

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

INGENIERÍA EN SISTEMAS DE INFORMACIÓN

Trabajo de Titulación como requisito previo para la obtención del título de Ingeniero en Sistemas de Información

ANÁLISIS DE SENTIMIENTOS EN REDES SOCIALES EN TORNO A CONTENIDOS DE VIOLENCIA INSTITUCIONAL

Autor: Antonio David Proaño Novillo

Director/a – Tutor/a: PhD Henry Nelson Roa Marin

Quito, 13 de agosto 2026

DEDICATORIA

El siguiente trabajo se lo dedico a todas aquellas personas que me han acompañado en este proceso de grado, especialmente a mi tutor de trabajo de titulación, el Dr. Henry Roa por su acompañamiento y aporte en el trabajo, a mi familia que me ha acompañado y me ha motivado a dar lo mejor de mí, y en especial a mi padre, el Dr. Freddy Proaño, quien fue la principal persona que estuvo al tanto de mi desarrollo profesional dentro y fuera de la universidad.

AGRADECIMIENTOS

Antes que nada, agradezco a Dios por no solo darme la vida, sino la oportunidad de formar parte de este viaje que ha sido mi carrera.

Quiero agradecer a toda mi familia por ser mi soporte en todo este proceso, mi padre, el Dr. Freddy Proaño por ser la persona que me alentó a entrar en la carrera de ingeniería en sistemas de la información y quien me presentó a mi segunda casa, la Pontificia Universidad Católica del Ecuador. A mi madre, Tatiana Novillo, quien me ha acompañado en todo el proceso, ha visto mis fortalezas y me ha enseñado a abrazarlas y mis debilidades para aprender de ellas y ser mejor. A mi hermano, el Ing. Santiago Proaño, quien ha sido mi amigo y compañero de carrera, acompañándome en todo momento y siempre estando ahí para mí. A mi hermana, Cristina Proaño, quien me ha apoyado desde el inicio de mi carrera y estoy seguro que lo seguirá haciendo toda mi vida. A mis tíos Mercedes Proaño y Miguel Zúñiga, quienes con su ayuda y con su experiencia previa me inspiraron a seguir este camino.

Por otro lado, me gustaría agradecer a las personas que me han guiado en este trabajo, especialmente mi tutor, el Dr. Henry Roa, quien con su conocimiento ha llevado a la creación de este trabajo y ha buscado excelencia en el mismo, gracias a su tutoría mis ideas pudieron ser plasmadas en este trabajo. Al Ing. Charles Escobar, con quien empecé a realizar este trabajo, dando forma a lo que en un futuro acabaría con el Dr. Roa. Al Dr. Salgado, quien me ha dado la oportunidad de trabajar varias veces a su lado. Al Ing. Miguel Ortiz, al Ing. Sebastián Tamayo, al

Ing. Edison Mora, al Dr. Rafael Melgarejo y a la Dra. Susana Masapanta, por todos los aportes y retroalimentación respectiva con la que fueron moldeando este trabajo final, y, en sí, a todos los ingenieros que me dieron una gran formación académica, gracias por su excelencia.

A mis compañeros de carrera, por estar a mi lado en todo el proceso, especialmente a mis compañeros Alfonso Chafla, Sebastián Jácome, Pamela Chacón y Omar Lozada, quienes con sus grandes habilidades me mostraron ser capaces de dar soluciones e ideas innovadoras, cada uno destacando en sus propias áreas con sus fortalezas únicas, mostrando más que nada que además de ser buenos profesionales futuros han sido buenos amigos y buenos seres humanos.

Por último, me gustaría agradecer al personal administrativo que me acogió en todos mis procesos en la universidad, especialmente a quienes considero mis jefas y guías, a Melisa Pazmiño y a Ana Lucía Andrade, gracias por ver grandeza en mí y por impulsarme a darlo todo en mis trabajos.

A todas las personas mencionadas anteriormente y muchas otras más que tambien estuvieron presentes en esta etapa, mil gracias por sus aportes y su compañía. Que la fuerza los acompañe.

RESUMEN

La violencia institucional, una de las problemáticas sociales con mayor relevancia en la actualidad, genera diariamente miles de opiniones en medios digitales; sin embargo, la falta de herramientas automatizadas para organizar y analizar esta información dificulta su estudio. Ante esto, en el siguiente trabajo investigativo se desarrolla un modelo adecuado para el análisis de sentimientos en Python sobre la plataforma Google Colab, con el propósito de clasificar las percepciones sociales expresadas en comentarios de Youtube sobre diferentes casos de violencia institucional, clasificando a estos comentarios en positivos, negativos y neutros. Para ello se recopilaron más de 30 mil comentarios entre 20 videos distintos de canales informativos, educativos y de opinión, los cuales fueron sometidos a un procesamiento adecuado y etiquetados. Posteriormente se entrenaron diversos modelos de entrenamiento clásicos y un modelo LLM para su comparativa, incluyendo diversas maneras de trabajar con la data. Una vez obtenidos sus resultados se generan visualizaciones dinámicas e interactivas del conjunto de datos tanto del rendimiento del modelo seleccionado como del conjunto de datos en general por medio de la herramienta Power BI, revelando percepciones predominantes de los usuarios frente a la violencia institucional, de esta manera aportando con una herramienta útil para la comprensión del impacto social de este fenómeno y para el análisis académico de la opinión pública en entornos digitales.

ABSTRACT

Institutional violence, one of the most socially relevant issues today, gives rise to thousands of opinions daily across digital media; yet, the absence of automated tools capable of organizing and analysing this information makes a considerable challenge to its study. In response to this, the following research work presents the development of an adequate sentiment analysis model built in Python on Google Colab platform, aimed at classifying the social perceptions present in Youtube comments regarding various cases of institutional violence, categorizing them as positive, negative and neutral. To make this, over 30 thousand comments were gathered from 20 different videos from informational, educational and opinion-based channels, all of which went through pre-processing and labelling. Subsequently, several classical machine learning along side a LLM model were trained for comparison, including different approaches on handling the data. Once the results were obtained, dynamic and interactive visualizations were produced covering both the performance of the selected model and the dataset as a whole through the tool Power BI, revealing predominant perceptions of users regarding institutional violence, thus delivering valuable resources for understanding the social impact of this phenomenon and for the academic analysis of public opinion within digital environments.

ÍNDICE

DEDICATORIA I

AGRADECIMIENTOS III

RESUMEN V

ABSTRACT V

ÍNDICE VI

TABLA DE ILUSTRACIONES X

TABLA DE TABLAS XII

1. CAPÍTULO I: INTRODUCCIÓN 1

1.1. Planteamiento del problema 1

1.2. Justificación 2

1.3. Objetivos 2

1.3.1. Objetivo General 2

1.3.2. Objetivos Específicos 3

1.4. Alcance 3

2. CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA 4

2.1. Violencia institucional 4

2.1.1. Concepto de violencia institucional 4

2.1.2. Tipos de violencia institucional 5

2.1.3. Impacto social de la violencia institucional 7

2.2. Lenguaje de Programación 8

2.2.1. Python 9

2.2.2. Librerías 9
2.3. Entorno de Desarrollo 12
2.3.1. Google Colab 12
2.3.2. Power BI 13
2.4. Análisis de sentimientos 14
2.4.1. Aplicación y clasificación de sentimientos 15
2.5. Procesamiento de Lenguaje Natural 15
2.5.1. Normalización y limpieza de texto 16
2.5.2. Representación vectorial de texto 16
2.5.3. TF-IDF y Fine-tuning 17
2.6. Aprendizaje automático 17
2.6.1. Modelos supervisados para clasificación de texto 18
2.6.2. Submuestreo, sobremuestreo y sobreajuste 20
3. CAPÍTULO III: METODOLOGÍA DEL ANÁLISIS 22
3.1. Enfoque metodológico de la investigación 22
3.2. Fuente de datos 23
3.2.1. Comentarios de la red social Youtube 24
3.2.2. Usos de APIs para la extracción 26
3.3. Metodología CRISP-DM 26
3.3.1. Fase 1: Entendimiento del negocio o caso de estudio 28

3.3.2. Fase 2: Entendimiento de los datos 29
3.3.3. Fase 3: Preparación y trata de los datos 30
3.3.4. Fase 4: Modelado 30
3.3.5. Fase 5: Evaluación 31
3.3.6. Fase 6: Despliegue o implementación 31
4. CAPÍTULO IV: DESARROLLO DEL ANÁLISIS DE SENTIMIENTOS 32
4.1. Fase 1: Preparación del Entorno de Desarrollo y del plan 32
4.2. Fase 2: Preparación y limpieza de datos 33
4.2.1. Fase 3: Normalización de texto 35
4.2.2. Eliminación de duplicados y nulos 36
4.2.3. Etiquetado de sentimientos 37
4.2.4. *Undersampling* del conjunto de datos 39
4.2.5. Nubes de palabras y bigramas 40
4.3. Fase 4: Modelado de análisis de sentimientos 42
4.3.1. Selección de modelos de inteligencia artificial 42
4.3.2. Entrenamiento de modelos 44
4.3.3. Modelado con BERT usando pysentimiento 50
4.4. Fase 5: Evaluación de resultados 54
4.4.1. Comparación de modelos y selección del modelo final 54
4.4.2. Comparación con revisión manual, LLMs y pysentimiento 58

4.5. Fase 6: Estadísticas dinámicas del modelo final de predicción 61

4.5.1. Mapeo y gráficos de comentarios 61

4.5.2. Visualización con Power BI..... 66

5. CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES 69

5.1. Conclusiones 69

5.2. Recomendaciones..... 71

6. BIBLIOGRAFÍA 73

7. ANEXOS 78

Anexo A. Dataset sin procesamiento en Kaggle 78

Anexo B. Datasets clasificación Manual y de Gemini en Kaggle 79

Anexo C. Documento CRISP-DM en Google Colab..... 79

Anexo D. Visualización de gráficos en Power BI..... 79

TABLA DE ILUSTRACIONES

Figura 1 Ciclo Fases CRISP-DM..... 28

Figura 2 Funcionamiento youtube-comment-downloader 34
Figura 3 Revisión Dataset unido de comentarios 35
Figura 4 Dataframe comentarios_limpios con normalización de texto 36
Figura 5 Verificación y eliminación de duplicados y nulos 37
Figura 6 Clasificación de sentimientos sin balanceo 39
Figura 7 Clasificación de sentimientos balanceado 40
Figura 8 Nube de palabras general previos al entrenamiento 41
Figura 9 Bigramas generales previos al entrenamiento 42
Figura 10 TF-IDF y separación en train y test 43
Figura 11 Matriz de confusión árboles de decisión 45
Figura 12 Matriz de confusión Random Forest 47
Figura 13 Matriz de confusión SVM 48
Figura 14 Matriz de confusión Regresión Logística 50
Figura 15 Tokenización con BERT 51
Figura 16 Épocas modelo BERT 52
Figura 17 Matriz de confusión BERT 54
Figura 18 Grid Comparativo de Modelos 56
Figura 19 Distribución de sentimientos BERT+pysentimiento 57
Figura 20 Prueba datos sintéticos BERT+pysentimiento 58
Figura 21 Comparación pysentimiento vs Manual vs Gemini 60
Figura 22 Comparación pysentimiento vs BERT+pysentimiento 61
Figura 23 Nube de palabras comentarios positivos BERT-pysentimiento 63
Figura 24 Nube de palabras comentarios negativos BERT-pysentimiento 64

Figura 25 Nube de palabras comentarios neutros BERT-pysentimiento 65

Figura 26 Dashboard de visualización del rendimiento de los modelos evaluados 67

Figura 27 Dashboard de visualización del conjunto de datos 68

TABLA DE TABLAS

Tabla 1 Fuente de datos de videos 24

Tabla 2 Clasificación pysentimiento 38

Tabla 3 Métricas Árboles de decisión 45

Tabla 4 Métricas Random Forest 46

Tabla 5 Métricas SVM 48

Tabla 6 Métricas Regresión Logística 49

Tabla 7 Métricas BERT 53

Tabla 8 Métricas modelos 55

Tabla 9 Clasificación BERT+pysentimiento 57

1. CAPÍTULO I: INTRODUCCIÓN

1.1. Planteamiento del problema

Las personas en redes sociales generan información con diversas percepciones sobre la violencia institucional, pero la falta de estructura de los datos y de herramientas automatizadas para comprender u organizar estos sentimientos expresados constituye un problema. Es por esto que se plantea en este trabajo realizar un análisis de sentimientos que clasifique los comentarios para facilitar su comprensión y organización.

En función de esta problemática se plantea la siguiente pregunta principal de investigación:

- ¿De qué manera el modelado de un análisis de sentimientos puede ayudar a identificar y clasificar las percepciones sobre la violencia institucional a partir de datos recolectados?

Y las siguientes preguntas secundarias:

- ¿Qué tipo de sentimientos predominan en los comentarios relacionados a la violencia institucional?
- ¿Qué tan consistente es la clasificación de sentimientos al analizar datos de diversos orígenes?

1.2. Justificación

La violencia institucional es una problemática social relevante en nuestro día a día que influye en que en diversas redes sociales y en espacios de conversación, las personas den sus opiniones, testimonios y muestren emociones acerca de este tema y de cómo afecta a la sociedad. Por esta razón, obtener estas opiniones y comentarios para realizar un análisis de sentimientos mediante modelado con inteligencia artificial permite identificar patrones emocionales predominantes. Este proyecto busca aplicar dicho modelo para así aportar al análisis social del fenómeno de la violencia institucional y para entrenar algoritmos con este propósito.

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar un modelo de análisis de sentimientos que permita clasificar las percepciones sobre la violencia institucional a partir de los datos recopilados.

1.3.2. Objetivos Específicos

1. Recopilar datos textuales a través por medio de redes sociales.

2. Preparar los datos para su análisis y uso.

3. Comparar y evaluar diferentes modelos de inteligencia artificial en función del mejor desempeño en el análisis de sentimientos.

1.4. Alcance

Este proyecto de titulación tiene como alcance la recolección, el procesamiento y el análisis de los datos textuales relacionados únicamente con la violencia institucional, obtenidos de comentarios en redes sociales, en este caso de "Youtube". Se desarrollará un modelo de análisis de sentimientos en "Python" en "Google Colab", presentando el trabajo de investigación con la metodología "CRISP-DM". El proyecto contempla la comparación entre varios modelos de inteligencia artificial (modelos de clasificación) y la entrega de un modelo final, evaluado y entrenado para la predicción de sentimientos, junto con la visualización y las estadísticas de los resultados del análisis de sentimientos. No se incluirá la implementación en un entorno productivo ni la integración con algún sistema externo, sino una visualización de los resultados en *dashboards* mediante "Power BI".

2. CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA

2.1. Violencia institucional

2.1.1. Concepto de violencia institucional

Según Ámbito Jurídico (2023), la violencia institucional es un fenómeno ampliamente conocido, que comprende todas las conductas, ya sea por acción u omisión, de agentes pertenecientes a instituciones públicas o privadas que terminan vulnerando con agresiones efectivas los derechos de las personas. Este tipo de violencia no siempre es manifestada de forma directa o evidente, esta tambien se puede expresar por medio de prácticas normalizadas en nuestra sociedad actual como la negligencia, el trato desigual e incluso vulnerando a personas o entidades las cuales no necesariamente se encuentran indefensas. En este sentido, la violencia institucional no se limita únicamente a un solo ámbito, sino que abarca diversas formas, sea violencia de género, racismo, clasismo, asimismo, como la vulneración de derechos institucionales en general, siendo una expresión de discriminación estructural o de vulneración que afectan la dignidad y el bienestar de las personas.

2.1.2. Tipos de violencia institucional

Los tipos de violencia institucional en si no son clasificables, esto debido a la complejidad de la temática tratada, sin embargo, se pueden observar casos y patrones según la forma en la que se vulneran los derechos y los grupos afectados de acuerdo a la ley; por regla, todos parten de un mismo problema, el uso indebido de poder o negligencia dentro de estructuras institucionales que deberían proteger a las personas.

Uno de los casos más evidentes, y que normalmente se expone en medios como redes sociales, es la violencia de género en instituciones, que ocurre cuando mujeres, hombres o personas de identidades diversas reciben un trato desigual debido a su género o preferencia sexual, y estos casos suelen ser ignorados en sus denuncias. Desafortunadamente, este tipo de violencia se normaliza en prácticas cotidianas, como la falta de credibilidad hacia las víctimas, la minimización de los derechos denunciados e incluso el silenciar las voces que piden justicia al degradarlas, difamarlas y quitarles importancia.

Otro ámbito importante en sí mismo es la violencia racial, cultural y/o xenófoba en instituciones, que se manifiesta en la discriminación basada en la etnia, la cultura o la nacionalidad de la persona afectada. Esto puede reflejarse en el acceso desigual a servicios públicos, en prejuicios dentro del sistema judicial, en la vulneración de los derechos humanos internacionales o en la exclusión social de ciertos grupos. De manera similar, el clasismo institucional afecta a personas de bajos recursos, quienes muchas veces reciben menor atención

de calidad, enfrentan barreras burocráticas y se ven segmentados debido a la exclusión de oportunidades por su condición de vida o económica.

También existe la violencia psicológica institucional, que se presenta cuando las personas son sometidas a tratos humillantes, intimidación, indiferencia o negligencia emocional por parte de funcionarios o autoridades. A esto se suma la violencia física institucional, principalmente en contextos de uso excesivo de la fuerza o el maltrato evidente; relacionado con esto, se encuentra la violencia simbólica, que consiste en discursos, normas o prácticas que refuerzan estereotipos y desigualdades, legitimando la discriminación de manera indirecta

Uno de los casos particularmente graves se da a través de las violaciones de carácter sexual en instituciones, donde adultos o menores de edad son víctimas de abuso sexual por parte de figuras de autoridad. Aquí no solo se trata de abuso, sino también de encubrimiento o de falta de acción por parte de la institución, lo que permite que se produzcan perjuicios en estos casos. El tipo de violencia anterior se sitúa en la misma línea que la violencia sexual institucional en general, que incluye acoso, abuso, insinuaciones o explotación en entornos institucionales.

Por otro lado, existe la violencia administrativa o burocrática, que ocurre cuando los procesos institucionales son tan ineficientes o inaccesibles que vulneran derechos institucionales, nacionales e incluso internacionales. Este tipo de violencia no se da únicamente hacia otras personas, sino también hacia otras instituciones, y ocurre cuando una entidad obstaculiza, deslegitima o afecta el funcionamiento de otra, generando impactos indirectos en la sociedad.

Asimismo, la violencia por omisión se da cuando las instituciones no actúan ante situaciones que requieren intervención, dejando desprotegidos a los afectados.

A nivel más amplio, también se tiene la violencia institucional estructural, que está integrada en normas, políticas y malas prácticas del sistema, reproduciendo desigualdades de forma constante, lo que incluye la violencia por parte de instituciones gubernamentales al incumplir tratados o vulnerar los derechos humanos, ya sea mediante políticas restrictivas, falta de garantías o ausencia de protección a personas, a representantes de entidades institucionales o a las propias instituciones en sí.

2.1.3. Impacto social de la violencia institucional

En los últimos años el fenómeno de la violencia institucional ha adquirido mayor visibilidad como consecuencia del auge de plataformas digitales y redes sociales, como la que se utilizará en este trabajo, en las que los individuos encuentran espacios para compartir sus experiencias, denunciar abusos o violaciones de sus derechos, y, de este modo, generar conciencia colectiva sobre los temas tratados. Estas herramientas nos han permitido evidenciar patrones de comportamientos o hechos anormales, que se pueden evidenciar dentro de las instituciones o por parte de las mismas, así como la posibilidad de presionar mediante cambios o reformas que nos garanticen una atención más justa, equitativa y respetuosa de los derechos humanos. De esta manera, la exposición mediática no solo contribuye a la denuncia, sino

también impulsa el debate público y la construcción de mecanismos de rendición de cuentas para los afectados frente a las instituciones que no han respetado la armonía y el bien común.

A pesar de lo anteriormente mencionado, el tratar casos y temas sobre la violencia institucional no es cosa fácil, varias personas pueden tener diversas opiniones acerca de ciertos casos, los mismos que pueden ser controversiales y/o algunos que se podrían ver justificados ante la ley, es por eso que el uso de redes sociales facilita el dar una voz a todo aquello que tenga un punto de vista acerca del tema.

2.2. Lenguaje de Programación

Según la UNED (2021), un lenguaje de programación es un sistema formal de comunicación con orden, sintaxis y semántica diseñado para permitir que las personas puedan expresar instrucciones y algoritmos para interactuar con computadores, que interpretarán y ejecutarán estas instrucciones. Estos lenguajes cuentan con una sintaxis muy específica dependiendo de cuál sea usado, lo que determina la forma correcta en la que se deberá escribir para permitir desarrollar programas capaces de resolver problemas, automatizar procesos y manipular información de manera eficiente, y, aunque la semántica pueda cambiar en base al lenguaje que sea utilizado, normalmente se intenta seguir un esquema común de cómo tratar el código con sus diversas maneras de trabajar, sea orientados a objetos, funcionales o procedimentales, y entre estos encontramos uno de los lenguajes más versátiles en la actualidad, Python.

2.2.1. Python

Python es un lenguaje de programación versátil que nos permite desarrollar diferentes tipos de programas como aplicaciones web, análisis de datos, automatización, inteligencia artificial y software en general. Según la página oficial de Python (s.f), se define como un lenguaje de programación interpretado, de alto nivel y orientado a objetos, diseñado para ser fácil de leer y comprender gracias a su sintaxis sencilla y organizada. Además se caracteriza por facilitar el mantenimiento de código y acelerar procesos de desarrollo, lo que lo convierte en uno de los lenguajes más utilizados en la actualidad tanto en el ámbito académico y profesional por su fácil aprendizaje, entendimiento y por su versatilidad, ya que este se encuentra integrado en la gran mayoría de servicios como lenguaje predeterminado y porque cuenta con una gran cantidad de librerías para trabajar desde los programas más simples hasta los más complejos como sería la ciencia de datos.

2.2.2. Librerías

Las librerías o bibliotecas en el ámbito de programación son conjuntos de código previamente creados, ya sea por uno mismo o por otros programadores, que pueden ser reutilizados dentro del lenguaje de programación cuando sea compatible y permitido. Estas pueden ser desde funciones simples hasta muy complejas, evitando con ello la necesidad de desarrollar funciones desde cero, en razón de que pueden ser instalada e importadas directamente al proyecto para así facilitar y agilizar el uso en el código.

A continuación, se describe un detalle de las librerías que serán utilizadas en el desarrollo de este trabajo:

- **youtube-comment-downloader:** Según Bouman (s.f.), sirve para la obtención de comentarios de la plataforma de Youtube, descargándolos en el equipo donde se trabaje y en el formato que sea requerido, normalmente en formato “csv” o “json”.

- **pysentimiento:** Según finiteautomata (s.f.), permite analizar los sentimientos de los formatos entregados y trabajar con varios idiomas, siendo el principal y el que la carterita el idioma español.

- **transformers:** Es una librería especializada en modelos de lenguaje avanzados basados en arquitecturas “LLM”, como, por ejemplo, “BERT”, permitiendo comprender el contexto completo de la data tratada debido a la información previa con la que viene cargada.

- **pandas:** Es fundamental para el manejo y análisis de datos, esto debido a que permite cargar *datasets*, organizar la información en *dataframes* y realizar operaciones como filtrar, limpiar, agrupar y transformar datos.

- **numpy:** Enfocada en el cálculo numérico eficiente. Trabaja con arreglos y matrices normalmente, pero tambien sirve para trabajar con operaciones matemáticas y manipulación de datos numéricos.

- **nltk:** Especializada en procesamiento de lenguaje natural. Permite trabajar con texto realizando tareas como eliminar palabras vacías, “tokenizar”, lo cual es separar las palabras y darles un espacio, y normalizar texto.

- **wordcloud:** Permite generar nubes de palabras, en donde las palabras más frecuentes de nuestra data tratada aparecerán con mayor tamaño.

- **plotly:** Es una librería de visualización interactiva que permite generar gráficos dinámicos.

- **cufflinks:** Conecta pandas con “ploty”, facilitando la creación de gráficos interactivos directamente desde la data cargada.
- **matplotlib:** Genera la visualización de datos al crear gráficos estáticos como barras, líneas, histogramas, etc, siendo fundamental para la representación visual de los resultados de un análisis.

- **seaborn:** Es una librería de visualización estadística basada en “matplotlib” que facilita la creación de gráficos más estilizados y con un mejor diseño.

- **re:** Es usada para trabajar con expresiones regulares. Permite buscar, extraer y manipular patrones dentro de textos.

- **torch:** Trabaja con redes neuronales, tensores y procesamiento acelerado mediante “GPU”, incluyendo el uso de lenguajes avanzados.

- **scikit-learn/sklearn:** Es una librería de *machine learning* que proporciona herramientas para el procesamiento de datos y modelos para su entrenamiento y uso. Esta incluso incluye la herramienta pipeline, la cual sirve para ejecutar código de una manera secuencial y automatizada, sea con una tarea simple o compleja.

2.3. Entorno de Desarrollo

2.3.1. Google Colab

Según Google (s.f.), Google Colab es una plataforma de código abierto. Permite ejecutar código Python y *markdowns* con “Jupyter” en el navegador o mediante extensiones, sin necesidad de instalar programas en el computador e incluso accediendo de manera remota. Al ser un servicio basado en la nube, considerado un “PaaS” (*Platform as a Service*), nos ofrece acceso

a recursos virtuales como GPU o “TPU”, dependiendo del plan en uso, incluyendo un plan gratuito.

Esta plataforma está orientada principalmente al análisis de datos, al *machine learning* y a la inteligencia artificial, gracias a los servicios de computación en la nube que ofrece. Además, facilita el trabajo colaborativo, ya que varios usuarios con una cuenta de Google independientemente del plan que estén usando pueden editar y ejecutar los trabajos realizados en la plataforma, conocidos como *Notebooks*, todo ello de manera online y con guardados en puntos.

En este proyecto se utilizará esta plataforma para desarrollar y ejecutar el código necesario, ya sea para entrenar los modelos o para procesar los datos que se obtendrán, aprovechando la capacidad de utilizar recursos virtuales de forma gratuita.

2.3.2. Power BI

Según Microsoft (s.f), la empresa encargada de crear esta herramienta, Power BI es una plataforma de análisis empresarial que permite conectar distintas fuentes de datos, transformarlas y convertirlas en información visual mediante paneles, gráficos e informes interactivos. Esta herramienta que puede usarse de forma pagada, gratuita e incluso en la nube, está diseñada para facilitar la interpretación de grandes volúmenes de datos y apoyar la toma de decisiones en proyectos, permitiendo identificar patrones, tendencias y métricas importantes de manera rápida, dinámica y comprensible.

Además, Power BI, puede integrar información proveniente de bases de datos, archivos, servicios en la nube y múltiples plataformas digitales, ofreciendo un entorno colaborativo para compartir reportes y análisis en tiempo real. Debido a sus capacidades de visualización y de procesamiento de datos, esta herramienta es ampliamente utilizada en áreas de inteligencia de negocio, análisis de datos y ciencia de datos, siendo este último el caso de este proyecto.

2.4. Análisis de sentimientos

Según Universidad y Sociedad (2021), el análisis de sentimientos es una técnica basada en el procesamiento del lenguaje natural que permite identificar, clasificar e interpretar opiniones, emociones y percepciones presentes en textos como entrevistas, comentarios o documentos. Esta tiene como objetivo transformar la información subjetiva en resultados más organizados y analizables, reduciendo la influencia de los prejuicios o de las interpretaciones personales en una investigación.

Además, este análisis utiliza herramientas tecnológicas y métodos computacionales, como el *machine learning* y la visualización mediante gráficas, para determinar si los sentimientos expresados en un texto pueden clasificarse, lo que permite obtener conclusiones más precisas y objetivas a partir de grandes volúmenes de información cualitativa.

2.4.1. Aplicación y clasificación de sentimientos

Al poder usarse para clasificar grandes volúmenes de datos, el análisis de sentimientos es una de las técnicas más demandas por las empresas actuales, especialmente en redes sociales, ya que permite clasificar de manera simple comentarios complejos y subjetivos colocándolos en base al porcentaje en que pueden pertenecer a una clase, normalmente estas clases son Positivo, Negativo y Neutro. En la actualidad las redes sociales son parte del core de los negocios, por lo que entender, categorizar y controlar la información que se tiene sobre las opiniones acerca de cualquier producto de una empresa es fundamental, permitiendo a la empresa tener una Viena visión para la toma de decisiones y hacer una correcta inteligencia de negocio tanto a nivel operacional, estratégico y de marketing incluso.

2.5. Procesamiento de Lenguaje Natural

Según la revista TIA (2021), el "PLN" o Procesamiento de Lenguaje Natural es un campo de estudio de la inteligencia artificial que busca comprender cómo funciona el lenguaje en textos, su construcción y generación, permitiendo a los algoritmos interpretar, analizar y producir lenguaje natural. Además, facilita tareas como la traducción, generación de texto y comprensión de información escrita, transformando el lenguaje en datos procesables para su correcto exploración y tratamiento, siendo usado en casos de análisis de sentimientos.

2.5.1. Normalización y limpieza de texto

Normalizar y limpiar el texto es una etapa fundamental dentro del procesamiento de lenguaje natural, esto se debe a que permite preparar los datos antes de ser analizados por modelos u algoritmos en el análisis de sentimiento, ya que, los textos obtenidos de comentarios de redes sociales suelen contener errores ortográficos, palabras repetidas, enlaces, inconsistencias e incluso emojis, aunque estos últimos pueden ser interpretados dependiendo de la herramienta usada y la finalidad del proyecto. Este procedimiento se realiza mediante técnicas variadas como la imputación, convertir todas las palabras a minúsculas, reducir palabras e incluso eliminar vacíos o palabras poco relevantes, por lo que, de esta manera se logra reducir el ruido y mejorar el rendimiento de los modelos para el PLN.

2.5.2. Representación vectorial de texto

Esto se refiera a un proceso utilizado en el PLN que consiste en transformar palabras, frases o incluso documentos completos en valores numéricos que puedan sen interpretados por modelos de IA, esto debido a que los algoritmos no logran comprender texto directamente, por lo que un proceso de vectorización es importante para dar un valor matemático para su procesamiento correcto, permitiendo clasificar lo que sea trabajado y así que ser interpretado por el algoritmo.

2.5.3. TF-IDF y Fine-tuning

Según IBM (2023), *Term Frecuency – Inverse Document Frequency* (TF-IDF) o también conocido como *Bag of words* (BoW), es una técnica utilizada en casos de PLN que mide la importancia de una palabra dentro de un conjunto de datos de tipo texto. Esta analiza tanto la frecuencia con la que aparece una palabra o qué tan común es dentro de todos los datos, permitiendo identificar términos relevantes y transformar el texto en datos numéricos medibles y procesables, así los modelos comprenden la información analizada de mejor manera y las relaciones presentes en esta.

Otra de las técnicas importantes en estos casos según IBM (2024) es el *fine-tuning*, la cual es una técnica que, a diferencia de TF-IDF, toma modelos avanzados ya entrenados y los especializa utilizando nuevos datos etiquetados, teniendo una mayor dimensionalidad de los datos, mejorando su capacidad para comprender contexto, intención y significado dentro de tareas como el análisis de sentimientos, incluso llegando a tomar en cuenta el sarcasmo.

2.6. Aprendizaje automático

Según la revista Ingenaire (2021), el aprendizaje automático, o también conocido como *machine learning*, es un campo de la inteligencia artificial el cual permite que los algoritmos aprendan a partir de datos y experiencias sin necesidad de ser programados explícitamente para

cada acción. Además, busca que los algoritmos mejoren su rendimiento de forma automática al identificar patrones, lo que facilita tareas como la clasificación de datos, las predicciones y el procesamiento del lenguaje natural.

Dentro de este campo podemos encontrar el aprendizaje profundo, también conocido como *deep learning*, que emplea redes neuronales artificiales con múltiples capas para resolver problemas complejos como el reconocimiento de imágenes vectorizadas, la interpretación de texto y el análisis de grandes volúmenes de información.

2.6.1. Modelos supervisados para clasificación de texto

Según IBM (s.f), los modelos supervisados para la clasificación de texto son algoritmos de *machine learning* que, a diferencia de los modelos no supervisados, se entrenan con datos etiquetados o clasificados, en los que cada dato posee una etiqueta ya definida; por ejemplo, en el análisis de sentimiento de texto podrían ser las etiquetas positivo, negativo y neutro. Estos modelos aprenden de patrones y relaciones presentes en los datos para, posteriormente, clasificar la nueva información de manera automática y precisa, a través de técnicas como PLN, para mejorar la comprensión de los datos, y son ampliamente aplicados en tareas de ciencia de datos, como el análisis de sentimientos y la categorización de datos, al usar algoritmos dedicados específicamente a la clasificación. Algunos de los algoritmos supervisados considerados para este proyecto son:

- **Árboles de Decisión:** Según IBM (2021), los árboles de decisión son algoritmos para modelos supervisados de *machine learning* utilizados para clasificación y predicción de datos mediante una estructura similar a la de un árbol, donde cada rama representa una decisión basada en ciertas características de la información analizada. Este aprende patrones presentes en los datos para posteriormente poder clasificar nueva información automática y precisa, utilizando reglas y condiciones que facilitan la interpretación de los resultados obtenidos.

- ***Random Forest*:** Según IBM (s.f.), este se basa en la combinación de múltiples árboles de decisión para mejorar la precisión y estabilidad de las predicciones, esto usando diferentes subconjuntos de datos y posteriormente combinando sus resultados para obtener una clasificación más precisa y menos propensa al sobreajuste.

- ***Support Vector Machine* (SVM):** Según IBM (2023), este busca separar diferentes categorías mediante el uso de hiperplanos que permitan diferenciar las clases de la manera más precisa posible, esto al aprender patrones de los datos etiquetados para clasificar la nueva información, siendo especialmente útil en conjuntos de datos complejos o de alta dimensión.

- **Regresión Logística:** Según IBM (s.f.), sirve para predecir la probabilidad de que un dato pertenezca a una categoría específica a partir de datos anteriores, aprendiendo patrones y relaciones con las que posteriormente clasifica nueva información, y puede trabajar tanto con clasificación binaria como multiclase.

- ***Bidirectional Encoder Representations from Transformers*** **(BERT):** A diferencia de los algoritmos anteriores, que son considerados algoritmos clásicos, según IBM (2020), BERT es un algoritmo de modelos de lenguaje de gran tamaño , o también llamados LLM, de *deep learning* y de PLN para comprender correctamente el contexto y significado de las palabras dentro de un texto de manera bidireccional, es decir, analizando tanto las palabras anteriores como posteriores dentro de una oración, lo que permite interpretar mejor la intención y relaciones semánticas de los datos, y adicional a esto, puede contar con data previa para así poder incluso clasificar de manera correcta casos especiales y anormales en los textos, incluyendo a texto sarcástico.

2.6.2. Submuestreo, sobremuestreo y sobreajuste

Según IBM (s.f.), el "submuestreo", también conocido como *undersampling* o *downsampling*, es una técnica empleada en *machine learning* cuando existe un desbalance en los datos, esto ocurre cuando una clase tiene valores muy diferentes a la otra, por lo que en el caso del submuestreo consiste en equilibrar el conjunto de datos desbalanceados reduciendo la

cantidad de ejemplos de las clases mayoritaria, esto se hace al eliminarlos y este no debe ser con exactitud al número de la clase minoritaria, sino que simplemente se debe buscar que el conjunto de datos se encuentre relativamente balanceado para el entrenamiento de modelos.

Por otro lado, según IBM (s.f.), el "sobremuestreo", también conocido como *oversampling* o *upsampling*, consiste en aumentar la cantidad de datos en las clases minoritarias mediante la repetición de ejemplos existentes en el conjunto de datos, ya sea usando ejemplos aleatorios, usando los datos de mayor repetición o con la generación de datos sintéticos, permitiendo que el modelo aprenda de manera más equilibrada.

Estas dos técnicas utilizan habitualmente debido al "sesgo" de los datos, causado por la falta de clases igualadas, lo que desemboca en un sobreajuste. El "sobreajuste" u *overfitting*, según IBM (2021) es un problema de aprendizaje en los modelos de *machine learning* que ocurre cuando el modelo aprende demasiado de los datos de entrenamiento o cuando los datos desbalanceados sesgan al modelo hacia una variable específica. Esto puede deberse al aprendizaje de patrones irrelevantes o de ruido, lo que da lugar a buenos resultados únicamente con el conjunto de datos trabajado, pero a un bajo desempeño con nueva información. Esto provoca que el modelo no generalice correctamente, por lo que aplicar las técnicas anteriormente mencionadas o incluso técnicas como la normalización de los datos o la validación cruzada entre variables, mejora su capacidad de adaptación y precisión.

3. CAPÍTULO III: METODOLOGÍA DEL ANÁLISIS

3.1. Enfoque metodológico de la investigación

El trabajo se enfocará en el área de ciencia de datos, una disciplina que combina estadística, programación e inteligencia artificial para extraer conocimiento útil a partir de grandes volúmenes de datos, lo que permite identificar patrones, tendencias y relaciones que apoyan la toma de decisiones. El uso de *machine learning* en este proyecto permitirá automatizar las tareas realizadas, siendo estas la obtención de la data, la manipulación de la misma, el análisis de sentimientos, el entrenamiento del modelo, la comparativa del funcionamiento del modelo y la visualización de los resultados del análisis de sentimientos con el modelo entrenado.

El uso de la técnica de análisis de sentimientos sirve para identificar, extraer y clasificar la información no estructurada sobre opiniones, lo que resulta útil para este trabajo investigativo, debido a que con este se puede interpretar la percepción social sobre la violencia institucional a partir de comentarios obtenidos, clasificándolos en tres grupos de sentimiento, siendo Positivo, Negativo y Neutro, o para abreviarlos sería “POS”; “NEG” y “NEU.

Una vez el conjunto de datos ha sido clasificado y teniendo en cuenta que la data ha sido revisada, límpida y normalizada, se puede usar para entrenar los modelos seleccionados, y debido a la clasificación previamente mencionada, serían modelos supervisados de clasificación, pero habría que tomar en cuenta un posible balanceo de las clases para evitar *overfitting* debido a

que la data sobre la violencia institucional tiende a ir a comentarios negativos, por lo que se buscará usar una técnica para balanceo.

Y en cuanto a la visualización de los datos se buscará enseñar la data más importante, la cual en este caso será la comparación de modelos, conteo de aparición de palabras en el conjunto de datos y la clasificación despues del entrenamiento, ya que, con esta visualización dinámica se podrá trabajar en conclusiones y encaminar el entrenamiento para su futuro uso.

3.2. Fuente de datos

Para la obtención de datos se utilizarán únicamente datos provenientes de comentarios de la plataforma de videos Youtube, por ser una red social de fácil acceso y navegación, lo que le permite contar con una gran diversidad de usuarios de diferentes contextos sociales y culturales. Asimismo, la plataforma permite acceder públicamente a comentarios y opiniones generadas en distintos videos, facilitando la recolección de información relacionada con temas de violencia institucional y la percepción social, lo que servirá para construir un conjunto de datos, también conocido como *dataset*, amplio y variado, adecuado para el análisis de sentimiento, para el PLN y para el entrenamiento de los modelos.

3.2.1. Comentarios de la red social Youtube

Los comentarios se obtendrán de distintos videos relacionados con la violencia institucional, esto debido a que al trabajar con videos sobre un tema delicado no se puede encontrar en un solo video suficientes comentarios como para un estudio amplio, así que al trabajar con varios videos también se busca abarcar varios campos de la misma, como el abuso de autoridad, conflictos sociales, violencia policial, discriminación, abuso en general y problemáticas gubernamentales, publicados tanto en canales informativos y educativos como de opinión y podcast. Entre los videos utilizados se encuentran los de la siguiente tabla:

Tabla 1

Fuente de datos de videos

Video	**Autor**
LA ESCALERA DE LA VIOLENCIA	El Muro, 2011
CIDH condena "violencia institucional" del régimen de Maduro	CNN en Español, 2024
Contraloría anuncia fiscalización a polémico curso sobre "violencia institucional" (RD)	Bio Bio, 2025
El feminicidio infantil de Fátima: omisiones y violencia institucional	Rompeviento TV, 2020
Revelan que 1 de cada 3 delitos cometidos en estallido social fue la violencia institucional	Bio Bio, 2025

Fue detenido por un robo y apareció muerto... ¿violencia institucional?	Crónica TV, 2025
FUI ABUS4D4 POR 5 SOLDADOS DEL EJÉRCITO (CON: DALY PARRA) |Vos podés el podcast - EP 138|	VOS PODÉS, EL PODCAST!, 2024
HAY QUE ACABAR CON LA VIOLENCIA INSTITUCIONAL DE LOS JUECES; SE CORROMPEN CON MUCHA FACILIDAD	Jesús Escobar Tovar, 2025
¿Por qué existe la violencia?	CuriosaMente, 2018
Violencia institucional racista | Irene Monter	Podemos, 2025
Así fue minuto a minuto del operativo que llevó a la captura de Nicolás Maduro en Venezuela	Noticias Caracol, 2026
Un repartidor logra escapar de los militares en Chicago tras protestar contra ellos	El HuffPost, 2025
Detención violenta en Ciudad de México	Grupo Fórmula, 2025
Expulsan a sacerdote ‘influencer’ que denunció abuso sexual: la iglesia dice que no confirmó su caso	N+ UNIVISION, 2024
Video: Denuncian a gerente por maltrato laboral en call center	Al Rojo Vivo, 2022
Policías golpean a preso en Tijuana	El Universal, 2018
Exhiben a policías de Chimalhuacán golpeando a detenido	El Universal, 2018

Un policía disparó cartuchos de goma a jóvenes para dispersar un enfrentamiento	LA NACION, 2023
El Escuadrón de la Muerte en Ecuador	Cris Cabrera, s.f.
Así fue la entrada de la Policía de Ecuador a la embajada de México para detener a Jorge Glas	CNN en Español, 2024

3.2.2. Usos de APIs para la extracción

Se utilizará, como se mencionaba previamente, la librería "youtube-comment-downloader", la cual fue creada por egbertbouman y compartida por un repositorio en "Github", junto a toda la documentación respectiva para su uso, siendo de esta manera en la que la he encontrado, y entre los beneficios de esta librería tenemos que es de fácil uso, simplemente se tendrá que instalar en el entorno y llamar a los videos por medio de su enlace o link, además de que esta no necesita de una API key, la cual es un código único que se usa para autenticar y autorizar el acceso, lo que facilita su adopción por el público general.

3.3. Metodología CRISP-DM

Para el estudio consiguiente se ha decidido usar la metodología de trabajo CRISP-DM o *Cross Industry Standard Process for Data Mining*, la cual es una metodología utilizada en proyectos de minería de datos tal cual como lo dice su nombre, en *machine learning* y para

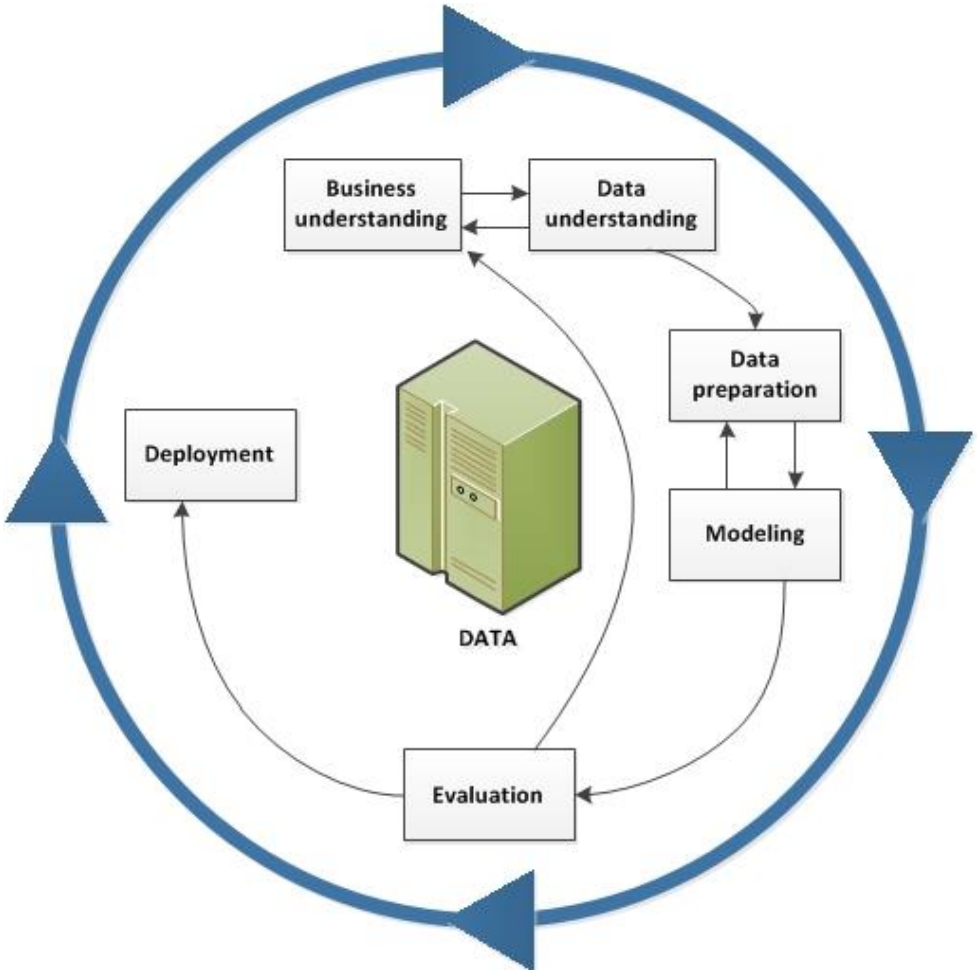
visualización de minería, lo que proporciona un marco para organizar y estructurar el desarrollo de modelos para el análisis de datos y la convierte en una de las metodologías más eficientes y con aprobación mundial para trabajos de ciencia de datos e inteligencia de negocio.

Según la página web de IBM (2021), CRISP-DM, como metodología "incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre tareas" (sección Conceptos básicos de ayuda de CRISP-DM, párr. 2). Debido a esto, se considera una metodología flexible y ampliamente utilizada en proyectos de PLN y de análisis de sentimientos.

Al trabajar con la metodología CRISP-DM debemos tener en cuenta que se divide en seis fases principales: la comprensión del negocio o del caso de estudio, la comprensión de los datos, la preparación y el tratamiento de los datos, el modelado, la evaluación y el despliegue o la implementación. Cada etapa garantiza que los datos sean tratados correctamente y que se tenga bien documentado de manera secuencial lo que se hizo para llegar a cada fase, incluso permitiendo trabajar con pipelines, aunque también nos permite el regresar entre ciertas fases de su ciclo, relacionándolas entre todas para hacer una evaluación de los modelos construidos y que respondan adecuadamente a los objetivos del trabajo o investigación, tal y como se observa en la Figura 1.

Figura 1

Ciclo Fases CRISP-DM



Nota. Recuperado de IBM (2021)

3.3.1. Fase 1: Entendimiento del negocio o caso de estudio

Al corresponder a la etapa inicial del proyecto, la fase 1 busca comprender el problema, las necesidades y los objetivos que se desean resolver mediante la minería de datos.

Incluye conceptos básicos como la comprensión comercial o trasfondo del estudio, la determinación de objetivos e indicadores clave de rendimiento (KPI), evaluación de la situación actual sobre el tema, definición de objetivos de minería y/o visualización y elaboración de un plan preliminar para el proyecto, por lo que su importancia radica en que establece una dirección clara para el desarrollo al identificar recursos, limitaciones, alcance, contexto y resultados esperados para la aplicación correcta de las técnicas y modelos.

3.3.2. Fase 2: Entendimiento de los datos

Esta fase consiste en recopilar, analizar y evaluar la información que se utilizará durante el proyecto, compuesta por la recopilación de datos iniciales, la descripción del *dataset* obtenido o creado, la exploración de la información, la verificación de la calidad de los datos y en casos específicos la estadística descriptiva. El objetivo principal es identificar si la data es apropiada para trabajar, ver sus características, patrones, posibles errores e inconsistencias, permitiendo comprender mejor la información antes de aplicar las técnicas con datos que si sean adecuados, confiables y útiles para el proyecto, dejando paso a que en la siguiente fase se trate la problemática encontrada en los datos.

3.3.3. Fase 3: Preparación y trata de los datos

La fase de preparación tiene como finalidad acondicionar la información obtenida en la fase anterior para que pueda ser utilizarse de manera eficiente en el modelado. En esta etapa se realiza la selección de los datos más relevantes, la identificación de variables a trabajar, la limpieza de la información, la revisión de la data mediante comparativa y el formato adecuado según lo que necesite el proyecto, lo que puede incluir, en ciertos casos, la construcción de nuevos datos o incluso la integración de distintas fuentes de información. Con este proceso, los datos adquieren una estructura más organizada, consistente y normalizada, por lo que mejora el rendimiento de las técnicas utilizadas y la alimentación de datos a los modelos.

3.3.4. Fase 4: Modelado

Debido a que en las fases anteriores se ha escogido la data y se la ha tratado, esta fase se dedica exclusivamente al modelado, al seleccionar la técnica que pueda ser usada, sea un modelado supervisado o no supervisado, los algoritmos seleccionados para su comparativa de rendimiento posterior, la generación de un diseño de comprobación del uso de los algoritmos al alimentarlos con la data, la generación de los modelos en sí y una pequeña evaluación previa a la fase 5, en donde solo se compararían aspectos generales.

3.3.5. Fase 5: Evaluación

Esta fase tiene como objetivo analizar y verificar si los modelos previamente entrenados cumplen correctamente con lo planteado y con el contexto de fases anteriores, escogiendo el modelo que presente mejores métricas y que represente de la mejor manera los objetivos del proyecto. Comprende la evaluación de los resultados con la selección del mejor modelo, su entrenamiento y el proceso de revisión del desarrollo hasta ese momento, permitiendo identificar posibles errores, limitaciones, interpretaciones erróneas o mejoras necesarias antes de implementar la solución final, garantizando que las conclusiones obtenidas sean confiables, útiles y que estén dentro del contexto para el caso de estudio planteado.

3.3.6. Fase 6: Despliegue o implementación

La etapa final corresponde al despliegue o implementación, en la que se prepara la utilización de los resultados evaluados, lo que permite planificar tanto el despliegue como el control, el costo y el mantenimiento, además de crear un informe final que documente todo el proceso realizado y su función. Este no necesariamente implica implementar un sistema completo en producción, ya que también puede consistir en una propuesta de implementación, en un caso hipotético de uso o en la visualización de los resultados del modelo mediante *dashboards*, por lo que presenta y comunica de manera clara el propósito de los resultados generados y del trabajo en sí.

4. CAPÍTULO IV: DESARROLLO DEL ANÁLISIS DE SENTIMIENTOS

4.1. Fase 1: Preparación del Entorno de Desarrollo y del plan

Para preparar el entorno, se usará Google Colab por su capacidad de uso en la nube, mediante un plan gratuito y eligiendo como servicio de hardware el entorno de ejecución con la opción de GPU T4, para que las tareas de procesamiento de código se distribuyan de manera eficiente y puedan ejecutarse con rapidez y facilidad. En cuanto al lenguaje a usar, será Python 3, ya que no solo es el lenguaje predilecto en Google Colab, sino que, por su versatilidad, también es apropiado para trabajos como el análisis de sentimientos. En la metodología de trabajo se aplicará CRISP-DM, ya que presenta sus fases de manera clara y organizada, lo que ayudará a mantener un código limpio. Además, el código será comentado para facilitar su comprensión, aplicando, dentro de las medidas posibles, las buenas prácticas de PEP8 para que esté escrito adecuadamente y sea legible.

Para aplicar la primera fase de CRISP-DM, se determinó el para qué de la investigación, colocando el contexto previo en el que se explica el fenómeno de la violencia institucional en redes sociales de manera breve para el lector, estableciendo que la violencia institucional es una problemática social relevante en nuestro día a día que influye a que en diversas redes sociales y espacios de conversación, las personas den sus opiniones, testimonios y muestran emociones acerca del tema y cómo afecta a la sociedad. Además se hace una evaluación del caso en la que se explica el propósito del trabajo, siendo este el extraer las

opiniones anteriormente mencionadas para realizar un análisis de sentimientos por medio de un modelado con inteligencia artificial en el que se permita identificar patrones emocionales predominantes y que se clasifique a los comentarios, esto con el fin de contribuir a la comprensión del impacto social de la violencia institucional, haciendo visible lo que las personas sienten y expresan frente a las instituciones, colocando esta información a servicio de la investigación académica y el análisis social. Posterior a esto, se coloca cómo se hará con un plan preliminar, con una visión a futuro, los riegos posibles dentro del proyecto como sería el sesgo, el alcance anteriormente mencionado, los posibles algoritmos a usar según el tipo de modelado, que por su naturaleza sería un modelado supervisado de clasificación, un pequeño glosario de términos mencionados anteriormente en el escrito, indicadores clave de rendimiento o "KPIs" para medir la eficiencia del proyecto y el objetivo de la visualización que se realizará despues del modelado.

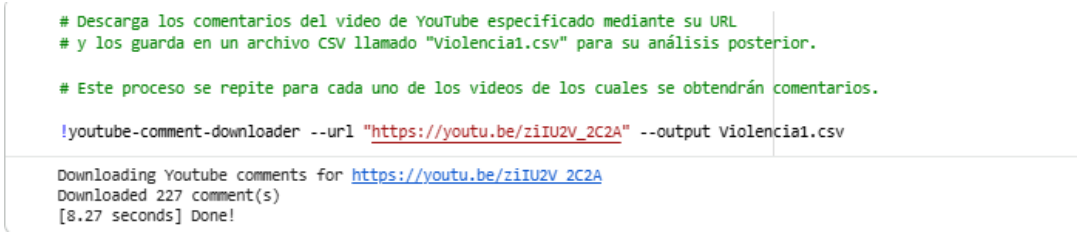
4.2. Fase 2: Preparación y limpieza de datos

Una vez establecido el plan en la fase 1, la fase 2 se cargará la data tratada para explorar la misma, en este caso se cargarán comentarios de los videos de Youtube que se mencionaron en la Tabla 1, explicando las fuentes de dónde se obtienen y se hará un *dataset* general formado por todos estos videos, por lo que, para poder empezar a trabajar en la programación para la obtención de los datos se cargarán e instalarán todas las librerías respectivas que se usarán en el proyecto, tal y como fueron mencionadas en las Librerías.

Los videos de Youtube se descargarán con la librería youtube-comment-downloader, proporcionando los enlaces de los videos y recibiendo en el código la retroalimentación necesaria para confirmar la descarga, mostrando información como el número de comentarios descargados y el tiempo de descarga como puede observarse en la Figura 2. Posterior a la descarga de los 20 videos, se combinarán los archivos obtenidos en un solo archivo tipo csv llamado "Violencia_youtube.csv", y se realizará una exploración de los datos del total de comentarios descargados previa a la limpieza, que en este caso son 35922 comentarios con una sola columna, como se puede observar en la Figura 3. Además, se subirá este archivo creado a "Kaggle", una plataforma web usada para ciencia de datos, donde se pueden encontrar repositorios para todo tipo de estudio, con la finalidad de que, si en un futuro alguien busca trabajar con data sobre violencia institucional, tenga un dataset de fácil acceso, aunque aún no esté limpiado ni normalizado.

Figura 2

Funcionamiento youtube-comment-downloader



Nota. Elaborado por Antonio Proaño

Figura 3

Revisión Dataset unido de comentarios

df_primero

	comentarios
0	text
1	like si te pidieron hacer una reflexión de esto
2	Yo lloré cuando le pegaron al perro
3	La violencia no es solo para a Muller pode ser...
4	¿quien mas viene por la tarea?
...	...
35918	Mexico es un narco-gobierno por eso no hacen...
35919	@GhostRaccon89 por fin ecuador ya basta de ta...
35920	@labuenaagricultura138
35921	Y no es asombro que Moreno, Lasso y Noboa teng...
35922	@GhostRaccon89dar asilo a delincuentes te par...

35923 rows × 1 columns

Nota. Elaborado por Antonio Proaño

4.2.1. Fase 3: Normalización de texto

Para empezar la fase 3, en donde se hará la trata de los datos, se normalizará el texto para reducir el ruido. Se eliminarán del dataset todos los prefijos, enlaces URL, menciones a usuarios con @, símbolos de hashtag, caracteres especiales como emojis y números en general; así, quedará un dataset más limpio y con los comentarios adecuados para ser estudiados. Además, se pasará todo el texto a minúsculas para una mejor manipulación de la data. Esta limpieza no se aplicaría sobre la columna previamente establecida de “comentarios”,

sino que se creará una nueva columna “comentarios_limpios”, visible en la Figura 4, esto con la finalidad de poder comparar los datos en ambas columnas.

Figura 4

Dataframe comentarios_limpios con normalización de texto

```
# Aplica la función limpiar_comentarios a cada valor de la columna "comentarios"
# y guarda el resultado en una nueva columna llamada "comentarios_limpios".
df_primero["comentarios_limpios"] = df_primero["comentarios"].apply(limpiar_comentarios)

# Aplica el pipeline de limpieza al DataFrame
df_primero = pipeline_limpieza.transform(df_primero)
```

```
#Analiza el dataframe con este añadido
df_primero
```

	comentarios	comentarios_limpios
1	like si te pidieron hacer una reflexión de esto	like si te pidieron hacer una reflexión de esto
2	Yo lloré cuando le pegaron al perro	yo lloré cuando le pegaron al perro
3	La violencia no es solo para a Muller pode ser...	la violencia no es solo para a muller pode ser...
4	¿quien mas viene por la tarea?	quien mas viene por la tarea
5	Más bien la cadena del abuso	más bien la cadena del abuso
...	...	...
35918	Mexico es un narco-gobierno por eso no hacen...	mexico es un narcogobierno por eso no hacen na...
35919	@GhostRaccon89 por fin ecuador ya basta de ta...	por fin ecuador ya basta de tanta injusticia a...
35920	@labuenaagricultura138	
35921	Y no es asombro que Moreno, Lasso y Noboa teng...	y no es asombro que moreno lasso y noboa tenga...
35922	@GhostRaccon89dar asilo a delincuentes te par...	asilo a delincuentes te parece muy inteligente...

35922 rows × 2 columns

Nota. Elaborado por Antonio Proaño

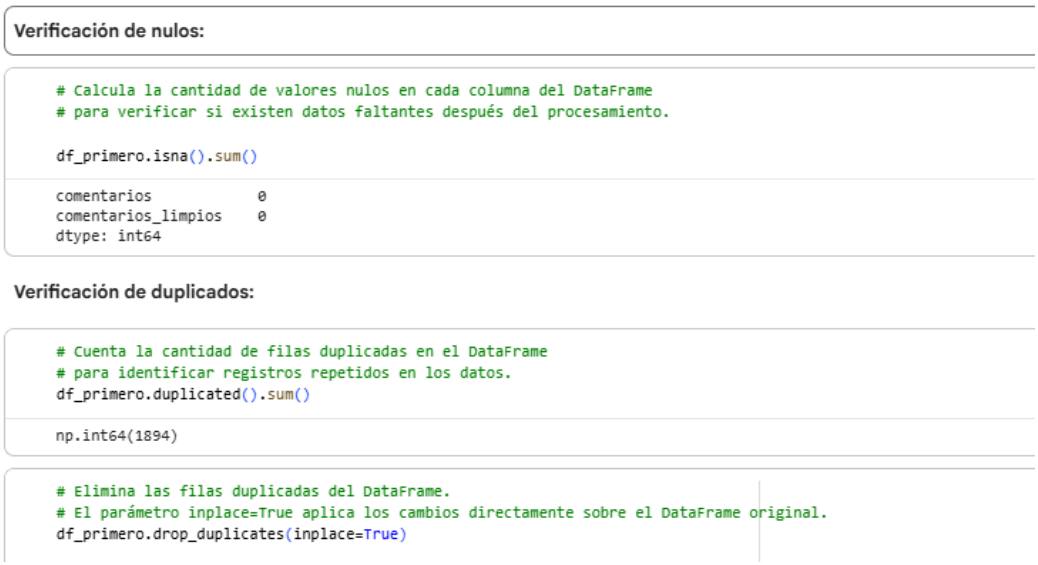
4.2.2. Eliminación de duplicados y nulos

Se hizo un análisis general para verificar datos nulos y duplicados sobre la columna de comentarios_limpios, y, aunque en este caso no se tienen datos nulos, se encontraron 1894

comentarios duplicados, por lo que se los eliminará, dejándonos con un conjunto de 34027 datos tal y como se observa en la Figura 5.

Figura 5

Verificación y eliminación de duplicados y nulos



Nota. Elaborado por Antonio Proaño

4.2.3. Etiquetado de sentimientos

De la librería “pysentimiento” se llama a “create_analyzer”, el cual crea un analizador de sentimientos configurado para idioma español que permitirá clasificar por medio de vectorización los textos como positivos, negativos o neutros, además de darles a estos una puntuación porcentual que representa el qué tanto pertenecen a cada clase, y tras hacer una

revisión de la librería respectiva para verificar su funcionamiento se hará la clasificación de los comentarios, añadiendo al conjunto de datos la columna “sentimiento” y la columna “score” para representar el puntaje.

Una vez clasificado el conjunto de datos se observará cuantos datos se encuentran distribuidos en cada clase, llamadas NEG, NEU y POS, teniendo como resultado 33301 comentarios clasificados a los que son adecuados para estas categorías, siendo 18606 negativos, 9717 neutros y 4978 positivos tal y como se observa en la Figura 6 y en la Tabla 2, una distribución realista de comentarios que se pueden encontrar en temas sobre la violencia institucional, guardando esto en un archivo que se lo llamará “clasificados_desbalanceado.csv” para una comparativa posterior al modelado.

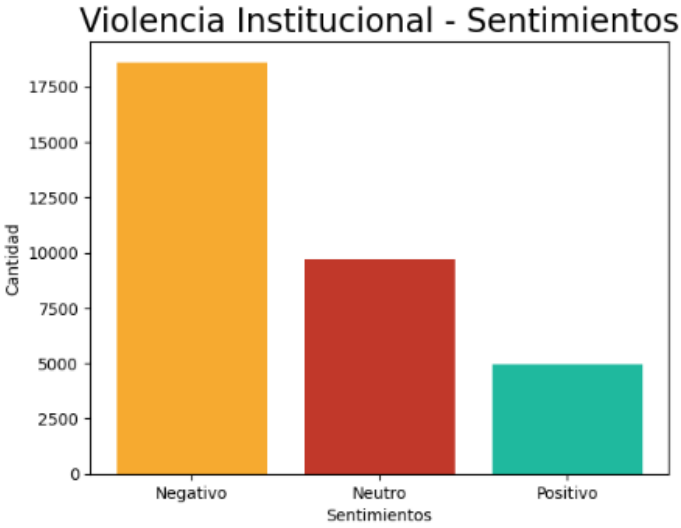
Tabla 2

Clasificación pysentimiento

Sentimientos	**Cantidad de comentarios**
Negativo/NEG	18606
Neutro/NEU	9717
Positivo/POS	4978
Totales	33301

Figura 6

Clasificación de sentimientos sin balanceo



df

	comentarios	comentarios_limpios	sentimiento	score
0	like si te pidieron hacer una reflexión de esto	like si te pidieron hacer una reflexión de esto	NEU	0.840430
1	Yo lloré cuando le pegaron al perro	yo lloré cuando le pegaron al perro	NEG	0.701408
2	La violencia no es solo para a Muller pode ser...	la violencia no es solo para a muller pode ser...	NEG	0.792927

Nota. Elaborado por Antonio Proaño

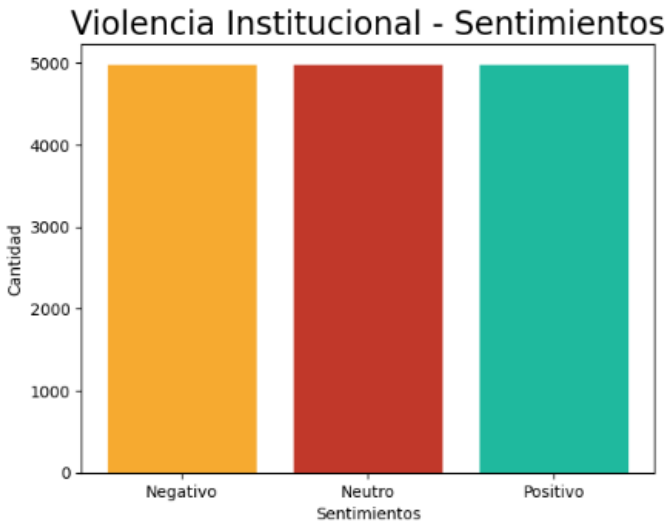
4.2.4. *Undersampling* del conjunto de datos

Debido a que, a pesar de ser lógico que, por el caso, el dataset tenga más negativos, se encuentra desbalanceado, por lo que, al entrenar el modelo, este va a tender a clasificar los comentarios hacia la clase mayoritaria, siendo en este caso los comentarios negativos. Por eso, se lo debe equilibrar. Se escogerá la técnica de *undersampling*, ya que es la más apropiada porque, a diferencia del oversampling, no requiere generar datos sintéticos. Se necesitará

igualar a las otras dos clases a la clase minoritaria, en este caso la clase de positivos, con 4978 comentarios, obteniendo un *dataset* final que se usará para el modelado y que queda con el mismo número de datos para todas las clases, con un total de 14934 comentarios, observando en balanceo correcto en la Figura 7.

Figura 7

Clasificación de sentimientos balanceado



Nota. Elaborado por Antonio Proaño

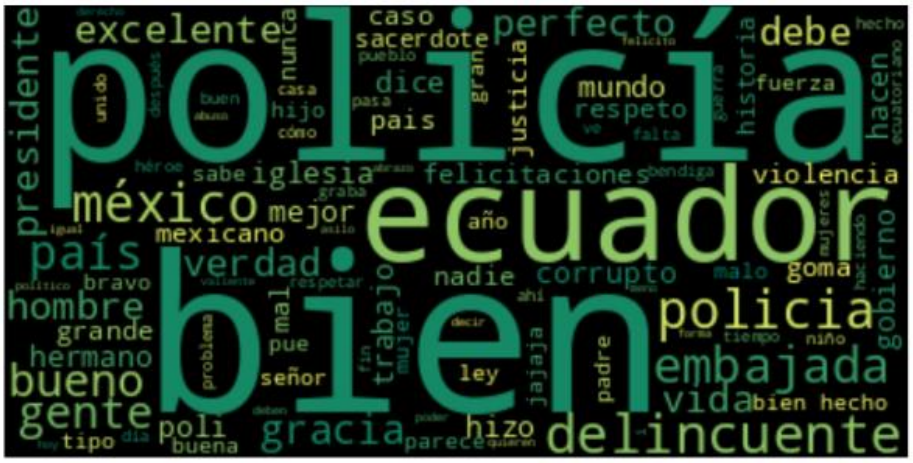
4.2.5. Nubes de palabras y bigramas

Por último, antes de pasar al entrenamiento de los modelos en la fase 4, se hará una revisión general del *dataset*, en donde se revisarán las palabras más usadas en cada una de las

clasificaciones como se puede observar en la Figura 8 y los bigramas del conjunto de palabras más usadas en el conjunto de datos en cada una de las clasificaciones como se observa en la Figura 9, esto con el fin de entender bien el contexto de los comentarios que se estudiarán posteriormente.

Figura 8

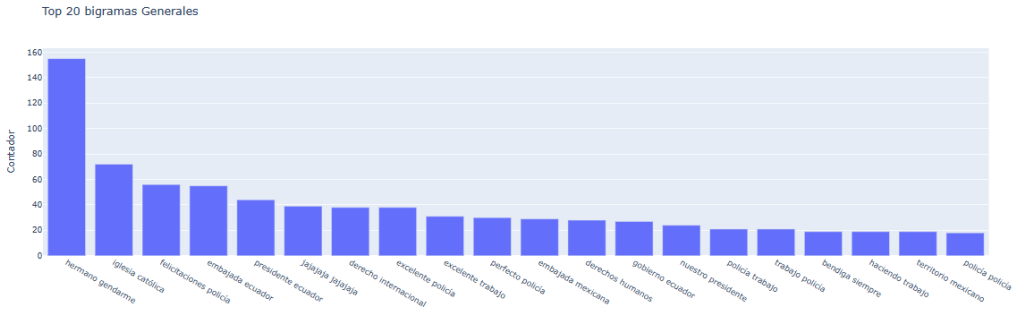
Nube de palabras general previos al entrenamiento



Nota. Elaborado por Antonio Proaño

Figura 9

Bigramas generales previos al entrenamiento



Nota. Elaborado por Antonio Proaño

4.3. Fase 4: Modelado de análisis de sentimientos

4.3.1. Selección de modelos de inteligencia artificial

Para la selección de los algoritmos a usar en la fase 4de CRISP-DM, se debe tener en cuenta que, debido a la naturaleza de los datos, lo más apropiado para el caso son modelos supervisados de clasificación, ya que los datos están previamente etiquetados en la columna sentimiento y lo que se busca en el trabajo es la clasificación correcta de la data subida, esto para entrenar un modelo apropiado de clasificación automática, por lo que, los algoritmos supervisados que serían considerados son los mencionados en el capítulo Modelos supervisados para clasificación de texto, siendo estos Árboles de decisión, *Random Forest*, SVM y Regresión Logística como modelos clásicos, usando la librería “scikit-learn”, y BERT como modelo moderno de LLM.

Para preparar los datos para el entrenamiento se usará TF-IDF para limitar el análisis a las 4000 palabras más relevantes, permitiendo vectorizar y transformar los datos en representaciones numéricas que puedan ser interpretadas por los algoritmos, y posteriormente se convertirá cada comentario en una matriz numérica, almacenando los sentimientos correspondientes a cada comentario. Otro de los pasos previos al entrenamiento y que es muy importante es el dividir la data en conjuntos de entrenamiento y prueba, siendo un 80% dedicado para entrenamiento y un 20% dedicado para comprobar el modelado, como se muestra en la Figura 10, esto debido a que se busca entrenar y evaluar los modelos para evitar sesgos y sobreajuste.

Figura 10

TF-IDF y separación en train y test

```
# Inicializa el vectorizador TF-IDF limitando a 4000 características (palabras más relevantes)
# Convierte el texto en una representación numérica para modelos de machine learning
tfidf = TfidfVectorizer(max_features=4000)

# Ajusta el vectorizador al texto y transforma los comentarios en una matriz numérica
# toarray() convierte la matriz dispersa en un arreglo denso
X = tfidf.fit_transform(df['comentarios_limpios']).toarray()

# Define la variable objetivo (etiquetas de sentimiento)
y = df['sentimiento']

# Define también X en texto puro para usar con los Pipelines
# Los Pipelines incluyen internamente el paso de vectorización TF-IDF
X_texto = df['comentarios_limpios']
```

```
#Dividir los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Dividir también los datos en texto para los Pipelines
# Los Pipelines incluirán internamente el paso de vectorización TF-IDF
X_train_texto, X_test_texto, y_train_texto, y_test_texto = train_test_split(
    X_texto, y, test_size=0.2, random_state=42, stratify=y
)
```

Nota. Elaborado por Antonio Proaño

4.3.2. Entrenamiento de modelos

- **Árboles de decisión:** Para el entrenamiento de este modelo se carga el *dataset* ya balanceado, usando la vectorización TF-IDF, lo que permite usar el texto en crudo. Una vez entrenado, se obtienen las predicciones mostradas por medio de una matriz de confusión en la Figura 11, en la que se puede observar los comentarios clasificados correctamente y los que el modelo pudo haber confundido, pero, adicionalmente, se calculan y se visualizan las métricas de evaluación ponderadas entre las tres clases: la precisión mide qué tan confiables o correctas son las predicciones, el *recall* mide que proporción de los casos reales de una clase se identificó correctamente en la predicción, el "F1-score" combina ambas en una sola métrica balanceada, el *accurracy* representa la exactitud en porcentaje total de comentarios clasificados sobre el conjunto de datos y el *support* que muestra cuantos comentarios están en cada clase . Los resultados obtenidos son un *accuracy* y un F1-score de 0.59, con una precisión general de 0.60, un valor bajo del umbral aceptado de 0.70, y un *recall* de 0.59 tal y como se observa en la Tabla 3; lo que indica que el modelo de árboles de decisión logra clasificar más de la mitad de los comentarios, aunque este tiene un margen considerable de mejora.

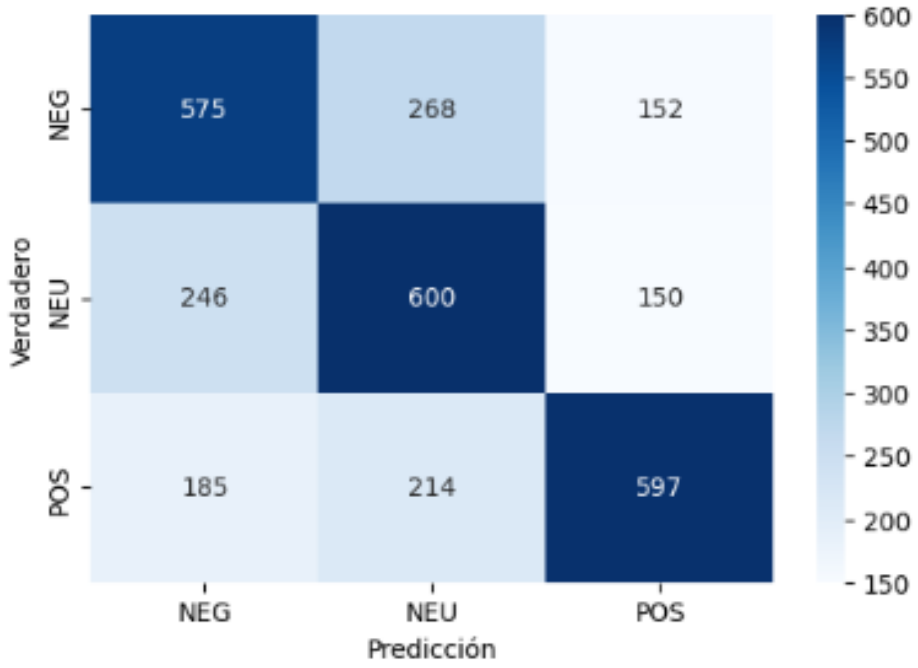
Tabla 3

Métricas Árboles de decisión

Clase	***Precision***	***Recall***	**F1-score**
NEG	0.57	0.58	0.57
NEU	0.55	0.60	0.58
POS	0.66	0.60	0.63
Total	0.60	0.59	0.59

Figura 11

Matriz de confusión árboles de decisión



Nota. Elaborado por Antonio Proaño

- ***Random Forest:*** En el caso del *random forest* se siguieron los mismos pasos, por lo que los resultados fueron los siguientes: una precisión de 0.68, un *recall* de 0.67 y, por ende, un F1-score y *accuracy* de 0.67 como se puede observar en la Tabla 4, además de la matriz de confusión en la Figura 12, un valor mayor comparado al obtenido con el modelo de árboles de decisión, pero aún no se encuentra sobre el umbral aceptable de 0.70.

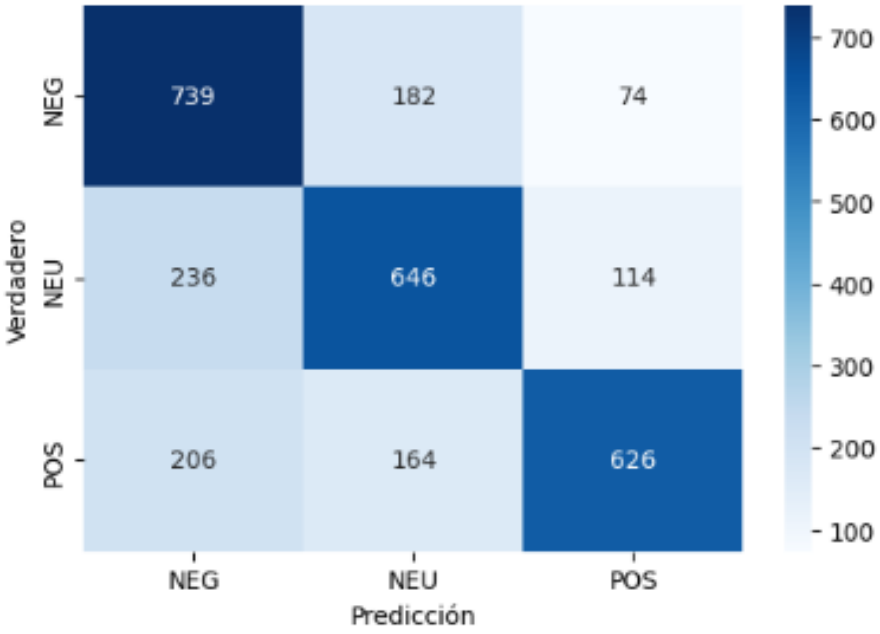
Tabla 4

Métricas Random Forest

Clase	***Precision***	***Recall***	**F1-score**
NEG	0.63	0.74	0.68
NEU	0.65	0.65	0.65
POS	0.77	0.63	0.69
Total	0.68	0.67	0.67

Figura 12

Matriz de confusión Random Forest



Nota. Elaborado por Antonio Proaño

- **SVM:** A diferencia de los dos modelos anteriores, este utiliza "SVC" (Support Vector Classifier) para trabajar directamente con las frecuencias del TF-IDF, por lo que, con los datos ya normalizados y representados en un espacio vectorial de alta dimensionalidad, busca el hiperplano óptimo que separe las clases mediante una frontera basada en los puntos más cercanos de cada clase. Con esto en cuenta, las métricas obtenidas fueron las siguientes: una precisión de 0.71 y un *recall* de 0.70, por lo que se tendría un F1-score y un *accuracy* de 0.70 como se ve en la Tabla 5,

además de la matriz de confusión en la Figura 13, un valor ya considerado aceptable en cuanto a clasificación, pero aún no es el más óptimo.

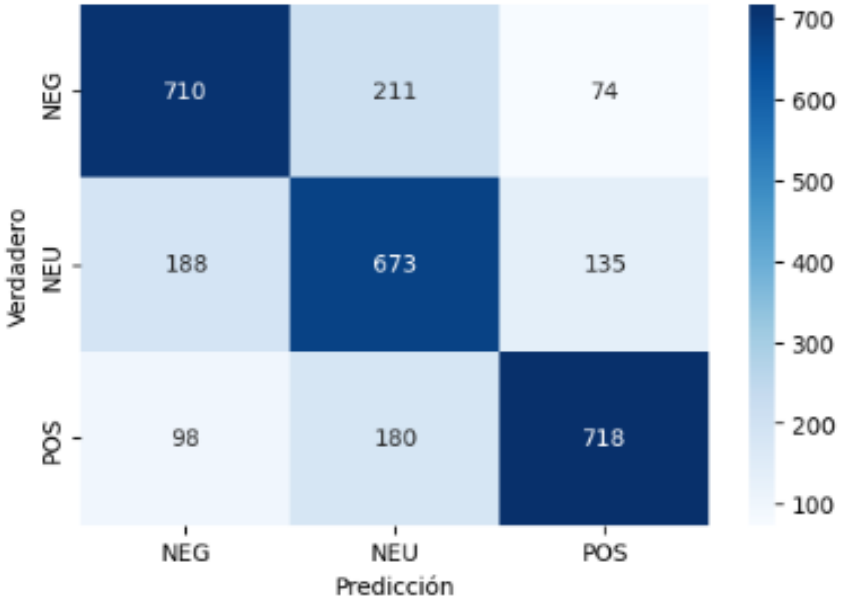
Tabla 5

Métricas SVM

Clase	***Precision***	***Recall***	**F1-score**
NEG	0.71	0.71	0.71
NEU	0.63	0.68	0.65
POS	0.77	0.72	0.75
Total	0.71	0.70	0.70

Figura 13

Matriz de confusión SVM



Nota. Elaborado por Antonio Proaño

- **Regresión Logística:** Entrenado similar a los dos primeros modelos, la regresión logística obtiene las siguientes métricas: una precisión de 0.72, un *recall* de 0.71, un *accuracy* de 0.71 y F1-score de 0.72 como se puede observar en la Tabla 6, además de la matriz de confusión en la Figura 14, por lo que, el modelo de regresión logística se encuentra con valores sobre el umbral de 0.70, siendo un modelo óptimo para su uso. Si bien estos resultados posicionan a este modelo como una alternativa sólida, la selección del modelo final no debe basarse únicamente en los modelos tradicionales, ya que existen enfoques más recientes y avanzados que podrían ofrecer un mejor rendimiento.

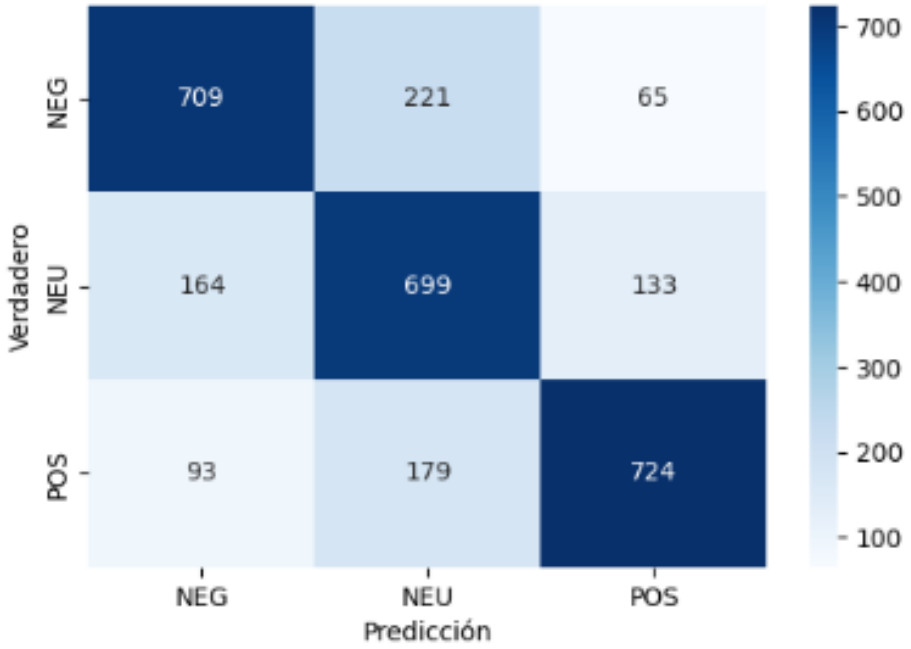
Tabla 6

Métricas Regresión Logística

Clase	***Precision***	***Recall***	**F1-score**
NEG	0.73	0.71	0.72
NEU	0.64	0.70	0.67
POS	0.79	0.73	0.75
Total	0.72	0.71	0.72

Figura 14

Matriz de confusión Regresión Logística



Nota. Elaborado por Antonio Proaño

4.3.3. Modelado con BERT usando pysentimiento

A diferencia de los modelos clásicos anteriores, según lo trabajado por Ingeniare (2021) y por Frogames formación (s.f.), BERT necesita una preparación de datos previa como se puede observar en la Figura 15, donde la clasificación de los sentimientos se convierte en valores numéricos, siendo POS representado con 0, NEU con 1 y NEG con 2, transformando los textos en tokens y usando la variable de español, ya que BERT cuenta con la capacidad de trabajar con varios idiomas. Posteriormente los datos se organizan en “DataLoaders”, que son

estructuras que dividen el conjunto de datos en lotes de muestras de forma ordenada, evitando cargar todos los datos en memoria simultáneamente al momento del entrenamiento, por lo que adicionalmente se vuelve a separar a los datos de entrenamiento y de prueba con los mismos porcentajes de 80% y 20%.

Figura 15

Tokenización con BERT

```
# BERT necesita etiquetas numéricas (0, 1, 2)
# Mapeamos POS=0, NEU=1, NEG=2
etiqueta_map = {'POS': 0, 'NEU': 1, 'NEG': 2}
etiqueta_inv = {0: 'POS', 1: 'NEU', 2: 'NEG'}

# Tomamos los textos y etiquetas del df clasificado
textos_bert = df['comentarios_limpios'].tolist()
etiquetas_bert = df['sentimiento'].map(etiqueta_map).tolist()

# División entrenamiento / prueba
X_train_bert, X_val_bert, y_train_bert, y_val_bert = train_test_split(
    textos_bert, etiquetas_bert, test_size=0.2, random_state=42
)

print(f"Entrenamiento: {len(X_train_bert)} muestras")
print(f"Validación: {len(X_val_bert)} muestras")


# Pipeline de preparación de datos para BERT
def preparar_datos_bert(df, etiqueta_map={'POS': 0, 'NEU': 1, 'NEG': 2}, test_size=0.2):
    textos = df['comentarios_limpios'].tolist()
    etiquetas = df['sentimiento'].map(etiqueta_map).tolist()
    X_train, X_val, y_train, y_val = train_test_split(
        textos, etiquetas, test_size=test_size, random_state=42
    )
    return X_train, X_val, y_train, y_val

pipeline_preparar_bert = Pipeline([
    ('preparar', FunctionTransformer(preparar_datos_bert, validate=False))
])

X_train_bert, X_val_bert, y_train_bert, y_val_bert = pipeline_preparar_bert.transform(df)
```

Nota. Elaborado por Antonio Proaño

El entrenamiento es realizado mediante fine-tuning para aprovechar el conocimiento lingüístico que BERT ya adquirió durante su pre-entrenamiento sobre grandes volúmenes de texto, por lo que únicamente se ajustará sus pesos para la tarea específica de clasificación de sentimientos, lo que permite que se capturen mejor las relaciones contextuales a diferencia de

los modelos tradicionales. Para este ajuste se utiliza el optimizador “AdamW”, que funciona como un mecanismo de corrección progresiva en donde en cada paso del entrenamiento calcula qué tan equivocado estuvo el modelo y determina la dirección de ajuste a partir de ahí. La tasa de aprendizaje obtenida es de 2e-5, lo que controla que tan grandes son los ajustes en cada paso, por lo que es un valor bajo que muestra una distorsión casi nula en el conocimiento de BERT, de esta manera siendo usada como se puede observar en la Figura 16.

Figura 16

Épocas modelo BERT

```
#Entrenamiento del modelo BERT
optimizer = AdamW(modelo_bert.parameters(), lr=2e-5)

# Trbaja con 3 épocas es suficiente para fine-tuning
EPOCHS = 3

for epoch in range(EPOCHS):
    modelo_bert.train()
    total_loss = 0

    for batch in train_loader:
        input_ids, attention_mask, labels = [b.to(device) for b in batch]

        optimizer.zero_grad()
        outputs = modelo_bert(input_ids=input_ids, attention_mask=attention_mask, labels=labels)
        loss = outputs.loss
        loss.backward()
        optimizer.step()

        total_loss += loss.item()

    avg_loss = total_loss / len(train_loader)
    print(f"Época {epoch + 1}/{EPOCHS} - Loss promedio: {avg_loss:.4f}")
```

```
Época 1/3 - Loss promedio: 0.7054
Época 2/3 - Loss promedio: 0.4599
Época 3/3 - Loss promedio: 0.3017
```

Nota. Elaborado por Antonio Proaño

El entrenamiento es realizado durante 3 épocas, donde cada época representa un recorrido completo por todo el conjunto de datos, siendo suficiente para el fine-tuning por su base sólida, obteniendo una precisión general de 0.78 y un *recall* de 0.77, por lo que BERT alcanza un *accuracy* y un F1-score de 0.77 tal y como se muestra en la Tabla 7, superando el umbral de 0.70 y teniendo una mejora sustancial de desempeño frente a los modelos tradicionales, y podemos corroborar la clasificación en la matriz de confusión en la Figura 17.

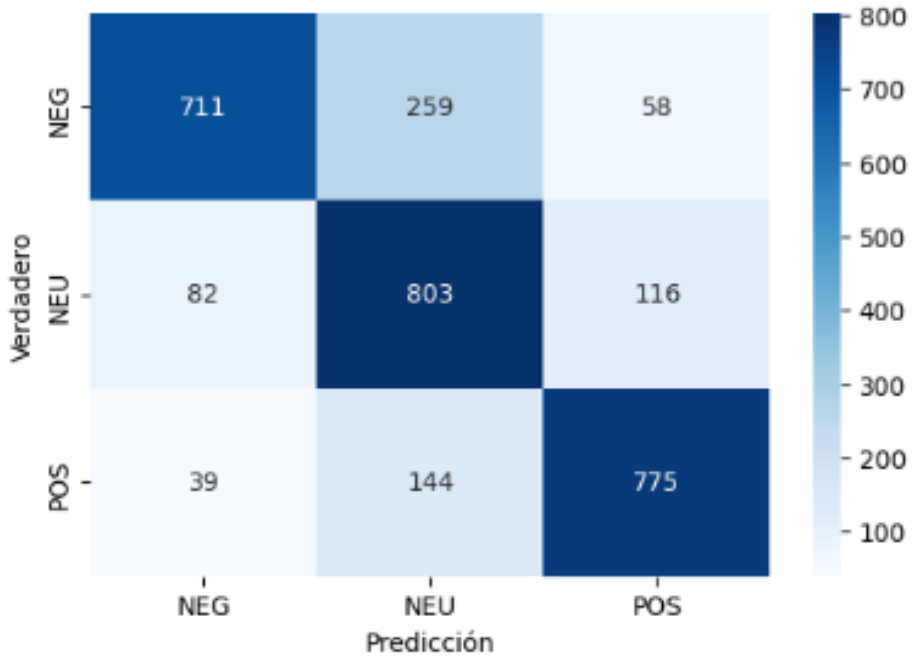
Tabla 7

Métricas BERT

Clase	***Precision***	***Recall***	**F1-score**
NEG	0.85	0.69	0.76
NEU	0.67	0.80	0.63
POS	0.82	0.81	0.81
Total	0.78	0.77	0.77

Figura 17

Matriz de confusión BERT



Nota. Elaborado por Antonio Proaño

4.4. Fase 5: Evaluación de resultados

4.4.1. Comparación de modelos y selección del modelo final

Una vez obtenidas las métricas de todos los modelos, en la fase 5 de CRISP-DM se realiza una comparación de los resultados obtenidos, haciendo un énfasis especial tanto en el F1-score como en el *accuracy*, así mostrando todos estos en un *grid* comparativo para poder seleccionar el más adecuado para el trabajo.

En esta gráfica en la Figura 18 y en la Tabla 8, se pudo observar que el modelo con métricas más bajas fue el de árboles de decisión, con un F1-score y un *accuracy* de 0.59; el resto de modelos con métricas ente 0.67 y 0.71, un rango aceptable, pero con posibles mejoras, y el que mejores métricas presenta sería BERT, con un F1-score y un *accuracy* de 0.77, lo que demuestra un mejor rendimiento constante en todas las métricas para el entendimiento del contexto y significado de las palabras dentro de los comentarios, ya que este no analiza cada termino de manera aislada como los modelos tradicionales que únicamente se basan en TF-IDF, interpretando tambien relaciones entre las mismas dentro la oración, tal y como lo hace pysentimiento, y gracias a su arquitectura *Transformer* entrenada previamente con grandes volúmenes de datos con fine-tuning tiene un mejor entendimiento de los casos atípicos o patrones lingüísticos mucho más complejos, por lo que el modelo escogido será BERT, usado con el fin de comprender lenguaje natural de manera estable para este caso de estudio sobre la violencia institucional.

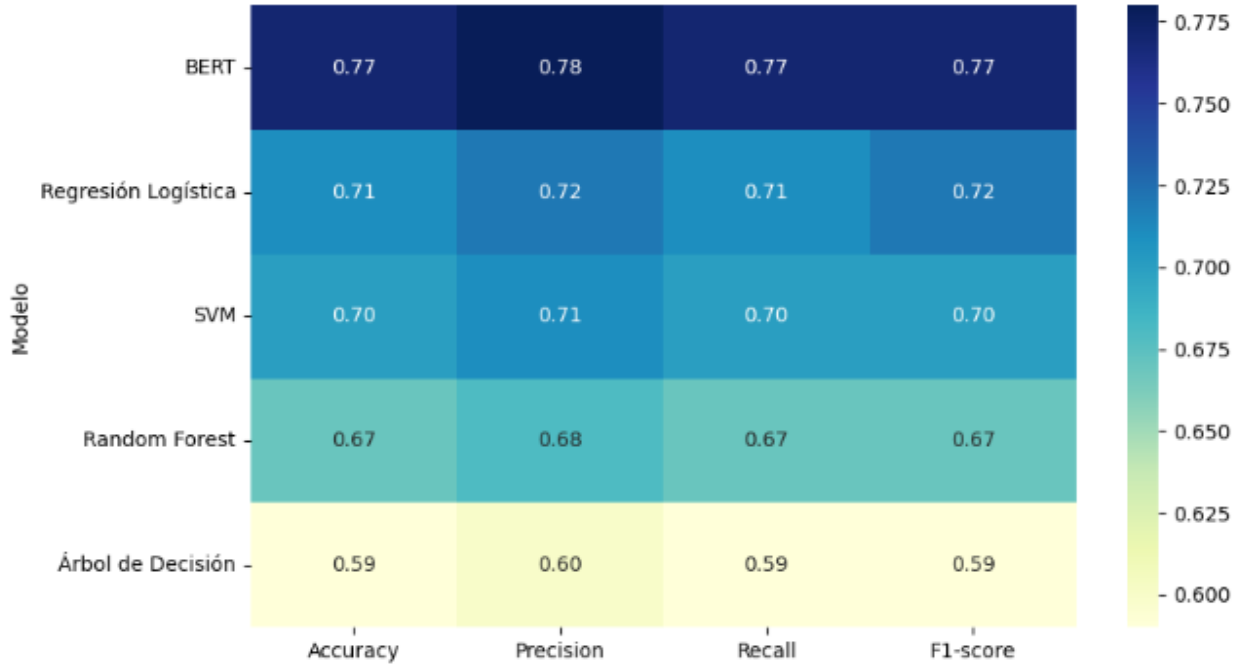
Tabla 8

Métricas modelos

Modelo	***Accuracy***	***Precision***	***Recall***	**F1-score**
BERT	0.77	0.78	0.77	0.77
Regresión Logística	0.71	0.72	0.71	0.72
SVM	0.70	0.71	0.70	0.70
Random Forest	0.67	0.68	0.67	0.67
Árboles de decisión	0.59	0.60	0.59	0.59

Figura 18

Grid Comparativo de Modelos



Nota. Elaborado por Antonio Proaño

Al escoger BERT como el modelo para el trabajo, se utilizó nuevamente el *dataset* no clasificado, esto con el fin de clasificar los comentarios con BERT usando pyentimiento, y, una vez clasificados 33301 comentarios se obtienen que son: 14044 negativos, 12833 neutros y 6424 positivos tal y como se puede observar en la Tabla 9, catalogando a diferencia del uso de solo pysentimiento muchos más neutros y positivos al entender los casos atípicos como el sarcasmo como se muestra en la Figura 19, algo muy común en comentarios sobre la violencia institucional debido a su naturaleza.

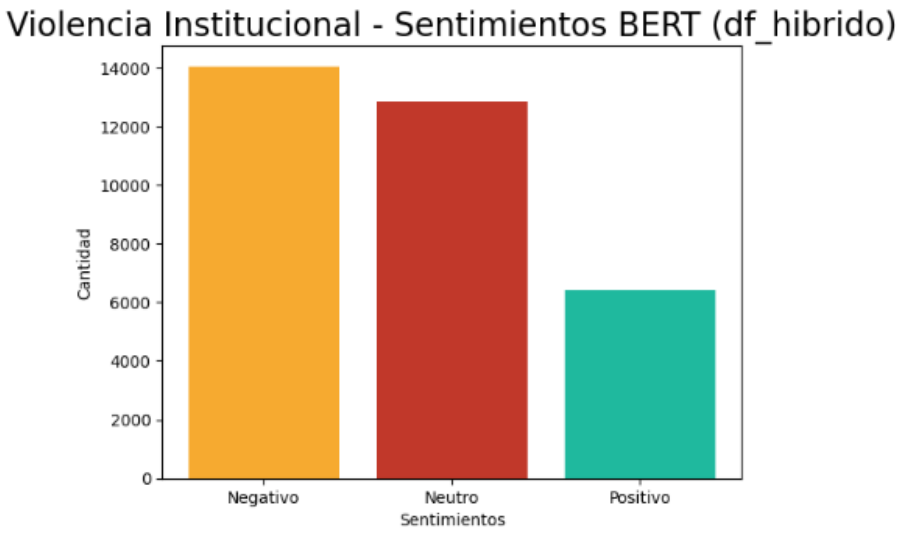
Tabla 9

Clasificación BERT+pysentimiento

Sentimientos	**Cantidad de comentarios**
Negativo/NEG	18606
Neutro/NEU	9717
Positivo/POS	4978
Totales	33301

Figura 19

Distribución de sentimientos BERT+pysentimiento




```
# Muestra el DataFrame df_hibrido con la columna de clasificación BERT añadida
df_hibrido
```

	comentarios	comentarios_limpios	sentimiento
0	like si te pidieron hacer una reflexión de esto	like si te pidieron hacer una reflexión de esto	NEU

Nota. Elaborado por Antonio Proaño

Adicional, con comentarios sintéticos como se observa en la Figura 20, se hizo una revisión del modelo para observar la existencia de un mal entrenamiento o de anomalías, pero este mostró ser funcional al clasificar correctamente los comentarios sintéticos, y con una revisión adicional a lo planteado en la fase 1 de CRISP-DM se pudo confirmar que el modelo cumple con lo propuesto con una predicción sobre el 75%.

Figura 20

Prueba datos sintéticos BERT+pysentimiento

```
# Pruebas
prediccion_bert("La policía abusó de su poder y nadie hizo nada")
prediccion_bert("Espero que las instituciones mejoren algún día")
prediccion_bert("Es una situación complicada para todos")
prediccion_bert("Bien me parece que los maten a esos corruptos")
prediccion_bert("Dios los bendiga")
```

```
Frase: La policía abusó de su poder y nadie hizo nada
Sentimiento predicho: NEG
Frase: Espero que las instituciones mejoren algún día
Sentimiento predicho: NEU
Frase: Es una situación complicada para todos
Sentimiento predicho: NEG
Frase: Bien me parece que los maten a esos corruptos
Sentimiento predicho: NEG
Frase: Dios los bendiga
Sentimiento predicho: POS
'POS'
```

Nota. Elaborado por Antonio Proaño

4.4.2. Comparación con revisión manual, LLMs y pysentimiento

Además de la comparación entre los distintos modelos, se realizó una evaluación complementaria mediante una revisión manual de los datos y el uso de LLM, esto con el fin de corroborar la funcionalidad de pysentimiento y posteriormente del modelo de BERT usando pysentimiento. Sin embargo, para aplicar estos métodos, especialmente la revisión manual, es

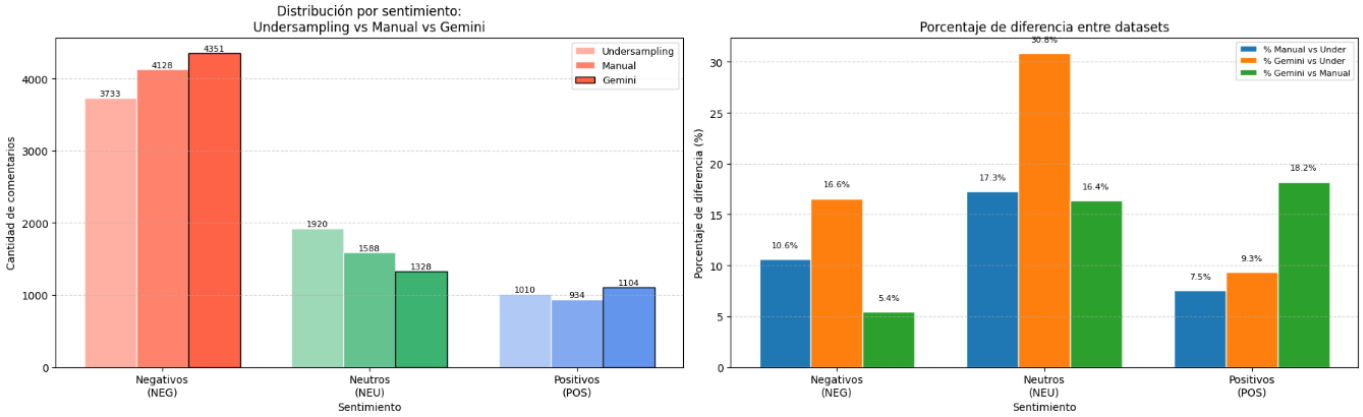
necesario contar con un conjunto de datos más manejable, ya que éste cuenta con un gran volumen de comentarios en el *dataset* original, y revisar cada comentario de forma individual manualmente habría requerido un esfuerzo bastante considerable y un aumento innecesario en el tiempo de trabajo, por lo que se aplicó un *undersampling* al conjunto de datos sin clasificar, obteniendo una muestra de 6806 comentarios, equivalente al 20% del tamaño del *dataset* original.

Una vez realizado el *undersampling*, el conjunto de datos resultante fue clasificado de tres maneras: analizado mediante diferentes modelos LLM, entre ellos "ChatGPT", "Claude" y "Gemini", todos con resultados similares, por lo que para la comparativa se seleccionó el modelo "Gemini 1.5 Flash" bajo un enfoque "Zero-Shot Sentiment Analisis", una técnica que permite clasificar el texto sin entrenar previamente al modelo con ejemplos, escogido debido a su capacidad de comprender el contexto y por su rápido procesamiento. La siguiente revisión fue realizada de manera manual, clasificando uno por uno todos los comentarios del dataset reducido, y, por último, se revisó al clasificar con pysentimiento.

Al comparar las distribuciones entre estos tres métodos en la Figura 21, se observó que las diferencias porcentuales entre cada clase fueron inferiores a un 50%, y en general menor a un 15%, por lo que esta similitud evidencia el funcionamiento correcto de pysentimiento.

Figura 21

Comparación pysentimiento vs Manual vs Gemini

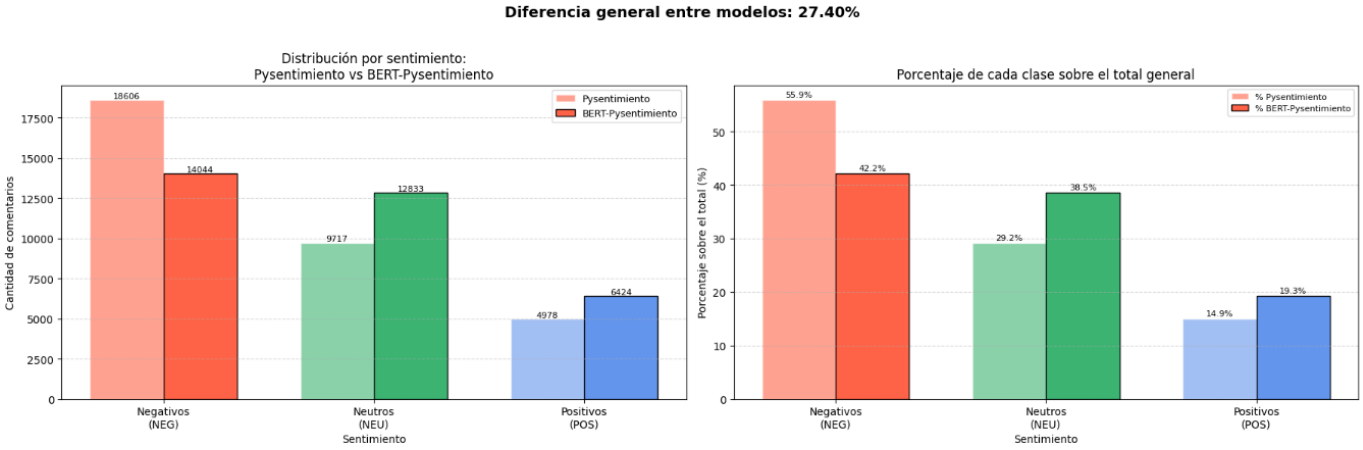


Nota. Elaborado por Antonio Proaño

Una vez realizada esta comprobación, se buscó comparar la clasificación general de pysentimientos con la del modelo entrenado de BERT usando pysentimiento, entendiendo que en este contexto hubo sesgos en los modelos previos al no tener en cuenta sarcasmo, o en el caso de la revisión manual fue el tener ciertas ideologías, se puede observar en la Figura 22 que el modelo híbrido de BERT usando pysentimiento clasifica de mejor manera los comentarios, con una diferencia porcentual general de 27,4 %, lo que muestra una mejora con el modelo entrenado ante pysentimieno, y por ende, ante los modelos anteriormente mencionados y la revisión manual.

Figura 22

Comparación pysentimiento vs BERT+pysentimiento



Nota. Elaborado por Antonio Proaño

4.5. Fase 6: Estadísticas dinámicas del modelo final de predicción

4.5.1. Mapeo y gráficos de comentarios

Con la clasificación mediante el modelo entrenado de BERT con pysentimiento, se buscó representar la data y sus estadísticas de forma visual para facilitar su interpretación. Para ello, se incorporaron dos tipos de visualizaciones: nubes de palabras y bigramas, ya que estas permiten identificar términos recurrentes y asociaciones frecuentes dentro de los comentarios, lo que permite relacionar los gráficos con el estudio de la violencia institucional. Adicional a esto, se realizó un proceso para no tomar en cuenta palabras comunes como conectores, conjunciones y preposiciones entre palabras con el fin de observar las palabras más relevantes

para el estudio. Sin embargo, este proceso no puede ser realizado en los *dashboards* de Power BI debido a que en esta herramienta se busca tener el menor sesgo en las gráficas.

Para comenzar, se generaron nubes de palabras específicas para cada una de las clases, las cuales resultan especialmente útiles para el análisis y por ende serán las que se usaran para generar *dashboards*, ya que permiten observar los términos más frecuentes en cada clase, mostrando con un tamaño mayor aquellas palabras que aparecen con más frecuencia en los comentarios, de esta manera, estas visualizaciones facilitan la identificación de los temas y conceptos predominantes en los comentarios clasificados como positivos, negativos y neutros.

En el caso de la nube de palabras de comentarios positivos en la Figura 23, en este estudio se obtuvieron palabras con mayor repetición en el tema de violencia institucional como “bien” y “policía”. Esto sugiere que, dentro de los comentarios existe una tendencia a destacar favorablemente el accionar de la policía con medidas relacionadas a la protección y de todo bien accionar de ciudadanos ante casos de violencia, por lo que la recurrencia de estos términos indica que se observa de manera positiva todo accionar o intervenciones orientadas al mantenimiento del orden, protección y justicia ante estos casos.

Figura 23

Nube de palabras comentarios positivos BERT-pysentimiento



Nota. Elaborado por Antonio Proaño

Para la nube de palabras de los comentarios negativos en la Figura 24 se obtuvo que, por la naturaleza del *dataset* trabajado gran parte de los comentarios hace referencia al contexto de la violencia dada dentro de los países, así como a la percepción de que las acciones institucionales no siempre son suficientes para enfrentar estas problemáticas, además de pedir justicia por toda violación a la seguridad dada por delincuentes dentro de cada caso, reflejando preocupaciones con la seguridad nacional y polémica política, un tema realmente relevante en la actualidad.

Figura 24

Nube de palabras comentarios negativos BERT-pysentimiento

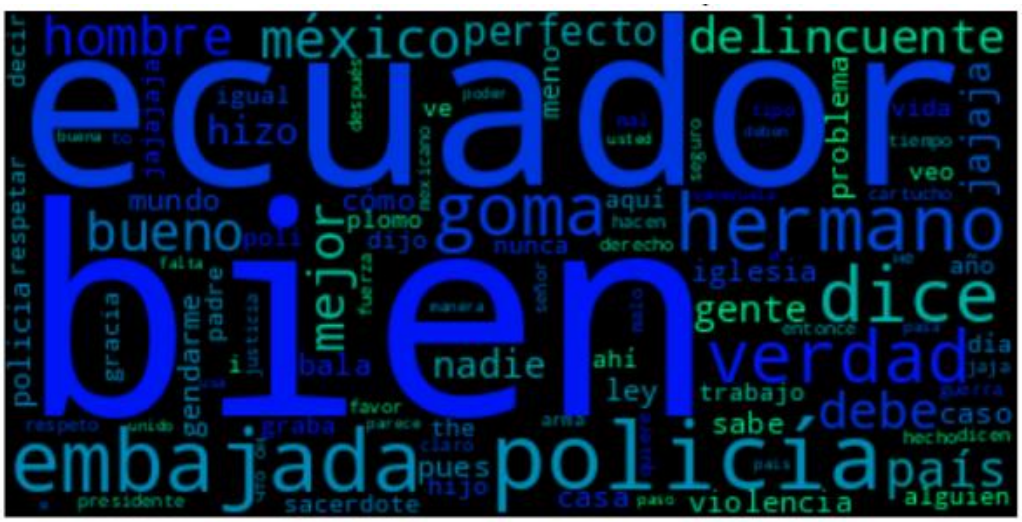


Nota. Elaborado por Antonio Proaño

La última nube de palabras muestra a todos aquellos comentarios catalogados como neutros, observando en la Figura 25 que los comentarios en esta clase se centran en describir las situaciones, acontecimientos y contextos relacionados a la violencia institucional, especialmente en Ecuador, sin expresar una postura ante los temas, manteniendo un tono predominante informativo o descriptivo de las diferentes situaciones, incluso comentarios sarcásticos al verse imparciales ante los temas.

Figura 25

Nube de palabras comentarios neutros BERT-pysentimiento



Nota. Elaborado por Antonio Proaño

En cuanto al análisis de bigramas, este se realizó sobre el *dataset* completo y no de manera separada por clases, con el objetivo de identificar las combinaciones de palabras más frecuentes dentro de los comentarios. Debido a que los comentarios neutros representan un 38,5% del *dataset*, algunas de las asociaciones más recurrentes reflejan expresiones coloquiales o satíricas utilizadas dependiendo de la polémica, pero, por otro lado, entre las combinaciones más usadas se encontraron casos como “iglesia católica” o “derecho internacional”, lo que evidencia una presencia significativa de referencias a instituciones o temáticas y sugiere que pueden ser problemáticas muy comunes vinculadas a discusiones sobre la violencia institucional.

4.5.2. Visualización con Power BI

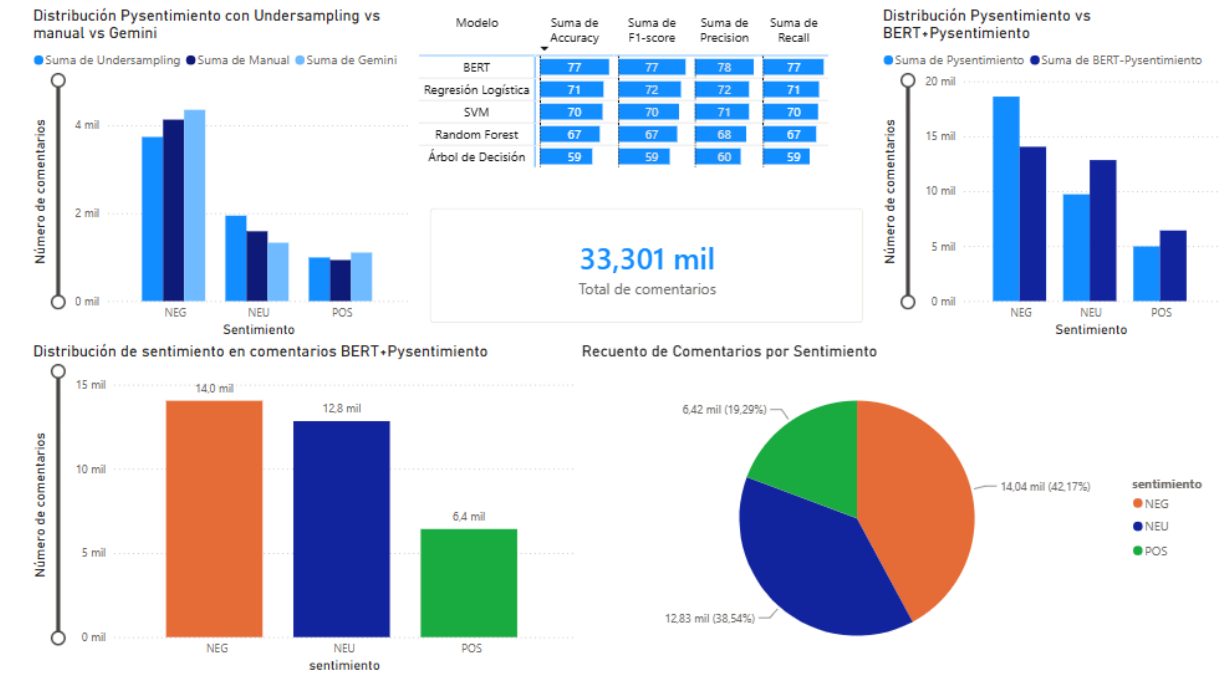
En la fase 6, tras conclusiones generales del archivo, de la elaboración del plan de trabajo posterior y la definición de *pipelines* con un flujo secuencial alrededor del trabajo con fines de una automatización hipotética futura, se generaron diversos archivos “CSV” que recopilan los datos necesarios para la construcción de *dashboards* dinámicos. Estos archivos contienen la información correspondiente a las gráficas consideradas más relevantes para el análisis, las cuales son: la comparativa de las métricas obtenidas de todos los modelos, la comparativa del funcionamiento de pysentimiento, la comparativa de pysentimiento contra el modelo híbrido de BERT, la distribución de la clasificación usando el modelo entrenado, las nubes de palabras con el modelo y el análisis de los bigramas, adicionalmente se incluyeron varias versiones diferentes del *dataset* hechos a lo largo del trabajo.

Una vez generados y descargados los archivos, se utilizó Power BI para la construcción de los *dashboards* finales, debido a su capacidad para desarrollar visualizaciones dinámicas e interactivas orientadas al análisis empresarial, la ciencia de datos y la inteligencia de negocio. Esta herramienta permite que empresas, instituciones o personas independientes, siendo este el caso, puedan interpretar la información mediante enfoques de análisis para la toma de decisiones o para el aprendizaje en general, facilitando la extracción de valor de los datos a partir de los resultados obtenidos en el caso de estudio.

El *dashboard* fue estructurado en dos páginas principales: La primera se enfocó en la visualización del rendimiento de los modelos evaluados como muestra la Figura 26, permitiendo comparar métricas del modelo entrenado y analizar su desempeño, mientras que la segunda visible en la Figura 27 estuvo destinada a la exploración del conjunto de datos, por lo que en esta se incorporó el complemento “Word Cloud” con el fin de obtener una representación similar a las nubes de palabras generadas previamente en el cuaderno de Google Colab. Con esta etapa se completaron todas las fases propuestas en CRISP-DM, concluyendo de esta manera el desarrollo integral del proyecto.

Figura 26

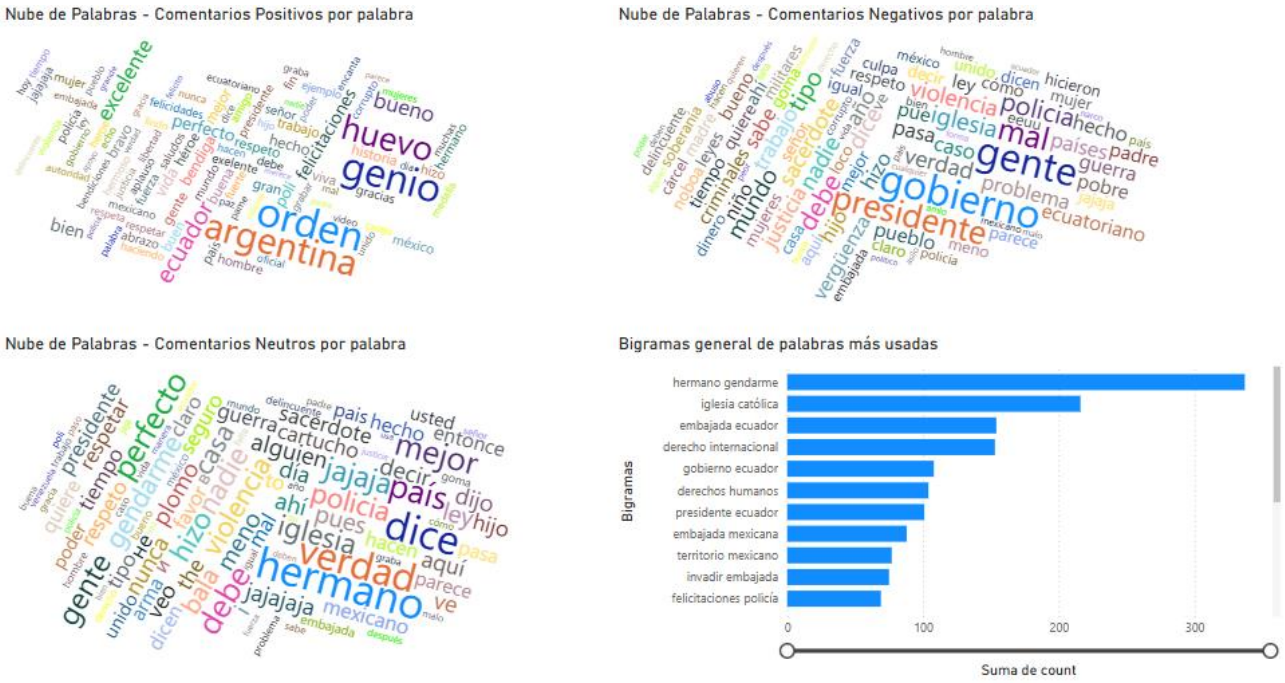
Dashboard de visualización del rendimiento de los modelos evaluados



Nota. Elaborado por Antonio Proaño

Figura 27

Dashboard de visualización del conjunto de datos



Nota. Elaborado por Antonio Proaño

5. CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Los resultados obtenidos demuestran la factibilidad de desarrollar sistemas automatizados y confiables para el análisis de sentimientos a partir de modelos más especializados, como BERT debido a sus métricas satisfactorias con un F1-score y un *accuracy* de 77%, y que los comentarios obtenidos de Youtube, al ser usados para entrenar estos modelos, ayudan a entender de mejor manera el fenómeno de la violencia institucional, esto al observar los resultados presentados en los *dashboards* por medio de una herramienta dinámica, siendo en este caso Power BI.

El uso de la metodología CRISP-DM permitió organizar el proyecto de forma estructurada y con capacidad de volver a cualquiera de los pasos anteriores si había la necesidad de algún cambio de enfoque o alguna corrección, además de poder observar si el planteamiento inicial del proyecto está siendo cumplido en las fases finales, desde la definición del problema hasta la presentación de resultados, facilitando un proceso claro y replicable.

Se puede observar que, en los modelos evaluados, BERT presentó un mayor desempeño al superar a los algoritmos clásicos como SVM o *Random Forest*, por lo que este

tipo de algoritmo LLM, considerado una red neuronal, está más preparado para tareas como el análisis de sentimientos, debido a su capacidad para comprender el contexto de las palabras y manejar expresiones complejas por su arquitectura transformer.

La comparativa entre el etiquetado manual, el realizado por un LLM como Gemini y por pysentimiento evidencia que existen resultados consistentes, validando el uso de estas herramientas, aunque el modelo entrenado con BERT muestra ventajas al no tener un gran sesgo a diferencia del resto, debido a las herramientas con las que trabaja para el entendimiento completo del contexto, un error que cometen todo el resto de metodologías para el análisis de sentimientos.

Por último, en cuanto a la data tratada, el análisis de más de 33 mil comentarios por medio de BERT entrenado con pysentimiento mostró que según la naturaleza de la problemática predominan los sentimientos negativos, representando cerca del 56% de los datos, lo que refleja una percepción mayoritariamente crítica frente a la problemática, siendo plataformas y redes sociales como Youtube espacios para poder criticar este tipo de comportamiento institucional y para dar una voz a todas las personas, sean victimas de los casos o sea simplemente para dar su opinión, con esto, al automatizar herramientas para el análisis de sentimientos podemos entender qué temáticas tienen relevancia dentro de la violencia institucional, los patrones emocionales dentro de los casos y se puede tener un accionar adecuado ante estos incidentes.

5.2. Recomendaciones

Se recomienda que, en un futuro, los resultados obtenidos en este trabajo sean automatizados dentro de sistemas, sea el caso de que se use por personas independientes o por empresas, para el análisis y comprensión más profunda del fenómeno de la violencia institucional. Los patrones emocionales identificados, la predominancia de sentimientos negativos y términos más frecuentes en cada clase constituyen información valiosa que puede orientar a investigadores, instituciones académicas y organizaciones sociales a facilitar la toma de decisiones informadas y el diseño de estrategias de intervención efectivas.

Debido a la naturaleza abierta del trabajo y a la automatización futura que se le puede dar al mismo, con el objetivo de mejorar la robustez del modelo, también se recomienda el ampliar el conjunto de datos, incorporando una mayor cantidad de comentarios de Youtube, incluyendo videos de distintos casos nuevos que puedan ocurrir en un futuro y que marquen una diferencia dentro del estudio de la violencia institucional.

Además, como una extensión posible de la alimentación del modelo, se recomienda integrar maneras de trabajar con otras redes sociales adicionales, como “X”, “Instagram” o “TikTok”, esto con el fin de ampliar las fuentes de las opiniones analizadas ya que cada plataforma cuenta con diferentes dinámicas para comunicarse y con diferentes reacciones ante el tema según los grupos de cada uno de estos espacios.

Se recomienda que, tal y como se hizo en este trabajo, según lo requiera el proyecto de ciencia de datos, análisis o investigación se adopte la metodología de CRISP-DM. Su organización en fases bien definidas y su flexibilidad para adaptarse a diferentes casos fueron de gran ayuda tanto en el desarrollo por mi parte como en el análisis del mismo por medio del archivo de Google Colab, ya que su facilidad de lectura y documentación la convierten en una herramienta altamente transparente y útil.

Finalmente, se recomienda el uso de este modelo en diversos casos para extender su campo, entrenando más allá del ámbito de la violencia institucional, aprovechando su capacidad de clasificar los sentimientos de textos y de entender su contexto. Dado que el modelo de BERT fue ajustado mediante el uso de pysentimiento y del fine-tunning con datos ya etiquetados, el modelo cuenta con el potencial para ser aplicado en el análisis de opinión en diversas aéreas de interés, sin limitar su posible uso y siendo una herramienta reutilizable de alto valor para proyecto de análisis de lenguaje natural.

6. BIBLIOGRAFÍA

Ajayi, D. (2020, 3 de diciembre). *Cómo los modelos BERT y GPT cambian las reglas del juego para el PLN*. IBM Think Insights. https://www.ibm.com/es-es/think/insights/how-bert-and-gpt-models-change-the-game-for-nlp

Al Rojo Vivo. (2022, 16 de marzo). *Video: Denuncian a gerente por maltrato laboral en call center* [Video]. YouTube. https://www.youtube.com/watch?v=EluQvduRJqM

Beltrán Beltrán, N. C., y Rodríguez Mojica, E. C. (2021). *Procesamiento del lenguaje natural (PLN) - GPT-3: Aplicación en la ingeniería de software*. Tecnología Investigación y Academia, 8(1), 18–37. https://revistas.udistrital.edu.co/index.php/tia/article/view/17323

Bergmann, D. (2024, 15 de marzo). *¿Qué es el fine-tuning?* IBM Think. https://www.ibm.com/es-es/think/topics/fine-tuning

Bio Bio. (2025, 18 de febrero). *Contraloría anuncia fiscalización a polémico curso sobre "violencia institucional"* [Video]. YouTube. https://youtu.be/BVyJ6IgUoso

Bio Bio. (2025, 17 de octubre). *Revelan que 1 de cada 3 delitos cometidos en estallido social fue la violencia institucional* [Video]. YouTube. https://youtu.be/wgw5HLzfugI

Bouman, E. (s.f.). *youtube-comment-downloader* [Repositorio de software]. GitHub. https://github.com/egbertbouman/youtube-comment-downloader/blob/master/README.md

Cabrera, C. (s.f.). *El Escuadrón de la Muerte en Ecuador* [Video]. YouTube. https://www.youtube.com/watch?v=q0DVmYW_MVY

CNN en Español. (2024, 15 de agosto). *CIDH condena "violencia institucional" del régimen de Maduro* [Video]. YouTube. https://youtu.be/fpy99qqR2bw

CNN en Español. (2024, 6 de abril). *Así fue la entrada de la Policía de Ecuador a la embajada de México para detener a Jorge Glas* [Video]. YouTube. https://www.youtube.com/watch?v=_UQOH5wDfB0

Crónica TV. (2025, 6 de agosto). *Fue detenido por un robo y apareció muerto... ¿violencia institucional?* [Video]. YouTube. https://youtu.be/3egMsFXuMvY

CuriosaMente. (2018, 23 de diciembre). *¿Por qué existe la violencia?* [Video]. YouTube. https://youtu.be/v0uaGnCJTu4

Díaz Ramírez, J. (2021). *Aprendizaje automático y aprendizaje profundo*. Ingeniare. Revista Chilena de Ingeniería, 29(2), 180. https://www.scielo.cl/scielo.php?pid=S0718-33052021000200180&script=sci_arttext

El HuffPost. (2025, 29 de septiembre). *Un repartidor logra escapar de los militares en Chicago tras protestar contra ellos* [Video]. YouTube. https://www.youtube.com/shorts/C5N3raXAWaw

El Muro. (2011, 30 de junio). *La escalera de la violencia* [Video]. YouTube. https://youtu.be/ziIU2V_2C2A

El Universal. (2018, 5 de noviembre). *Policías golpean a preso en Tijuana* [Video]. YouTube. https://www.youtube.com/watch?v=4lckAruKJKs

El Universal. (2018, 24 de octubre). *Exhiben a policías de Chimalhuacán golpeando a detenido* [Video]. YouTube. https://youtu.be/OAcqsVRSr5c

Escobar Tovar, J. (2025, 15 de abril). *Hay que acabar con la violencia institucional de los jueces; se corrompen con mucha facilidad* [Video]. YouTube. https://youtu.be/sEv3c_g2nVA

Frogames Formación. (s.f.). *Cómo usar BERT para análisis de sentimientos*. https://cursos.frogamesformacion.com/pages/blog/26-052-bert-para-analisis-de-sentimientos

Google. (s.f.). *Google Colaboratory*. https://developers.google.com/colab

Grupo Fórmula. (2025, 19 de diciembre). *Detención violenta en Ciudad de México* [Video]. YouTube. https://www.youtube.com/shorts/l8s_Qj__-DE

IBM. (s.f.). *Aprendizaje supervisado vs. no supervisado: ¿cuál es la diferencia?* IBM Think. https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

IBM. (2021, 17 de agosto). *Conceptos básicos de ayuda de CRISP-DM*. IBM Documentation. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

IBM. (2021, 15 de octubre). *¿Qué es el sobreajuste?* IBM Think. https://www.ibm.com/es-es/think/topics/overfitting

IBM. (s.f.). *¿Qué es random forest?* IBM Think. https://www.ibm.com/mx-es/think/topics/random-forest

IBM. (s.f.). *¿Qué es la regresión logística?* IBM Think. https://www.ibm.com/es-es/think/topics/logistic-regression

IBM. (s.f.). *¿Qué es el submuestreo?* IBM Think. https://www.ibm.com/es-es/think/topics/downsampling

IBM. (s.f.). *¿Qué es el sobremuestreo?* IBM Think. https://www.ibm.com/es-es/think/topics/upsampling

Kavlakoglu, E. (2021, 2 de noviembre). *¿Qué es un árbol de decisión?* IBM Think. https://www.ibm.com/es-es/think/topics/decision-trees

Kavlakoglu, E. (2023, 12 de diciembre). *¿Qué es una máquina de vectores de soporte?* IBM Think. https://www.ibm.com/es-es/think/topics/support-vector-machine

LA NACION. (2023, 24 de febrero). *Un policía disparó cartuchos de goma a jóvenes para dispersar un enfrentamiento* [Video]. YouTube. https://www.youtube.com/shorts/5cAgZrEPn1Y

López Condori, J. J., y Gonzales Saji, F. O. (2021). *Análisis de sentimiento de comentarios en español en Google Play Store usando BERT*. Ingeniare. Revista Chilena de Ingeniería, 29(3), 557–563. https://doi.org/10.4067/S0718-33052021000300557

Marketing Trending. (s.f.). *Google Colaboratory (Colab): Guía completa en español.* https://www.marketing-trending.com/google-colaboratory-colab-guia-completa-espanol/

Martín Villalba, C., Urquía Moraleda, A., y Rubio González, M. Á. (2021). *Lenguajes de programación* (2.ª ed.). UNED.

Microsoft. (s.f.). *Microsoft Power BI*. https://www.microsoft.com/es-es/power-platform/products/power-bi

Microsoft. (s.f.). *Tutorial: Integración de Análisis de texto y Power BI*. Microsoft Learn. https://learn.microsoft.com/es-es/azure/ai-services/language-service/key-phrase-extraction/tutorials/integrate-power-bi

Murel, J., y Kavlakoglu, E. (2023, 22 de diciembre). *¿Qué es bag of words?* IBM Think. https://www.ibm.com/think/topics/bag-of-words

N+ UNIVISION. (2024, 11 de marzo). *Expulsan a sacerdote 'influencer' que denunció abuso sexual: la iglesia dice que no confirmó su caso* [Video]. YouTube. https://www.youtube.com/watch?v=qj6zVD-WFbs

Noticias Caracol. (2026, 13 de enero). *Así fue minuto a minuto del operativo que llevó a la captura de Nicolás Maduro en Venezuela* [Video]. YouTube. https://www.youtube.com/watch?v=ICBc8mY6YaA

Perez, J. M., Furman, D., Alemany, L., y Luque, F. (s.f.). *pysentimiento: Una biblioteca de Python para análisis de sentimientos y tareas de PLN en español* [Repositorio de software]. GitHub. https://github.com/pysentimiento/pysentimiento

Podemos. (2025, 10 de diciembre). *Violencia institucional racista | Irene Monter* [Video]. YouTube. https://www.youtube.com/shorts/Gjs8nOZMMT8

Python. (s.f.). *Python Software Foundation.* (s.f.). Python.org. https://www.python.org

Rompeviento TV. (2020, 18 de febrero). *El feminicidio infantil de Fátima: omisiones y violencia institucional* [Video]. YouTube. https://www.youtube.com/watch?v=fxWEYYRqBkA

Rueda, N. (24 de enero de 2023). *Violencia institucional y el Estado como agresor.* Ámbito Jurídico. https://www.ambitojuridico.com/noticias/columnista-impreso/constitucional-y-derechos-humanos/violencia-institucional-y-el-estado

Urra Canales, M., y Jana Aguila, R. (s.f.). *Minería de datos y análisis de opiniones.* Universidad y Sociedad. https://rus.ucf.edu.cu/index.php/rus/article/view/2476

VOS PODÉS, EL PODCAST! (2024, 21 de agosto). *Fui abus4d4 por 5 soldados del ejército (con: Daly Parra)* | *Vos podés el podcast - EP 138* [Video]. YouTube. https://www.youtube.com/watch?v=FkkvsdryHXI&list=PLnJ9QY-iRWoPyZ6e1qeuUm7-ACDejgj7_&index=3

7. ANEXOS

Anexo A. Dataset sin procesamiento en Kaggle

El conjunto de datos descargado de los comentarios de Youtube presentes en la previa investigación, sin aplicar procesos de preprocesamiento, se lo cargó y se encuentra disponible en el siguiente enlace: https://www.kaggle.com/datasets/antonioproao/comentarios-de-youtube-de-violencia-institucional

Anexo B. Datasets clasificación Manual y de Gemini en Kaggle

Los conjuntos de datos generados a partir del proceso de clasificación manual y de la clasificación automatizada mediante Gemini se encuentran disponibles en el siguiente enlace: https://www.kaggle.com/datasets/antonioproao/violencia-institucional-manual-y-por-gemini

Anexo C. Documento CRISP-DM en Google Colab

El documento correspondiente a la metodología CRISP-DM desarrollada para el presente proyecto puede consultarse en el siguiente enlace: https://colab.research.google.com/drive/17MuTHc6sXzpRyM8MtTJNwXGiBt3erZqW?usp=sharing

Anexo D. Visualización de gráficos en Power BI

Los *dashboards* y las visualizaciones elaboradas en Power BI para la exploración y el análisis de los datos utilizados en esta investigación pueden consultarse en el siguiente enlace: https://app.powerbi.com/links/1wom_Dvp0A?ctid=54474fb3-6359-40c1-b726-5d56a7dd2976&pbi_source=linkShare&bookmarkGuid=1364c18e-c033-4e12-b5f3-86d917defda8