respecto-la-convencion-constitucional-en>.
- [Tran, 2020] Tran, K. (2020). Supercharge your python string with textblob.
- [Tufekci, 2017] Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.
- [UDGVirtual, 2024] UDGVirtual, C. (2024). Flourish: periodismo de datos a otro nivel - centro de formación en periodismo digital. Último acceso: 16 de marzo de 2024.
- [Urgilés et al., 2022] Urgilés, C. M. F., Urgilés, C. H. F., y Sacoto, A. S. Q. (2022). Análisis de sentimiento político en redes sociales, como instrumento para la toma de decisiones estratégicas pre electorales. *Pro Sciences: Revista de Producción, Ciencias e Investigación*.
- [Web, 2023] Web, V. (2023). Consecuencias de las redes sociales en la sociedad. Consultado el 16-11-2023.
- [Wiebe et al., 2005] Wiebe, J., Wilson, T., y Cardie, C. (2005). Annotating expressions of opinions and emotions in language. En *Language resources and evaluation*, volumen 39, pp. 165–210. Springer.
- [Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., y Brew, J. (2020). Transformers: State-of-the-art natural language processing. Consultado el 16-11-2023.
- [Álvarez J.; Mesa J.; González A., 2013] Álvarez J.; Mesa J.; González A., R. M. (2013). Metodologías para la realización de proyectos de data mining.