

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“ANÁLISIS DE SENTIMIENTOS CON TÉCNICAS DE
MACHINE LEARNING EN LA RED SOCIAL X RESPECTO A
INSTITUCIONES PÚBLICAS CHILENAS: CARABINEROS Y
CONGRESO NACIONAL”

Jael Toledo Parra

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL EN INFORMÁTICA

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

Agosto - 2024

DEDICATORIA

Le dedico este trabajo y este logro a los tres ángeles que tengo en esta vida, a mi padre Joel, que en paz descanse, a mi madre Gleni y a mi hermana Sandy. Las tres personas que más han influido en mi como persona, quienes confiaron plenamente en mis capacidades y me han apoyado en todo a lo largo de mi vida.

AGRADECIMIENTOS

Aparte de dedicarles este trabajo, quisiera agradecerle a mi familia, por todo lo que han hecho por mí, por su apoyo, comprensión y amor incondicional. A mi padre, por ser la persona que despertó mi curiosidad y las ganas de aprender. A mi madre, por guiarme siempre hacia lo correcto y enseñarme los excelentes valores que terminaron formando mi ética y moral. A mi hermana, por siempre estar para mí y por permitirme ser su modelo a seguir. Espero que se sientan orgullosos de mí. Quiero agradecer también a mi novio, Kevin, que se convirtió en mi incondicional y máximo apoyo en los últimos 8 años de mi vida. Quién ha estado conmigo en las buenas y en las malas, quién más me toleró en mis episodios de estrés y ansiedad que pasé en la carrera y me incitó a seguir adelante.

Le agradezco enormemente a mi profesor guía, Mauricio Olivares, por su gran paciencia, comprensión y empatía. También por guiarme en este gran hito de mi carrera como profesional.

Gracias a Greta, por llegar de manera imprevista a mi vida, a iluminarla y darle alegría, por desvelarse conmigo y acompañarme estos 5 años.

También agradecerle a mi grupo de amigos del colegio, quienes me han dado tanta felicidad y cariño, se han mantenido a mi lado en los buenos y malos momentos a lo largo de los años, me han brindado contención y un ambiente de lealtad y confianza que atesoro dentro de mí.

A mis amigos de la universidad. A las niñas por ser mi red de apoyo dentro de este rubro tan masculino, por su compañía en los excelentes y estresantes momentos que vivimos juntas en la carrera. A los chicos, por su buena voluntad y las grandes instancias de alegría que brindaron. Gracias a todos porque sin ellas y ellos estos 6 años no hubieran sido lo que fueron.

RESUMEN

Resumen—El presente trabajo de memoria trata del análisis de sentimientos realizado sobre las opiniones de chilenos en X (Twitter) sobre las instituciones de Carabineros y el Congreso Nacional, en los periodos de tiempo que albergan el estallido social y el posterior primer plebiscito constitucional. Este análisis de sentimientos se ha basado en técnicas de aprendizaje automático, las cuales usan modelos para el procesamiento del lenguaje natural. Los resultados de este análisis y este trabajo resaltan la disminución de la opinión negativa de un 81 % a un 58 % y el aumento de la opinión neutra y positiva, en menor medida, en el caso de Carabineros. En el caso del Congreso Nacional, los resultados indican que los sentimientos de las opiniones se han mantenido en esos 3 años, con valores entre el 73 %, el 20 % y 6 % para los sentimientos negativos, neutros y positivos, respectivamente. Este trabajo de investigación pretende encontrar las razones de estos resultados y el porqué estas opiniones han cambiado de esta manera a lo largo de esos 3 años.

Palabras Clave— Carabineros, Congreso, análisis de sentimientos, procesamiento del lenguaje natural.

ABSTRACT

Abstract— This work deals with the sentiment analysis of the opinions by Chileans on X (Twitter) about the institutions of Carabineros and the National Congress, in the time periods that host the social outburst and the subsequent first constitutional plebiscite. This sentiment analysis has been based on machine learning techniques, which use models for natural language processing. The results of this analysis and this work highlight the decrease of negative opinion from 81 % to 58 % and the increase of neutral and positive opinion, to a lesser extent, in the case of Carabineros. In the case of the National Congress, the results indicate that the sentiments of opinions have been maintained in those 3 years, with values between 73 %, 20 % and 6 % for negative, neutral and positive sentiments, respectively. This research work aims to find the reasons for these results and why these opinions have changed in this way over these 3 years.

Keywords— Carabineros, Congress, sentiment analysis, natural language processing.

GLOSARIO

BERT: *Bidirectional Encoder Representations from Transformers*.

BNB: *Bernoulli Naive Bayes*.

BETO: *BERT en español*.

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

GB: *Gigabyte*.

GNB: *Gaussian Naive Bayes*.

IA: *Inteligencia artificial*.

KDD: *Knowledge Discovery in Databases*.

LN: *Lenguaje natural*.

LP: *Lenguaje de programación*.

ML: *Machine Learning*.

MNB: *Multinomial Naive Bayes*.

MSE: *Mean Squared Error*.

NB: *Naive Bayes*.

NLG: *Natural Language Generation*.

NLU: *Natural Language Understanding*.

PLN: *Procesamiento del lenguaje natural*.

RF: *Random Forest*.

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

SVM: *Support Vector Machines*.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
GLOSARIO	V
ÍNDICE DE FIGURAS	VIII
ÍNDICE DE TABLAS	VIII
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	2
1.1 Órganos y servicios públicos a analizar	5
1.1.1 Congreso Nacional de Chile	6
1.1.2 Carabineros de Chile	7
1.2 Factor sociológico de las opiniones	7
1.3 Impacto de la solución	8
1.4 Objetivos	8
1.4.1 Objetivo General	8
1.4.2 Objetivos Específicos	8
CAPÍTULO 2: MARCO CONCEPTUAL	10
2.1 Procesamiento del lenguaje natural (PLN)	10
2.1.1 Entendimiento del Lenguaje Natural (NLU)	10
2.1.2 Generación del Lenguaje Natural (NLG)	11
2.1.3 Aplicaciones del Procesamiento del Lenguaje Natural	11
2.2 Análisis de Sentimientos	12
2.2.1 Enfoque basado en léxico	14
2.2.2 Enfoque basado en aprendizaje automático (ML)	15
2.2.3 Análisis de sentimientos y X (Twitter)	15
2.3 Machine Learning	15
2.3.1 Aprendizaje no supervisado	16
2.3.2 Aprendizaje supervisado	17
2.3.3 Naive Bayes	18
2.3.4 Random Forest (RF)	19
2.3.5 Support Vector Machines (SVM)	20
2.3.6 BERT (Bidirectional Encoder Representations from Transformers)	21
2.4 Métricas de evaluación para análisis de sentimientos	22
2.5 Metodologías para el análisis de los datos	25
2.5.1 CRISP-DM (Cross Industry Standard Process for Data Mining)	25
2.5.2 KDD (Knowledge Discovery in Databases)	26
2.5.3 SEMMA (Sample, Explore, Modify, Model and Assess)	27

2.5.4	Comparación entre las tres metodologías	27
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN		29
3.1	Selección de la metodología a utilizar	29
3.2	Extracción de los datos	30
3.3	Selección de los datos	34
3.4	Preprocesamiento de los datos	35
3.5	Transformación de los datos	38
3.6	Minería de datos	38
3.6.1	Entrenamiento del modelo basado en Naive Bayes	39
3.6.2	Entrenamiento del modelo basado en BETO	40
CAPÍTULO 4: ANÁLISIS DE LOS RESULTADOS Y VALIDACIÓN DE LA SOLUCIÓN		43
4.1	Análisis de validación de los entrenamientos	43
4.1.1	Multimonial Naive Bayes	43
4.1.2	BETO	47
4.1.3	Selección de los modelos	53
4.1.3.1	Dataset Carabineros	53
4.1.3.2	Dataset Congreso Nacional	54
4.2	Análisis de sentimientos	55
4.2.1	Análisis de sentimientos: Carabineros	55
4.2.2	Análisis de sentimientos: Congreso	57
4.3	Análisis de emociones	59
4.3.1	Análisis de emociones: Carabineros	59
4.3.2	Análisis de emociones: Congreso	61
4.4	Análisis de frecuencia de palabras y hashtags	63
4.4.1	Frecuencia de palabras y hashtags: Carabineros	63
4.4.2	Frecuencia de palabras y hashtags: Congreso	67
4.5	Interacciones del Tweet y Sentimientos	71
4.5.1	Carabineros	71
4.5.2	Congreso	72
4.6	Interpretaciones del análisis	72
4.6.1	Interpretaciones: Carabineros	74
4.6.2	Interpretaciones: Congreso	75
CAPÍTULO 5: CONCLUSIONES		77
TRABAJO FUTURO		80
REFERENCIAS BIBLIOGRÁFICAS		81

ÍNDICE DE FIGURAS

1	Carabineros, pacogate	2
2	Carabineros, usando la fuerza	3
3	Congreso, cafetería	3
4	Congreso, Pamela Jiles corriendo como <i>Naruto</i>	4
5	Carabineros, Cadem indica que la aprobación de Carabineros ha subido . . .	4
6	Niveles del análisis de sentimientos	13
7	Técnicas utilizadas para el análisis de sentimientos	14
8	Técnicas de Machine Learning	16
9	Aprendizaje supervisado y no supervisado	17
10	Flujo de trabajo de las técnicas de aprendizaje supervisado	18
11	Ejemplo de Random Forest	20
12	Ejemplo de Support Vector Machine	21
13	Ejemplo de una matriz de confusión, con datos ficticios	24
14	Metodología a seguir	30
15	Datos extraídos de un Tweet por Webscraping usando Selenium	32
16	Datos extraídos de un Tweet por Webscraping usando Selenium en un CSV .	32
17	Métricas resultantes dataset de Carabineros por etiqueta	44
18	Matriz de confusión dataset de Carabineros	45
19	Métricas resultantes dataset del Congreso por etiqueta	46
20	Matriz de confusión dataset del Congreso	47
21	Métricas resultantes dataset Carabineros, modelo BETO	49
22	Matriz de confusión dataset Carabineros, modelo BETO	50
23	Métricas resultantes dataset Congreso, modelo BETO	52
24	Matriz de confusión dataset Congreso, modelo BETO	53
25	Análisis de sentimientos de la totalidad de Tweets extraídos sobre Carabineros	56
26	Análisis de sentimientos de la totalidad de Tweets extraídos sobre el Congreso	58
27	Visualización emociones en tweets de Carabineros	61
28	Visualización emociones en tweets del Congreso	62
29	Palabras más frecuentes en tweets de Carabineros, 2019	64
30	Palabras más frecuentes en tweets de Carabineros, 2022	64
31	Palabras más frecuentes en tweets de Carabineros, sin temporalidad	65
32	Palabras más frecuentes en tweets del Congreso, 2019	67
33	Palabras más frecuentes en tweets del Congreso, 2022	68
34	Palabras más frecuentes en tweets del Congreso, sin temporalidad	69
35	Sentimientos hacia Carabineros y Congreso	73
36	Sentimientos hacia Carabineros y Congreso, separados por año	74

ÍNDICE DE TABLAS

1	Comparación de las 3 metodologías	28
2	Tiempos y cantidad de Tweets extraídos utilizando Snsrape	33

3	Ejemplo de Tweets extraídos utilizando Snsrape	34
4	Ejemplo de Tweet después de la limpieza de datos	37
5	Ejemplo de Tweets extraídos utilizando Snsrape	38
6	Tiempos de ejecución del modelo Multinomial Naive Bayes por dataset . . .	43
7	Métricas modelo Multinomial Naive Bayes para set de datos de Carabineros .	44
8	Métricas modelo Multinomial Naive Bayes para set de datos del Congreso . .	45
9	Variables utilizadas para BETO	48
10	Métricas por iteración del entrenamiento de datos de Carabineros	48
11	Métricas finales del entrenamiento de datos de Carabineros, modelo BETO .	48
12	Métricas por iteración del entrenamiento de datos del Congreso	51
13	Métricas finales del entrenamiento de datos del Congreso, modelo BETO . .	51
14	Comparación métricas del entrenamiento de datos de Carabineros	53
15	Comparación tiempos de ejecución de ambos modelos. Carabineros	54
16	Comparación métricas del entrenamiento de datos del Congreso	54
17	Comparación tiempos de ejecución de ambos modelos. Congreso	55
18	Tiempos de clasificación de sentimientos en dataset de Carabineros	55
19	Clasificación de sentimientos. Tweets Carabineros	56
20	Clasificación de sentimientos en valores porcentuales. Tweets Carabineros . .	56
21	Tiempos de clasificación de sentimientos en dataset del Congreso	57
22	Clasificación de sentimientos. Tweets Congreso	57
23	Clasificación de sentimientos en valores porcentuales. Tweets Congreso . . .	58
24	Tiempo de clasificación de emociones por Pysentimiento	59
25	Emociones en tweets de Carabineros, por año y sentimiento	60
26	Emociones en tweets del Congreso, por año y sentimiento	61
27	Frecuencia de hashtags en tweets de Carabineros, año 2019	66
28	Frecuencia de hashtags en tweets de Carabineros, año 2022	67
29	Frecuencia de hashtags en tweets del congreso, año 2019	70
30	Frecuencia de hashtags en tweets del congreso, año 2022	70
31	Correlación entre las interacciones y el sentimiento del tweet, Carabineros .	72
32	Correlación entre las interacciones y el sentimiento del tweet, Congreso . . .	72

INTRODUCCIÓN

Este trabajo de memoria, titulado “Análisis de Sentimientos con Técnicas de Machine Learning en la Red Social X Respecto a Instituciones Públicas Chilenas: Carabineros y Congreso Nacional” se contextualiza en algunos de los últimos años, que han sido de los más polémicos en Chile con respecto a la política. Los años correspondientes al estallido social, 2019 y el año del primer plebiscito consitucional, 2022.

Los hechos sucedidos en el país en los últimos años han dado iniciativa para que cada vez se formen más opiniones políticas. Y con el aumento de las redes sociales y el acceso a internet, el ambiente es idóneo para emitir opiniones sobre alguna entidad o tópico en los distintos medios sociales del internet.

El objetivo principal de este estudio es lograr un análisis de las opiniones emitidas en la red social de X (ex Twitter) de los ciudadanos del país, con respecto a dos de las principales instituciones de la nación, con técnicas de aprendizaje automático y procesamiento del lenguaje natural, como lo es el análisis de sentimientos.

La solución que se propone en el documento, está basada en técnicas de aprendizaje automático, modelos de clasificación de sentimientos y uso del área y técnicas del procesamiento del lenguaje natural. Con ello, se busca clasificar las opiniones que emiten las personas en X, durante los meses entre septiembre y noviembre de los años 2019 y 2022, que hacen referencia a las dos instituciones de interés.

Esta solución propuesta se va construyendo a medida se avanza en la lectura de esta memoria. Comenzando con el Capítulo 1: Definición del Problema, que inicia con la contextualización de la problemática, el panorama de las instituciones a analizar y los objetivos de este trabajo. En el Capítulo 2: Marco Conceptual, se detalla la base teórica sobre la cual se sustenta el trabajo, basada en literatura relevante para el estudio, definiendo teóricamente, las técnicas, métricas y modelos a utilizar en la solución. En el Capítulo 3: Propuesta de Solución, se plantea la solución a la problemática y se sigue una metodología escogida para abordar el problema. En el Capítulo 4: Análisis de los Resultados y Validación de la Solución, se completa la propuesta de solución, en donde esta misma se valida y se procede a hacer otros tipos de análisis, formando interpretaciones sobre los resultados. Por último, se encuentran las Conclusiones del trabajo realizado, en dónde se revisa el aporte de la investigación y el cumplimiento de los objetivos de la memoria.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

El estallido social marcó un antes y un después en Chile. Como bien es conocido, el aumento de 30 pesos en la tarifa del pasaje del transporte público santiaguino fue la chispa que prendió todo. El malestar acumulado evidenció “a una sociedad chilena con alto nivel de desigualdad social, segregación educacional, territorial... un país moderno, pero con problemas estructurales en lo más profundo de su configuración”[Sanhueza *et al.*, 2021].

En ese momento los indicadores económicos situaban a Chile como uno de los países más estables del continente, pero mirando en retrospectiva, el país estaba bajo una enorme crisis silenciosa. Los motivos del por qué sucedió son muchos y variados. Tal y como se mencionó anteriormente, la desigualdad social fue una de las más grandes causas, también se considera la calidad de la educación y salud, el sistema de pensiones y la falta de representatividad y confianza en las instituciones del país.

Bajo los dos últimos puntos, gran parte de los chilenos desconfiaba de algunas instituciones como Carabineros y sentía falta de representatividad en la política, sobre todo en el parlamento chileno. El caso de la policía chilena previamente al estallido no era tan crítico como lo fue durante el transcurso, ya que la desconfianza apuntaba directamente al conocido caso de corrupción “Pacogate”, pero no a sus métodos de trabajo. Sin embargo, durante y después del estallido, bajo ese mismo contexto, se pudo evidenciar un uso desmedido de la fuerza, que terminaron en graves denuncias de violaciones a los derechos humanos [Orchard *et al.*, 2022], por lo que la opinión y la forma de ver a la policía chilena de un gran número de los ciudadanos del país cambió negativamente.



Figura 1: Carabineros, pacogate
Fuente: Buscador de Google



Figura 2: Carabineros, usando la fuerza
Fuente: Diario Constitucional

Con respecto al Congreso, la baja participación en elecciones estos últimos años evidenciaba la crisis de representatividad que los chilenos tienen en la política, la demora en la legislación y el espectáculo mediático que se daba en ocasiones aumentaba esta mala percepción.



Figura 3: Congreso, cafetería
Fuente: CNN Chile



Figura 4: Congreso, Pamela Jiles corriendo como *Naruto*
Fuente: La Tercera 2020

Últimamente, con el notable aumento de la inseguridad y la delincuencia, se requiere la acción de Carabineros con más urgencia, la legitimidad que la sociedad chilena le reconoce a carabineros ha aumentado [Orchard *et al.*, 2022]. Por el lado del congreso, con la rechazada propuesta de la constitución, se espera que el parlamento legisle más activamente algunos proyectos de ley que sean de carácter social.

PLAZA PÚBLICA | 06 NOVIEMBRE 2022

Aprobación de Carabineros sube 19pts durante lo que va del 2022 y llega a 73%, el nivel más alto de los últimos 6 años



Figura 5: Carabineros, Cadem indica que la aprobación de Carabineros ha subido
Fuente: Cadem

Dado lo anterior, surge la problemática de conocer, comparar y analizar la opinión que tienen los chilenos sobre ambas instituciones, esto bajo periodos de interés, los cuales corresponden a los meses alrededor de la revuelta y a los meses alrededor del tercer aniversario del

estallido, el cual comprende también el tiempo en el que se realizó el plebiscito constitucional para evaluar la primera propuesta de constitución (2022).

Por otro lado, una opinión puede ser expresada en muchos lugares y contextos, pero para el desarrollo de este trabajo se considerará la red social de X (ex Twitter), ya que es una buena alternativa para poder practicar la libertad de expresión mediante opiniones textuales y porque es una de las redes sociales más utilizadas en Chile; 2.9 millones de personas es la audiencia chilena total que podría emitir opiniones en la red social [Medina, 2022].

El análisis se realizará bajo la creación y aplicación de un modelo de análisis de sentimientos, de manera de llegar a resultados que puedan vislumbrar lo que sucedió con la opinión de los chilenos con respecto a las instituciones públicas de Carabineros y Congreso en los periodos de interés.

1.1. Órganos y servicios públicos a analizar

Para comprender y conocer qué son y qué hacen los organismos públicos en Chile, tenemos que conocer el concepto de Estado.

Un Estado es una unidad política y administrativa que rige sobre un territorio y es independiente de otros Estados [SII, sf], además posee ciertos organismos y servicios de gobierno que apoyan en la política y la administración del territorio. Existen dos tipos de Estado: el simple, es aquel Estado que centraliza el poder político, todo el territorio debe estar bajo los derechos y deberes de un solo archivo constitucional; y el federal, donde la administración y la política son descentralizadas y se separan por regiones, es decir, existen gobiernos de carácter regional.

En el caso de Chile, el Estado se divide en 3 poderes: el ejecutivo, que contiene al presidente, en la actualidad Gabriel Boric Font, y los ministerios; el legislativo, poder que ejerce el Congreso Nacional del País; y el judicial, representado por la Corte Suprema.

La Ley Orgánica de las Bases Generales de la Administración del Estado indica en el primer artículo que:

La Administración del Estado estará constituida por los Ministerios, las Intendencias, las Gobernaciones y los órganos y servicios públicos creados para el cumplimiento de la función administrativa, incluidos la Contraloría General de la República, el Banco Central, las Fuerzas Armadas y las Fuerzas de Orden y Seguridad Pública, los Gobiernos Regionales, las Municipalidades y las empresas públicas creadas por ley [ley, 2001].

Y el artículo 28 de la misma ley define los servicios públicos como:

Los servicios públicos son órganos administrativos encargados de satisfacer necesidades colectivas, de manera regular y continua. Estarán sometidos a la dependencia o supervigilancia del Presidente de la República a través de los respectivos Ministerios, cuyas políticas, planes y programas les corresponderá aplicar [ley, 2001].

Bajo ambas citas se puede evidenciar que el Congreso Nacional de Chile y Carabineros son servicios públicos bajo el Estado.

1.1.1. Congreso Nacional de Chile

El Congreso Nacional de Chile es el órgano que se encarga del poder legislativo del Estado. Es bicameral, está compuesto por la cámara de diputados y la de senadores. Dentro de las actividades más importantes se encuentran *"ejercer la representación de la ciudadanía, concurrir a la formación de las leyes conjuntamente con el Presidente de la República y, en el caso de la Cámara de Diputados, fiscalizar los actos del Gobierno"*[BCN, sfb].

Se debe tener en cuenta que la asignación de los senadores y diputados es por elección popular, estos se escogen cada 8 y 4 años, respectivamente, en las mismas votaciones por el presidente de la república. Por parte de la cámara de diputados, en la actualidad la conforman 155 personas y la de senadores es compuesta por 50, haciendo que el Congreso en su totalidad tenga 205 personas legislando, las cuales también serán foco del estudio en la extracción de opiniones.

- **Cámara de Diputados:** Los integrantes de la cámara baja duran 4 años en su cargo y pueden ser reelectos. Para su elección se dividen los 155 cupos de diputados en los 28 distritos, donde cada distrito tiene entre 3 y 8 diputados.

La composición de la cámara consiste en 67 diputados de Oficialismo (Alianza de Gobierno), 7 de apoyo externo, 77 de Oposición (Chile Vamos) y 4 que no entran en ninguna de estas dos categorías. La presidenta actual es Karol Cariola, perteneciente al Partido Comunista de Chile[Diputados, sf].

Aparte de la función legisladora que tiene, puede tener atribuciones tales como acordar temas relevantes del país con el Presidente, citar a ministros para cuestionarlos acerca de su trabajo y fiscalizar acciones del gobierno de turno mediante comisiones [BCN, sfa].

- **Senado:** Los senadores duran 8 años en su cargo y también pueden ser reelectos, pero solo una vez. Para su elección se toman en cuenta las 15 circunscripciones, donde cada una tiene distintos cupos. Ya que los senadores permanecen en el cargo durante 8 años, el senado debe actualizarse de manera parcial cada 4 años.

La composición de la cámara alta consiste en 18 senadores de Oficialismo, 3 de apoyo externo, 27 de Oposición y 2 que no pertenecen a ninguna de las anteriores.

El presidente del senado es José García Ruminot, del Partido Renovación Nacional [Senado, sf].

De la misma manera que la Cámara de Diputados, deben legislar y también tiene atribuciones especiales tales como: rehabilitación de la ciudadanía, dar o no el consentimiento al presidente de la república en ciertas ocasiones especiales, inhabilitar al presidente por alguna enfermedad física o mental que le impida ejercer sus funciones, entre otras [BCN, sfc].

1.1.2. Carabineros de Chile

Carabineros de Chile es una institución pública policial que se encarga de la fuerza de orden y seguridad del país junto a la Policía de Investigaciones (PDI). Está bajo el cargo del Ministerio del Interior, ya que es este el que tiene a cargo la seguridad pública del país.

La institución fue fundada el 27 de abril de 1927 por el Coronel del Ejército, Carlos Ibáñez del Campo. A la actualidad tiene distintas divisiones y direcciones, cada una encargada de distintos aspectos relativos a la seguridad y el orden del país. Algunas direcciones son: Dirección General de Carabineros a cargo del General Director Ricardo Yáñez, Dirección de Derechos Humanos, Dirección de Justicia, Dirección de Control de Drogas, entre otras [Carabineros, sf].

Tiene como principal misión *"brindar seguridad a la comunidad en todo el territorio nacional mediante acciones prioritariamente preventivas, apoyadas por un permanente acercamiento a la comunidad"* [ChileAtiende, sf].

1.2. Factor sociológico de las opiniones

Hace algunos años atrás, las grandes masas de personas casi no tenían injerencia en la decisiones políticas [Roucek, 2016] y sus opiniones no tenían grandes repercusiones, solo se quedaban siendo opiniones. Tomando en cuenta la popularización de ejercer la democracia y la masificación del internet y con ello, de las redes sociales, las masas han logrado manifestar sus opiniones con más facilidad, llegando al alcance de poder viralizarlas y encontrar un montón de otras personas que las puedan compartir o no.

Lo mencionado anteriormente, hace que la persona que opina en redes sociales se sienta respaldada por la comunidad del internet y pueda opinar libremente. También hace que pueda conocer las opiniones de otros y dejar que estas permeen sobre las propias. Todo esto más la viralización intrínseca de las redes sociales más populares hoy en día crean el ambiente perfecto para que se forme un pensamiento colectivo sobre algo o alguien.

Algunos de los tópicos de las opiniones más encontradas en redes sociales incluyen la política, ya que esta está presente y afecta directamente en la vida de las personas. En el caso chi-

leno, con mayor razón en los últimos años, se ha evidenciado un aumento en la cantidad de opiniones sobre instituciones relacionadas a la política del país. Estos grandes pensamientos colectivos han perjudicado o beneficiado la imagen de ciertos órganos públicos, tales como Carabineros y Congreso.

En ocasiones solo basta navegar en redes sociales u observar la televisión abierta para caer en cuenta del poder que tienen las opiniones junto a las redes sociales sobre la imagen pública de personas o instituciones.

1.3. Impacto de la solución

La solución propuesta, un análisis de sentimientos en las opiniones de los chilenos con respecto a Carabineros y el Congreso, puede beneficiar significativamente a tales instituciones.

Con el análisis que se realizará en esta memoria estos órganos públicos pueden estudiar la manera de recuperar la confianza de la sociedad y mejorar su imagen. Como consecuencia del estudio a realizar, instituciones bien valoradas, bien vistas, confiables y eficientes en su trabajo logran demostrar una solidez y estabilidad en la política, la seguridad y la legislación del país, siendo esto un impacto positivo a toda la sociedad.

1.4. Objetivos

1.4.1. Objetivo General

- Desarrollar un análisis de las opiniones de los ciudadanos del país sobre las instituciones públicas de Carabineros y el Congreso en la subárea del Procesamiento del Lenguaje Natural de Análisis de Sentimientos.

1.4.2. Objetivos Específicos

- Elaborar un dataset apropiado para entrenar el algoritmo a diseñar bajo la extracción de tweets sobre opiniones de Carabineros y Congreso en periodo Septiembre-Noviembre 2019 y Septiembre-Noviembre 2022.
- Diseñar modelo de análisis de sentimientos basado en técnicas de machine learning tales como Naive Bayes, BERT, entre otros, que clasifique las opiniones, de manera que se pueda hacer una comparativa entre ambos periodos e Instituciones.
- Evaluar la calidad del modelo diseñado aplicando las métricas de evaluación más estándar en los análisis de sentimientos (Accuracy, Precision).

- Identificar relaciones entre el sentimiento de la opinión con otros tipos de análisis, para determinar las razones de la clasificación del sentimiento.

CAPÍTULO 2

MARCO CONCEPTUAL

En este capítulo, se presentará el marco conceptual que contiene todo el fundamento teórico para la realización de la memoria. Los puntos a revisar son los siguientes:

- Procesamiento del lenguaje natural
- Análisis de Sentimientos
- Machine Learning
- Metodología para el análisis de los datos

2.1. Procesamiento del lenguaje natural (PLN)

El lenguaje natural (LN) es un sistema de comunicación que se basa en conjuntos de reglas, símbolos, sonidos, gestos o signos para poder transmitir y recibir información de otro individuo. Es el lenguaje que es desarrollado y practicado por seres humanos.

Este lenguaje se diferencia del lenguaje de las computadoras porque este último es artificial y es creado por motivos informáticos y computacionales.

El procesamiento del lenguaje natural es un subárea de la Inteligencia Artificial (IA) que estudia la manipulación de lenguajes naturales con herramientas computacionales, esto para hacer que las computadoras entiendan declaraciones o palabras y frases escritas en lenguajes humanos [Khurana *et al.*, 2023]. La conexión entre los lenguajes naturales y los de programación (LP) se dan en este contexto ya que enlazan el LN con las máquinas.

El PLN se puede dividir en dos partes que se encargan de distintas actividades dentro del procesamiento: el Entendimiento del Lenguaje Natural (NLU: Natural Language Understanding) y la Generación de Lenguaje Natural (NLG: Natural Language Generation).

2.1.1. Entendimiento del Lenguaje Natural (NLU)

El entendimiento del lenguaje natural es una subdisciplina del PLN que se centra en la comprensión del significado del lenguaje natural por parte de la máquina, utilizando como medio a la inteligencia artificial. Es decir, la máquina toma la capacidad de poder interpretar el lenguaje humano y lo puede analizar extrayendo conceptos, entidades, emociones o palabras clave [Khurana *et al.*, 2023]. Con esto va estructurando los datos y formando las relaciones

que se dan entre las palabras y frases. Los humanos hacen esto de forma natural, pero el entendimiento del lenguaje natural lo que busca es que de cierta manera la máquina logre replicar ciertos procesos cognitivos (del humano) para entender el lenguaje natural, así una persona puede “comunicarse” con la computadora utilizando el lenguaje natural. Cabe destacar que el NLU aplica para texto y discurso (voz).

2.1.2. Generación del Lenguaje Natural (NLG)

La segunda subdisciplina del PLN se basa en la capacidad de la máquina de poder producir y formular correctamente palabras, frases, oraciones, párrafos o discursos de manera automática y que tengan sentido en el lenguaje natural a partir de datos, tal y como si lo hubiese hecho un humano. En NLG, la máquina hace lo opuesto a NLU, en el entendimiento analiza y comprende lo que el lenguaje humano plantea y en la generación tiene que ser capaz de decidir como formular frases o discursos coherentes.

2.1.3. Aplicaciones del Procesamiento del Lenguaje Natural

El procesamiento del lenguaje natural puede ser utilizado en varias aplicaciones o áreas y entre ellas hay algunas que se usan en el día a día o al menos cotidianamente.

- Traducción de idiomas: Debido a la globalización y a la presencia del internet en casi todo el mundo, en la actualidad se está más en contacto con grandes cantidades de datos e información. Estos pueden estar en cualquier idioma y cada uno tiene sus propias reglas gramaticales y estructuras. La traducción de idiomas traduce frases, palabras o textos completos de un idioma a otro automáticamente. Buen ejemplo de estas aplicaciones son Google Traductor, Bing Traductor o DeepL. Estos traductores utilizan inteligencia artificial o redes neuronales para el aprendizaje automático ya que el mayor desafío que presentan es no caer en traducir una frase “palabra a palabra”, sino mantener el significado de la oración original intacto [Khurana *et al.*, 2023].
- Categorización de texto: La categorización de texto se encarga de clasificar textos en distintas clases. Se utiliza inteligencia artificial para entrenar algoritmos que se encargan de clasificar. Tiene variadas aplicaciones, tales como el filtrado de spam de los correos, análisis de sentimientos de un texto y la moderación de contenidos en redes sociales.
- Extracción de información y resumen: Día a día hay más datos e información en la web, con esto nace la necesidad de poder extraer las ideas principales de cada texto o poder resumir grandes artículos o archivos, esto para poder entregar información que sea relevante para el usuario.

- Asistentes de voz: En la actualidad los asistentes de voz son muy utilizados, estos brindan apoyo al usuario o realizan ciertas acciones por ellos [Khurana *et al.*, 2023]. Para poder asistir deben poder comprender los requerimientos que les pide el usuario. También son capaces de entablar conversaciones. Algunos de los ejemplos más conocidos mundialmente son Siri, la asistente de los dispositivos Apple; Alexa, de Amazon; Cortana, de Windows; el asistente de Google o Bixby de Samsung.

2.2. Análisis de Sentimientos

El análisis de sentimientos o minería de opinión (que por lo general se ocupan como conceptos sinónimos) es un área del análisis de texto y categorización de texto, que a su vez son áreas del procesamiento del lenguaje natural. En el análisis de sentimientos se estudian las opiniones o emociones de las personas acerca una entidad o tópico [Medhat *et al.*, 2014].

Con el creciente uso del internet, redes sociales y foros en línea en las últimas dos décadas ha incrementado también la cantidad de opiniones y toda la valiosa información que se puede extraer de ellas en la web. Es por eso que para numerosas, entidades, empresas, gobiernos o estados y organizaciones se ha vuelto tan importante y tan utilizado el análisis de sentimientos. Esto para poder monitorear la opinión pública y tomar decisiones beneficiosas en base a ella [Wankhade *et al.*, 2022].

La diferencia entre la minería de opinión con el análisis de sentimientos es que la primera extrae y analiza las opiniones sobre algo o alguien [Medhat *et al.*, 2014], mientras que el análisis de sentimientos busca clasificar textos según el sentimiento que predomina en él. Las clasificaciones por lo general corresponden a “Positivo”, “Negativo” y “Neutral”.

Algunos de los más grandes desafíos que tiene el análisis de sentimientos y a su vez el procesamiento de lenguaje natural están relacionados al texto informal, al contexto del texto u opinión a analizar, al sarcasmo, a la ironía, a los modismos y también al idioma [Wankhade *et al.*, 2022] e incluso la geografía, ya que ciertas palabras o frases pueden tener distintas connotaciones en países o regiones que se hable el mismo idioma.

En el análisis de sentimientos hay cuatro niveles: análisis de sentimientos de documentos, de sentencias, de frases y de aspectos (ver Figura 6).

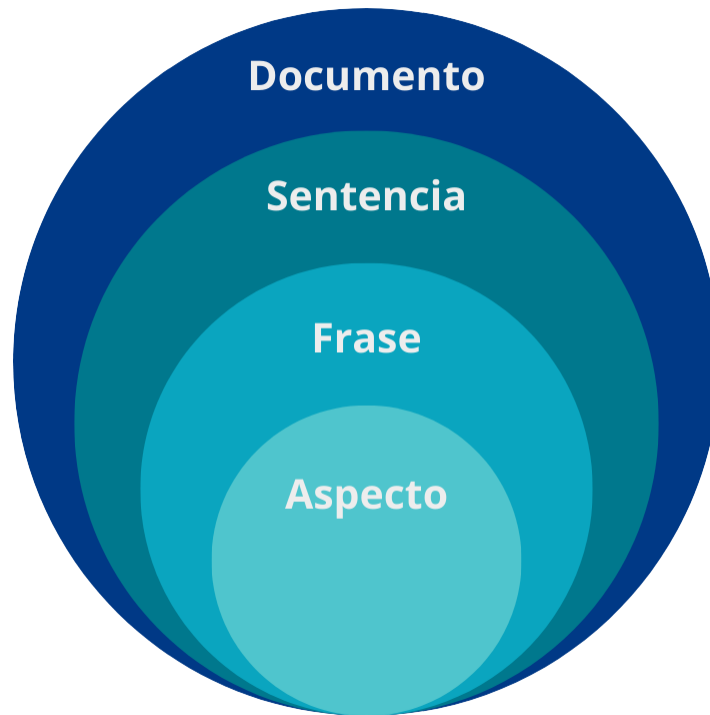


Figura 6: Niveles del análisis de sentimientos
Fuente: Elaboración propia basado en [Wankhade *et al.*, 2022]

- **Análisis de sentimiento a nivel de documento:** El nivel que analiza una cantidad de texto más extensa. Acá el análisis se realiza en la totalidad de un documento, por lo que la categorización se le da al documento entero. Aplica por ejemplo al sentimiento en el que se clasificaría un libro o un capítulo de este.
- **Análisis de sentimiento a nivel de sentencia u oración:** En el segundo nivel más extenso se analizan las oraciones de un texto. En este nivel el sentimiento de cada sentencia se determina de manera independiente respecto de las otras oraciones. Es utilizado cuando el documento tiene más de una clasificación involucrada, o la polaridad resulta ser neutra (dependiendo de como el modelo de clasificación trabaja los neutros).
- **Análisis de sentimiento a nivel de frase:** Se trata de la clasificación de la polaridad o sentimiento en una frase. Va un paso más allá de las oraciones, ya que ellas pueden estar formadas por algunas frases.
- **Análisis de sentimiento a nivel de aspecto:** Es el nivel más particular del análisis de sentimientos. Una frase u oración puede tener varios aspectos y un sentimiento o polaridad ligado a ellos. En este nivel de análisis hay que tener más factores en cuenta, tal como el contexto, el tópico del texto y quién lo escribe.

A continuación se presentan las técnicas que se pueden emplear para el análisis de sentimientos (ver Figura 7).

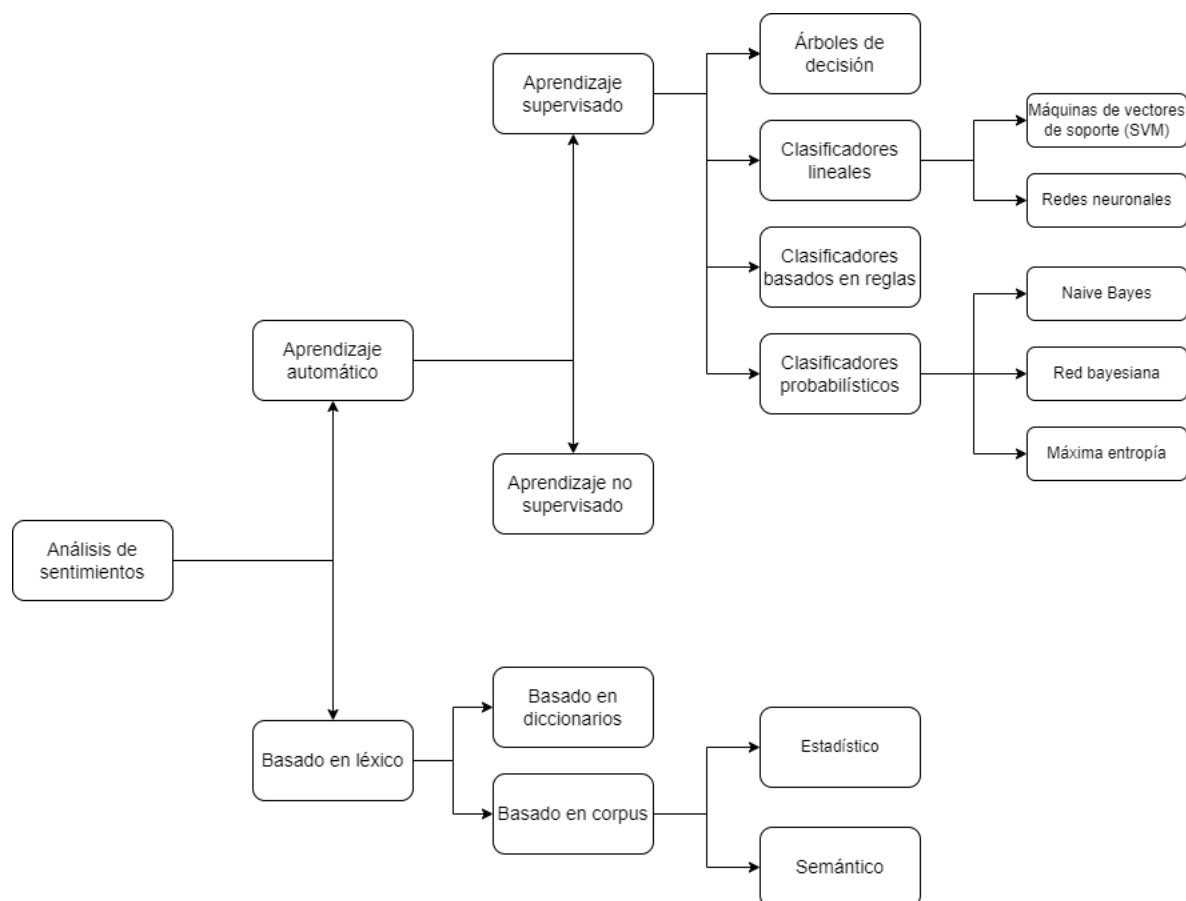


Figura 7: Técnicas utilizadas para el análisis de sentimientos
Fuente: Elaboración propia basado en [Medhat *et al.*, 2014]

2.2.1. Enfoque basado en léxico

Las técnicas de análisis de sentimientos basadas en léxico utilizan colecciones de palabras o frases (llamadas tokens en la comunidad que estudia el análisis de sentimientos) que tienen puntuaciones de polaridad: positivo, neutral o negativo. Estas puntuaciones podrían ser fijas o no. Para las puntuaciones fijas se utiliza el “1” para los tokens positivos, “0” para los neutros y “-1” para los negativos. También como se mencionó anteriormente, estas puntuaciones podrían no ser fijas y oscilar entre estos valores, significando lo mismo pero con valores intermedios, por ejemplo, si se tiene el valor 0.4 en un token quiere decir que este oscila entre lo neutro y lo positivo, siendo más cercano a lo positivo, ya que el consenso de la literatura del análisis de sentimientos considera neutros los valores entre -0.1 y 0.1

Este enfoque tiene ventajas y desventajas. Algunas de las ventajas más destacables es que

es un enfoque sencillo y no necesita de entrenamiento para poder clasificar sentimientos. La principal desventaja viene de su gran fortaleza, ya que al ser sencillo cuando se necesita un análisis más profundo este enfoque podría no ser el mejor. Otra desventaja que resalta es que se limita a la puntuación de los tokens y no analiza el contexto detrás del texto y como se ha mencionado anteriormente el sentimiento de un texto depende mucho del contexto en el que se desarrolla.

Dentro del enfoque basado en léxico está el basado en diccionarios y el basado en corpus. El enfoque basado en diccionarios utiliza una colección de palabras clasificadas con cierta puntuación de polaridad y para el análisis del texto se van buscando estas palabras o sus sinónimos y antónimos en la gran colección de palabras categorizadas para poder puntuarlas con el mismo valor o el valor opuesto para el caso de los antónimos. Por el otro lado, el enfoque basado en corpus emplea patrones semánticos y sintácticos para poder clasificar el sentimiento o polaridad de la oración y su orientación en un corpus grande[Wankhade *et al.*, 2022].

2.2.2. Enfoque basado en aprendizaje automático (ML)

Los técnicas de Machine Learning se pueden utilizar para analizar y clasificar sentimientos en texto. Estas usan algoritmos y modelos que deben aprender sobre patrones o clasificaciones de sentimientos a partir de datos clasificados (aprendizaje supervisado) o no clasificados (aprendizaje no supervisado), esta es la gran división dentro de las técnicas de aprendizaje automático y se verán con más detalle en una de las próximas secciones.

2.2.3. Análisis de sentimientos y X (Twitter)

La razón por la que se escogió X para el desarrollo de la memoria se debe a que es la red social que prefieren los usuarios que quieren opinar sobre cualquier temática. Otras redes sociales se descartaron porque su contenido es principalmente multimedia y son mayormente ocupadas para entretenimiento. Los Tweets (en este trabajo se seguirán denominando Tweets aunque ahora sean llamados Posts en X, para evitar confusiones), son la elección popular para realizar análisis de texto [Sailunaz y Alhajj, 2019], actualmente posee un número limitado de 280 caracteres para los usuarios normales sin suscripción.

2.3. Machine Learning

El *Machine Learning* (ML) o aprendizaje automático es un área de la inteligencia artificial en la que se pretende hacer que las computadoras aprendan por si solas mediante técnicas o algoritmos, sin ser programadas específicamente para eso. El aprendizaje se da cuando se entrena al algoritmo con un conjunto de datos.

Existen diferentes tipos de aprendizaje en el ML, tales como los que se presentan a continuación:

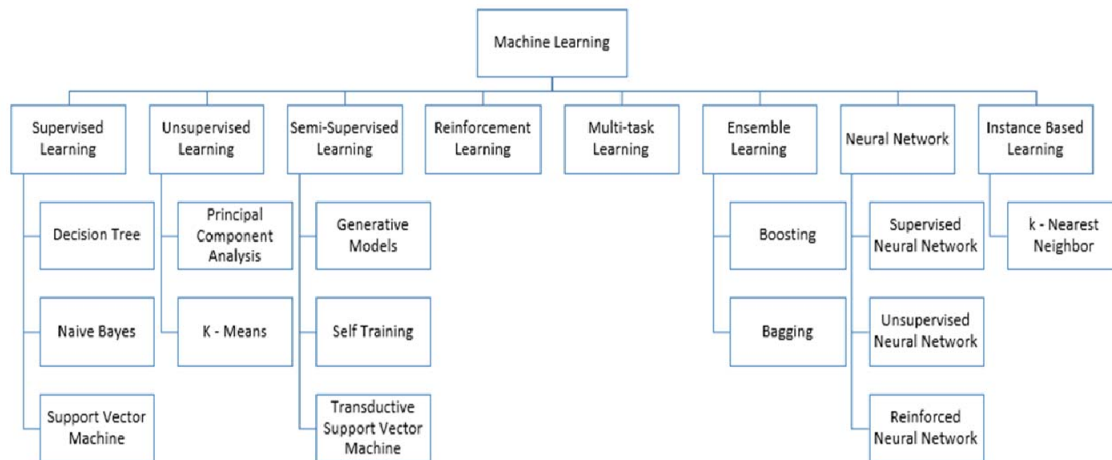


Figura 8: Técnicas de Machine Learning
Fuente: [Mahesh, 2019]

El tipo de aprendizaje semi-supervisado es una combinación del aprendizaje supervisado y no supervisado. El aprendizaje por refuerzo (*Reinforcement learning*) es cuando el software debe tomar ciertas acciones para poder aumentar la noción de premio. *Multitasking learning* es un tipo de aprendizaje en donde se hacen diferentes tareas al mismo tiempo. *Ensemble learning* es cuando a través de varios modelos se trata de solucionar algún problema de inteligencia artificial. *Neural Networks* se basa en una serie de algoritmos que trata de imitar el funcionamiento de las neuronas. Por último, el aprendizaje basado en instancias se trata de varias técnicas de clasificación que generan un etiquetado del dato basado en el parecido que tiene con sus vecinos (en los datos de entrenamiento) [Mahesh, 2019].

Para efectos de esta memoria, nos centraremos en las técnicas de aprendizaje supervisado.

2.3.1. Aprendizaje no supervisado

A diferencia del anterior, en este tipo de aprendizaje se entrena el algoritmo con datos sin etiquetar y no necesita ayuda externa. Debe ser el mismo algoritmo el que encuentre patrones en el conjunto de datos [Ramos y Freire, 2021]. Es usado principalmente para *clustering* [Mahesh, 2019].

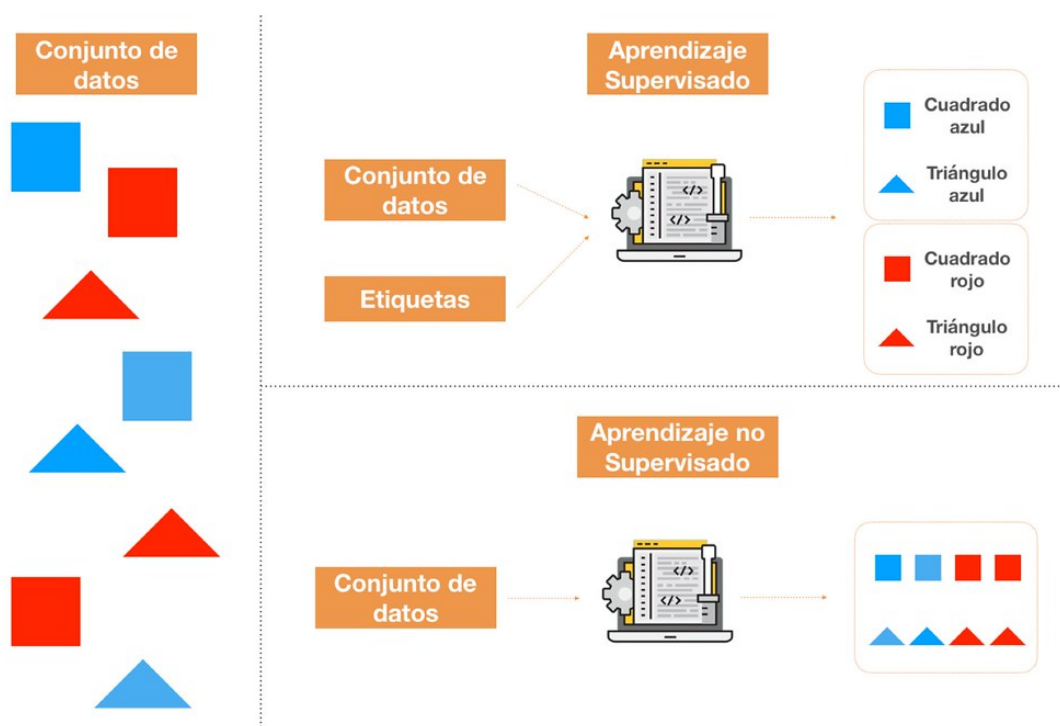


Figura 9: Aprendizaje supervisado y no supervisado
Fuente: [Gonzalez, 2020]

2.3.2. Aprendizaje supervisado

En este tipo de aprendizaje los algoritmos necesitan apoyo externo, se entrena al algoritmo con datos que han sido previamente etiquetados de manera que logra distinguir que elementos caen en la clasificación y son “similares” a los etiquetados. La totalidad de los datos que se tienen a disposición deben ser divididos en datos de entrenamiento y datos de prueba. Como *output* se tiene una predicción o una clasificación de la data [Mahesh, 2019]. Estos algoritmos o modelos siguen un flujo de trabajo como el de la Figura 10.

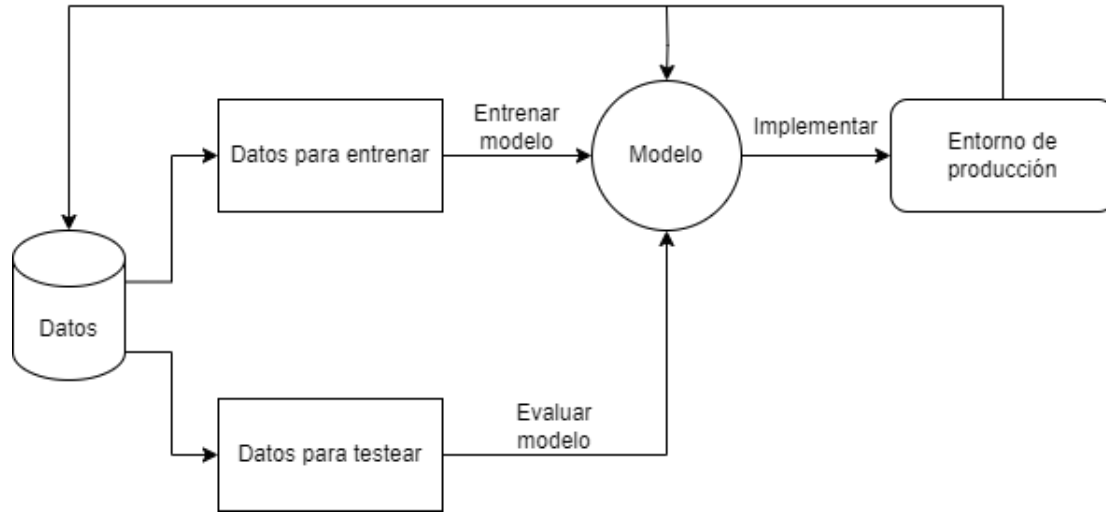


Figura 10: Flujo de trabajo de las técnicas de aprendizaje supervisado
Fuente: Elaboración propia basado en [Mahesh, 2019]

A continuación se presentarán algunas técnicas o algoritmos de interés para el desarrollo de esta memoria.

2.3.3. Naive Bayes

Se nombra de esta manera por el creador del teorema de Bayes, Thomas Bayes. Este representa un método de aprendizaje supervisado y también un método estadístico para la clasificación. Asume un modelo probabilístico y es capaz de resolver problemas que requieran diagnosticar o predecir [Latha, 2019]. Es el clasificador más simple que se puede encontrar en la literatura y estudios, clasifica rápidamente y se utilizan datos etiquetados para ello. Se le denomina “ingenuo” (naive) porque supone la independencia de las características entre ellas, es decir, que las características no afectan sus valores unas a otras. Está basado en el teorema de Bayes que establece la probabilidad de un evento teniendo en consideración nueva información que podría afectar a la probabilidad de que ocurra el evento.

La fórmula del Teorema se expresa como:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (1)$$

A pesar de que Naive Bayes supone que las características son independientes entre sí, esta suposición se cumple escasas veces. Aún así, el algoritmo funciona bien en muchos problemas de clasificación y es altamente recomendado como una solución simple en la literatura.

En análisis de texto, Naive Bayes tiene algunas variantes que podrían ser más útiles en ciertos contextos.

- Gaussian Naive Bayes (GNB): Esta variante asume que los datos son continuos y que sus características siguen una distribución normal.
- Multinomial Naive Bayes (MNB): MNB se utiliza para datos discretos. Se utiliza comúnmente en las clasificaciones de textos donde los datos corresponden a la frecuencia de las palabras.
- Bernoulli Naive Bayes (BNB): La variante Bernoulli se asimila al Multinomial pero es utilizado cuando la clasificación es binaria.

2.3.4. Random Forest (RF)

Random Forest es un algoritmo de aprendizaje automático que utiliza árboles de decisión. Este algoritmo se ocupa para clasificación y regresión. La predicción de alto grado del RF se da cuando una predicción ocurre varias veces en el árbol de decisión [T.K. *et al.*, 2021].

Para el algoritmo de bosque aleatorio se seleccionan muestras del conjunto de datos original (estas deben ser aleatorias) y se construyen variados árboles de decisión, en cada nodo del árbol se selecciona algún subconjunto aleatorio de características de los datos y se busca la mejor entre ellas para poder dividir los datos. Entonces así, cada árbol puede entrenarse con una muestra distinta del dataset y así se van seleccionando las mejores divisiones del subconjunto de las características. Luego los combina para poder obtener una predicción más precisa.

Ya que el algoritmo es útil para clasificación y regresión la diferencia radica en que para la clasificación se utiliza la clase que más árboles predicen y para la regresión se usa el promedio de las predicciones de todos los árboles.

Entre sus ventajas está la flexibilidad que lo caracteriza, ya que funciona bien cuando faltan datos, también es capaz de manejar grandes conjuntos de datos con altas dimensiones y este modelo suele ser más preciso que utilizar solo un árbol de precisión.

En la Figura 11 se puede visualizar un esquema de su funcionamiento.

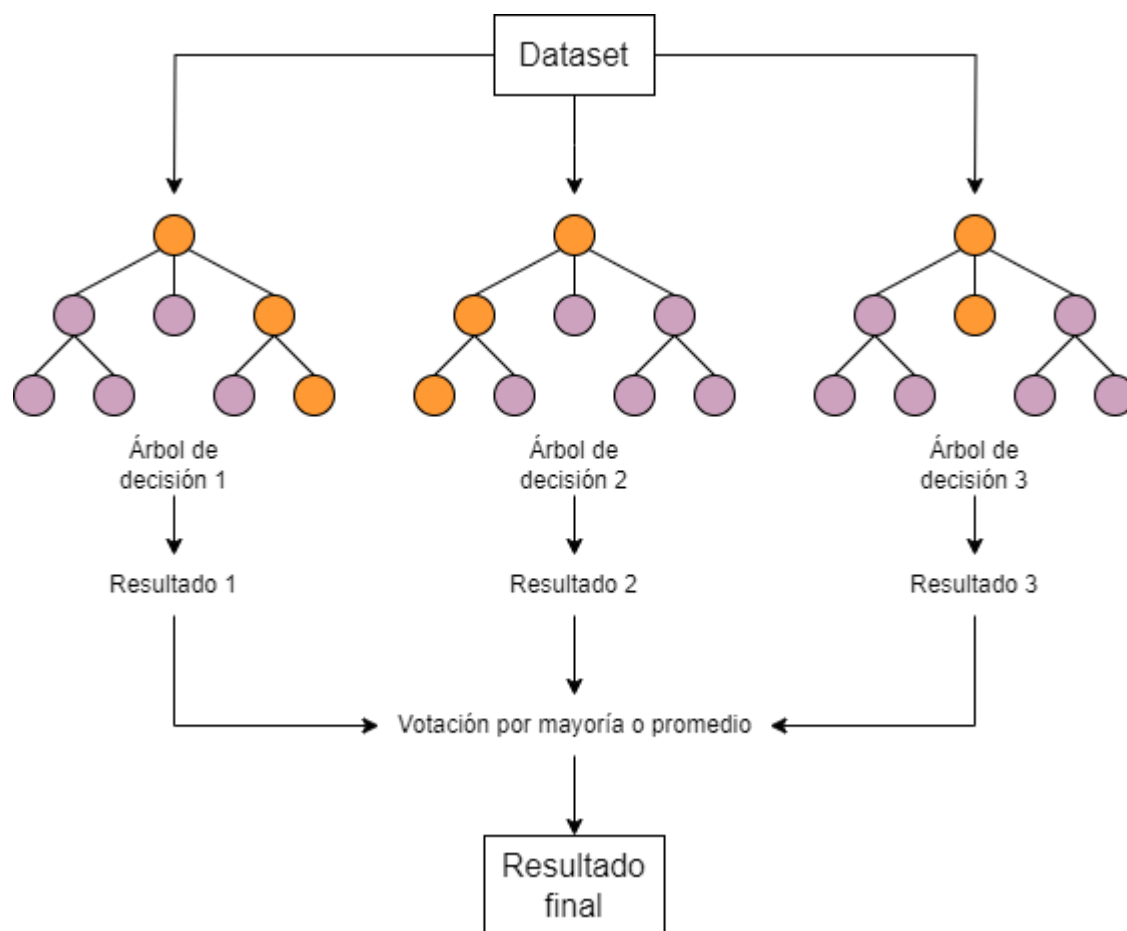


Figura 11: Ejemplo de Random Forest
Fuente: Elaboración Propia

2.3.5. Support Vector Machines (SVM)

Support Vector Machines es uno de los algoritmos de aprendizaje supervisado más utilizados. Este algoritmo se usa para clasificación y regresión. Clasifica los datos que recibe dibujando un plano sobre ellos, la mejor clasificación es la mejor brecha en los lados de un plano [T.K. *et al.*, 2021]. También se pueden utilizar planos multidimensionales cuando los datos así lo requieran. Para predecir la clase de un dato una vez entrenado el algoritmo, se determina de qué lado del plano se encuentra el punto. En la Figura 12 se puede ver una ilustración del algoritmo para un mejor entendimiento.

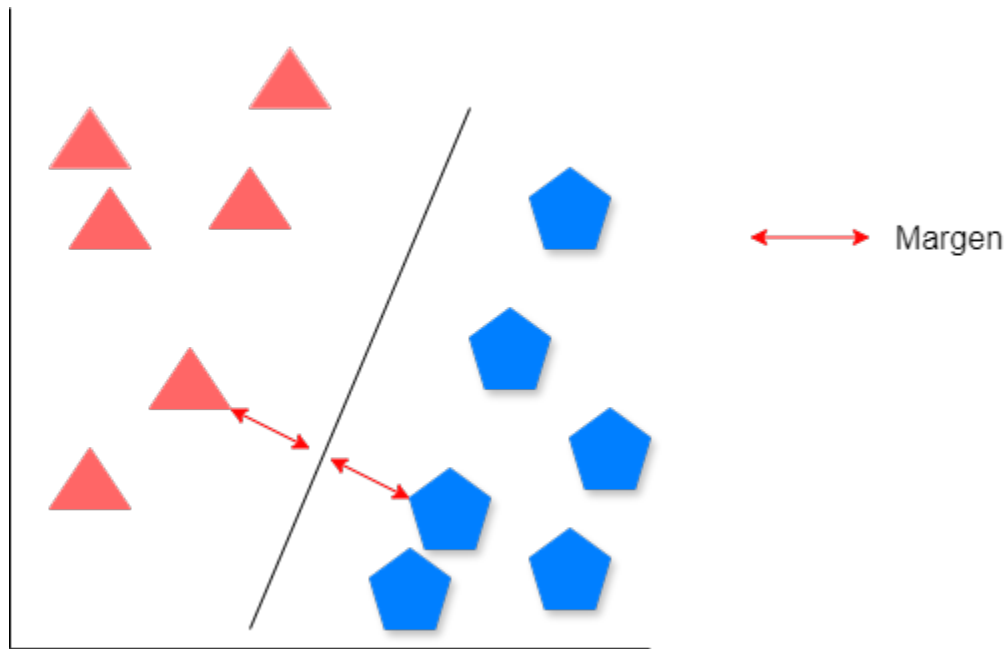


Figura 12: Ejemplo de Support Vector Machine
Fuente: Elaboración Propia

Este modelo es eficiente en espacio multidimensionales y también cuando tiene pocos ejemplos. Por otro lado, a medida que crece el tamaño del conjunto de datos, pierde eficacia, tiempo y memoria.

2.3.6. BERT (Bidirectional Encoder Representations from Transformers)

Es un modelo de procesamiento del lenguaje natural basado en transformadores creado por Google. Ha sido uno de los más grandes avances en el estudio del procesamiento del lenguaje natural. A diferencia de las técnicas anteriormente revisadas, este no es un modelo de aprendizaje supervisado, BERT se considera como un modelo de aprendizaje autosupervisado (Self-Supervised).

Para entender mejor BERT, hay que conocer que son los Transformadores (Transformers). Un transformador es una arquitectura de redes neuronales que fue presentada por [Vaswani *et al.*, 2017] en el renombrado artículo "Attention Is All You Need" en el año 2017. Son capaces de manejar dependencias en el texto y realizar entrenamiento de forma paralela.

Un transformador está conformado por un codificador y decodificador. Pero en el modelo de BERT solo se utiliza el codificador.

BERT se caracteriza por ser un modelo bidireccional, es decir, lee el texto en ambas direcciones.

nes al mismo tiempo, esto permite que se pueda entender el contexto de mejor manera, ya que al analizar una palabra se consideran también las palabras que le rodean.

BERT ha sido pre entrenado en inglés pero hay versiones para su funcionamiento en español. Es un modelo que ha sido entrenado con textos sin etiquetar, las etiquetas de datos se generan automáticamente y son utilizadas en las iteraciones posteriores al pre entrenamiento, por eso se le denomina aprendizaje autosupervisado. Se diferencia de los modelos recurrentes o secuenciales ya que permite que todos los tokens que recibe se procesen de forma paralela [Rosenbrock *et al.*, 2021].

Una de las versiones en español más conocida es BETO, desarrollado por el Departamento de Ciencias de la Computación Universidad de Chile, es muy similar a BERT pero este se ha pre entrenado con textos en español [Cañete *et al.*, 2020].

2.4. Métricas de evaluación para análisis de sentimientos

Las métricas de evaluación o validación son fundamentales para poder medir numéricamente la efectividad de los algoritmos de análisis de sentimientos o de clasificación en general.

Para explicar las fórmulas se definirán las siguientes variables:

- VP: Verdadero positivo. Cuando el modelo clasifica correctamente un positivo en los datos de prueba.
- VN: Verdadero negativo. Es cuando el modelo categoriza de manera correcta un negativo en los datos de prueba.
- FP: Falso positivo. Corresponde a un falso positivo cuando el modelo clasifica incorrectamente un dato como positivo.
- FN: Falso negativo. Es un falso negativo cuando el modelo categoriza de manera incorrecta un dato como negativo.

Las métricas más utilizadas en la literatura son:

- Exactitud: Exactitud o Accuracy es la proporción que se da entre las predicciones que son correctas y el total de todas las predicciones realizadas.

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

- Precisión: La precisión es la métrica que mide la proporción entre las predicciones correctas positivas sobre el total de todas las predicciones positivas hechas por el modelo.

$$\text{Precision} = \frac{VP}{VP + FP}$$

- Recall: Corresponde a la proporción de positivos reales que fueron clasificados de manera correcta por el modelo.

$$\text{Recall} = \frac{VP}{VP + FN}$$

- F1 Score: El Puntaje F1 es descrito como la media armónica de las métricas de Recall y precisión, así entrega un nuevo puntaje que es capaz de equilibrar ambas métricas.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Estas fórmulas son útiles para cuando el problema de clasificación es binario. Cuando se debe clasificar con más de dos clases, estas métricas se pueden ampliar usando técnicas de promediado. Las más comunes son los promediados micro y macro [Kumarl, 2023].

- Exactitud: Cuando son más de dos clases esta métrica se puede extender sumando los verdaderos positivos de todas las clases, esto dividido entre el total de las muestras.

$$\text{Exactitud} = \frac{\sum_{i=1}^C VP_i}{\text{Total de Muestras}}$$

- Precisión:

- Precisión Micro: La precisión con el promediado micro se calcula de manera que se sumen los verdaderos y falsos positivos de todas las clases. Luego se aplica la fórmula de precisión.

$$\text{Precisión}_{micro} = \frac{\sum_{i=1}^C VP_i}{\sum_{i=1}^C (VP_i + FP_i)}$$

- Precisión Macro: Se promedia la precisión de cada clase.

$$\text{Precisión}_{macro} = \frac{1}{C} \sum_{i=1}^C \frac{VP_i}{VP_i + FP_i}$$

- Recall:

- Recall Micro: Se suman los verdaderos positivos y falsos negativos de todas las clases. Después se debe aplicar la fórmula de recall.

$$\text{Recall}_{micro} = \frac{\sum_{i=1}^C VP_i}{\sum_{i=1}^C (VP_i + FN_i)}$$

- Recall Macro: Se promedia la métrica recall de cada clase.

$$\text{Recall}_{macro} = \frac{1}{C} \sum_{i=1}^C \frac{VP_i}{VP_i + FN_i}$$

■ F1 Score:

- F1 Score Micro: El puntaje F1 micro se calcula con la precisión y recall micro, siguiendo la fórmula.

$$F1\text{-score}_{micro} = 2 \cdot \frac{Precisión_{micro} \cdot Recall_{micro}}{Precisión_{micro} + Recall_{micro}}$$

- F1 Score Macro: Se calcula de tal manera que se promedien los F1 Score de todas las clases.

$$F1\text{-score}_{macro} = \frac{1}{C} \sum_{i=1}^C 2 \cdot \frac{Precisión_i \cdot Recall_i}{Precisión_i + Recall_i}$$

- Matriz de confusión: Otra métrica que es común en este tipo de problemas es la matriz de confusión. Esta se basa en una matriz o tabla que muestra las verdaderas etiquetas y las predicciones incorrectas.

Valores Verdaderos	Positivo	1254	83	145
	Neutro	124	1189	132
	Negativo	95	149	1237
		Positivo	Neutro	Negativo
		Valores Predichos		

Figura 13: Ejemplo de una matriz de confusión, con datos ficticios

Fuente: Elaboración Propia

- **Loss:** Función de pérdida. Esta función es una medida cuantitativa de cuánto se desvían las predicciones hechas por el modelo de las etiquetas reales. Este se utiliza como feedback para que el algoritmo de inteligencia artificial a entrenar ajuste algunos parámetros o valores internos, tal y como los pesos de las redes neuronales. Todo esto con el objetivo de que logre mejorar el rendimiento. El rango de valores que debe tomar depende del escenario, pero siempre es ideal que el valor se acerque a 0. En problemas binarios valores menores a 0,2 son considerados buenos. Pero en problemas multiclase depende del número de clases, idealmente debe ser un valor bajo mientras lo permita la clasificación multiclase.

- **Mean Squared Error (MSE):** Este tipo de función de pérdida se suele utilizar en problemas de regresión. Consiste en la suma de las diferencias al cuadrado entre las entradas del vector de predicción y el real. Esta suma se divide por la longitud de los vectores.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Cross Entropy Loss:** Para problemas de clasificación, donde las etiquetas son de carácter categórico, esta es la función de pérdida más común. Consiste en la comparación de la distribución de probabilidad de las etiquetas predichas con la distribución de la probabilidad de las etiquetas verdaderas. La fórmula es la siguiente:

$$CE = - \sum_{i=1}^N y_{true_i} \cdot \log(y_{pred_i})$$

Donde N corresponde a la cantidad de clases.

2.5. Metodologías para el análisis de los datos

El análisis de datos en general, independiente de su tipo, es gran parte de lo que se realiza en la ciencia y minería de datos. Esto implica la extracción de información desde grandes volúmenes de datos. Para poder ordenar y guiar cada etapa del proceso de análisis de datos existen varias metodologías. Entre las más utilizadas y reconocidas están CRISP-DM, KDD y SEMMA, cada una con un enfoque distinto.

2.5.1. CRISP-DM (Cross Industry Standard Process for Data Mining)

La metodología CRISP-DM es utilizada ampliamente en el contexto de los negocios. Esta se compone de 6 partes.

- **Entendimiento del negocio:** La primera etapa se centra en entender los objetivos que conlleva el proyecto y la comprensión de los requerimientos y requisitos desde la perspectiva del negocio. Así, con todo esto, se puede idear un plan cuyo diseño permita alcanzar los objetivos propuestos.
- **Entendimiento de los datos:** En esta etapa se recopilan los datos a utilizar y se realizan actividades que permitan conocer e interiorizar los datos. También, en esta fase se identifican problemas de la calidad de los datos y se obtiene información preliminar de ellos.
- **Preparación de los datos:** La etapa de la preparación de los datos contempla todas las acciones que se deben realizar para poder dejar el dataset de una manera apropiada para su análisis y las etapas posteriores.
- **Modelado:** En la etapa de modelado se aplican las técnicas de modelados seleccionadas para el propósito del análisis. La idea de esta etapa es probar algoritmos y encontrar parámetros que sean óptimos.
- **Evaluación:** En esta etapa se evalúan los algoritmos y técnicas de modelado de la etapa anterior para asegurarse de que se cumplan los objetivos del negocio.
- **Implementación:** Se implementa el modelo realizado en los procesos del negocio.

2.5.2. KDD (Knowledge Discovery in Databases)

La metodología KDD consta de un proceso más amplio y completo. Esta utiliza la minería de datos para su realización y tal como lo dice su nombre, extrae el conocimiento desde grandes volúmenes de datos, como lo es una base de datos [Azevedo y Santos, 2008]. Está compuesta de 5 etapas:

- **Selección de los datos:** La etapa inicial de la metodología se centra en seleccionar los datos según los requerimientos del proyecto o estudio a realizar y crear un dataset apropiado para el análisis que se llevará a cabo.
- **Preprocesamiento de los datos:** En esta etapa se deben limpiar y preprocesar los datos, así como resultado se obtienen unos datos consistentes y apropiados.
- **Transformación de los datos:** Se transforman los datos en un formato adecuado para el análisis. Se aplican técnicas que reducen la dimensión de los datos.
- **Minería de datos:** En esta etapa se deben aplicar técnicas de minería de datos (tales como métodos estadísticos, algoritmos de aprendizaje automático, entre otros) para buscar patrones en ellos, de esta manera se pueden descubrir ciertas relaciones que se encontraban invisibles u ocultas entre los datos.

- **Evaluación:** Se evalúan los resultados de la etapa anterior con distintas métricas para que posteriormente estos se puedan interpretar.

2.5.3. SEMMA (Sample, Explore, Modify, Model and Assess)

La metodología SEMMA fue desarrollada por el instituto SAS [Azevedo y Santos, 2008]. Este sirve para guiar los proyectos que tratan de minería de datos.

- **Muestreo (Sample):** Esta etapa puede ser opcional. Trata de escoger una buena muestra de la totalidad de los datos, así se puede conocer información esencial de estos mismos de manera preliminar. La muestra debe ser de un tamaño que permita manipular los datos rápidamente.
- **Explorar (Explore):** En esta etapa se analizan los datos de una forma exploratoria, pueden utilizarse visualizaciones para identificar relaciones y patrones.
- **Modificar (Modify):** Se preparan, transforman y modifican los datos para su posterior análisis. Esta etapa incluye la limpieza y normalización de los datos.
- **Modelar (Model):** Se aplican las técnicas y algoritmos de modelados correspondientes.
- **Evaluar (Assess):** Se evalúa el desempeño y funcionamiento del modelo siguiendo distintas métricas, como por ejemplo la precisión.

2.5.4. Comparación entre las tres metodologías

Entendiendo las fases de cada una de las metodologías, se podría decir que tienen etapas parecidas o equivalentes. Esto se destaca más en la comparación de la metodología KDD y SEMMA, ya que ambas constan de 5 fases principalmente, que son en esencia casi lo mismo cada una: Selección equivale a Muestreo, Preprocesamiento a Exploración, Transformación a la etapa de Modificar, Minería de datos con el Modelar y la Evaluación con su igual de SEMMA [Azevedo y Santos, 2008].

La metodología CRISP-DM es algo distinta a las dos mencionadas anteriormente, pero de igual forma tiene algunas etapas que cumplen con funciones parecidas a las etapas de KDD y SEMMA. Por ejemplo el Entendimiento de los datos de CRISP-DM podría compararse con las etapas de Selección y Preprocesamiento de KDD y el Muestreo y la Exploración de SEMMA, la Preparación de los datos equivale a la Transformación de los datos (KDD) y a Modificación de los datos SEMMA [Azevedo y Santos, 2008]. Lo mismo ocurre con las etapas de Modelado y la Evaluación.

Con esta comparación (Tabla 1) es evidente que las metodologías cumplen con etapas parecidas pero que cada una de ellas tiene un enfoque distinto según el proyecto o estudio a realizar.

CRISP-DM	KDD	SEMMA
Entendimiento del negocio	Pre KDD	--
Entendimiento de los datos	Selección	Muestreo
Entendimiento de los datos	Preprocesamiento	Exploración
Preparación de los datos	Transformación	Modificar
Modelado	Minería de datos	Modelar
Evaluación	Evaluación	Evaluación
Implementación	Post KDD	--

Tabla 1: Comparación de las 3 metodologías
Fuente: Elaboración Propia, basado en [Azevedo y Santos, 2008]

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

En el presente capítulo se expondrá de manera exhaustiva la propuesta de solución planteada para poder trabajar y abordar el problema formulado en el Capítulo 1. A partir del análisis del marco teórico que ha sido expuesto en el capítulo anterior, se detalla una propuesta escogiendo una de las metodologías descritas anteriormente, para luego recorrer todas las fases del proceso, desde la selección de los datos hasta la implementación de los algoritmos. Como resultado de esta propuesta de solución, se tiene la construcción de dos modelos de análisis de sentimientos, uno con enfoque sencillo y el otro con un enfoque más complejo.

3.1. Selección de la metodología a utilizar

Para comenzar a trabajar y usando como guía el marco conceptual, se debe definir qué metodología se usará en el desarrollo de este capítulo, de esta forma se sigue un orden idóneo para el trabajo propuesto que permita cumplir con los objetivos y el enfoque de la problemática.

De las tres metodologías planteadas para el análisis de los datos se considera que la más apropiada es KDD, ya que CRISP-DM tiene un enfoque completamente distinto al estar orientada a negocios (podría servir si el estudio lo realizara una empresa u organización) y SEMMA considera una exploración previa de los datos y hasta visualizaciones o gráficas antes de ir al análisis en sí, siendo un tipo de preanálisis. KDD propone un orden de las etapas que se ajusta más a la investigación y análisis que se desea hacer y aparte va directo al asunto, sin hacer análisis exploratorios más extensos de lo necesario justo después de seleccionar los datos.

Entonces, con KDD hay que seguir el orden de seleccionar los datos, preprocesarlos, transformarlos, modelarlos y luego evaluarlos para su posterior interpretación.

La selección de datos se comenzará mediante el filtrado y extracción de tweets de interés, que traten de las entidades a analizar: Carabineros y Congreso, en el periodo de tiempo definido por el problema. Se formarán dos datasets distintos, uno por cada entidad. Posterior a esto, se limpiarán y transformarán los datos dejándolos aptos para entregarlos al modelo a entrenar. Se buscarán parámetros óptimos para que el modelo sea lo suficientemente “bueno” para cumplir con las validaciones y luego para poder clasificar todos los demás tweets que se extrajeron.

En la Figura 14 se puede visualizar la metodología a seguir para analizar el sentimiento de los tweets a extraer.

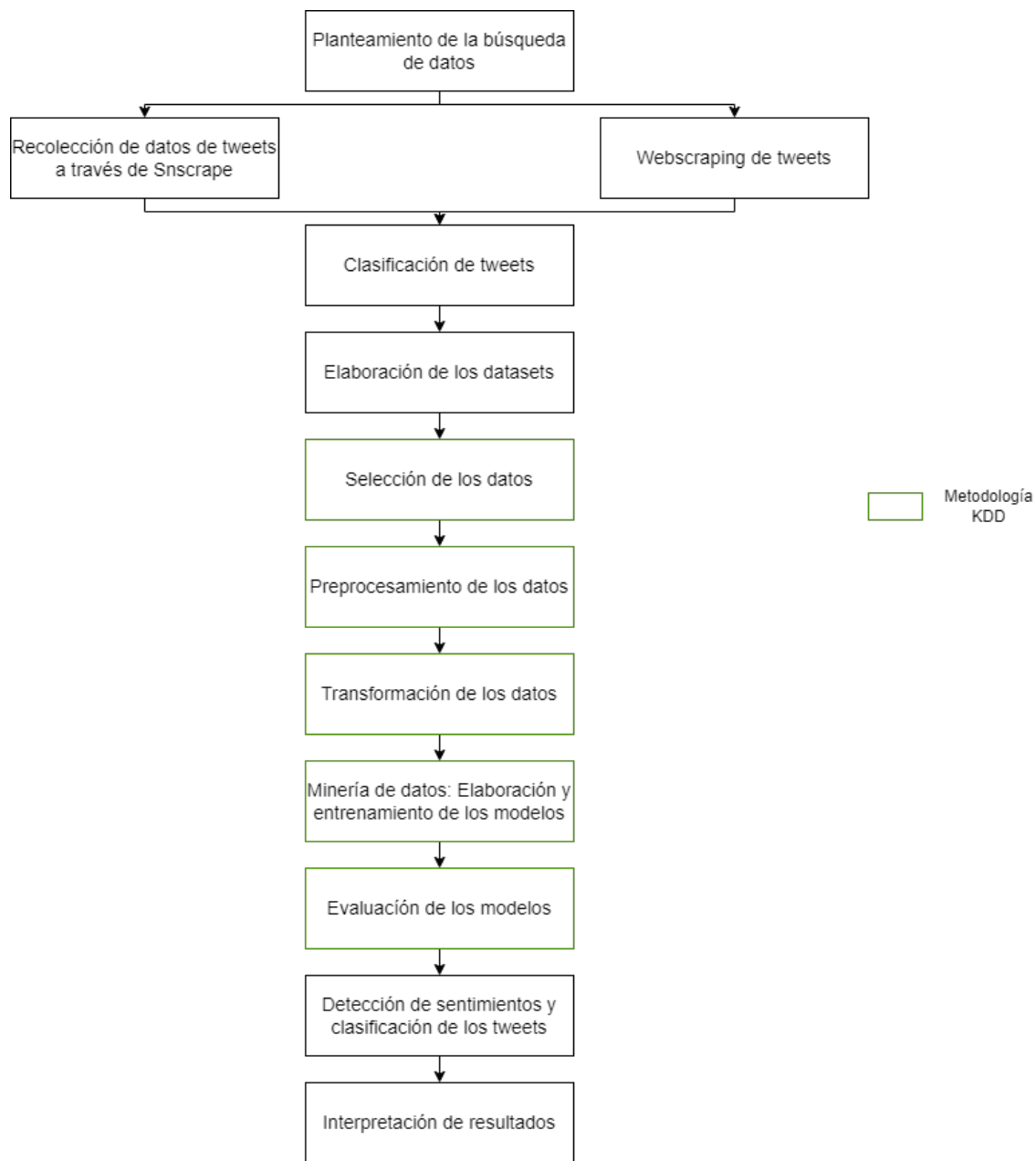


Figura 14: Metodología a seguir
Fuente: Elaboración propia

3.2. Extracción de los datos

En la fecha en la que se extrajeron los datos, *Twitter* (ahora *X*) estaba pasando por una serie de cambios debido a la llegada de Elon Musk al mandato de la red social. Anteriormente, cuando aún era *Twitter*, poseían una API de acceso público, en la cual los usuarios podían ex-

traer información para diferentes fines, siendo uno de los más populares el fin investigativo.

Esta API permitía la existencia de muchas opciones para extraer información desde Twitter en grandes cantidades, como por ejemplo librerías escritas en diferentes lenguajes de programación que eran creadas y mantenidas por los mismos usuarios e investigadores. Para el propósito de esta memoria se planeaba utilizar una de ellas, la librería de extracción de Tweets más usada en el lenguaje Python, *Snsrape*. Aparte de poder extraer información de Twitter, esta librería tiene más opciones de redes sociales, tales como *Facebook*, *Instagram*, *Reddit*, *Telegram*, entre otras.

Uno de los desafíos que se formó en la etapa de extracción de Tweets fue que la administración de Elon Musk comenzó a cerrar la API original que se mantenía aún en Twitter, comenzando con uno de los endpoints disponibles que utilizaba la librería *Snsrape*, esto para poder implementar una nueva versión de la API, la cual comenzaría a ser de pago y con numerosas limitaciones en comparación a su antigua versión. Uno de los grandes cambios aparte del pago, era la cantidad de Tweets que se podían manejar en la nueva API, que son números considerablemente menores. Esto dificultó la extracción masiva de Tweets que se planeaba, por lo que se tuvo que recurrir a otros métodos mientras se trataba de buscar una solución que permitiera volver a usar la API de twitter.

El método utilizado fue el Webscraping, que es una técnica que se utiliza para poder extraer información desde cualquier sitio web. Esta técnica implica el uso de programas computacionales que simulan ser un humano navegando por la página web. El lenguaje utilizado para esta solución fue Python y se utilizó Selenium que es una herramienta de pruebas de automatización de código abierto, muy conveniente para extraer información en páginas de contenido dinámico.

Con Selenium en Python se elaboró un script que simulaba ser un humano utilizando el navegador Google Chrome, donde primero se dirigía a la página de Login de Twitter y se ingresaban las credenciales de un usuario creado específicamente para este propósito. Luego se iniciaba sesión y se redirigía a la página principal del usuario, en donde el script buscaba en la barra de búsqueda los siguientes comandos:

- **'congreso de chile (senador OR senadora OR senadores OR diputado OR diputada OR diputados OR camara OR parlamentario OR parlamentarios OR congreso OR chile OR nacional) lang:es until:2019-09-02 since:2019-09-01'**: Este ejemplo de comando de búsqueda recopila todos los tweets que obligatoriamente contengan las palabras "congreso de chile" y que opcionalmente estén formados por cualquiera de las palabras que están dentro del paréntesis. Estos tweets se filtran por el idioma y la fecha indicada. Este es solo un ejemplo en una fecha, se debe considerar que el estudio contempla los meses de septiembre a noviembre del año 2019 y 2022, por lo que se aplicaron más comandos con las fechas correspondientes para cumplir con el tiempo a analizar.

Se desistió de buscar tweets que contengan nombres de parlamentarios, ya que son

155 personas en su totalidad. Aparte de ser un filtrado extenso, este podía ocasionar datos de mala calidad por alcance de nombres o apellidos en tweets que no estuvieran contextualizados con el Congreso de Chile.

- **'(carabinero OR carabineros OR carabinera OR carabineras) lang:es until:2019-09-02 since:2019-09-01'**: En el caso de la policía chilena, el comando de búsqueda recopilaba cualquier tweet que estuviera conformado por las palabras carabinero, carabineros, carabinera o carabineras. Estos tweets cumplían la condición de estar escritos en español y comprendían las mismas fechas que el comando del Congreso, por lo que se buscaron las demás fechas con el mismo comando solo alterando el filtro.

De la búsqueda de estos comandos, se capturaban los primeros 1000 tweets al día que se encontraban. La información que guardaba en una estructura de lista consistía en el nombre, nombre de usuario (@), la fecha de publicación, a quién le estaba respondiendo (si es que correspondía), la cantidad de retweets, la cantidad de likes y el tweet en sí (ver Figura 15). Al terminar de extraer los 1000 tweets, el script escribía toda la información guardada en la lista de almacenamiento en un archivo CSV.

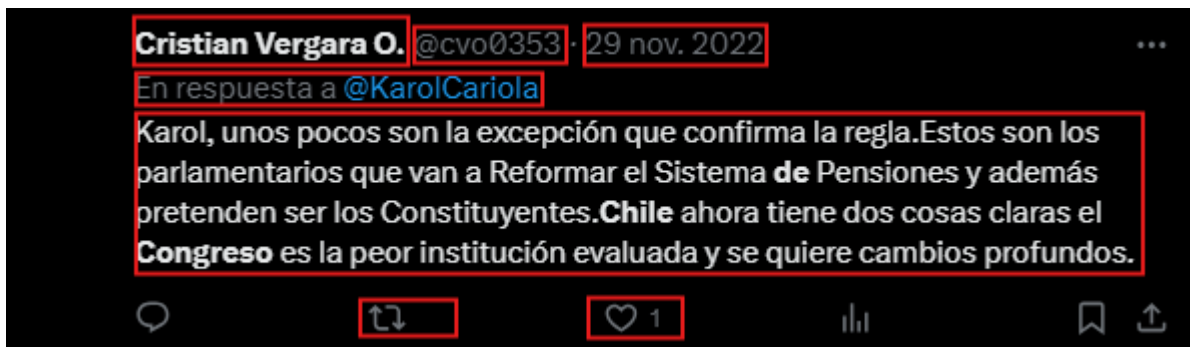


Figura 15: Datos extraídos de un Tweet por Webscraping usando Selenium
Fuente: Elaboración propia

Name	Username	Date	Responding	Retweets	Likes	Tweet
Cristian Vergara O.	@cvo0353	2022-11-29T16:58:27	"Replying to @KarolCariola"		1	Karol, unos pocos son la excepción que confirma la regla. Estos son los parlamentarios que van a Reformar el Sistema de Pensiones y además pretenden ser los Constituyentes. Chile ahora tiene dos cosas claras el Congreso es la peor institución evaluada y se quiere cambios profundos.

Figura 16: Datos extraídos de un Tweet por Webscraping usando Selenium en un CSV
Fuente: Elaboración propia

La cantidad de 1000 tweets para extraer por día se definió según dos factores: probando el tiempo de demora del script y qué cantidad de tweets podrían ser suficientes para armar un dataset por cada entidad y que sobren tweets para después poderlos clasificar. Extraer

1000 tweets de esta manera demoraba un promedio de 13 minutos si es que no habían problemas de conexión de internet o de memoria y permitía hasta 180000 tweets de cada entidad (tomando un enfoque optimista ya que habían algunos días en los que no habían 1000 tweets que cumplieran con ese filtrado y el script acabó antes), considerando 1000 tweets diarios por 6 meses: desde septiembre hasta noviembre de los años 2019 y 2022.

Con esta técnica de scraping se logró completar la extracción de los 6 meses de datos de la entidad de Carabineros en un transcurso de 9 días. Paralelamente a esta actividad, los creadores y mantenedores de Snsrape lograron encontrar una solución temporal para volver a utilizar la librería. La nueva administración de Twitter aún no eliminaba todos los endpoints por lo que pudieron utilizar los endpoints aún disponibles hasta que los cerraran. Aprovechando esta situación se logró utilizar Snsrape para poder extraer todos los tweets posibles de Carabineros y los que aún faltaban de la entidad del Congreso.

La extracción de datos utilizando Snsrape es más sencilla, el script usado es más simple y corto, a diferencia de la utilización de Selenium. Se comienza almacenando las queries de búsqueda de Twitter, tal y como las usó Selenium, pero como la extracción es más rápida con la nueva librería se considera una query por año por cada entidad, resultando en 4 queries a utilizar. Con cada query, Snsrape obtenía todos los ítem que calzaban con ese filtrado y los guardaba en una estructura de lista, para después crear un dataframe con todos los datos y convertirlo a un archivo de formato CSV, estos pasos se realizaron con las 4 queries, por lo que se elaboraron 4 archivos de valores separados por coma: 2 de cada entidad y dentro de eso, 1 de cada año.

La ventaja de poder utilizar Snsrape fue valiosa, ya que se redujeron los tiempos de extracción considerablemente y ya no era necesario tener el límite de 1000 tweets diarios, por lo que podía extraer todos los tweets que encontraba que pertenecieran al filtrado.

Entidad	Tiempo (hh:mm)	Tweets 2019	Tweets 2022	Cantidad Total de Tweets
Carabineros	07:27	586.177	154.680	740.857
Congreso	00:33	45.922	35.626	81.548

Tabla 2: Tiempos y cantidad de Tweets extraídos utilizando Snsrape
Fuente: Elaboración Propia

Los elementos del Tweet capturados por la librería fueron los siguientes:

- ID: Número identificativo incremental del tweet extraído.
- Date: Fecha y hora del tweet.
- User: Nombre de usuario (@).
- Tweet: Cuerpo del tweet.

- ReplyCount: Cantidad de respuestas al tweet.
- RetweetCount: Cantidad de retweets.
- LikeCount: Cantidad de likes.
- QuoteCount: Cantidad de citados que tuvo el tweet.
- ViewCount: Cantidad de vistas que tuvo el tweet.
- Hashtags: Lista de los hashtags incluidos en el tweet.
- Vibe: Vibra del tweet, corresponde a un tipo de estado que el usuario pudo incluir en el tweet.

Se puede ver un ejemplo de los tweets extraídos en la Tabla 3.

Ítem	Dato
ID	53
Date	2019-09-07 23:35:20+00:00
User	G_Cereceda
Tweet	"#EstadioCooperativa Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros."
ReplyCount	0
RetweetCount	2
LikeCount	2
QuoteCount	0
ViewCount	
Hashtags	['EstadioCooperativa']
Vibe	

Tabla 3: Ejemplo de Tweets extraídos utilizando Snsrape
Fuente: Elaboración Propia

Afortunadamente, se logró extraer estos Tweets dentro de la ventana de tiempo que estuvo disponible esta solución temporal, que fue de una semana aproximadamente.

3.3. Selección de los datos

Una vez que se obtuvieron los tweets, se debían seleccionar los datos y construir el dataset a utilizar por cada entidad. La construcción de estos dataset constaba de etiquetar una

porción de los datos extraídos dentro de las tres clasificaciones: positivo, neutro y negativo. La cantidad a etiquetar no podía ser fija, esto por la específica razón de que se necesitaba construir un dataset equilibrado, que tuviera cantidades parecidas de las tres etiquetas.

El procedimiento que se utilizó para etiquetar los datos fue completamente manual, con un script desarrollado en Python que facilitó la lectura de los tweets y el etiquetado de estos. Este script iba leyendo al azar los datos de los archivos CSV en los que se encontraban almacenados todos los tweets y los mostraba por consola para que la persona encargada de etiquetar ingresara un input según la clasificación que le otorgaba al tweet. Estos inputs eran '0' para el tweet considerado negativo, '1' para el neutro y '2' para el positivo. No se utilizaron los típicos números para el análisis de sentimientos (-1, 0, 1) debido a que escribir un número negativo por consola entorpecía el trabajo, de esta manera fue más ágil. A pesar de que el etiquetador consideraba como input el 0, 1 y 2 para la categorización, el script se encargaba de escribir -1, 0 y 1 en su lugar.

Para elaborar el dataset de la entidad del Congreso se etiquetaron datos hasta tener una cantidad balanceada de las tres clasificaciones, esto resultó en una clasificación de alrededor de 15000 tweets en 2 semanas utilizando algunas horas del día. De estos 15000 tweets se formó un dataset de aproximadamente 6000 tweets, que consistía en casi 2000 tweets catalogados en cada clasificación.

El dataset de Carabineros tomó más trabajo y más tiempo de etiquetado debido a la gran cantidad de datos extraídos. El tiempo que demoró elaborar el dataset fue de 7 semanas, en la que se etiquetaron cerca de 60000 tweets, esto ocupando varias horas al día, casi todos los días. De todos estos datos se logró armar un dataset de aproximadamente 30000 tweets, con alrededor de 10000 tweets de cada categorización.

3.4. Preprocesamiento de los datos

La etapa del preprocesamiento de los datos es una de las más relevantes para el correcto funcionamiento del modelado a posterior y para realizar un buen análisis de los datos, ya que se pretende dejarlos de una manera adecuada para su transformación.

Para comenzar con el preprocesamiento de los set de datos, se ha desarrollado un script detallado para realizar la limpieza de los datos, esto para poder preprocesar y normalizar el contenido de los tweets de manera sistemática. El script, implementado en Python, está conformado por una secuencia de funciones que se aplican una tras otra al tweet en sí. Cada una de estas funciones toma el output de la función anterior como input, así se puede asegurar una limpieza secuencial y organizada en distintos pasos.

El proceso de limpieza empieza con la eliminación de las direcciones web, las URLs, que están presentes en el tweet. Luego, el tweet es procesado por una función que remueve los hashtags y los emojis. Los hashtags del tweet están guardados en una lista en el campo de-

nominado con el mismo nombre. Si no se limpiaran, estos podrían interferir con el análisis, ya que introduciría ruido en los datos.

El siguiente paso consiste en la eliminación de menciones a otros usuarios. Este paso es crucial en la limpieza, ya que las menciones podrían afectar la polaridad del tweet, especialmente si estos aluden a opiniones o contextos que son diferentes a la clasificación del tweet en cuestión.

Posterior a eso, se eliminan los números, signos de puntuación y tildes. Con esto se prepara al texto para la etapa siguiente, que consiste en convertir todo el texto a minúsculas para que este sea uniforme.

Como últimas etapas de limpieza, se remueven los espacios adicionales y los saltos de línea extras, para que el texto sea una entrada coherente para el modelo. Además de esto, se eliminan las “stopwords”, que corresponden a palabras que por sí solas no tienen polaridad y no aportan información de carácter relevante, algunos ejemplos son los artículos y ciertos conectores, entre otros.

Finalmente, se realiza la lematización a las palabras, transformándolas a su forma base, tales como los verbos, que se dejan en su infinitivo, esto para facilitar un mejor análisis del texto.

Cabe destacar que en esta etapa de preprocesamiento de los datos se preprocesan todos los tweets que se van a someter al análisis de sentimientos una vez se tenga el modelo a utilizar, aparte de preprocesar también el dataset de cada entidad.

Siguiendo el ejemplo de la tabla 3, el tweet después de la limpieza se conformaría de la siguiente manera:

Después de	Resultado
Tweet original	#EstadioCooperativa Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros
Eliminar URLs	#EstadioCooperativa Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros.
Eliminar menciones	#EstadioCooperativa Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros.
Eliminar Hashtags y emojis	Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros.
Eliminar números	Un vándalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimógenos en un sector lleno de familias. Un descriterio impresentable. Con las familias se envalentonan, pero cuando tienen q enfrentar narcos se cagan enteros.
Eliminar signos de puntuación y tildes	Un vandalo subido en los techos pone en jaque Carabineros que no haya nada mejor q soltar gases lacrimogenos en un sector lleno de familias Un descriterio impresentable Con las familias se envalentonan pero cuando tienen q enfrentar narcos se cagan enteros
Convertir a minúsculas	un vandalo subido en los techos pone en jaque carabineros que no haya nada mejor q soltar gases lacrimogenos en un sector lleno de familias un descriterio impresentable con las familias se envalentonan pero cuando tienen q enfrentar narcos se cagan enteros
Eliminar espacios extra	un vandalo subido en los techos pone en jaque carabineros que no haya nada mejor q soltar gases lacrimogenos en un sector lleno de familias un descriterio impresentable con las familias se envalentonan pero cuando tienen q enfrentar narcos se cagan enteros
Eliminar stopwords y lematizar	vandalo subido techos pone jaque carabineros mejor soltar gases lacrimogenos sector lleno familias descriterio impresentable familias envalentonan enfrentar narcos cagan enteros

Tabla 4: Ejemplo de Tweet después de la limpieza de datos
Fuente: Elaboración Propia

3.5. Transformación de los datos

Para la transformación de los datos, se definió una actividad de eliminación de elementos del dato que no eran de utilidad para el estudio, con esto se reduce la dimensionalidad de los datos y el análisis se enfoca solamente en la información relevante para el estudio.

Al igual que todo el trabajo previo, se utiliza Python y con ello la biblioteca Pandas, que permite manipular información para su procesamiento.

Snsrape extrae 11 datos considerando el tweet, pero no se utilizarán todos. En un análisis exploratorio de los datos, se encontró que en la mayoría de tweets extraídos los datos de "ViewCount" y "Vibe" se encontraban vacíos. Por lo que estas dos columnas fueron las primeras candidatas a eliminar. Luego, también se decidió eliminar las columnas de "ReplyCount" y "QuoteCount" ya que para el estudio a realizar no es relevante la cantidad de citas y respuestas.

Al igual que la etapa anterior, en este paso se transforman todos los tweets de los que se disponen, incluyendo el dataset que será utilizado para entrenar y el conjunto de datos que será dispuesto como entrada al análisis de sentimientos.

Con esto, después de la transformación el tweet se conforma de la siguiente manera:

Ítem	Dato
ID	53
Date	2019-09-07 23:35:20+00:00
User	G_Cereceda
Tweet	vandalo subido techos pone jaque carabineros mejor soltar gases lacrimogenos sector lleno familias descriterio impresentable familias envalentonan enfrentar narcos cagan enteros
RetweetCount	2
LikeCount	2
Hashtags	['EstadioCooperativa']

Tabla 5: Ejemplo de Tweets extraídos utilizando Snsrape
Fuente: Elaboración Propia

3.6. Minería de datos

La cuarta etapa dentro de la metodología KDD es la minería de datos, es una de las más cruciales y en esta se realiza la extracción de información valiosa a través de diferentes algoritmos y técnicas de análisis que logran detectar patrones o relaciones entre los datos. La información que brindan estos algoritmos es información que siempre ha estado en los datos pero que se vuelve visible debido a estos.

En el caso de esta memoria, esta fase trata de entrenar un modelo o algoritmo y llegar a parámetros óptimos que sean suficientes para poder validar el modelo, para luego utilizarlo y clasificar la totalidad de los tweets que se han extraído.

Se decidió entrenar dos modelos para hacer una comparación de cuál resultó mejor y escoger ese para poder clasificar todos los tweets. Los elegidos fueron Naive Bayes y el basado en redes neuronales, BERT. El criterio de elección consistió en elegir un modelo que sea sencillo y simple y uno que sea más complejo, de esta manera podemos comparar cual se desempeña mejor en cada dataset.

El orden que se seguirá en esta sección será entrenar al modelo de Naive Bayes con los dos datasets y luego entrenar a BERT.

3.6.1. Entrenamiento del modelo basado en Naive Bayes

Para entrenar el algoritmo Naive Bayes se utilizó Google Colab, que tiene distintos entornos de ejecución. El plan utilizado es Colab Pro, específicamente el tipo de ejecución CPU, que permite utilizar entornos de ejecución altos en memoria RAM.

Las características del entorno de ejecución de Colab Pro durante el entrenamiento con Naive Bayes fueron las siguientes:

- Sistema Operativo: Linux basado en Debian.
- Procesador (CPU): El procesador utilizado por Colab Pro varía con el tiempo y la disponibilidad, pero estos siempre son CPUs con múltiples núcleos y modernos.
- Memoria RAM: El entorno de High RAM proporciona más memoria que el entorno común. Hay disponible una cantidad de 51 GB, siendo 50 GB utilizables aproximadamente.
- Almacenamiento: Hay disponible una cantidad de 225 GB, pero este almacenamiento es temporal.

El script utilizado para entrenar a Naive Bayes es corto y sencillo. Se utilizó la librería de Python, Sklearn. Como los datos fueron preprocesados y transformados con anterioridad, estos están casi listos para poder ingresar al modelo. Lo primero que se realiza es la vectorización de los tweets transformados, de esta manera el texto se transforma en vectores numéricos. Se utilizó la herramienta “CountVectorizer”, este vectorizador construye un vocabulario de todas las palabras que aparecen en el set de datos, luego cuenta cuantas veces aparece cada palabra del vocabulario que armó.

A raíz de esto se tienen dos matrices, la primera a la que llamaremos X y la segunda a la que llamaremos Y; X son los tweets vectorizados e Y son las etiquetas de tales tweets.

Después de tener las dos matrices, se dividen los datos en datos de entrenamiento y de prueba, utilizando un 20 % de los datos para probar el modelo. Por lo que se forman cuatro matrices: X_{train} , X_{test} , y_{train} , y_{test} .

Al tener las matrices listas, se invoca la variante a utilizar, para este tipo de datos es conveniente usar una variante que sea ideal para datos discretos, ya que hay que clasificar los tweets entre 3 clasificaciones: Positivo, Neutro o Negativo. Debido a esto, la variante más idónea es Multinomial Naive Bayes.

Se entrena Multinomial Naive Bayes con las matrices de entrenamiento. Todo el proceso descrito dura aproximadamente 2 minutos para cada set de datos, sumando los tiempos de cada paso.

Algorithm 1 Resumen del algoritmo de entrenamiento de un modelo Naive Bayes para análisis de sentimientos

- 1: **Importar librerías necesarias:** Importar Sklearn
 - 2: **Cargar dataset:** Leer archivo CSV usando Pandas
 - 3: **Vectorizar texto:** Usar CountVectorizer para convertir texto en matrices numéricas
 - 4: **Dividir dataset:** Usar train_test_split para dividir datos en conjuntos de entrenamiento y prueba
 - 5: **Entrenar modelo:** Crear modelo Naive Bayes usando MultinomialNB y entrenar el modelo con datos de entrenamiento
 - 6: **Evaluar modelo:** Predecir etiquetas en datos de prueba y calcular métricas y estadísticas
-

3.6.2. Entrenamiento del modelo basado en BETO

El entrenamiento del modelo basado en BETO, al igual que el modelo de Naive Bayes, se realizó en Google Colab, con el mismo plan, Colab Pro. Pero con la diferencia de que este modelo se entrenó utilizando el entorno de ejecución T4, que permite usar unidades de procesamiento gráfico (GPU) de buen rendimiento.

Así, las características del entorno de ejecución del entrenamiento del modelo basado en BETO fueron:

- Sistema Operativo: Linux basado en Debian.
- Procesador (CPU): El procesador utilizado por Colab Pro varía con el tiempo y la disponibilidad, pero estos siempre son CPUs modernos y con múltiples núcleos.
- Memoria RAM: El entorno posee una cantidad de memoria RAM de 12.7 GB utilizables.
- GPU RAM: El entorno T4 tiene disponibles una cantidad de 15 GB de procesamiento gráfico.

- Almacenamiento: La cantidad de almacenamiento aprovechable es de 180 GB aproximadamente.

Ya que el modelo basado en BETO es más complejo que un acercamiento basado en Naive Bayes por las características propias de cómo está elaborado, el script utilizado es más largo y complicado. Se utilizó el lenguaje de programación Python y variadas bibliotecas de códigos, tales como: “Transformers”, que contiene los modelos y las diferentes funciones que se deben utilizar en el entrenamiento; “Torch”, para el uso de la computación de tensores a través de procesamiento gráfico (GPU); “Sklearn”, para la división de los dataset de entrenamiento y prueba y “Pandas”, para el análisis y manejo de los datos.

En el código del entrenamiento se comienza definiendo una semilla aleatoria, para que sea utilizada por el modelo; la cantidad de clases a utilizar, que son 3, una clase por cada clasificación; se deben definir las variables “MAX_LEN” y “BATCH_SIZE” que contienen los valores del tamaño máximo de palabras por texto que leerá el modelo y el tamaño de los paquetes de tweets que ingresarán al modelo. Luego se leen los datos con Pandas y se utiliza PyTorch para el uso de GPU durante todo el entrenamiento.

Una vez realizadas todas las configuraciones iniciales, se debe importar el modelo preentrenado a utilizar, en este caso este corresponde a BETO en minúsculas: “bert-base-spanish-wwm-uncased”, para inicializar con este mismo el Tokenizador que convertirá las palabras de los tweets en tokens.

Luego, se debe crear la clase del dataset a utilizar, la cual contiene las características principales del Tweet y un necesario para la clase: el tweet en sí mismo, el sentimiento de su clasificación, el tamaño máximo de la cantidad de palabras del tweet y el tokenizador. Como única función de la clase, se crea a la codificación necesaria para que los datos sean introducidos a BETO, esta codificación se encarga principalmente de añadir los tokens de clasificación, separación y los tokens vacíos para rellenar al tamaño máximo en caso de que los tokens que resulten del tweet sean menos que este tamaño. En esta misma codificación, se le indica que no debe prestar atención a estos tokens vacíos, ya que no contienen información de utilidad.

Ya realizada la clase, se debe crear una función DataLoader, que lea los datos. Esta función toma en cuenta los datos tokenizados, el Batch Size y la cantidad de “Workers”: que son procesos que se trabajarán en paralelo, es decir, cada Worker tomará una cantidad de datos del tamaño del Batch Size y leerá estos datos paralelamente con los demás Workers. Junto a esto, se separa el dataset en datos de entrenamiento y prueba, en una proporción del 80 % y 20 % y se crean dos dataloader, uno para los datos del entrenamiento y otro para los de prueba.

Ya listos los preparativos de los datos para ingresarlos al modelo, se debe crear este último. Se define una clase de red neuronal, a la cual se le inserta el modelo preentrenado y se le agrega una capa adicional de red neuronal, correspondiente a la capa del entrenamiento que se realizará, personalizando el modelo ya preentrenado para que funcione adecuadamente con los datos seleccionados. En esta clase también se crean las funciones necesarias para la

conexión de las capas de redes neuronales entre las iteraciones del entrenamiento.

Teniendo la clase definida, se debe llevar el modelo a la GPU y proceder al entrenamiento. Se fija la cantidad de iteraciones del entrenamiento del modelo y se trabaja con el optimizador “AdamW”, que mejora la estabilidad y el rendimiento en el entrenamiento de redes neuronales en modelos grandes. Luego, se crean dos funciones, una para el entrenamiento y otra para la prueba. Estas funciones son las encargadas de orquestar todas las etapas del entrenamiento y la prueba. Hacen el paso y la carga de los datos en el modelo; cuentan las predicciones hechas, las pérdidas; fijan los parámetros requeridos y asocian el optimizador.

Finalmente, teniendo todo preparado para el entrenamiento como tal, se debe trabajar y ejecutar el entrenamiento de manera iterativa para que aprenda, esto con todos los pasos y etapas que fueron mencionadas anteriormente. En cada iteración el modelo aprende, por lo que teniendo una secuencia de iteraciones se puede asegurar que con los parámetros correctos, el modelo aprende cada vez más.

Algorithm 2 Resumen del algoritmo de entrenamiento de BETO para análisis de sentimientos

- 1: **Importar librerías necesarias:** Importar Transformers, Torch, Numpy, Sklearn, Pandas
 - 2: **Inicializar parámetros:** Establecer RANDOM_SEED, MAX_LEN, BATCH_SIZE
 - 3: **Configurar dispositivo:** Establecer el dispositivo para usar GPU
 - 4: **Cargar dataset:** Leer archivo CSV con Pandas
 - 5: **Tokenización:** Cargar tokenizer de BETO
 - 6: **Definir clase de dataset personalizado:** Preparar tweets y datos
 - 7: **Dividir dataset:** Usar train_test_split para dividir datos en entrenamiento y prueba
 - 8: **Crear DataLoaders:** Crear DataLoaders para manejar los datos de entrenamiento y prueba
 - 9: **Definir modelo de clasificación:** Extender nn.Module para construir un modelo BERT con capas adicionales de redes neuronales para la clasificación
 - 10: **Configurar optimizador:** Definir AdamW
 - 11: **Definir funciones de entrenamiento y evaluación:**
 - 12: **function** train_model
 - 13: Entrenar el modelo en un iteración
 - 14: **end function**
 - 15: **function** eval_model
 - 16: Evaluar el rendimiento en el conjunto de prueba
 - 17: **end function**
 - 18: **Entrenar el modelo:**
 - 19: **for** iteración in iteraciones **do**
 - 20: Entrenar y evaluar el modelo
 - 21: Mostrar métricas y estadísticas
 - 22: **end for**
-

El último paso de la metodología KDD se revisará con profundidad en el siguiente capítulo.

CAPÍTULO 4

ANÁLISIS DE LOS RESULTADOS Y VALIDACIÓN DE LA SOLUCIÓN

En el presente capítulo se revisarán los resultados del entrenamiento de los dos modelos elaborados y sus métricas de evaluación para su validación. Luego, se define el modelo más conveniente según estas métricas, para ser utilizado en la implementación de la clasificación de los sentimientos de los tweets recogidos. Con esto, se analiza la opinión de la sociedad sobre Carabineros y el Congreso entre los años 2019 y 2022.

4.1. Análisis de validación de los entrenamientos

A continuación se analizarán los resultados de los cuatro entrenamientos realizados y se compararán entre sí.

4.1.1. Multimonial Naive Bayes

Una vez ya completado el proceso de entrenamiento del modelo Multinomial Naive Bayes, se observó que el tiempo requerido para el entrenamiento fue notablemente breve. Esta eficiencia en el tiempo se puede atribuir en gran medida al entorno de ejecución que fue utilizado, el proporcionado por Google Colab Pro, que dispone de una alta cantidad de memoria RAM.

Dataset	Cantidad Tweets en Entrenamiento	Tiempo (mm:ss)	Exactitud (Accuracy)
Carabineros	30.000	04:29	0,664583
Congreso	5.925	00:46	0,642592

Tabla 6: Tiempos de ejecución del modelo Multimonial Naive Bayes por dataset

Fuente: Elaboración Propia

A pesar de ocupar el mismo modelo, las métricas resultantes varían entre los datasets. La diferencia entre los números no es exagerada pero si es relevante. En el caso de Carabineros de Chile, los valores de las métricas son los siguientes:

Métrica	Valor
Exactitud	0.6646
Precisión	0.6752
Recall	0.6646
F1	0.6603

Tabla 7: Métricas modelo Multinomial Naive Bayes para set de datos de Carabineros
Fuente: Elaboración propia

También, se calcularon las métricas correspondientes a las etiquetas. Estas estadísticas se evaluaron con los acercamientos micro, macro y ponderados. Esto debido a que la clasificación era multiclase y no binaria.

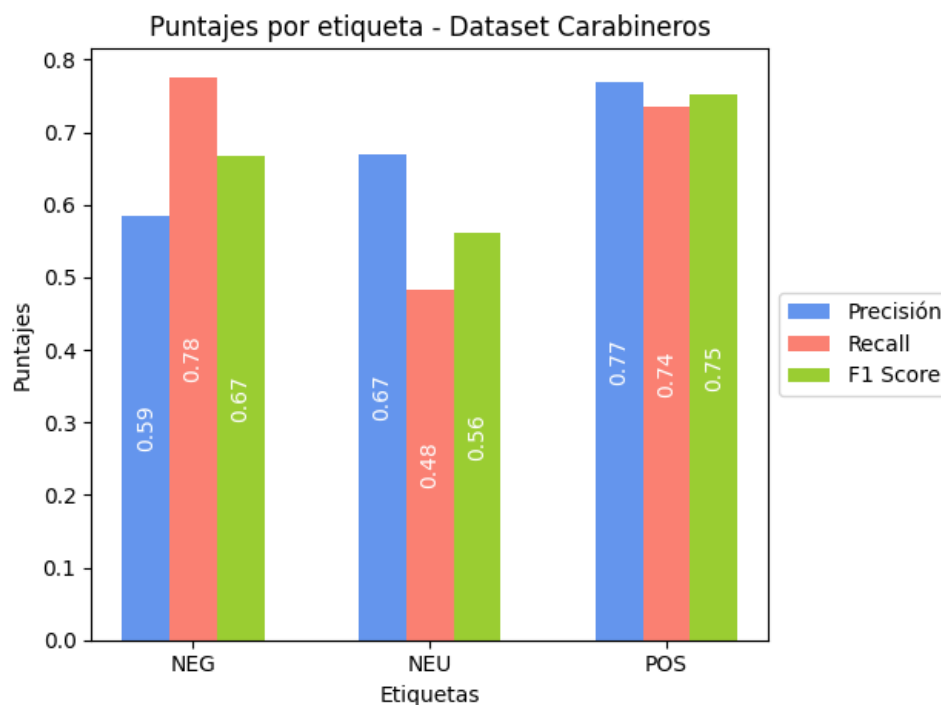


Figura 17: Métricas resultantes dataset de Carabineros por etiqueta
Fuente: Elaboración propia

Se logra evidenciar diferencias en los números entre las 3 clasificaciones, siendo la etiqueta “Positivo” (POS) la que tiene mejor desempeño y “Neutro” (NEU) la que tiene peor cantidad.

Con la matriz de confusión se puede visualizar la cantidad exacta de tweets bien y mal catalogados durante la etapa de la prueba en el entrenamiento.

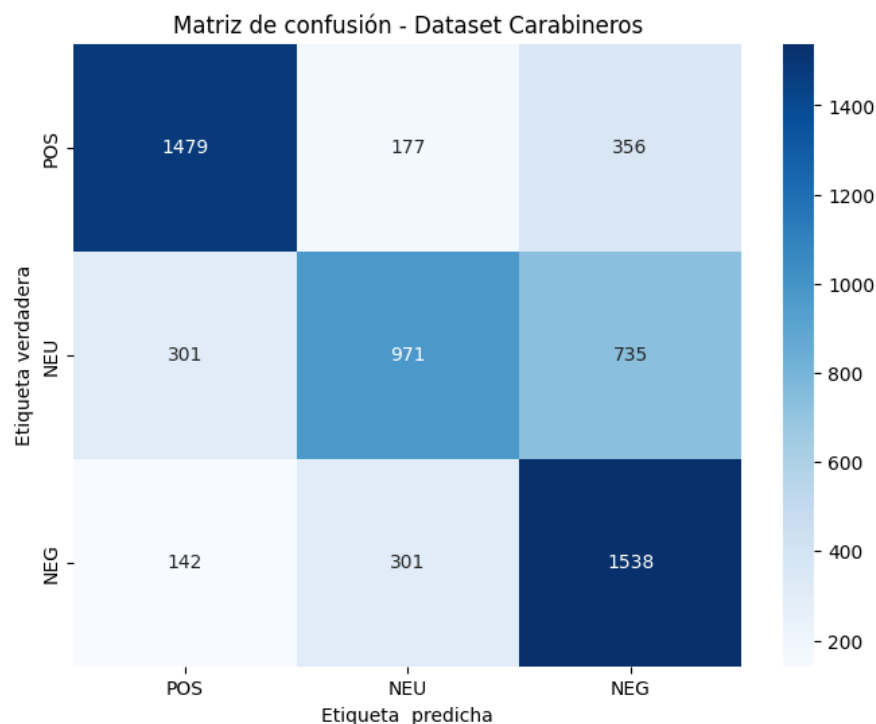


Figura 18: Matriz de confusión dataset de Carabineros
Fuente: Elaboración propia

Los valores presentes en la matriz siguen los lineamientos de los números de las métricas, ya que se demuestra que el desempeño de la clasificación de los neutros es bajo, teniendo casi la mitad de los datos de prueba mal etiquetados, es decir, una cantidad cercana a la mitad de los tweets neutros fue catalogado como falso positivo o falso negativo.

Continuando con el caso del dataset del Congreso, las métricas porcentuales son levemente más bajas.

Métrica	Valor
Exactitud	0,6426
Precisión	0,6430
Recall	0,6426
F1	0,6407

Tabla 8: Métricas modelo Multinomial Naive Bayes para set de datos del Congreso
Fuente: Elaboración propia

A su vez, las métricas de las etiquetas se calculan de la misma manera que en el dataset de Carabineros. Y estas resultan distintas al caso anterior, en donde en términos generales, el peor desempeño lo obtuvo la etiqueta “Neutro” (al igual que el dataset de Carabineros),

pero el mejor desempeño lo tuvo la clasificación “Negativa”, siendo bastante cercano a la clasificación “Positiva”.

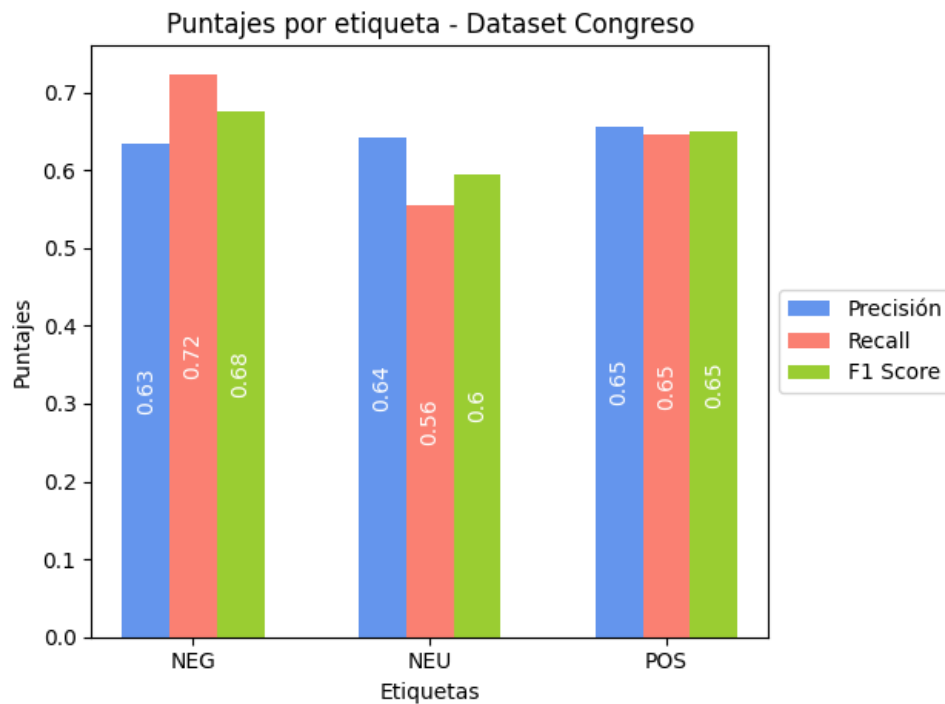


Figura 19: Métricas resultantes dataset del Congreso por etiqueta
Fuente: Elaboración propia

La matriz de confusión permanece similar al caso anterior, no teniendo mayor diferencia. Se presenta el mismo fenómeno de la clasificación de los datos Neutros.

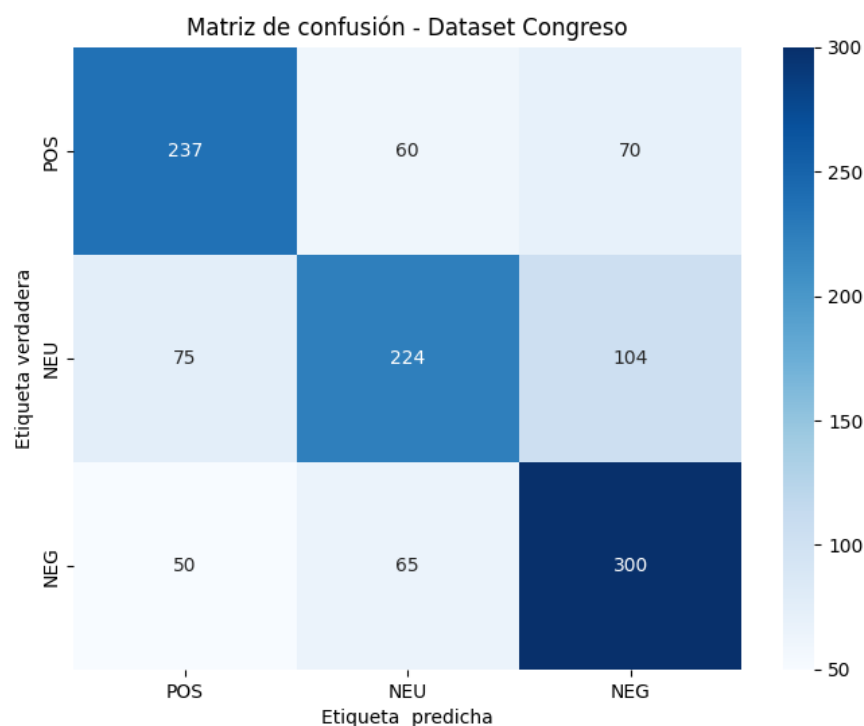


Figura 20: Matriz de confusión dataset del Congreso
Fuente: Elaboración propia

4.1.2. BETO

Como primera acción a realizar para entrenar el modelo basado en BETO, se deben definir las variables que utiliza el algoritmo.

Estos valores fueron elegidos después de realizar variados entrenamientos probando diferentes valores. El entrenamiento con el dataset de Carabineros y Congreso se realizó 14 y 26 veces respectivamente. Como el conjunto de datos del Congreso es más pequeño, se probó primero con él, ya que el entrenamiento era considerablemente más breve, pudiendo demorar entre 15 minutos a 2 horas aproximadamente según los valores de las iteraciones (Epochs) y Batch Size. A diferencia del caso de los datos de Carabineros, ya que al ser más grande el conjunto de datos, estos entrenamientos variaron de 1 hora a 3 horas aproximadamente según el valor de sus parámetros.

Después de terminar todos esos intentos, se logró dar con valores que permitían que el modelo funcionara con un buen desempeño para los dos dataset. Los valores escogidos se encuentran a continuación:

Variable	Carabineros	Congreso
Random Seed	37	15
Max Len	100	100
Batch Size	64	64
Workers	4	4
Learning Rate (AdamW)	1e-6	1e-6
Epochs	5	10

Tabla 9: Variables utilizadas para BETO
Fuente: Elaboración propia

Cabe destacar que al ser un entrenamiento iterativo, se obtienen métricas de desempeño en cada etapa. Estas son útiles para poder conocer cómo se desarrolla el entrenamiento durante su ejecución, de una forma numérica. Se diseña de esta manera para tener una noción sobre la necesidad de modificación de las variables y parámetros durante el entrenamiento, así se puede interrumpir el aprendizaje del modelo cuando se estime conveniente. Los valores de las métricas para los datos de Carabineros por cada Epoch se encuentran en la tabla 10.

Epoch	Entrenamiento		Prueba		Tiempo Ejecución (hh:mm:ss)
	Exactitud	Loss	Exactitud	Loss	
1	0,7976	0,4988	0,8517	0,3727	00:13:04
2	0,8524	0,3737	0,8598	0,3511	00:26:08
3	0,8611	0,3499	0,8632	0,3415	00:39:12
4	0,8684	0,3341	0,8663	0,3382	00:52:16
5	0,8721	0,3273	0,8660	0,3354	01:05:20

Tabla 10: Métricas por iteración del entrenamiento de datos de Carabineros
Fuente: Elaboración propia

El modelo ha tardado en entrenarse 1 hora con 5 minutos y 20 segundos, demorando 13 minutos aproximadamente por cada epoch. Ya que los valores de las métricas de las epochs demuestran un buen aprendizaje y desempeño del modelo, se procede a calcular las típicas métricas al final del entrenamiento, para así poder comparar con el modelo Multinomial Naive Bayes en el análisis que se realizará posteriormente.

Métrica	Valor
Exactitud	0,8661
Precisión	0,8663
Recall	0,8661
F1-Score	0,8657

Tabla 11: Métricas finales del entrenamiento de datos de Carabineros, modelo BETO
Fuente: Elaboración propia

Al igual que el caso del otro modelo realizado, se procedió a hacer el cálculo de las métricas por etiqueta y elaborar la matriz de confusión, las cuales se encuentran presentes en las figuras 21 y 22.

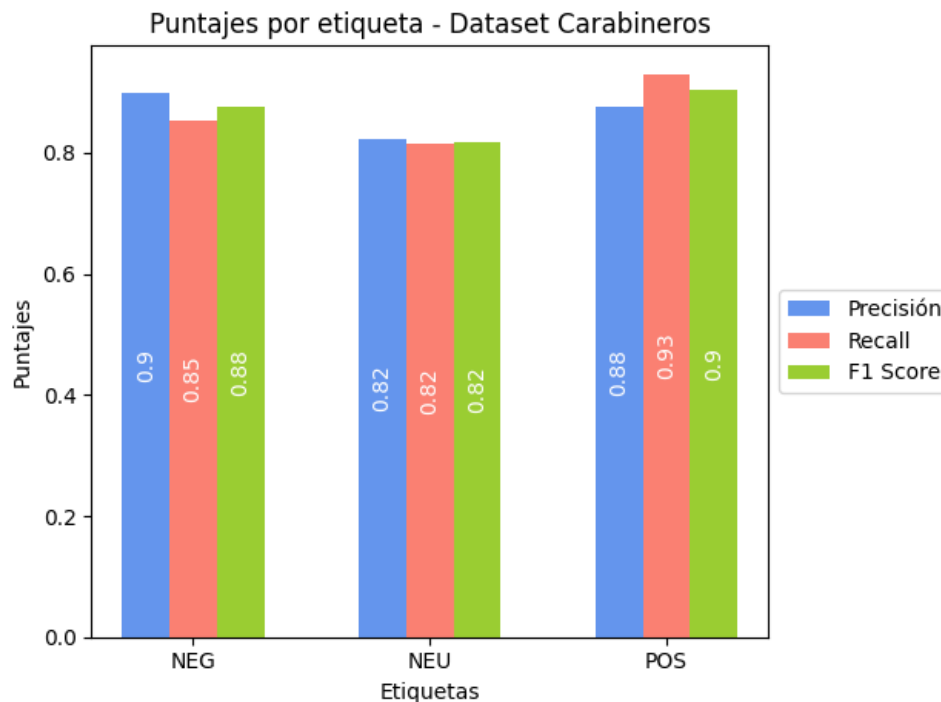


Figura 21: Métricas resultantes dataset Carabineros, modelo BETO
Fuente: Elaboración propia

La visualización recién presentada permite apreciar detalladamente el rendimiento del modelo en función de los puntajes de las métricas de clasificación. Se observa que los valores que se obtuvieron para cada una de las 3 clases son, en general, bastante satisfactorios. Se destaca notablemente el desempeño en la clasificación de la etiqueta positiva, esto evidenciado por sus puntajes altos. Todo esto indica, que el modelo tiene una buena capacidad de rendimiento para poder identificar y clasificar de manera precisa los tweets que tienen sentimientos positivos.

Por otro lado, a pesar de que los puntajes que corresponden a la clasificación neutro son más bajos en comparación a los valores obtenidos para la etiqueta positiva o negativa, los resultados siguen siendo buenos y reflejan un buen rendimiento.

Resumiendo, la visualización muestra que el modelo presenta un buen desempeño a manera general, con un rendimiento especialmente fuerte en la identificación de tweets positivos.

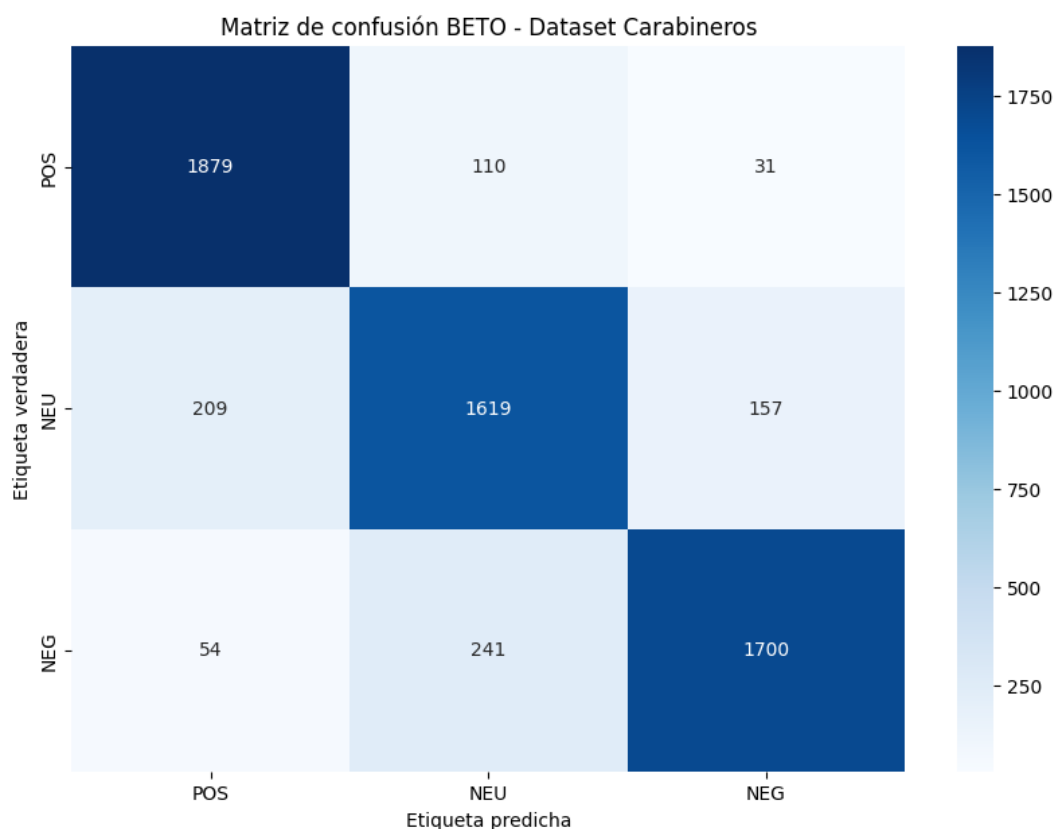


Figura 22: Matriz de confusión dataset Carabineros, modelo BETO
Fuente: Elaboración propia

La matriz de confusión confirma que la clasificación positiva es la que mejor categoriza el modelo, ya que el primer cuadrante es el que tiene el coloreado más fuerte .

Para el caso del dataset del Congreso Nacional, se precisaron de un número de 10 iteraciones para lograr alcanzar buenos valores de Exactitud y de la función de pérdida. Están levemente por debajo de los valores del dataset de Carabineros. El modelo se interrumpe a la iteración 10, ya que el valor de la función de pérdida en el set de prueba comenzaría a aumentar levemente y esto es justamente lo que se tiene que evitar. Con esta interrupción, se logra obtener unos valores equilibrados de Exactitud y Loss para poder evaluar las otras métricas de interés.

Epoch	Entrenamiento		Prueba		Tiempo Ejecución (hh:mm:ss)
	Exactitud	Loss	Exactitud	Loss	
1	0,6969	0,7203	0,7931	0,50127	00:08:53
2	0,8096	0,4760	0,8147	0,4469	00:17:46
3	0,8331	0,4230	0,8234	0,4188	00:26:39
4	0,8477	0,3883	0,8279	0,4072	00:35:32
5	0,8593	0,3643	0,8313	0,4007	0:44:25
6	0,8642	0,3527	0,8353	0,3970	00:53:18
7	0,8672	0,3422	0,8390	0,3936	01:02:11
8	0,8729	0,3310	0,8407	0,3913	01:11:04
9	0,8752	0,3292	0,8417	0,3908	01:19:57
10	0,8720	0,3280	0,8430	0,3912	01:28:50

Tabla 12: Métricas por iteración del entrenamiento de datos del Congreso
Fuente: Elaboración propia

El entrenamiento del modelo para el set de datos del Congreso se demoró 1 hora con 28 minutos y 50 segundos en ejecutarse en su totalidad. Cada epoch con su entrenamiento y prueba tardó casi 9 minutos en realizarse.

Las demás métricas correspondiente a precisión, recall y puntaje F1 resultaron bastante parecidas con valores buenos y aceptables. Todas cercanas a un valor del 84 %.

Métrica	Valor
Exactitud	0,8407
Precisión	0,8399
Recall	0,8407
F1-Score	0,8399

Tabla 13: Métricas finales del entrenamiento de datos del Congreso, modelo BETO
Fuente: Elaboración propia

Luego, se calcularon los puntajes de las métricas por clase en el set de datos del Congreso.

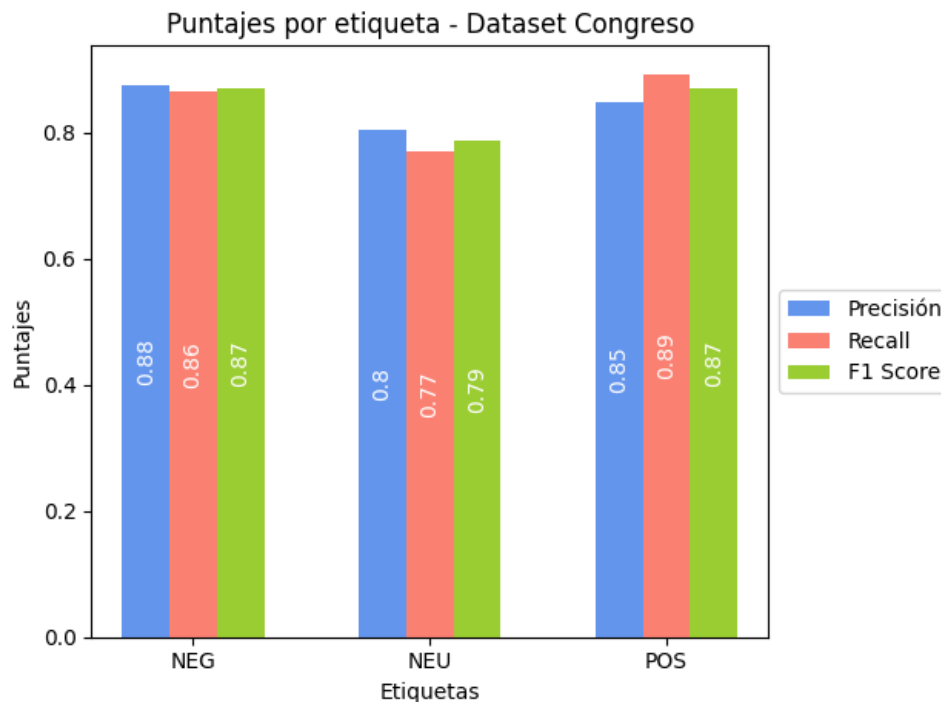


Figura 23: Métricas resultantes dataset Congreso, modelo BETO
Fuente: Elaboración propia

La visualización sugiere y confirma el buen desempeño globalmente y en cada clase. Se repite el fenómeno de la clasificación neutra, que tiene valores numéricos buenos pero por debajo de las clasificaciones positivas y negativas, y estas, a su vez tienen valores similares, implicando que el modelo se comporta similar al identificar y clasificar los tweets con sentimiento negativo y positivo.

En la matriz de confusión a continuación se puede revisar el detalle de la clasificación del set de prueba del Congreso. Se confirma que los verdaderos positivos y negativos son mejores clasificados, demostrado por el color.

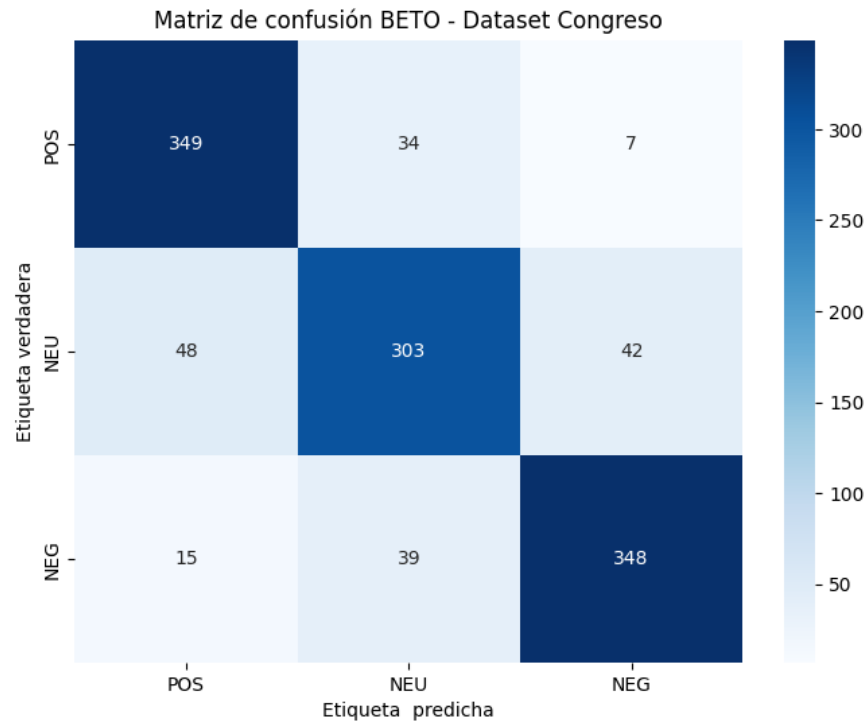


Figura 24: Matriz de confusión dataset Congreso, modelo BETO

Fuente: Elaboración propia

4.1.3. Selección de los modelos

Para la selección del modelo en cada set de datos, se compararon los valores numéricos de las métricas de clasificación. Esto implica que el modelo seleccionado presenta un mejor desempeño que el que no fue elegido, ya que sus métricas cuantitativamente son mejores.

4.1.3.1. Dataset Carabineros En el caso del set de datos de Carabineros si se compara una a una la métrica a evaluar, es notable que los valores numéricos del entrenamiento del modelo BETO son superiores a los valores de las métricas de Multinomial Naive Bayes, superándolo en un 20 % en todas las métricas.

Métrica	Valor Multinomial Naive Bayes	Valor BETO
Exactitud	0,6646	0,8661
Precisión	0,6752	0,8663
Recall	0,6646	0,8661
F1	0,6603	0,8657

Tabla 14: Comparación métricas del entrenamiento de datos de Carabineros

Fuente: Elaboración propia

Los tiempos de ejecución del entrenamiento varían enormemente entre los dos modelos, el que demora menos es Multinomial Naive Bayes. El tiempo de ejecución de BETO se explica por la naturaleza del mismo, el modelo de por sí es más complejo que un MNB: el modelo de BETO es un modelo de aprendizaje autosupervisado basado en entrenamiento de redes neuronales con inteligencia artificial.

Tiempo (hh:mm:ss) Multinomial Naive Bayes	Tiempo (hh:mm:ss) BETO
00:04:29	01:05:20

Tabla 15: Comparación tiempos de ejecución de ambos modelos. Carabineros
Fuente: Elaboración propia

A pesar de que el tiempo de ejecución de BETO sea casi 14 veces más largo que el de MNB, se escogerá el algoritmo BETO para poder clasificar todos los datos extraídos desde Twitter para el set de datos de Carabineros. Seleccionar un buen modelo es sustancial para la solución del problema planteado en el primer capítulo. Es por eso que se opta por seguir los valores de los puntajes de las métricas y desestimar el tiempo de ejecución, ya que la demora total del entrenamiento de BETO es un tiempo normal y común dentro de los problemas de este tipo.

4.1.3.2. Dataset Congreso Nacional Para el caso del dataset del Congreso Nacional, también se comparan una a una las métricas definidas en el marco conceptual. En todos los casos, los puntajes del entrenamiento BETO son superiores al entrenamiento de Multinomial Naive Bayes, casi en un 20 % en todas las métricas disponibles.

Métrica	Valor Multinomial Naive Bayes	Valor BETO
Exactitud	0,6426	0,8407
Precisión	0,6430	0,8399
Recall	0,6426	0,8407
F1	0,6407	0,8399

Tabla 16: Comparación métricas del entrenamiento de datos del Congreso
Fuente: Elaboración propia

Al igual que el caso anterior, el entrenamiento con el modelo basado en redes neuronales demoró una cantidad considerablemente mayor en comparación con el tiempo del modelo MNB. Las razones son las mismas, el modelo es mucho más complejo por sí mismo. Aparte, se debe considerar que en este dataset las Epoch fueron 10, a diferencia del dataset anterior que fueron 5. Una iteración demora un tiempo por sí sola. En este caso, cada una demoraba aproximadamente 9 minutos. Si se hubiera considerado la misma cantidad de iteraciones que en el otro dataset, el tiempo habría sido de al menos la mitad, pero el objetivo de subirlas fue mejorar los puntajes de las métricas y el valor de la función de pérdida.

Tiempo (hh:mm:ss) Multinomial Naive Bayes	Tiempo (hh:mm:ss) BETO
00:00:46	01:28:50

Tabla 17: Comparación tiempos de ejecución de ambos modelos. Congreso
Fuente: Elaboración propia

Con todos estos antecedentes revisados, el modelo de clasificación a escoger para categorizar los tweets del Congreso de Chile es BETO. También se procede a desestimar el tiempo de ejecución en favor de priorizar los mejores valores de las métricas de clasificación.

4.2. Análisis de sentimientos

En el inciso anterior se seleccionaron los dos modelos BETO creados para los dos dataset. Una vez finalizado el entrenamiento, se guarda el modelo para ambos casos.

Se creó una función para clasificar el sentimiento de los tweets. Esta función sigue todos los pasos previos que necesita el modelo para recibir inputs: se inicializa el codificador, se lleva el cómputo a la GPU de Google Colab y se comienzan a ingresar los tweets al modelo clasificador. Todos estos pasos descritos corresponden a los dos casos. Se debe considerar que el trabajo de un dataset se encuentra en un Notebook de Google Colab distinto al otro.

4.2.1. Análisis de sentimientos: Carabineros

Tal y como se mencionó en el inciso anterior, el modelo escogido para realizar la labor de clasificar la totalidad de los tweets extraídos y preprocesados es BETO.

La cantidad de tweets clasificados y el tiempo que demoró se encuentran en la siguiente tabla:

Volumen de Tweets Carabineros	Tiempo de Clasificación (hh:mm:ss)
710.857	15:55:03

Tabla 18: Tiempos de clasificación de sentimientos en dataset de Carabineros
Fuente: Elaboración propia

En este gran grupo de datos se encontraban los tweets de las dos fechas de interés, estos se separarán por fecha posteriormente para realizar análisis.

Con esta medición de tiempo, se tiene que el modelo demoró en clasificar todos los tweets casi 16 horas, sin interrupción. Con un promedio de 12,4 tweets por segundo.

Los resultados del análisis de sentimientos se encuentran en la tabla. Estos números incluyen la totalidad de tweets extraídos, es decir, los tweets del dataset y los que se dejaron para que sean inputs del modelo de clasificación. Cabe destacar que, los tweets del dataset de entrenamiento no fueron sometidos como input al modelo de clasificación.

Año Tweet	Positivo	Neutro	Negativo	Total
2019	25.646	84.520	476.011	586.177
2022	15.075	49.446	90.159	154.680

Tabla 19: Clasificación de sentimientos. Tweets Carabineros
Fuente: Elaboración propia

En la tabla 20 se pueden encontrar los valores porcentuales resultantes de la clasificación de sentimientos y en la figura 25 se encuentra una visualización para la comprensión visual de estos datos.

Año Tweet	Positivo %	Neutro %	Negativo %
2019	4,38 %	14,42 %	81,21 %
2022	9,75 %	31,97 %	58,29 %

Tabla 20: Clasificación de sentimientos en valores porcentuales. Tweets Carabineros
Fuente: Elaboración propia

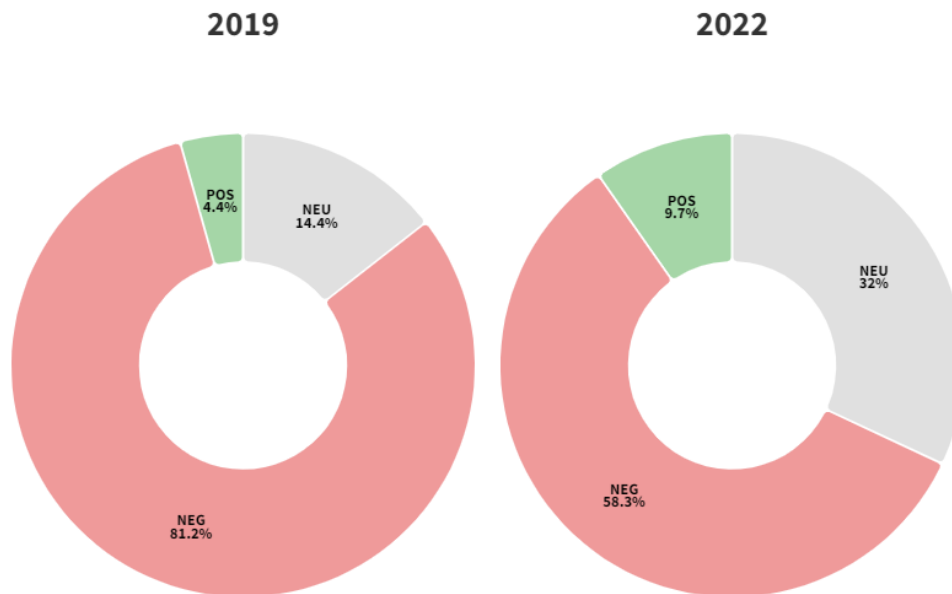


Figura 25: Análisis de sentimientos de la totalidad de Tweets extraídos sobre Carabineros
Fuente: Elaboración propia

Con la clasificación de estos tweets, se puede evidenciar que las opiniones negativas fueron las predominantes durante los dos períodos de tiempo considerados. En el 2019, de los datos considerados de septiembre a noviembre, un 81 % corresponden a opiniones de carácter negativo, con un 14,4 % de opiniones neutras y tan solo un 4,4 % de opiniones positivas.

En cambio, en el 2022 se reduce la cantidad total de opiniones extraídas desde la red social, pero las proporciones cambian. En este año, desde septiembre a noviembre, el orden de las clasificaciones de tweets se mantienen, la que predomina es la negativa, siguiéndole la neutra y por último la positiva. Pero, se logra evidenciar una considerable baja de opiniones negativas, pasando de un 81 % en 2019, a un 58,3 % en 2022. Con ello, aumentan las proporciones de las opiniones neutras y positivas, teniendo un 32 % en las neutras del 2022 y un 9,7 % en las opiniones positivas.

Es decir que, la proporción de mejora de las opiniones sobre Carabineros en Twitter, es de casi el doble de la proporción del año 2019 en los tweets neutros y positivos.

4.2.2. Análisis de sentimientos: Congreso

Para el análisis de sentimiento de los datos del Congreso, el modelo entrenado seleccionado fue BETO. El volumen de datos a analizar del Congreso es considerablemente menor al caso anterior. Por lo que la clasificación de sentimientos de estas opiniones demoró mucho menos. En esta instancia se clasificaron todos los datos extraídos que no forman parte del dataset de entrenamiento.

Volumen de Tweets Congreso	Tiempo de Clasificación (hh:mm:ss)
75.623	02:10:37

Tabla 21: Tiempos de clasificación de sentimientos en dataset del Congreso
Fuente: Elaboración propia

El modelo BETO se demoró aproximadamente 2 horas en clasificar los tweets dispuestos. Con un promedio de 9,7 tweets clasificados por segundo.

Para la estadística del análisis de sentimientos, se consideraron todos los tweets. Los que forman parte del dataset de entrenamiento y los que se ingresaron como input al modelo de clasificación BETO. Esto es para considerar la totalidad de los tweets extraídos y que el número estadístico sea el real que se encontró en Twitter.

Año Tweet	Positivo	Neutro	Negativo	Total
2019	2.690	9.742	33.490	45.922
2022	2.125	7.319	26.182	35.626

Tabla 22: Clasificación de sentimientos. Tweets Congreso
Fuente: Elaboración propia

En la tabla 23 se pueden revisar los números de tweets clasificados por etiqueta en valor porcentual y en la figura 26 se pueden visualizar estos números.

Año Tweet	Positivo %	Neutro %	Negativo %
2019	5,86 %	21,21 %	72,93 %
2022	5,96 %	20,54 %	73,49 %

Tabla 23: Clasificación de sentimientos en valores porcentuales. Tweets Congreso
Fuente: Elaboración propia

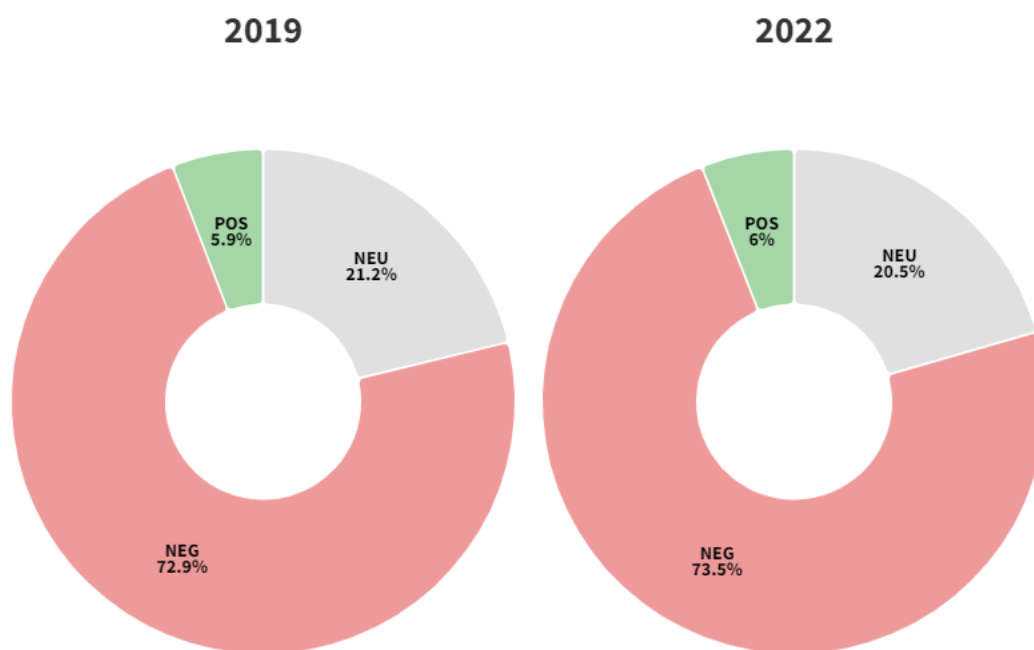


Figura 26: Análisis de sentimientos de la totalidad de Tweets extraídos sobre el Congreso
Fuente: Elaboración propia

Teniendo los tweets clasificados para la entidad del Congreso, se puede visualizar que las proporciones entre los dos años no cambian significativamente, es más, son casi las mismas. En esta extracción de datos también sucedió que hacia el año 2022 se encontraron menos Tweets que en el año 2019, pero aún así las proporciones son las mismas.

Esto significa que los sentimientos de las opiniones de los usuarios de Twitter chilenos se mantuvieron en estos 3 años. La percepción de esta entidad en la red social no muestra ninguna mejoría o empeoramiento.

4.3. Análisis de emociones

Para el análisis de emociones se consideró entrenar un modelo que pueda clasificar la emoción del Tweet, pero esta idea fue descartada a la brevedad. Entrenar un modelo para cada dataset, que ahora identificara emociones particulares era un trabajo aún más largo del que fue hacerlo con sentimientos negativos, positivos y neutros. Esto implicaba la idea de tomar los tweets extraídos y comenzar a clasificarlos manualmente, buscando emociones para luego armar un dataset equilibrado. La idea era desafiante debido al tiempo y complejidad. Armar un dataset bueno para someterlo a un modelo de entrenamiento para análisis de emociones requiere que sea equilibrado y ya que las emociones principales son alrededor de 6 o 7 (según la literatura que se considere), había que armar dos dataset con cantidades similares de tweets de cada emoción. La clasificación es más abierta por lo que, lógicamente, no se tiene la cantidad necesaria de tweets de cada emoción.

Aparte, otro problema es identificar la emoción del texto como tal. Etiquetas como negativo, neutro o positivo son más amplias y fáciles de identificar para un etiquetador, a diferencia de clasificar emociones.

Debido a estas razones se comenzó a buscar candidatos para poder analizar las emociones con modelos o bibliotecas que ya estuvieran disponibles. Para el idioma del inglés hay varias opciones, pero se necesita un algoritmo que trabaje los textos en español. Con esto, se llegó a la biblioteca Pysentimiento [Pérez *et al.*, 2023]. Esta librería está basada en Transformers y tiene variados análisis de procesamiento del lenguaje natural, tales como análisis de sentimiento, análisis de emociones, detección de discursos de odio, de ironía, entre otros.

El análisis de emociones con Pysentimiento considera 7 emociones: enojo, disgusto, miedo, alegría, otro (que según los creadores de la biblioteca es neutro), tristeza y sorpresa.

Se procedió a armar un script que leyera todos los Tweets del dataframe y llamara a la función de Pysentimiento encargada de clasificar emociones. El tiempo que demoró en clasificar los tweets por entidad se encuentra en la tabla 24

Entidad	Cantidad Tweets	Tiempo de clasificación (hh:mm:ss)
Carabineros	740.857	14:17:54
Congreso	81.548	01:52:11

Tabla 24: Tiempo de clasificación de emociones por Pysentimiento
Fuente: Elaboración propia

4.3.1. Análisis de emociones: Carabineros

Los resultados del análisis de emociones con Pysentimiento para los tweets de Carabineros se encuentran en la tabla 25. Estos se muestran separados por año y por clasificación de

sentimiento.

Sentimiento	Enojo	Disgusto	Miedo	Alegría	Otro/Neutro	Tristeza	Sorpresa
2019 NEG	62,05 %	0,03 %	0,33 %	0,20 %	32,68 %	3,96 %	0,74 %
2019 NEU	5,13 %	0 %	0 %	2,20 %	91,77 %	0,56 %	0,34 %
2019 POS	3,21 %	0 %	0 %	54,22 %	40,96 %	1,61 %	0 %
2022 NEG	65,45 %	0,06 %	0,02 %	0,36 %	29,48 %	4,41 %	0,21 %
2022 NEU	3,70 %	0 %	0 %	3,51 %	92,20 %	0,39 %	0,19 %
2022 POS	2,98 %	0 %	0 %	57,95 %	37,75 %	1,32 %	0 %

Tabla 25: Emociones en tweets de Carabineros, por año y sentimiento

Fuente: Elaboración propia

Se tiene que para las clasificaciones negativas de cada año, la emoción que predomina es el enojo, lo cual es esperable ya que esta es una emoción negativa. Luego, la que tiene el segundo mayor porcentaje es la emoción neutra y le sigue la tristeza, con menos de 5 unidades porcentuales en cada año.

Para las clasificaciones neutras de cada año, la emoción que más abunda es la neutra también, con valores que sobrepasan el 90 %. En esta clasificación no hay otra emoción que destaque.

Y por último para las clasificaciones positivas de cada año, la emoción que ocupa el primer lugar es la alegría, emoción de carácter positivo, rondando los 55 % y 58 %. La segunda más frecuente es la neutra con casi un 40 % en ambos años. Las demás emociones no pasan las 4 unidades porcentuales.

Esta revisión de valores se puede revisar en la visualización presente en la figura 27.

Emociones en Tweets de Carabineros

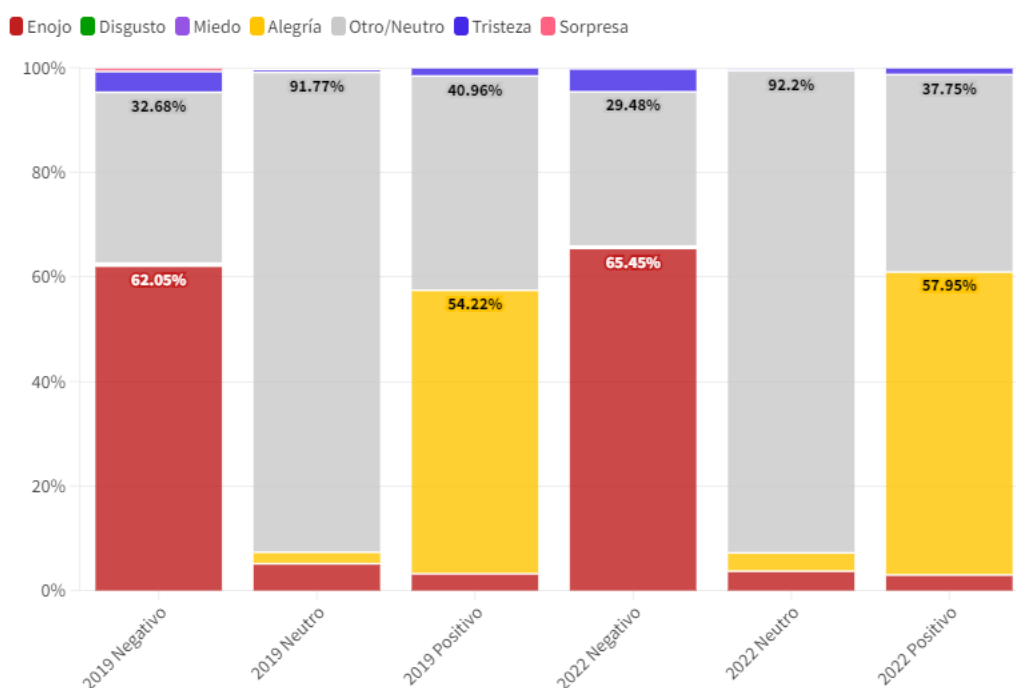


Figura 27: Visualización emociones en tweets de Carabineros

Fuente: Elaboración propia

4.3.2. Análisis de emociones: Congreso

Las emociones presentes y en qué medida están incluidas en los tweets sobre la entidad del Congreso se encuentran en la tabla 26. En esta tabla se encuentran los valores porcentuales de la presencia de las emociones en los tweets.

Sentimiento	Enojo	Disgusto	Miedo	Alegría	Otro/Neutro	Tristeza	Sorpresa
2019 NEG	61,11 %	0,08 %	0,09 %	0,35 %	35,31 %	2,85 %	0,21 %
2019 NEU	2,36 %	0 %	0 %	2,31 %	95,08 %	0,22 %	0,03 %
2019 POS	1,67 %	0 %	0 %	39,67 %	58,33 %	0,33 %	0 %
2022 NEG	61,75 %	0,15 %	0,06 %	0,36 %	34,88 %	2,55 %	0,25 %
2022 NEU	1,65 %	0 %	0 %	3,20 %	94,85 %	0,29 %	0,01 %
2022 POS	1,18 %	0 %	0 %	41,55 %	56,89 %	0,38 %	0 %

Tabla 26: Emociones en tweets del Congreso, por año y sentimiento

Fuente: Elaboración propia

Parecido al caso de las emociones de los tweets de Carabineros, se tiene que en la clasifica-

ción de sentimiento negativo predomina una emoción negativa; en la etiqueta de neutro, la emoción más frecuente es la neutra y en la categoría positiva una emoción de este mismo carácter es la que más abunda.

Para el caso de las tweets negativos, en ambos años, la emoción más presente en los tweets es el enojo, con valores entre el 61 % y 62 %. Seguido por la emoción neutra, que está presente en valores cercanos al 35 %, sin ninguna otra que resalte, con valores que son del orden de una unidad porcentual.

En los tweets neutros solo destaca la emoción neutra, con el 95 % aproximadamente en los dos años evaluados.

Por último, en los tweets positivos se da un fenómeno especial y diferente a lo esperado. La emoción que resalta en ambos años no es una de carácter positivo, es la emoción o categorización otro/neutro. En esta instancia, la alegría queda por segundo lugar con valores cercanos al 40 %.

Emociones en Tweets de Congreso

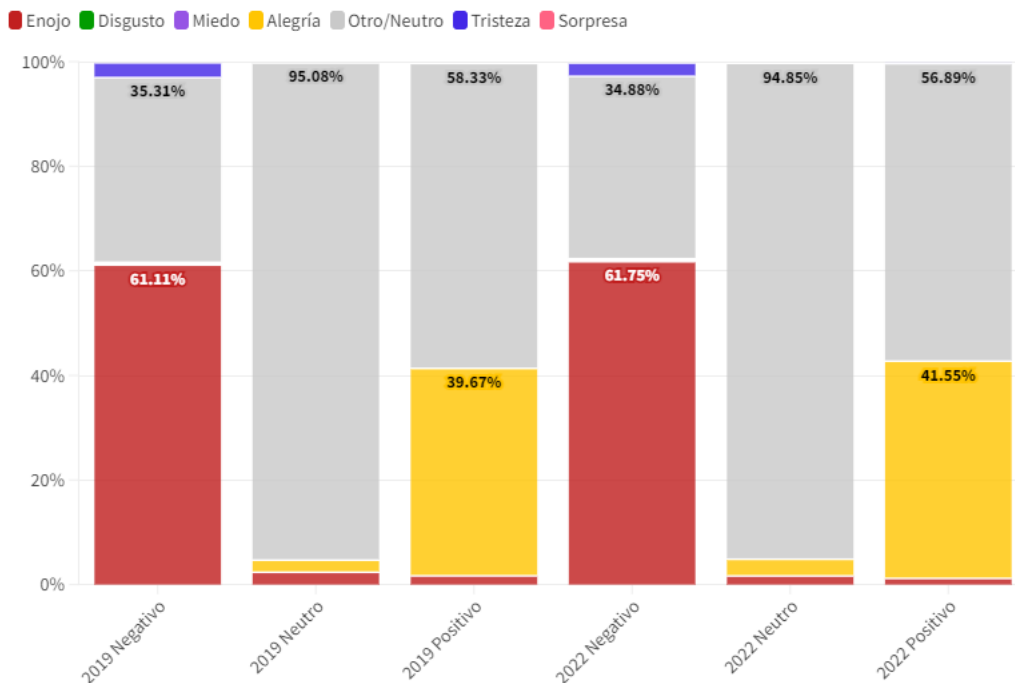


Figura 28: Visualización emociones en tweets del Congreso

Fuente: Elaboración propia

4.4. Análisis de frecuencia de palabras y hashtags

El análisis de frecuencia de palabras es una técnica básica pero muy utilizada en el procesamiento del lenguaje natural. Permite conocer cuáles son las palabras que más se repiten en un texto dado. Es sumamente útil para poder identificar qué palabras son las más comunes o importantes en un texto, esto entrega información sobre el tema y las características del texto.

El análisis de la frecuencia de los hashtags consiste en la misma idea, detectar cuáles son los hashtags más utilizados en las publicaciones de una red social.

Si bien la idea es la misma, estos análisis difieren en el cómo. Para analizar las frecuencias de las palabras se debe trabajar sobre el tweet como tal, y para analizar los hashtags, se debe examinar la lista de Hashtags que contiene cada tweet, ya que estos fueron eliminados del tweet en la limpieza de los datos.

Un tweet puede tener escrito varias veces el mismo hashtag, pero el análisis a realizar considera solo una repetición del mismo hashtag por tweet. Es decir, el tweet puede tener varios hashtags pero para el estudio se cuentan los que son distintos.

Para este análisis, se utilizó la librería Pandas y Matplotlib para las visualizaciones.

Los tweets y todos los campos de interés estaban almacenados en un dataframe de Pandas. Se elaboró un script que contaba la frecuencia de cada palabra en los tweets y los guardaba en una variable de tipo Counter, que es un tipo de diccionario que guarda las frecuencias de las variables. Así, se crean 5 Counter que ingresarán como input a Matplotlib para la generación de la visualización.

4.4.1. Frecuencia de palabras y hashtags: Carabineros

Para el caso de Carabineros en el año 2019, entre los meses de septiembre y noviembre, las palabras más utilizadas en los tweets fueron “sí”, “están”, “gente”, “violencia”, “militares”, “piñera”, “ahora”, “delincuentes”, “gobierno”, entre otras. La mayoría de palabras presente en la visualización son consideradas de sentimiento negativo según la clasificación del tweet, ya que el sentimiento que predomina en tal fecha y tal dataset, según lo revisado en los incisos anteriores, es el negativo. Lo mismo se podría decir de la emoción del enojo.

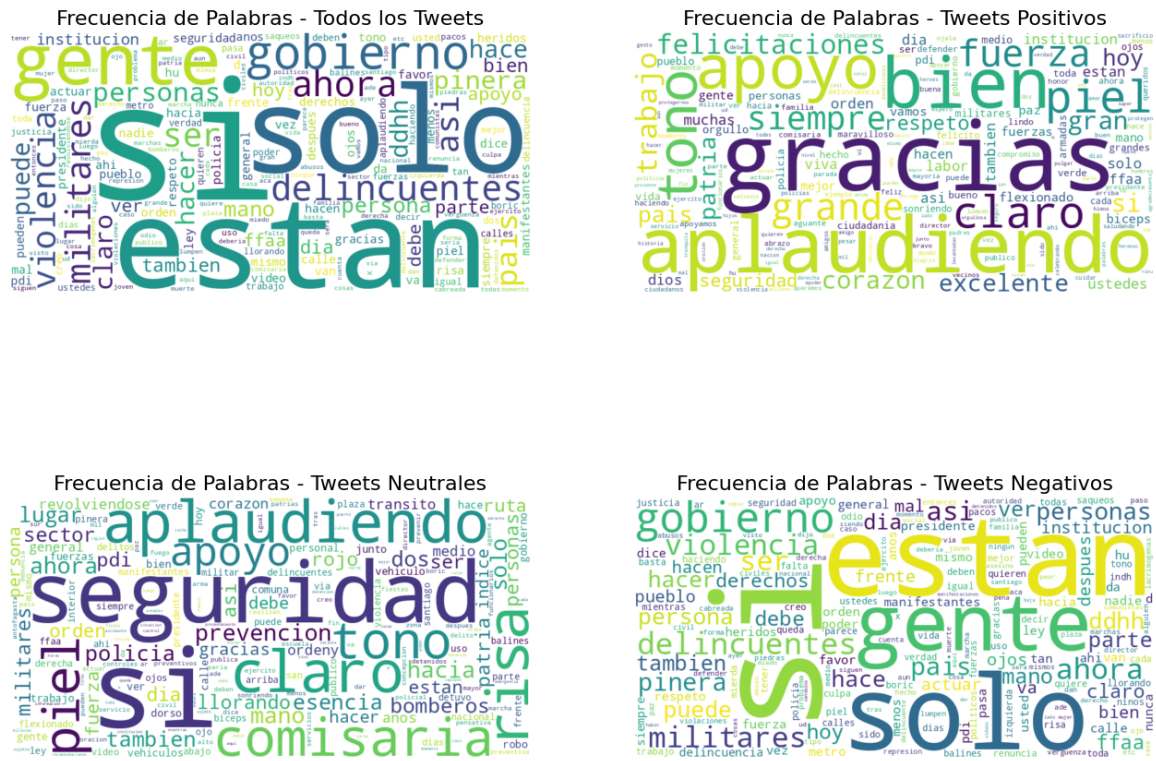


Figura 31: Palabras más frecuentes en tweets de Carabineros, sin temporalidad
Fuente: Elaboración propia

También, en ambos años algunas de las palabras más usadas fueron “piñera” y “boric”, coincidiendo con el año en el que ambas personas tomaron el rol de presidente de la república. Se podría inferir, que en su mayoría, los dos apellidos están relacionados mayoritariamente al sentimiento negativo y neutro, ya que son los dos sentimientos que más presentes están en los tweets de ambas temporalidades.

En general, en la frecuencia de palabras de todos los tweets, las palabras que más destacan corresponden a palabras que podrían ser neutras, como “gente”, “si”, “solo”. En las palabras de tamaño medio, que representan palabras que igualmente destacan pero en menor medida, se encuentran palabras que se asocian a sentimientos negativos, tales como “delincuentes” y “violencia”.

Las palabras más utilizadas en los tweets positivos de Carabineros, podrían estar relacionadas con el agradecimiento y el apoyo a su labor. Estas son “gracias”, “apoyo”, “bien” y “aplaudiendo”, todas reflejan sentimientos positivos y de gratitud.

Las palabras más presentes en los tweets neutrales son palabras que no necesariamente son preconcebidas con algún sentimiento en especial, como “seguridad” y “comisaría”.

En los tweets negativos, obviando las palabras “si”, “están”, “gente” y “solo”, se presentan en

medidas similares palabras que podrían ser preconcebidas como negativas, tales como las recientemente revisadas, “delincuentes” y “gobierno”.

Continuando con el análisis de frecuencia de hashtags, se encontraron los hashtags más usados en general, independientemente si estos eran escritos en tweets positivos, neutrales o negativos. Se consideró el sentimiento predominante el sentimiento del cual era la mayoría de los tweets relacionados al hashtag.

Un punto a considerar es que se obviaron las palabras que pertenecían al filtrado de Twitter, tales como Carabineros y Carabineras, ya que estas eran las que estaban presente en todos los tweets.

En los hashtags más populares de los meses de interés del año 2019, plena época del estallido social, se encuentran hashtags de apoyo a Carabineros, siendo los primeros 3 con más frecuencia, todos positivos: “GraciasCarabineros”, con una utilización de 2.175 veces; “YoApoyoACarabineros”, con una frecuencia de 271 veces y “GraciasCarabinerosDeChile”, con 213 apariciones. También, hay hashtags relacionados al estallido social, como lo son “PazParaChile” y “Chile Despertó”, que indican un sensación de anhelo de unidad y de manifestación, respectivamente.

Hashtag	Sentimiento Predominante	Frecuencia
GraciasCarabineros	Positivo	2.175
YoApoyoACarabineros	Positivo	271
GraciasCarabinerosDeChile	Positivo	213
Bienvenidos13	Negativo	165
gracias	Positivo	153
Temuco	Negativo	128
Chile	Negativo	101
Antofagasta	Neutro	85
PazParaChile	Positivo	71
Chile Despertó	Negativo	62

Tabla 27: Frecuencia de hashtags en tweets de Carabineros, año 2019

Fuente: Elaboración propia

Entre los hashtags más utilizados en el año 2022, se encuentran algunos que aparecieron más en tweets neutros, tales como “Antofagasta”, “Temuco”, “CooperativaContigo”, “Iquique”, “HolaChileLaRed”, que corresponden a nombres de ciudades chilenas y medios informativos. Se podría inferir que los tweets que hablaban sobre hechos objetivos, utilizaban el hashtag de su ciudad o del medio informativo de su preferencia.

Todos los hashtags positivos del 2022 están relacionados con las fuerzas de orden del país y con Carabineros. El más utilizado “ParadaMilitar2022”, hecho que se lleva a cabo el 19 de septiembre y en el que participa Carabineros; “YoApoyoACarabinerosDeChile” y “Carabine-

En los tweets neutros se habla en su mayoría del presidente y todo lo relacionado a la constitución: las palabras más frecuentes son “presidente”, “nueva”, “constitución”, “constituyente”, “nacional” y “proyecto”. Palabras que podrían estar presentes en opiniones neutrales o tweets sobre hechos objetivos.

En los tweets negativos del Congreso, se encuentran palabras tales como “sí”, que de por sí esta es de sentimiento positivo, pero que en este contexto se utiliza como supuesto o sarcasmo e ironía; “solo”; “país”; “gobierno”. Se desprende que, las entidades del Gobierno y del Congreso son vistas de manera negativa entre los opinantes del Congreso.



Figura 34: Palabras más frecuentes en tweets del Congreso, sin temporalidad

Fuente: Elaboración propia

En el análisis de la frecuencia de los hashtags, solo se encontraron hashtags negativos predominantes, en ambos años.

En el 2019, los hashtags, todos negativos, estaban relacionados con medios de comunicación: como “Bienvenidos13” y “LlegoTuHoraTVN”; con el estallido social como tal y la petición de una nueva constitución y un proceso constituyente. Todas estas opiniones y peticiones en tweets, se realizaban desde un sentimiento negativo, arrastrando la molestia que hizo que la sociedad estallara.

Hashtag	Sentimiento Predominante	Frecuencia
Chile	Negativo	462
Bienvenidos13	Negativo	117
MesaCentral	Negativo	62
NoalaAsambleConstituyente	Negativo	43
Congreso	Negativo	39
LlegoTuHoraTVN	Negativo	31
ChileDesperto	Negativo	30
AsambleaConstituyente	Negativo	21
NuevaConstitucion	Negativo	17

Tabla 29: Frecuencia de hashtags en tweets del congreso, año 2019

Fuente: Elaboración propia

Los hashtag más frecuentes del año 2022 también son todos negativos. Se destacan los que están relacionados con la constitución y el proceso constituyente. El 4 de septiembre de ese año se rechazó la primera propuesta constitucional, por lo que algunos opinantes comenzaron a pedir un nuevo proceso bajo el hashtag “NuevoProcesoConstituyenteYa”, con 67 apariciones. También se observa un hashtag contrario a este, “RechazoOtroProceso”, con las opiniones negativas de las personas que se oponían a otro proceso constituyente.

Se destacan hashtags de procesos legislativos como el “SextoRetiro” y “TPP11”, los cuales estaban presentes en tweets negativos. Para el caso del sexto retiro, este considera opiniones negativas de personas que estaban a favor y en contra de este mismo. Con el TPP11, la mayoría de tweets hacían alusión a una petición de rechazo de esta legislación.

Un caso interesante plantea el hashtag “YoEstoyConKarolCariola”, el hashtag como tal, podría ser clasificado como positivo, pero la mayoría de las opiniones en las que estuvo presente eran de carácter negativo. Lo que implica que, detractores de esta idea utilizaban el hashtag para opinar negativamente de este tópico o de la diputada como tal.

Hashtag	Sentimiento Predominante	Frecuencia
Chile	Negativo	167
NuevoProcesoConstituyenteYa	Negativo	67
BASTA	Negativo	64
ContigoCHV	Negativo	45
SextoRetiro	Negativo	35
RechazoOtroProceso	Negativo	32
TPP11	Negativo	27
YoEstoyConKarolCariola	Negativo	14

Tabla 30: Frecuencia de hashtags en tweets del congreso, año 2022

Fuente: Elaboración propia

4.5. Interacciones del Tweet y Sentimientos

En el análisis de sentimientos de opiniones de redes sociales se puede analizar si existe correlación entre el sentimiento de la opinión y algún campo que pertenezca al dato. En este caso, como se trabaja con tweets, se podría llegar a encontrar alguna relación entre las interacciones del tweet y el sentimiento de este.

La correlación indica si dos campos están relacionados o no. El valor de la correlación, r , indica lo siguiente:

- Si $r = -1$: Significa que hay una correlación perfecta negativa, es decir, que las variables son contrarias.
- Si $r = 0$: En este caso no hay relación entre las dos variables.
- Si $r = 1$: Significa que hay una correlación perfecta.

Para poder analizar este aspecto, se crea una matriz de correlación con ayuda de la biblioteca utilizada a lo largo de este trabajo, Pandas.

En este caso se debe codificar el sentimiento del tweet. Esto se realizó con los números típicos de los sentimientos; -1, para los negativos; 0, para los neutros; y 1 para los positivos. Luego, los sentimientos codificados se ingresan a la función de correlación.

4.5.1. Carabineros

En el caso de Carabineros, los resultados de la matriz de correlación indican que hay una correlación fuerte entre la cantidad de Likes y Retweets.

Pero no sucede lo mismo con la relación entre el Sentimiento y las interacciones. El valor entre el Sentimiento y RetweetCount es cercano a 0, y también sucede lo mismo con el valor de la correlación entre Sentimiento y LikeCount, es decir, las variables no están relacionadas entre sí.

Esto sugiere que, para los datos de Carabineros, el sentimiento de la opinión del tweet es trivial en la ocasión en que un usuario interactúa con el tweet. Es decir, no necesariamente porque un tweet sea negativo, positivo o neutro quiere decir que tendrá altos o bajos números de Likes o Retweets. Ocurre lo contrario con la correlación entre los Likes y los Retweets, cuando hay Likes es probable que hayan Retweets y viceversa.

	RetweetCount	LikeCount	Sentimiento
RetweetCount	1,000000	0,877627	-0,005430
LikeCount	0,877627	1,000000	0,014426
Sentimiento	-0,005430	0,014426	1,000000

Tabla 31: Correlación entre las interacciones y el sentimiento del tweet, Carabineros
Fuente: Elaboración propia

4.5.2. Congreso

Para los tweets del Congreso sucede igual que con los datos de Carabineros. Las variables LikeCount y RetweetCount si están correlacionadas, ya que su valor r es muy cercano a 1.

Los sentimientos de los tweets del Congreso no están relacionados con las interacciones que recibió la opinión.

Este hallazgo significa que, el hecho de que un tweet sea de un cierto tipo de sentimiento no determina la cantidad de Likes o Retweets. Por lo que el sentimiento no es un determinante para la interacción que recibe el tweet.

	RetweetCount	LikeCount	Sentimiento
RetweetCount	1,000000	0,904640	-0,014713
LikeCount	0,904640	1,000000	-0,005718
Sentimiento	-0,014713	-0,005718	1,000000

Tabla 32: Correlación entre las interacciones y el sentimiento del tweet, Congreso
Fuente: Elaboración propia

4.6. Interpretaciones del análisis

En el transcurso de esos 3 años entre el 2019 y el 2022 se percató de una disminución de los tweets generados por los usuarios con respecto a Carabineros y el Congreso. Esto se debe a que en plena época de estallido social los twitteros se encontraban activamente publicando y opinando sobre los hechos que estaban aconteciendo en el país, fue un hito que marcó un antes y un después en la nación. En cambio, en el 2022 no sucedió un hito tan grande como lo fue el estallido, lo más cercano fue el plebiscito del 4 de septiembre y la reciente crisis de seguridad que se instauró en el país después de la pandemia.

La época del estallido fue caótica, sucedían eventos simultáneamente y muy seguidos de tiempo, por lo que los tweets también tomaban estas características. En el 2022 no había una necesidad de estar publicando tweets tan frecuentemente, porque si bien, sucedieron eventos importantes, estos fueron sostenidos en el tiempo (proceso constituyente, plebiscito

y crisis de seguridad), lo que llevaba a opinar sobre las instituciones del país con menos frecuencia.

En ambos casos, de Carabineros y del Congreso de Chile, los tweets fueron hacia la baja en el 2022. En el caso global, tomando en cuenta todas las opiniones publicadas con respecto a las dos entidades en ambos años, se tiene que las proporciones de los sentimientos son parecidas para las dos instituciones, como se puede visualizar en la figura 35.

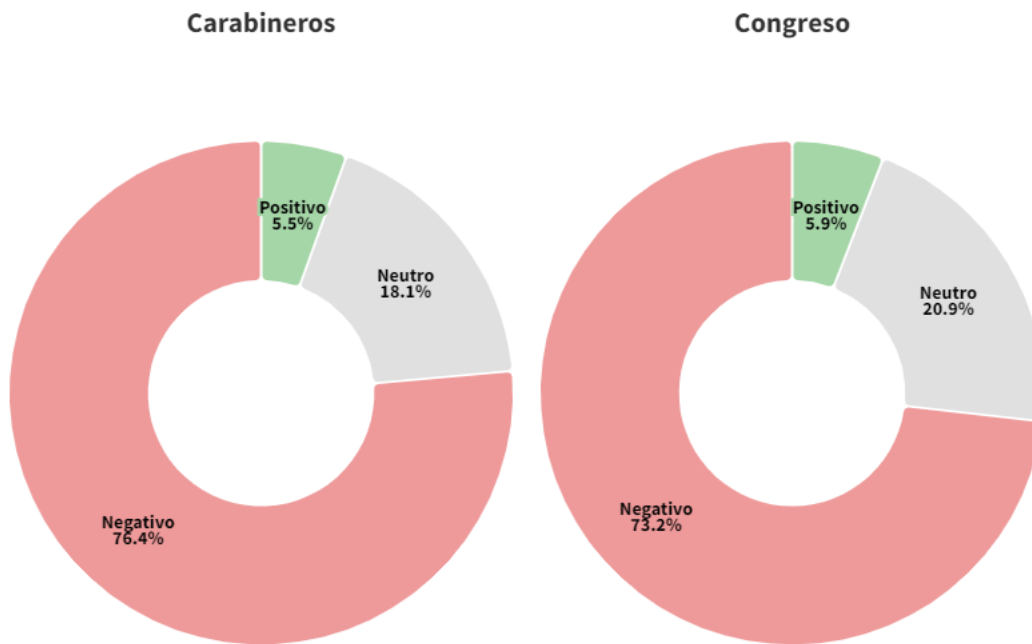


Figura 35: Sentimientos hacia Carabineros y Congreso
Fuente: Elaboración propia

Considerando todos los tweets, de ambos años, Carabineros cuenta con opiniones negativas en un 76 %, neutras en un 18 % y las positivas corresponden a un 5 %. El Congreso del país tiene opiniones negativas en un 73 %, neutras en un 20 % y positivas en un 5 %. Son instituciones que fueron percibidas similarmente, sin considerar la temporalidad el tweet.

En el caso de los sentimientos de cada entidad por año, se tiene que en el 2019, Carabineros era percibido más negativamente por las opiniones de twitter, con un valor de un 81 % de opiniones negativas, a diferencia del Congreso, que contó con un 73 %. Este último, tiene más tweets con sentimiento neutro y positivo que Carabineros, proporcionalmente hablando.

En el año 2022, Carabineros logró disminuir en gran cantidad la proporción de tweets con opiniones negativas y aumentar el sentimiento neutro y positivo, no así el Congreso. En este

año, la institución chilena más mal vista fue el Congreso, con un 74 % de tweets negativos, versus un 58 % de tweets con sentimiento negativo de Carabineros. Este análisis se puede visualizar en la figura 36.

Comparación de sentimientos de opiniones sobre Carabineros y Congreso

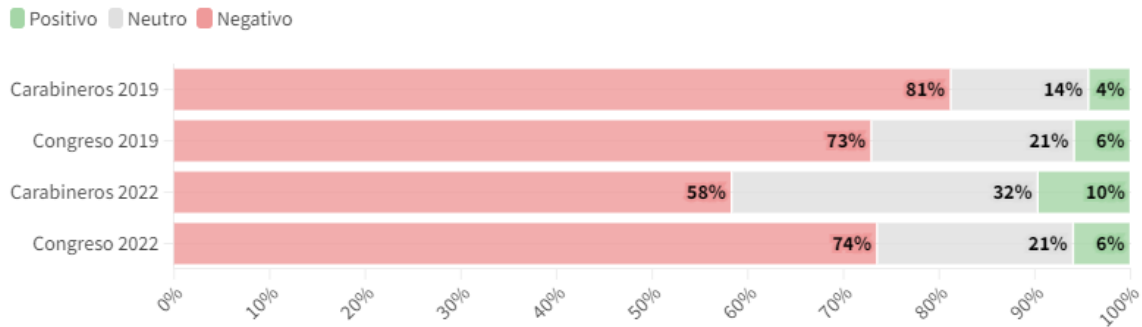


Figura 36: Sentimientos hacia Carabineros y Congreso, separados por año
Fuente: Elaboración propia

4.6.1. Interpretaciones: Carabineros

En el análisis de sentimientos, se demostró que la entidad de Carabineros logró ver una disminución en las opiniones negativas entre los años de 2019 y 2022, y algo de aumento en las opiniones neutras y positivas. Estas diferencias corresponden a que en el periodo del estallido social, en el 2019, Carabineros formó una imagen negativa entre los opinantes de Twitter, debido principalmente a las polémicas y a los métodos utilizados para instaurar el orden en un tiempo tan caótico. En cambio, en el 2022, con la reciente crisis de seguridad que se fue instaurando en el país después de la pandemia, los opinantes consideraron tomar opiniones de apoyo sobre la policía chilena, la sociedad manifestaba el apoyo, gratitud y la necesidad del actuar de Carabineros en contra la inseguridad. Esto queda demostrado con las palabras y hashtags más utilizados en cada año. Donde en el 2019, las palabras hacían alusión a la violencia perpetrada por la policía, mientras que en el 2022, estas se referían más a la seguridad y la delincuencia del país.

La emoción más presente en los tweets de Carabineros es el enojo, lo cuál es consecuente con la etiqueta de la mayoría de los tweets, que es negativa. El enojo o ira es una emoción que se asocia con lo negativo, con la pérdida de control, con la preocupación o ansiedad frente a algo o alguien. Se condice con la situación del 2019, donde los opinantes manifestaban injusticias y mal actuar por parte de Carabineros, y también, en el 2022, donde la inseguridad logra despertar esta gran preocupación de la sociedad.

Un caso particular ocurre en el caso de Carabineros, que a pesar de que la mayoría de los tweets en ambos años son negativos, los hashtags más utilizados estaban presentes mayor-

mente en tweets positivos. Esto significa que las personas que apoyan a Carabineros forman más opiniones organizadas en hashtags, que las personas que están en contra de la policía o piensan mal de ella. Estas últimas, suelen publicar sin hashtags organizados.

Por último, como resultado entre la correlación de los sentimientos de los tweets y las interacciones de estos, los opinantes de Carabineros en Twitter, dan su opinión y reciben interacciones con sus publicaciones por otros motivos, no porque el tweet publicado sea de un sentimiento negativo o positivo. Hay tweets negativos, neutros y positivos con varias interacciones, pero otros que no tienen, o que tienen valores muy bajos.

4.6.2. Interpretaciones: Congreso

Siguiendo los resultados del análisis de sentimientos. La entidad del Congreso mantuvo su reputación entre los opinantes de Twitter en los dos años considerados para el estudio. Este sostenimiento de los sentimientos de las opiniones del Congreso se debe a una molestia que la sociedad siempre ha tenido con la institución. Un parlamentario recibe muchos beneficios y altos sueldos en comparación con una persona normal en el país, lo mínimo que espera la sociedad es que se generen políticas que mejoren el bienestar del país y la calidad de la vida de las personas y es ahí en donde se encuentra el fallo, la sociedad y los opinantes no creen que los parlamentarios trabajen en favor de los intereses colectivos del pueblo. Se visualizó en el análisis de frecuencia de palabras y hashtags, que gran parte de la temática de los tweets del Congreso estaban relacionados con la petición de un proceso constituyente o con legislaciones que se encontraban activas en aquellos momentos. En general, esta percepción negativa sobre el Congreso Nacional, es un fenómeno que siempre se ha dado en la sociedad chilena y que el Parlamento como tal, no la logrado solucionar.

En los tweets del Congreso, la emoción que predomina es el enojo, hecho que se da por la naturaleza de los sentimientos de los tweets, que son negativos en su mayoría, para los dos años, 2019 y 2022. Este enojo presente en la opinión se asocia con lo negativo, con la impotencia y desconfianza que sienten los opinantes frente al Congreso, donde no se cree que hagan un trabajo serio, pensando en el país. Las personas desconfían porque creen que el parlamento siempre ha velado por intereses políticos, no comunes. Por eso se podría explicar que se mencionen tantas veces las palabras “presidente”, “gobierno”, “constitución”, “derecha” o “política”.

Un fenómeno particular que se dio en los tweets del Congreso es que los opinantes, solían hablar de otros tópicos también, en los mismos tweets, relacionados con el Congreso, pero no perteneciente a lo mismo, como lo fue todo el caso de una nueva constitución y un proceso constituyente.

Finalizando, las etiquetas de los sentimientos del tweet del Congreso no están correlacionados con la cantidad de interacciones que tienen los usuarios en sus publicaciones. Ocurre lo mismo que en el caso de las opiniones de Carabineros. Que un tweet sea de un sentimiento

no quiere decir que este tendrá más interacciones con los demás usuarios de la red social.

CAPÍTULO 5

CONCLUSIONES

En este trabajo de memoria, se ha desarrollado un exhaustivo trabajo de análisis de sentimientos sobre las opiniones de X (Ex Twitter) sobre las instituciones de Carabineros y Congreso Nacional, durante los periodos de los meses de septiembre y noviembre de los años 2019 y 2022, periodos de interés que contienen el estallido social chileno y el primer plebiscito constitucional después de la revuelta.

Con la creación de modelos basados en técnicas de aprendizaje automático, se logró realizar una comprensión de la posterior clasificación de sentimientos en los tweets recopilados, donde se consideraron métricas de evaluación y validación para la selección del modelo encargado de clasificar la totalidad de los tweets extraídos y seleccionados para el estudio.

Este trabajo se basó y estructuró según la metodología KDD, que está compuesta de cinco pasos para realizar análisis de los datos: la selección, preprocesamiento, transformación, minería de datos y como paso final, la evaluación e interpretación de los resultados.

Previo a la etapa de selección de los datos, se realizó una extracción de estos, en dónde se presentó un inmenso desafío para la realización de la memoria. Las bibliotecas candidatas para realizar la extracción en masa de tweets no funcionaban debido a los cambios en la directiva y políticas de X. La misma red social comenzó a cerrar los endpoint de los cuales se alimentaban estas bibliotecas para la extracción de los tweets. En el transcurso de este desafío se decidió crear un script de automatización que simulara ser una persona real navegando en X para la extracción de los datos. Sin embargo, esta solución no era tan eficiente y conveniente como lo era la biblioteca que se había seleccionado para este propósito, Snsrape. Mientras se trabajó con el script encargado de extraer tweets, los creadores y mantenedores de la librería Snsrape lograron encontrar una solución temporal que permitió el uso con normalidad de la biblioteca por un periodo de tiempo cercano a una semana. En esta ventana de tiempo se aprovechó de extraer todos los tweets que cumplían con los filtros de palabras, fechas e idioma definidos de las entidades de Carabineros y el Congreso.

En la fase de selección de los datos, se construyeron dos dataset equilibrados para poder entrenar a los modelos posteriormente, uno de cada entidad, clasificando manualmente sus sentimientos, durante algunas semanas. Con las fases de limpieza y transformación de los datos se dejaron los datos de los tweets de manera adecuada para que estos pudieran ser entradas de los modelos. En la minería de datos se entrenaron dos modelos por dataset: el modelo probabilístico Multinomial Naive Bayes, la versión utilizada para la clasificación multiclase; y el modelo BETO, que es basado en BERT pero para procesamiento del lenguaje natural en español.

Una vez entrenados ambos modelos en ambos datasets, se procede a la fase de evaluación de los resultados. En esta etapa de la metodología KDD, se calculan las métricas de validación

de cada uno de los modelos entrenados con los dos datasets. Los valores de estas métricas indicaron que el desempeño del algoritmo Multinomial Naive Bayes era algo deficiente, a diferencia del rendimiento del modelo BETO, que según las métricas de validación, resultó ser un buen modelo para ambos datasets, con una ventaja aproximada de un 20 % para cada métrica.

Los motivos de los regulares puntajes de validación del modelo MNB, se deben a diferentes razones y son por la naturaleza misma del modelo. Este asume independencia entre las palabras, lo cual se cumple rara vez en un texto, al ser un modelo probabilístico, esto puede llevar a un mal cálculo de las probabilidades y por lo tanto, a una mala clasificación. Naive Bayes no es capaz de correlacionar de buena manera las palabras del texto. Otra explicación es que los datos no siguieron la distribución de datos multinomial del modelo y esto afectó a las métricas de evaluación.

En ambos modelos se dio que los puntajes de las clasificaciones neutras eran algo más bajas que los puntajes de las métricas de las clasificaciones positivas y negativas, esto se explica por el simple hecho de que el sentimiento neutro es una clasificación intermedia entre una positiva y negativa, lo que lleva a los modelos a equivocarse un poco más en la categorización de este sentimiento. Aún así, la diferencia de las métricas del sentimiento neutro no eran exageradamente más bajas que las demás.

Con todo esto como antecedente, el modelo seleccionado para realizar análisis de sentimientos sobre la totalidad de los tweets que se disponían fue BETO, para ambas entidades. Una vez clasificados los tweets con BETO, estos dieron resultados sobre la percepción de la opinión de las personas de Twitter con respecto a Carabineros y el Congreso Nacional.

Después de tener los tweets clasificados se realizó un análisis con técnicas de procesamiento del lenguaje natural. Tales como análisis de emociones dentro de los tweets, frecuencia de palabras, hashtags y la correlación entre las interacciones de los tweets y la clasificación de sentimiento de este mismo.

Los resultados de la policía chilena arrojaron que las opiniones sobre la institución en Twitter eran mayoritariamente negativas en ambos años, con aproximadamente, un 81 % de tweets negativos, 14 % de tweets neutros y un 4 % de tweets positivos en el año 2019. Aún así, indican una transformación notable en la percepción pública hacia Carabineros, reflejando una disminución en la proporción de las opiniones negativas del 81 % en 2019 al 58 % en 2022 y el aumento de las neutras y positivas en este mismo año. Este cambio sugiere que los ciudadanos son capaces de reevaluar sus posturas frente a las instituciones en función de los cambios del contexto, particularmente en relación con sus preocupaciones sobre la seguridad personal. Esto manifiesta que las percepciones sociales se adaptan ante las situaciones que afectan la calidad de vida de la ciudadanía. Aunque las opiniones que predominaban en los resultados de Carabineros eran negativas, se destacó que los adherentes de la policía chilena fueron los que más usaron los hashtags a su favor, evidenciando el apoyo de una parte de los opinantes.

El Congreso como tal, mostró una estabilidad de carácter preocupante en su imagen. Las proporciones de sentimientos en los tweets se mantuvieron casi igual en los 3 años desde el 2019 al 2022, con valores cercanos al 73 % aproximadamente. Esto se puede interpretar como un indicador de desconfianza persistente en el desempeño del Congreso, lo que indica la necesidad a nivel país de reformas más profundas para abordar las inquietudes de los ciudadanos en temas de representatividad y la naturaleza del trabajo legislativo. También se destaca que el proceso constitucional, aunque no era llevado a cabo por el Congreso, terminó afectando las opiniones de las personas sobre la entidad. Esto se evidencia por el análisis de las frecuencias de las palabras y hashtags, que muestran que en gran parte de los tweets que hablaban del Congreso, también estaba presente el proceso constitucional como tópico.

Ya que el sentimiento predominante en ambos años, en las dos entidades era el negativo, se pudo evidenciar que la mayoría de las emociones eran las negativas y las neutras. Siendo el enojo la emoción desde la cual más se manifestaron todas las opiniones, generalmente hablando.

Se podría pensar que ciertas opiniones de un sentimiento podrían tener más interacciones por parte de otros usuarios en la red social, pero esto no es así en estos casos. Las opiniones negativas no obtienen interacciones porque son negativas y lo mismo ocurre con las positivas.

Finalmente, este trabajo marca una iniciativa para que las instituciones nacionales analicen por qué la gente opina lo que opina sobre las entidades del país y estudien las razones detrás de las opiniones negativas o positivas, con el objetivo de mejorar el servicio que brindan hacia el país y mejorar la percepción de la sociedad. Se destaca la importancia de la confianza pública en las instituciones gubernamentales. La mejora en la percepción de Carabineros puede servir como un modelo para que otras entidades busquen recuperar la confianza y el respaldo de la ciudadanía. Por el contrario, la falta de evolución en la evaluación y percepción del Congreso resalta la urgencia de poder implementar políticas que refuercen la transparencia con la que trabajan estas instituciones y la rendición de cuentas hacia la ciudadanía, aspectos que son completamente esenciales para poder fortalecer la relación existente entre las instituciones y los ciudadanos.

Estudiar las opiniones de las personas podría haber sido útil para entender fenómenos como lo fue el estallido social en su momento, o las propuestas constitucionales fallidas. Una sociedad que opine mayoritariamente de manera negativa sobre los entes que están a su servicio indica que hay un descontento con cómo se llevan política y socialmente los asuntos de un país.

En consecuencia de todo esto, el valor de las redes sociales en el estudio de las opiniones resalta como estas mismas y en particular X (ex Twitter) se convirtieron en herramientas sumamente valiosas para poder medir la opinión pública sobre algo en tiempo real. Estos análisis de sentimientos son provechosos para que las instituciones puedan obtener información crítica para evaluar el impacto de sus políticas y también mejorar la comunicación con la ciudadanía, creando así, un espacio más accesible y receptivo para el diálogo.

TRABAJOS FUTUROS

Durante este trabajo de memoria, se ha logrado desarrollar un análisis de sentimientos de tweets provenientes de X (ex Twitter), sobre dos instituciones públicas chilenas: Carabineros de Chile y el Congreso de Chile. A pesar del logro de los objetivos, hay diversos trabajos que podrían realizarse en un futuro sobre el procesamiento del lenguaje natural que traten de tópicos o instituciones del país, considerando el contexto social actual.

- **Actualización de los datos:** Los datos recolectados en esta memoria corresponden a años pasados: 2019 y 2022, por lo que es una buena dirección lograr tomar datos de la actualidad (2024). Con la percepción de inseguridad elevada en el país, se podría teorizar que Carabineros ha logrado aumentar aún más su buena percepción entre la sociedad chilena.
- **Procesamiento del lenguaje natural en otras redes sociales sobre instituciones chilenas:** Actualmente la extracción de información de X es limitada por todos los cambios que ha sufrido la red social en los últimos años, esto saca el análisis de X para poder realizar una ampliación hacia otras redes sociales relevantes hoy en día, como lo puede ser Tiktok o Instagram, que han aumentado considerablemente su popularidad en el país.
- **Analizar opiniones en varias redes sociales a la vez:** Otra dirección que podría enriquecer el estudio de las opiniones de los ciudadanos del país sobre las instituciones públicas es la recolección de datos de varias redes sociales y analizarlos desde otros puntos de vista. Se podrían estudiar las diferencias entre las tendencias en cada red social y el alcance que esto tiene.
- **Incluir emojis en el análisis de texto:** En los últimos años, debido al acceso de redes sociales mediante el uso de smartphones, los emojis se han integrado a la forma de comunicación textual de las personas, por lo que se incluyen en las opiniones que se emiten por redes sociales. Estos caracteres especiales podrían aportar información valiosa al análisis del texto.

REFERENCIAS BIBLIOGRÁFICAS

- [ley, 2001] (2001). *Ley Órgánica de las Bases Generales de la Administración del Estado*, volumen DFL 1.
- [Azevedo y Santos, 2008] Azevedo, A. y Santos, M. (2008). Kdd, semma and crisp-dm: A parallel overview. pp. 182–185.
- [BCN, sfa] BCN (s.fa). Cámara de diputados - formación cívica - biblioteca del congreso nacional de chile.
- [BCN, sfb] BCN (s.fb). El poder legislativo - formación cívica - biblioteca del congreso nacional de chile.
- [BCN, sfc] BCN (s.fc). Senado - formación cívica - biblioteca del congreso nacional de chile.
- [Carabineros, sf] Carabineros (s.f). Organigrama carabineros.
- [Cañete et al., 2020] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., y Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. En *PML4DC at ICLR 2020*.
- [ChileAtiende, sf] ChileAtiende (s.f). Chileatiende - carabineros de chile.
- [Diputados, sf] Diputados, C. (s.f). Cámara de diputadas y diputados.
- [Gonzalez, 2020] Gonzalez, L. (2020). Diferencia entre aprendizaje supervisado y no supervisado.
- [Khurana et al., 2023] Khurana, D., Koli, A., Khatter, K., y Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.
- [KumarI, 2023] KumarI, A. (2023). Micro-average, macro-average, weighting: Precision, recall, f1-score.
- [Latha, 2019] Latha, I. (2019). Sentiment analysis tool using machine learning algorithms. pp. 14791–14794.
- [Mahesh, 2019] Mahesh, B. (2019). Machine learning algorithms - a review. pp. 381–386.
- [Medhat et al., 2014] Medhat, W., Hassan, A., y Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.
- [Medina, 2022] Medina, K. R. (2022). Situación digital de chile en 2021/2022.
- [Orchard et al., 2022] Orchard, M., Gerber, M., Bahamondes, J., Figueiredo, A., y Sáez, L. (2022). ¿cómo se relaciona usted con carabineros?

- [Pérez *et al.*, 2023] Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., y Martínez, M. V. (2023). pysentimiento: A python toolkit for opinion mining and social nlp tasks.
- [Ramos y Freire, 2021] Ramos, G. y Freire, D. (2021). Análisis de sentimientos de datos de redes sociales usando técnicas de machine learning. Tesis de máster.
- [Rosenbrock *et al.*, 2021] Rosenbrock, G., Trossero, S., y Pascal, A. (2021). Técnicas de análisis de sentimientos aplicadas a la valoración de opiniones en el lenguaje español.
- [Roucek, 2016] Roucek, J. S. (2016). La sociología de la opinión pública. *Revista Mexicana de Opinión Pública*, 21:151–163.
- [Sailunaz y Alhajj, 2019] Sailunaz, K. y Alhajj, R. (2019). Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, 36:101003.
- [Sanhueza *et al.*, 2021] Sanhueza, R., Asencio, G., y Fernandez, G. (2021). *Chile de Hoy y del Mañana*. EDITORIAL ACADEMICA ESPAN.
- [Senado, sf] Senado (s.f). Senadores.
- [SII, sf] SII (s.f). La administración pública y tu profesión.
- [T.K. *et al.*, 2021] T.K., B., Annavarapu, C. S. R., y Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395.
- [Vaswani *et al.*, 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., y Polosukhin, I. (2017). Attention is all you need. En Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., y Garnett, R., editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.
- [Wankhade *et al.*, 2022] Wankhade, M., Rao, A. C. S., y Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.