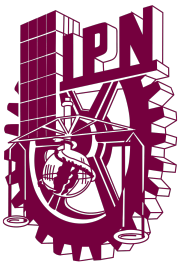

Instituto Politécnico Nacional

Centro de Investigación en Computación

T E S I S

Análisis semántico y clasificación de reportes delictivos georeferenciados

que para obtener el título de

Maestría en Ciencias de la Computación

presenta

Ing. Carolina Palma Preciado

Directores de tesis

Dra. Ana María Magdalena Saldaña Pérez
Dr. Grigori Sidorov

Ciudad de México
Noviembre 2022

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 02 de marzo del 2021

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Ordinaria No 12 celebrada el día 21 del mes diciembre de 2020, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	PALMA	Apellido Materno:	PRECIADO	Nombre (s):	CAROLINA

Número de registro: B 2 0 0 4 3 2

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Análisis semántico y clasificación de reportes delictivos georeferenciados"

Objetivo general del trabajo de tesis:

Desarrollar una metodología para el procesamiento de datos textuales provenientes de fuentes heterogéneas, que permita aplicar técnicas de PLN sobre los datos para su posterior clasificación por medio de aprendizaje automático y su visualización geoespacial. Como caso de estudio se trabajará con datos relacionados a eventos delictivos en un área de estudio determinada.

2.- Se designa como Directores de Tesis a los profesores:

Directora: **Dra. Ana María Magdalena Saldaña Pérez** 2º. Director: **Dr. Grigori Sidorov**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis

Dra. Ana María Magdalena Saldaña Pérez

2º. Director de Tesis

Dr. Grigori Sidorov

Aspirante

Carolina Palma Preciado

Presidente del Colegio

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

Dr. Marco Antonio Moreno Ibarra

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:40 horas del día 11 del mes de noviembre del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: **Centro de Investigación en Computación** para examinar la tesis titulada: **"Análisis semántico y clasificación de reportes delictivos georeferenciados"** del (la) alumno (a):

Apellido Paterno:	PALMA	Apellido Materno:	PRECIADO	Nombre (s):	CAROLINA

Número de registro: B 2 0 0 4 3 2

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 2% de similitud. **Se adjunta reporte de software utilizado.**

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** ☐ **NO** ☒ **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*
Las similitudes encontradas en el texto son mínimas y corresponden a porciones de texto debidamente referenciadas.

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** ☒ **SUSPENDER** ☐ **NO APROBAR** ☐ la tesis por **UNANIMIDAD** ☒ o **MAYORÍA** ☐ en virtud de los motivos siguientes:

Cumple con los requisitos para una tesis de Maestría.

COMISIÓN REVISORA DE TESIS

Dra. Ana María Magdalena Saldaña Pérez
Director de Tesis

Dr. Ildar Batyshin

Dra. Olga Kolesnikova

Dr. Grigori Sidorov
2° Director de Tesis

Dr. Francisco Hiram Calvo Castro

Dr. Marco Antonio Moreno Ibarra

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día **16** del mes de **Noviembre** del año **2022**, el (la) que suscribe **Carolina Palma Preciado** alumno(a) del programa **Maestría en Ciencias de la Computación** con número de registro **B200432**, adscrito(a) al **Centro de Investigación en Computación** manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección de la **Dra. Ana María Magdalena Saldaña Pérez** y el **Dr. Grigori Sidorov** y cede los derechos del trabajo intitulado **Análisis semántico y clasificación de reportes delictivos georeferenciados**, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo **c.palma.p0@gmail.com**. Si el permiso se otorga, el usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Carolina Palma Preciado

Resumen

La información sobre actos delictivos es de gran importancia para la sociedad ya que permite identificar zonas de riesgo o con un alto índice delictivo, este tipo de conocimiento no sólo proporciona información relevante sobre la seguridad de su entorno a la población, sino que también puede servir como herramienta para determinar qué zonas de la ciudad necesitan mayor atención. El presente trabajo trata datos de diferentes fuentes de información como datos abiertos y redes sociales para crear un conjunto de datos que con ayuda de técnicas de PLN y modelado geoespacial se pueda clasificar un modelo de aprendizaje automático y realizan representaciones visuales de los casos reportados mediante mapas coropléticos y de calor para identificar las zonas más afectadas por la delincuencia. Los algoritmos de línea base utilizados fueron SVM y Random Forest con diferentes representaciones de características N-gramas y *embeddings*, fue SVM con el *embedding* USE Multilingüe que se logró el mejor resultado con un valor F1 de 0.78. Sin embargo, se también se propuse el uso de modelos de aprendizaje profundo *transformer* pre-entrenados basados en BERT, los cuales alcanzaron un mejor desempeño. BETO obtuvo un valor F1 de 0.86 superando así la línea base.

Abstract

The information on criminal acts is of great importance to society as it allows to identify areas at risk or with a high crime rate, this type of knowledge not only provides relevant information about the safety of their environment to the population, but can also serve as a tool to determine which areas of the city need more attention. The present work treats data from different information sources such as open data and social networks to create a dataset that with the help of PLN techniques and geospatial modeling to classify a machine learning model and perform visual representations of the reported cases through choropleth and heat maps to identify the areas most affected by crime. The baseline algorithms were SVM and Random Forest with different feature representations N-grams and embeddings, it was SVM with embedding multilingual USE that achieved the best result with a F1-score of 0.78. However, we also proposed the use of pre-trained deep learning transformer models based on BERT, which achieved a better performance. BETO obtained an F1 value of 0.86, thus surpassing the baseline.

Agradecimientos

Este trabajo se realizó con la ayuda y apoyo de mi familia, en especial mis padres que siempre estuvieron dándome ánimos y me impulsaron a continuar con mis estudios.

A mis amigos y compañeros que estuvieron atentos y me ayudaron a lo largo de la maestría.

A mis asesores la Dra. Magdalena y el Dr. Grigori por sus comentarios, enseñanzas y atención durante el desarrollo de este trabajo.

A mi comité constituido por el Dr. Marco, Dr. Ildar, Dr. Hiram y la Dra. Olga por sus observaciones y sugerencias que ayudaron a mejorar el trabajo.

Al Centro de Investigación en Computación por la oportunidad de pertenecer y estudiar en el centro para formarme profesionalmente, así como también a los doctores que impartieron las materias cursadas las cuales ayudaron a obtener el conocimiento necesario.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por impulsar el desarrollo de la ciencia y por el financiamiento a través del apoyo económico brindado para la realización de esta investigación.

Contenido

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

Uno de los principales objetivos de las ciudades en desarrollo, es la seguridad de sus ciudadanos puesto que este aspecto social como muchos otros afectan a la población en general. Según el código penal federal de los Estados Unidos Mexicanos un delito es el *acto u omisión que sancionan las leyes penales* que se divide en tres tipos instantáneo, permanente y continuado (Cámara de Diputados, 2021). Aunque es un tema en el cual existe constante trabajo y atención por parte del gobierno es un tema recurrente ya que en muchos lugares aún no se ha podido contener.

En el presente trabajo de investigación se propone la combinación de diferentes ramas de la computación para el procesamiento de datos abiertos oficiales y datos de redes sociales que permitan conocer y visualizar los eventos delictivos que aquejan a la Ciudad de México.

El uso de diferentes procesos en conjunto, en este caso en específico Procesamiento de Lenguaje Natural (PNL), análisis geoespacial y aprendizaje automático se pueden usar para ayudar a la toma de decisiones sobre seguridad.

Al momento, aunque existen varias fuentes de datos en donde consultar sobre la delincuencia (redes sociales, periódicos, blogs, portales gubernamentales, entre otras), estas no siempre presentan de manera clara y concisa la información que se desea conocer. Es por ello que en la metodología propuesta, se analizan los datos de informes sobre delitos ocurridos en el área de estudio establecida realizados por la Fiscalía General de Justicia (FGJ) y un conjunto de tuits obtenidos durante el periodo 2021, para así estudiar información obtenida de múltiples fuentes mediante el uso de herramientas que permitan de manera general clasificar reportes y de esta forma tener un mejor conocimiento sobre los casos delictivos que están ocurriendo en un área determinada, con la intención que al integrar varios conjuntos de datos se genere un panorama más completo.

1.1 Planteamiento del problema

Actualmente en México el tema de inseguridad es uno de los aspectos sociales más relevantes. Siguiendo la premisa de permitir que las personas tengan acceso a los datos salvaguardados por el gobierno en repositorios de datos cuyo contenido refiere a aspectos de seguridad, que ahora están al alcance de los ciudadanos. sin embargo, dichos datos no son fáciles de interpretar si no se poseen conocimientos previos del tema que describen.

Uno de los aspectos que más preocupan a las personas, son los eventos criminales que ocurren a su alrededor, por lo que tener una herramienta de consulta, que de manera clara indique las diferentes áreas y puntos de peligro, representa una opción para que la población evite riesgos, asociada a que esta sea clara y confiable.

No obstante, las fuentes y herramientas de datos no siempre siguen las mismas metodologías de recolección, registro o clasificación de datos, por lo que, al realizar una integración de datos, se pueden generar inconsistencias o carecer de registros y campos, debido a que los datos son registrados con diferente nomenclatura, idioma, formato, catalogo, entre otros factores.

Por lo que, en ocasiones es necesario generar un proceso específico para extraer dichos datos de fuentes heterogéneas, y crear así un corpus a partir de diferentes fuentes, en donde se

recaben datos de recurrencia delictiva. En el caso particular del presente trabajo de investigación, se busca generar un corpus específico sobre de recurrencia delictiva que contenga casos de robo, asalto, daños, abuso, lesiones, extorsión, secuestro, homicidio, entre otros, desde una perspectiva geoespacial.

Entre las fuentes datos que resaltan útiles para la investigación, son las plataformas digitales desarrolladas por el gobierno, entre ellas se encuentran los mapas de delitos registrados del Instituto Nacional de Estadística y Geografía (INEGI), así como el portal de datos de Ciudad de México, en donde se muestran reportes de seguridad, delitos y víctimas de la FGJ de manera general (total de reportes, índice delictivo por estado y percepción de seguridad urbana), además de investigaciones científicas donde se generaran y analizan datos para detectar zonas de riesgo mediante mapas de calor.

Sin embargo, pese a los resultados obtenidos por dichas plataformas, aún quedan imprecisiones que deben subsanarse para permitir que la geolocalización de los eventos reportados mejore, así como la clasificación de los eventos descritos. Una forma de conseguirlo podría ser empleando el análisis semántico y el procesamiento de lenguaje natural en el estudio de reportes escritos que describen eventos relacionados con la inseguridad.

Por lo antes expuesto, un elemento a implementar en la presente investigación es el uso de técnicas de análisis semántico con el uso de procesamiento de lenguaje natural para clasificación de reportes delictivos, y generación de una metodología para su integración en una base de datos, que al final permitan detectar zonas o puntos conflictivos. Como caso de estudio se propone la Ciudad de México.

1.2 Justificación

En la Ciudad de México el índice delictivo es alto, tan solo en el mes de marzo del 2022 fueron registrados 20,655 reportes según la Secretaria Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP) y se iniciaron 16,786 investigaciones por la FGJ, con la iniciativa del gobierno sobre datos abiertos se propuso dar a conocer de manera libre los datos generados a partir de estos. Aunque estos datos están a disposición de la población necesitan ser procesados para su interpretación, puesto que su forma actual no brinda información concreta sobre lo que está pasando en la ciudad, de igual manera es posible que no todos los incidentes sean registrados o detectados por cada entidad, por lo que se necesitarían comparar e integrar los datos para brindar información más completa.

En este proyecto se plantea una metodología basada en el uso conjunto de procesamiento de lenguaje natural (PLN), análisis geoespacial, y aprendizaje automático para analizar datos relacionados a reportes delictivos, cuyo resultad sea útil en la toma de decisiones sobre seguridad del área analizada.

A pesar de que varias fuentes de datos en donde consultar sobre la delincuencia (redes sociales, periódicos, blogs, portales gubernamentales), estas no siempre presentan de manera clara y concisa la información que se desea conocer.

En esta propuesta se analizarán los datos de informes sobre delitos ocurridos en el área de estudio establecida, para integrar información de múltiples fuentes mediante el uso de PLN y análisis espacial que permitan de manera general tener un mejor conocimiento sobre los

casos delictivos que están ocurriendo con la intención de que, al integrar varios conjuntos de datos se genere un corpus que permita generar resultados de clasificación automática derivados de la implementación de modelos de aprendizaje automático.

1.3 Hipótesis

Es posible que mediante el uso de técnicas de minería texto se recopilen textos georeferenciados provenientes de fuentes abiertes que, con ayuda del análisis semántico y modelos de aprendizaje automático se puedan clasificar reportes delictivos, de tal forma que se produzca un conjunto de datos que permita modelar geoespacialmente los hallazgos obtenidos y poder identificar las zonas con mayor índice delictivo en el área de estudio.

1.4 Objetivos

A continuación, se presentan los objetivos que rigen la investigación.

1.4.1 Objetivo general

El objetivo general es el siguiente:

Procesar datos provenientes de redes sociales para identificar si un texto es un reporte delictivo mediante técnicas de PLN para su posterior clasificación por medio de algoritmos de aprendizaje automático y seleccionar reportes oficiales de datos abiertos para comparar patrones en los datos mediante modelado geoespacial. Como caso de estudio se trabajará con datos referentes a eventos delictivos de la Ciudad de México.

1.4.2 Objetivos específicos

Los objetivos específicos son los siguientes:

- Recolectar y analizar datos relacionados a eventos delictivos ocurridos en el área de estudio a partir de redes sociales para generar un conjunto de datos.
- Seleccionar y analizar datos abiertos para comparar los resultados de las investigaciones oficiales con los obtenidos en redes sociales.
- Procesar los datos por medio de herramientas de PLN.
- Entrenar modelos de aprendizaje automático supervisado para clasificar de forma binaria los datos en reporte y no reporte.
- Analizar geoespacialmente los datos procesados para la identificación de lugares en el que ocurren los sucesos y con ello, las zonas con mayor índice delictivo.

1.5 Alcances y limitaciones

En este trabajo la integración de información a partir de reportes delictivos obtenidos de diferentes fuentes de datos abiertos son procesados con un enfoque de procesamiento de lenguaje natural y geolocalización, con el fin de estructurar un conjunto de datos que describa las características de los eventos delictivos que ocurren en el área de estudio para la cual se propone la Ciudad de México, aunque la metodología propuesta puede ser replicada para diferentes áreas de estudios siempre y cuando se cuente con los datos.

Debido a la necesidad de estructurar el conjunto de datos a partir de información recolectada de diferentes fuentes, existen ciertas limitantes una de ellas es el proceso de limpieza, análisis y etiquetado por parte de un equipo para clasificar los datos recabados que serán usados para entrenar algoritmos de aprendizaje supervisado, el segundo son las credenciales requeridas para minar el texto y los recursos que estas proporcionen. Por último, la cantidad de datos que sean utilizables, el corpus a generar requiere que los reportes detectados estén georeferenciados lo cual puede ocasionar que gran parte de la información obtenida sea descartada.

1.6 Organización de la tesis

Este trabajo se ordena de la siguiente manera: en el Capítulo 2 se muestran los trabajos previos para dar un panorama del estado del arte respecto al tema de investigación, el Capítulo 3 se describen los conceptos, modelos y teorías los cuales forman parte del marco teórico. En el Capítulo 4 se describe la metodología propuesta y se explican las diferentes etapas que la componen. Posteriormente en el Capítulo 5 designado a la implementación se detallan los pasos realizados como la recopilación de información y los procesos que se llevaron a cabo hasta la aplicación y evaluación de los algoritmos seleccionados. El Capítulo 6 muestra en detalle los resultados obtenidos. Finalmente, en el Capítulo 7 se presentan las conclusiones y discusión con un resumen de los hallazgos, así como las observaciones y propuestas para trabajos futuros.

Capítulo 2. Estado del arte

El estudio de los casos delictivos es un tema de interés continuo, por lo que existen diferentes enfoques y metodologías que son usadas para el análisis de este problema, ya sea mediante detección de patrones, estudio de tendencia o clasificación de texto. En esta sección se presenta los trabajos relaciones al área de interés como base para el desarrollo del tema.

2.1 Información de acceso libre y redes sociales

La obtención de datos es un punto importante al realizar cualquier tipo de estudio, debido a que muchas agencias no dan a conocer la información recopilada o analizada, lo que ocasiona la necesidad de encontrar fuentes de acceso libre como: datos abiertos y algunas redes sociales. Sobre estas últimas podemos resaltar que, aunque las redes sociales no son de código abierto, permiten acceder a los datos que son generados por sus usuarios para fines de estudio, un ejemplo es Twitter.

Explorando patrones delictivos en la Ciudad de México (Exploring crime patterns in Mexico City)

Una de las aplicaciones para analizar datos de reportes delictivos, es el detectar sus patrones, en este sentido Piña-García y Ramírez-Ramírez (2019), realizaron un estudio de casos delictivos a partir de reportes oficiales elaborados por las entidades gubernamentales de seguridad pública (Datos Abiertos) y de redes sociales (Twitter y Google Trends), aunque el trabajo no manejo una metodología de integración de datos, utiliza una combinación de métodos para procesar la información de las diferentes fuentes.

Los investigadores lograron la comparación entre la información obtenida de las diferentes fuentes, para detectar su posible uso como estrategia en el cálculo de estimaciones de ocurrencia de crimen y detectar patrones espacio-temporales de datos sobre el delito.

Si bien los datos manejados en el estudio, fueron clasificados de manera manual, los autores presentaron diferentes situaciones a tener en cuenta como la metodología utilizada para manejar los datos, destacando que aunque se obtuvo acceso a los informes generados durante el periodo de 2013-2019, estos no fueron usados debido al cambio en los criterios aplicados para clasificar los reportes, por lo que solo fueron analizados los registros de los reportes policiales generados de enero del 2013 a septiembre del 2016.

De igual manera para el manejo de los datos recolectados de las redes sociales, se implementaron procesamiento de datos, los cuales se filtraron por crimen, y se omitieron los datos con información insuficiente, el corpus final obtenido de esta sección para Twitter fue de 9.7%.

En un análisis general de los datos, destaca el uso de un mapa coroplético para mostrar una aproximación de las áreas con mayor caso de reportes, haciendo énfasis en que este no hace distinción de la gravedad de los casos, pero distingue la ocurrencia de estos en el espacio, también se mencionan las posibles áreas sin delitos aparentes, que puedan ser generadas debido a la falta de información en esas zonas y no por su alta seguridad (Ver Figura 1).

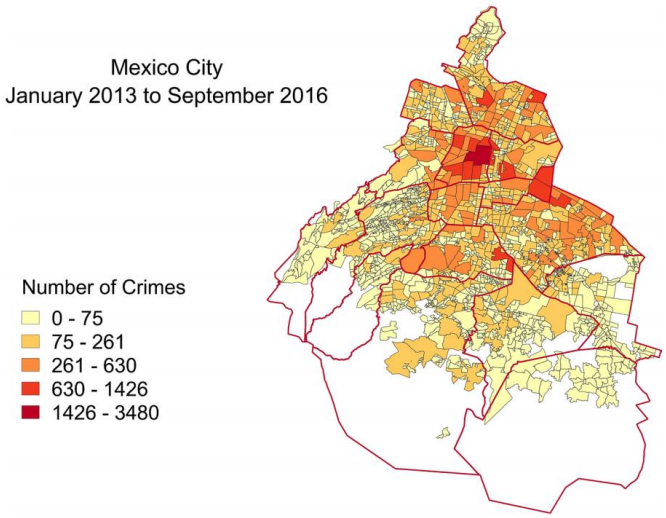

Figura 1. Número de delitos por barrios segmentados en la Ciudad de México (Piña-García & Ramírez-Ramírez, 2019).

Por último, para la predicción de patrones el trabajo utilizo modelos de análisis estadísticos como ARIMA y ARIMAX, así como métodos de análisis de regresión como LASSO. De forma general los datos obtenidos revelan que, si bien el uso de fuentes de datos como redes sociales no generan una muestra representativa suficiente, son útiles para comprender patrones sobre los delitos cometidos. Lo cual generó un indicio de la importancia de obtención de un corpus a partir de diferentes fuentes.

Mejorando la detección de organizaciones criminales en México usando ML y NLP (Enhancing the detection of criminal organizations in Mexico using ML and NLP)

Un enfoque diferente al presentado anteriormente, fue el uso de PLN y datos georeferenciados aplicado por Osorio y Beltran (2020), generaron una base de datos georreferenciada originada a partir de reportes de diferentes fuentes, entre ellas: periódicos digitales, artículos tanto nacionales como locales y de agencias gubernamentales.

Para detectar la presencia de violencia en México se creó una metodología de cuatro pasos; como primer paso se usó un rastreador web para recolectar información, después se implementaron clasificadores de aprendizaje automático que logran identificar los datos pertinentes al estudio de caso y que pasan posteriormente a un preprocesamiento para ser normalizados, el tercer paso consiste en la codificación de eventos para determinar la presencia de organizaciones criminales, esto consiste en aplicar PLN supervisado para extraer los elementos de los eventos (fuente, acción, objetivo, fecha, ubicación) y como paso final está la visualización de los datos mediante sistemas de información geográfica (SIG), lo cual permite la representación de tendencias temporales y espaciales.

La conclusión de este trabajo fue que tanto el uso de aprendizaje automático y herramientas PLN son elementos que pueden ser aplicados para la obtención y manejo de datos precisos de información geo-codificada. La base de datos final fue usada para la representación visual, en donde por medio de una aplicación se presentan análisis dinámicos de mapas de calor de territorios criminales (Ver Figura 2).

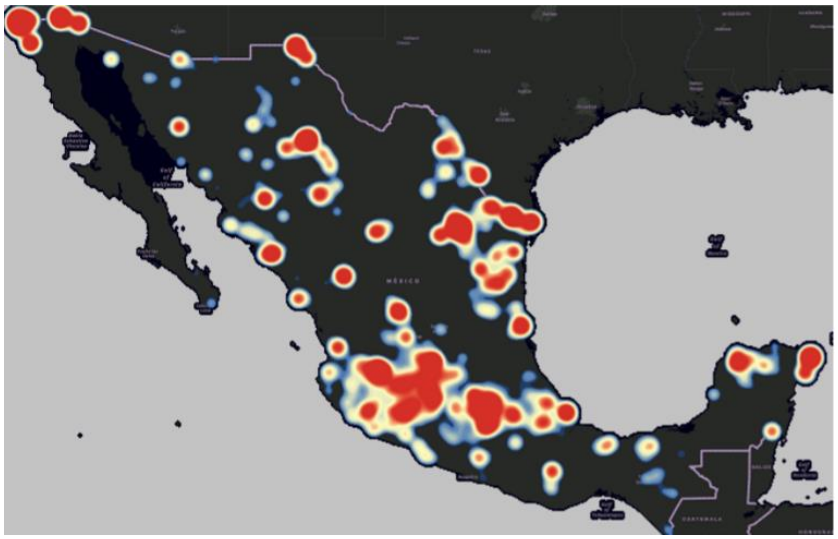

Figura 2. Distribución espacial de organizaciones criminales (Osorio & Beltran, 2020).

Propuesta de integración semántica de los datos sobre la delincuencia de la Ciudad de México (A proposal for semantic integration of crime data in Mexico City)

Por otro lado, Carrillo-Brenes *et al*. (2020), proponen un enfoque diferente para el manejo de los datos mediante una metodología ontológica (Neon) para explorar los informes de delitos y como estos se pueden asociar con otro tipo de datos. El estudio implementa los informes realizados por la fiscalía general de la Ciudad de México y la información recabada por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), mediante un proceso de disputa, el cual consiste en estandarizar datos incompletos o desarmonizados mediante diferentes etapas (identificar, extraer, limpiar e integrar).

Aunque identificaron 292 tipos de crimines, solo manejaron los más comunes con 27 categorías finales (amenazas, daño de propiedad, denuncias, tráfico de drogas, robo de vivienda, fraude, robo de conductores de vehículos, por mencionar algunos), para el manejo de los datos propusieron el uso de 15 variables sobre el delito algunas de ellas fueron: fecha, tipo de crimen, agencia, suburbio, calles, geo punto, entre otras.

La metodología usada se compone de tres partes: la construcción otológica que describe el delito junto con los datos socioeconómicos, seguido de la construcción de consultas para extracción conocimiento y por último la identificación de patrones. Los autores describen que la aplicación de este tipo de modelo permite contestar a preguntas como ¿Cuántos crímenes se han registrada?, ¿En dónde y cuándo ocurrieron? Como parte final, se creó un cubo de datos a partir de un archivo RDF, el cual contiene los datos obtenidos (Ver Figura 3), para realizar consultas representadas en mapas.

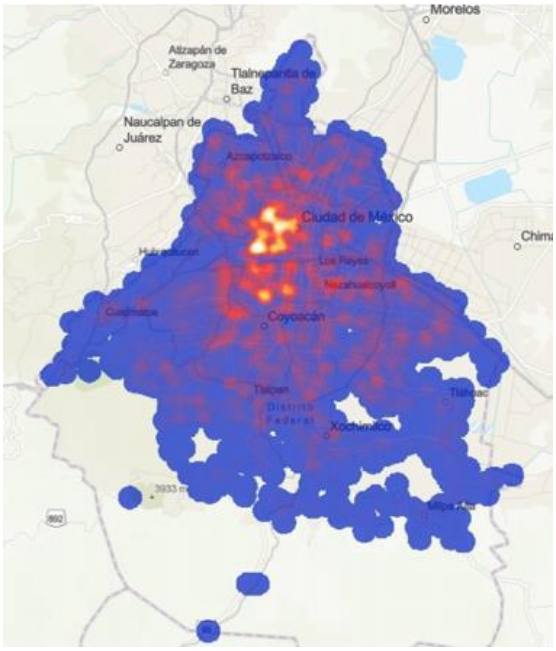

Figura 3. Distribución del crimen en la Ciudad de México (Carrillo-Brenes et al., 2020).

Análisis delictivo usando información de fuentes abiertas (Crime analysis using open source information)

El análisis de reportes delictivos sobre todo de información de código abierto puede ayudar a las agencias gubernamentales a reducir los crímenes, el estudio realiza por Nizamani *et al*. (2015), resaltan la importancia de utilizar este tipo de datos para detectar los hechos delictivos que ocurren en un área de interés. Para realizar el proceso de análisis aplicaron métodos no supervisados de minería de datos, específicamente técnicas de asociación (Apriori) y agrupación.

La metodología aplicada consiste en cuatro etapas: extracción de información requerida, preprocesar la información de acuerdo a la tarea, analizar información y extraer conclusiones. Si bien la metodología propuesta contempla el uso de cualquier información disponible de acceso libre el estudio utilizo oraciones de encabezado recuperados de una colección de comunicados de prensa del FBI en Albuquerque Ciudad en Nuevo México, se procesaron los titulares para extraer las frecuencias de términos con los cuales se generaron nubes de palabras para distinguir los términos que más destacan en las noticias.

Para el análisis de asociación se utilizaron las frecuencias de los términos anteriormente calculados para realizar un análisis de conglomerados de los términos donde devuelve los términos asociados, a partir de este estudio se creó un grupo jerárquico de términos de noticias el cual muestra la relación de los términos de las noticias. Los autores concluyeron que el aplicar este tipo de técnicas ayudan a obtener pistas sobre temas relevantes que acontecen en un área, donde los datos recopilados dan suficiente información sobre los delitos que ocurren.

2.2 Clasificación de texto de Twitter

Los estudios relacionados a la clasificación de texto para la detección de reportes delictivos están en gran parte relacionada con la clasificación de tuits ya que los recursos obtenidos de redes sociales son bastos. En particular los datos extraídos de Twitter han sido utilizados en diferentes investigaciones para tratar de detectar actos delictivos que estén siendo reportados por los usuarios en diferentes lugares.

Detección de tuits delictivos y no delictivos a través de Twitter (Detection of crime and non-crime tweets using Twitter)

El trabajo realizado por Sharma *et al*. (2018), aplican diversas técnicas de PLN en conjunto para crear un sistema capaz de recopilar datos de Twitter y preprocesar dichos datos. El sistema consta de cuatro módulos principales: recopilación de datos, minería de sentimiento, minería de datos y clasificación de salida.

Los datos recolectados provienen de Twitter API, se colectan en formato objeto lo cual ayuda a la selección de ciertos atributos importantes como: usuario, lugar, fecha, número de retuit, entre otros. Antes de iniciar la segunda etapa que consiste en minería de sentimientos los autores resaltan la necesidad de aplicar un preprocesamiento a los datos, en este caso consiste en remover las palabras auxiliares y repetidas, eliminar datos ruido (URLs, palabras cifradas, entre otros) y tokenización.

Una vez realizado el paso de limpieza de los datos aplicaron extracción de características y normalización, al final crearon un conjunto de entrenamiento y prueba denominados: Data training y train module. El algoritmo aplicado fue un clasificador bayesiano ingenuo, donde si el algoritmo identifica que el tuit está relacionado con delitos es puede ser clasificado en 4 tipos de crímenes: delito contra la persona, delito contra la propiedad, delito contra la patria, otros.

Con las técnicas utilizadas de aprendizaje maquinan con ayuda de técnicas de extracción de características de tutis por medio de PLN los autores lograron entrenar un modelo a partir de un algoritmo negación como algoritmo principal con buen desempeño, obtuvieron 93% para exactitud, 90.24% en recall y 92.50% en precisión.

Análisis y clasificación de tuits delictivos (Analysis and classication of crime tweets)

El trabajo realizado en la India por Lal *et al*. (2019), revela una correlación entre el uso de minería de texto con el uso de procesamiento de lenguaje natural y su implementación como herramienta para detección, análisis y clasificación de reportes delictivos generados de mensajes presentados en Twitter, por lo que, el implementar técnicas de PLN permite la identificación y clasificación de manera automática de *grandes cantidades de textos.*

El modelo usado está basado en minería de texto y empleo de cuatro algoritmos de aprendizaje automático para la clasificación: Naive Bayesian, Random Forest, J48 y ZeroR. Los reportes o textos de contenido a clasificar fueron delimitados a 10 tiempos de crimines: criminal, policía, fraude, crimen, robo, hurto, violencia, acoso y violación (Ver Figura 4).

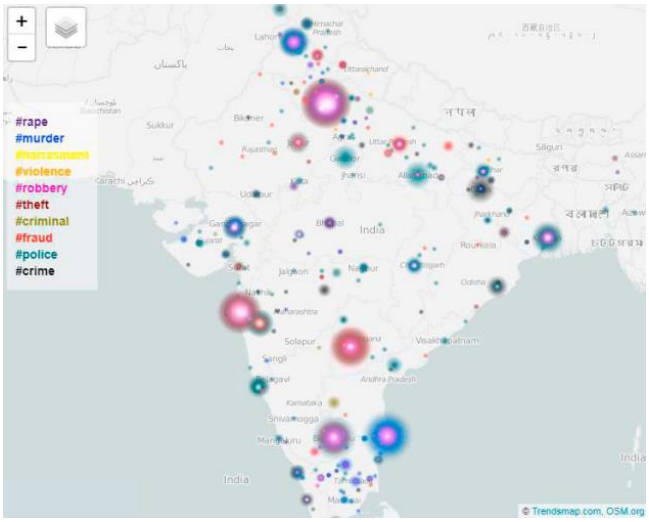

Figura 4. Mapa de calor que muestra una gran presencia de tweets relacionados con el crimen en India (Lal et al., 2019).

Para el manejo de la información, los autores destacaron la importancia del preprocesamiento de los datos al emplear técnicas de segmentación de sentencias, eliminación de palabras vacías, puntaciones, manejo de acrónimos y uso de frecuencia de términos.

El algoritmo de clasificación que obtuvo el mejor resultado fue Random Forest con 98.1%, lo que permitió llegar a dos conclusiones importantes; la primera, que el usar datos de reportes criminales de redes sociales puede ser benéfico y que la implementación de minería de texto es una opción viable para la clasificación de delitos.

Detección y clasificación de delitos a partir de publicaciones de Twitter en árabe mediante minería de texto (Detecting and classifying crimes from arabic Twitter posts using text mining)

Asimismo, AL-Saif y Al-Dossari (2018) aplican minería de texto para detectar y clasificar delitos en Twitter, el conjunto recabado incluye tuits de 2013 a 2016 obteniendo al final 8,000 tuits de lenguaje árabe el contiene dos tipos de texto la primera mitad sobre noticias de diferentes delitos y la segunda mitad con noticias de política, salud, entre otros. Para la recolección de datos se estableció una metodología de cuatro etapas, la primera contempla la recolección de datos, en la segunda se realiza el preprocesamiento y extracción de características, después se construye el modelo (aprendizaje maquina) y por último se analiza y visualizan los resultados obtenidos.

En este caso en particular, se entrenaron cuatro algoritmos de clasificación: SVM, DT (C4.5), CNB y KNN, si bien SVM alcanzo los mejores resultados con 91.55 de exactitud, las técnicas utilizadas para preprocesar los datos influyeron en el entrenamiento de los modelos. Los autores mencionan la aplicación de técnicas de extracción de características como: *stemming*, análisis morfológico y uso de n-gramas (aunque se realizaron pruebas con diferentes n-gramas de 2 a 5, el mejor resultado se obtuvo con trigramas para el entrenamiento del algoritmo SVM).

Debido a que el trabajo clasifica distintos tipos de delitos que ocurren en los países árabes se necesitó definir estos actos, en el trabajo se recalca la importancia de distinguir los tipos de

delitos que pueden ocurrir en otras partes del mundo ya que las clasificaciones pueden variar, los datos se clasificaron en 10 tipos de delitos.

Clasificación de delitos penales a partir de datos de Twitter mediante minería de reglas de asociación de clases (Classification of criminal crimes from data Twitter using class association rules mining)

Similarmente Gemasih et al. (2019) realizaron una clasificación de tuits recopilados de Twitter durante el periodo de enero-octubre de 2016. Los autores indican que la información transmitida por los usuarios de Twitter, pueden llegar a contener algo relacionado con su entorno incluyendo temas de delito, donde dicha información puede servir para clasificar y conocer tendencias de criminalidad.

Al igual que otros trabajos mencionados, este estudio aplica minería de datos con la diferencia de aplicar de manera particular Clasificadores basados en Reglas de Asociación (CARs). El método utilizado busca clasificar un texto respecto a una clase: delito contra la vida, moral, libertad de personas, propiedad/mercancía, relacionado con drogas, fraude y delitos contra la integridad física/corporal.

Si bien, se presenta un sistema para realizar el proceso de clasificación, cada etapa es descrita a detalle desde la recolección de los datos, etiquetado y preprocesamiento hasta el entrenamiento del modelo de clasificación, predicción de las clases y visualización.

El algoritmo aplicado fue Apriori, se obtuvo que para el número de reglas generadas cuanto mayor sea el mínimo valor de confianza y soporte las reglas serán menores, al revés para el caso contrario. Al final encontraron que, la precisión más alta (96%.) se obtiene con un valor de confianza mínimo de 80% y un soporte mínimo de 6%.

Análisis de la incidencia delictiva del fuero común mediante la red social Twitter

Por su parte, Hernández *et al*. (2020) recabaron tuits a partir de un término asociado o palabra clave (hashtag) con la finalidad de analizar la percepción de inseguridad e incidencia delictiva del fuero común en la República Mexicana, el trabajo se centra en crear información en forma de diccionarios y por medio de aplicación de herramientas de PLN realizaron minería de opiniones específicamente análisis de sentimientos con la librería de Python TextBlob.

Los tuits recolectados se obtuvieron durante un periodo de cuatro meses enero-abril del 2020, las autoras implementaron dos formas de recolectar los datos, la primera es usando la herramienta *Vicinitas* que permite realizar un seguimiento hashtags en Twitter y la segunda usando la librería Tweepy de Python, en total aplicaron 28 palabras de búsqueda para utilizar como filtros y mencionan que la etapa de limpieza de datos que implica seleccionar, eliminar duplicados y caracteres especiales entre otros es una parte importante del proceso.

Al final de 300,000 tuits se seleccionaron 128,000 ya que estos cumplían con las características necesarias, en el trabajo se señala que algunos tuits recopilados incluían hechos que ocurrieron fuera del país, estos no se incluyeron y solo se tomaron en cuenta los que acontecieron en los estados de México.

Si bien el trabajo se centra en el análisis delictivo de asaltos, propone una metodología para realizar la de extracción y recolección de datos a partir de Twitter. Aunque de momento solo se centra en el análisis de sentimientos se menciona una posible aplicación de Sistemas de Información Geográficos para mostrar de manera visual zonas de frecuencia delictiva.

Predicción de delitos mediante un enfoque híbrido de análisis de sentimientos basado en las representaciones del codificador bidireccional de los transformers (Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers)

El trabajo realizado por (Boukabous & Azizi, 2022) se enfoca en la importancia de seguridad inteligente la cual se encarga de recolectar y organizar datos referentes a ataques en el ciber espacio. Los autores utilizan un nuevo método hibrido basado en lexicón, el cual implica caracterizar la polaridad de los textos ya sea en negativos, positivos o neutrales y aprendizaje automático. Este método realiza principalmente un análisis de sentimiento también conocido como minería de opiniones.

El trabajo presenta una metodología de seis pasos, el primero consiste en recolectar los datos usando la API de Twitter, el segundo paso denominado enriquecimiento utilizó la librería de NLTK para dividir las sentencias en palabras (tokens) y etiquetarlos con ayuda de la categoría gramatical (parts of the speech, POS) y cinco diccionarios. El tercer paso consistió en la clasificación de los datos, usando las etiquetas obtenidas en el paso anterior se calculó el puntaje de cada sentimiento, un tuit positivo se consideró normal y un tuit negativo como relacionado a crimen.

Antes de entrenar el modelo de aprendizaje profundo se realizó un preprocesamiento de datos para limpiar los datos antes de ingresarlos al modelo. Se eliminaron URLs, números, menciones de usuarios, retuits, emoticonos, repetición de letras y palabras auxiliares. El quinto paso consistió en construir y entreno el modelo, la arquitectura usada fue un BERT. Finalmente, se utilizaron diferentes métricas para evaluar el modelo (exactitud, perdida, precisión, recall, valor F1 y matriz de confusión).

Del total de datos recolectados (70,000), 27,000 se etiquetaron como relacionados a crimen y 43,000 como normales, los autores resaltan que no fue posible recolectar datos históricos debido a las limitaciones de los permisos dados por la API. Este conjunto fue divido en 80:20 para crear el conjunto de entrenamiento y prueba, respectivamente. El modelo fue entrenado con ayuda de las librerías de Keras y Tensorflow las cuales permiten crear y entrenar modelos de aprendizaje profundo de una manera más fácil.

Los autores indican que los resultados obtenidos superan al estado del arte en la tarea de detección de delitos, ya que alcanzaron una exactitud de 94.91% y un valor F1 de 93% para la clase relacionada a crimen y 95% para la clase normal.

Capítulo 3. Marco teórico

En este capítulo se presentan los conceptos necesarios para el desarrollo de la investigación, se mencionan técnicas, herramientas y procesos implementados durante el estudio como la aplicación de minería de texto, datos abiertos, preprocesamiento de datos a partir de técnicas de PLN, entre otros.

3.1 Datos

En la actualidad la generación de datos a incrementado gracias a la cantidad de recursos disponibles en Internet y al uso de diferentes tecnologías. Esto ha causado que existan nuevas formas de recolección como: colaboración abierta distribuida, datos abiertos, redes sociales, entre otros, lo que ha ayudado enriquecer las bases de datos y con ello la creación de un conjunto de datos.

3.1.1 Datos abiertos

Los datos abiertos han sido de gran ayuda en los últimos años, puesto que brindan de manera libre y gratuita datos que pueden ser utilizados en diversas áreas. De acuerdo con Rouder (2015), uno de los puntos críticos que brinda este tipo de recursos es la disponibilidad de tener los datos en crudo (sin procesar) a disposición de cualquier persona. Asimismo, el gobierno de la República mexicana describe a los datos abiertos como información pública disponible a cualquier persona, en formato técnico y legal, de tal forma que se puedan usar para cualquier tipo de necesidad.

Existen múltiples fuentes de datos abiertos, entre ellas resultan relevantes las pertenecientes a organizaciones públicas y gubernamentales, en particular con el nuevo compromiso de los gobiernos a actuar con transparencia se han implementado compromisos para establecer un Gobierno Abierto, el cual contempla los datos abiertos.

Datos abiertos gubernamentales

Como se mencionó anteriormente la transparencia de los datos abiertos ha creado un incremento en la disponibilidad de datos abiertos del gobierno, lo que ha resultado en la creación de portales de información (Attard et al., 2015). En México, existe el portal del Gobierno de México donde se encuentran 40,727 conjuntos de datos de 280 instituciones (datos.gob.mx), también existen portales en los que se presenta exclusivamente información sobre la entidad federativa que lo proporciona, como es el caso de la Ciudad de México (datos.cdmx.gob.mx).

Entre de los sectores que cuentan con conjuntos de datos se encuentran: cultura y turismo, desarrollo, economía, educación, seguridad y justicia, salud, entre otros. La consolidación de los datos depende depende de la institución; cabe resaltar que en algunos conjuntos los datos están georeferenciados, sin embargo, representan la minoría de los conjuntos. El contar con bases de datos georeferenciados permite que estos puedan ser procesados geoespacialmente, donde la información espacial puede ser usada para realizar un análisis relacionado con lo que ocurre en un territorio (Quarati et al., 2021), que después se visualice en forma de mapa.

3.1.2 Redes sociales

El uso del Internet ha creado oportunidades para el desarrollo de aplicaciones y sistemas, en consecuencia, muchas de estas plataformas producen grandes cantidades de datos a través de publicaciones realizadas por los usuarios, entre las cinco redes sociales más usadas en México se encuentran: WhatsApp, Facebook, Instagram, Tiktok, Twitter (Statista, 2022).

Twitter

Es una red social de microblogging gratis que permite realizar intercambios cortos de manera rápida y fácil de diferentes temas, la plataforma cuenta con alrededor de 396.5 millones de usuarios en el mundo en la cual se envían aproximadamente 500 millones de tutis al día (Ltd, 2022).

Los mensajes breves publicados por los usuarios producen una gran cantidad de datos que se utilizar para realizar estudios como aquellos relacionados a comprender intereses y preferencias de un grupo de personas, organizaciones o comunidades; en muchas ocasiones estos datos facilitan el llevar acabo un análisis social (Hansen et al., 2011).

Uno de los aspectos más importantes del uso de Twitter es la estructuración de conjuntos de datos ya sea para el análisis de sentimientos, detección de crimen, análisis de opciones, entre otros. Twitter permite extraer los textos de publicaciones por medio credenciales o llaves que se adquieren en la plataforma de desarrollador, aunado a la basta documentación disponible que ayuda a crear consultas específicas a las necesidades del problema.

Si bien existen varias plataformas privadas y de uso libre para realizar la etapa de recolección, específicamente para el lenguaje de programación Python la más conocida es Tweepy, no obstante, también existe la librería Twarc. Ambas permiten realizar consultas, sin embargo, la estructura y filtros con los que cuenta cada una de ellas puede diferir.

3.2 Minería de texto

La extracción de datos de texto puede provenir de varias fuentes como lo son los sitios Web, bibliotecas digitales, servicios de noticias, redes sociales, entre otros. Aggarwal (2015) contempla tres pasos como parte del proceso de minería de texto:

> **Recopilación de datos**: esta etapa incluir el uso de distintas herramientas tanto de software como de hardware, lo cual depende del problema y los datos requeridos, los datos suelen ser almacenados en una base de datos.
>
> **Extracción de características y limpieza de datos**: después de la recolección es necesario revisar los datos extraídos, ya que en ocasiones están codificados en un formato específico y se ocupa cambiar a otro para facilitar su uso. En esta etapa se eliminan o corrigen los datos faltantes, erróneos o con ruido, se eligen los atributos relevantes para el proceso y se guardan en un conjunto estructurado.
>
> **Procesamiento analítico y algoritmos**: en la última etapa se diseñan los métodos analíticos más adecuados al problema a resolver y a los datos obtenidos en el primer paso.

Por su parte Chakraborty et al. (2013) mencionan un método de cuatro etapas, en el cual se incluyen tanto la minería de datos como el descubrimiento de conocimiento:

Recopilación de datos: como su nombre lo menciona en la primera fase se recolectan los datos textuales necesarios para el análisis.

Análisis y transformación de texto: esta etapa incluye PLN para extraer, limpiar y crear un diccionario de palabras, con las cuales se identifican entidades, eliminar palabras auxiliares, entre otras. Este proceso también considera la representación numérica del texto utilizando técnicas como análisis semántico latente (LSA), indexación semántica latente (LSI) y el modelo de espacio vectorial.

Filtrado de texto: dada la cantidad de textos que llega a contener un conjunto de datos es necesario filtrarlo para determinar cuáles son los textos relevantes, una forma de realizar este paso, es usar palabras clave. Este filtrado de términos modifica la matriz de frecuencia de término.

Minería de texto: se aplican algoritmos de agrupación, clasificación, análisis de asociación, entre otros.

3.3 Técnicas de preprocesamiento de texto

En el Procesamiento de Lenguaje Natural (NLP) existen múltiples pasos que resultan útiles en el procesamiento de los datos antes de que estos sean utilizados para el entrenamiento de los modelos, con el fin de mejorar el desempeño de los algoritmos (Krohn et al., 2019; Sidorov, 2019; Vajjala et al., 2020; Manning et al., 2008).

Tokenizar: separa el texto por palabras.

Convertir el texto a minúsculas: utilizar minúsculas ayuda a homogenizar el texto, el mantener palabras con letras mayúsculas hace que una palabra pueda ser representada de dos formas, por ejemplo, NLP y nlp, esto genera un espacio vectorial con más dimensiones.

Remover palabras auxiliares: son palabras que aportan poco significado y que ocurren con mayor frecuencia, algunos ejemplos son: de, con, en, la, sin, otro, que, entre otras.

Remover caracteres especiales: se eliminan símbolos especiales como los signos gramaticales de exclamación, paréntesis, arroba, entre otros.

N-gramas: la frecuencia de ciertas palabras suele ser alta, cuando esto ocurre y se requiere tener secuencias de palabras más significativas el aplicar n-gramas ayuda ya que ciertas secuencias aparezcan menos veces en el texto, lo cual puede ayudar a reflejar una mejor representación del texto en comparación con un los unigramas.

Stemming: se eliminan sufijos para simplificar una palabra a su forma base de tal forma que las variantes de esta palabra se representen de la misma manera.

Lematización: utiliza un vocabulario y análisis morfología de las palabras con la intención de eliminar terminaciones flexivas que cambia una palabra a su forma base (lema).

Eliminar enlaces de página web, saltos de línea, signos de puntación, entre otros: en ocasiones los textos contienen enlaces, espacios vacíos, emoticones, entre otros,

estos no aportan significado al texto por lo que en la mayoría de los casos son descartados.

3.4 Balanceo de datos

Un problema que llega a existir en los conjuntos de datos es el desbalance, esto indica que entre las clases de los datos una de ellas es considerablemente mayor a la otra u otras, dependiendo si es un caso binario o multiclase. Para resolver este inconveniente existen diferentes técnicas de remuestreo como el submuestreo, sobremuestreo o en dado caso el aumento de datos. El primero reduce la clase mayoritaria al tamaño de la clase menor, mientras que el segundo crea datos similares para aumentar la cantidad de datos finales, el tercero crea datos artificiales a partir de los datos originales con el mismo fin de ampliar la muestra de los datos (Ver Figura 5).

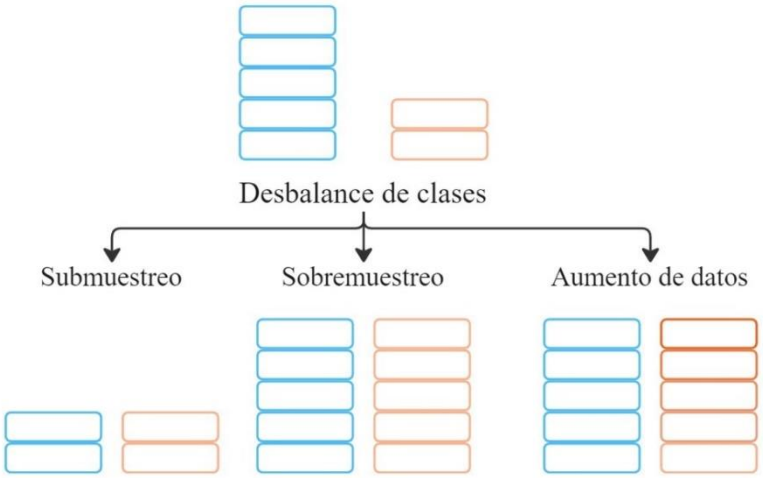

Figura 5. Técnicas para tratar conjuntos desbalanceados.

3.4.1 Aumento de datos

Aplicar procesos de aumento de los datos permite como su nombre lo indica, aumentar la cantidad de datos disponibles para el entrenamiento de manera artificial sin la necesidad de recopilar nuevos datos a través de la manipulación de los datos existentes, con la intención de mejorar el desempeño de un modelo en tareas de clasificación de texto (Tidke, 2022). Su aplicación es necesaria cuando no hay disponibilidad de grandes cantidades de datos, ya sea por el idioma, la fuente, entre otras, donde la recolección de datos requiere recursos y se convierte en una tarea laboriosa (Wong, 2021).

Existen diferentes técnicas para aplicar aumento de datos; en su trabajo Giridhara et al. (2019) explican cinco de ellas, resalta la necesidad de mantener la estructura gramatical y contextual del texto en el momento de generar nuevas sentencias automáticamente.

Palabras similares: busca palabras similares que obtengan el mejor puntaje y remplaza las palabras.

Sinónimos: se cambian palabras de manera aleatoria por un sinónimo (se limitan a sustantivos, adjetivos y adverbios), al remplazar estas palabras se generan nuevas oraciones, por ejemplo: vivienda – casa – hogar.

Interpolación: se obtiene los vecinos más cercanos de una palabra en el espacio vectorial, luego se encuentra el centroide de estos y se calcula el nuevo vector a partir del centroide obtenido y la palabra original, después de realizar el cálculo se sustituye por la incrustación original (*embedding*) y se crea una nueva lista de palabras para la oración.

Extrapolación: parecido al método anterior, se calcula el nuevo vector conforme a un centroide y el *embedding* de la palabra original.

Ruido aleatorio: a diferencia de los métodos anteriores, coloca perturbaciones a las incrustaciones de palabras (para cada palabra añadida agrega ruido gaussiano).

Por su parte Bayer et al. (2022) menciona la traducción de ida-vuelta como una forma de aumentar datos mediante modelos de traducción, donde una frase o documento es traducido a un idioma diferente al original (ida) y después se traduce de regreso al idioma inicial (vuelta), esta técnica se basa en que la traducción de texto suele variar debido a la complejidad del lenguaje, lo que crea diferentes resultados.

Si bien existen métodos son aplicados por sí mismos, se han desarrollado otras técnicas que aplican un proceso más robusto como Easy Data Augmentation (EDA), la cual utiliza cuatro procedimientos diferentes: sustitución por sinónimos, inserción aleatoria, intercambio aleatorio y eliminado aleatorio (Wei & Zou, 2019).

3.4.2 Submuestreo

El submuestreo consiste remuestrear el conjunto de datos de tal forma que se reduzca la muestra de la clase mayoritaria hasta tener la misma cantidad que la clase minoritaria. La aplicación más sencilla de esta técnica es elegir de manera aleatoria datos de la clase con mayor número de elementos hasta tener el mismo número de datos que la clase menor. También existen otras técnicas basadas en tomeks' links, centroides de agrupaciones, entre otros (Mohammed et al., 2020).

3.4.3 Sobremuestreo

Por su parte, el sobremuestreo aumenta la cantidad de datos de la clase minoritaria ya sea produciendo muestras nuevas o parecidas. Uno de los métodos más conocidos de sobremuestreo es SMOTE, utiliza la interpolación de la agrupación de las muestras para crear instancias sintéticas en lugar de utilizar la técnica de remplazo (Mohammed et al., 2020).

3.5 Inteligencia artificial

El campo de inteligencia artificial (IA) es cada vez más amplio pues su implementación es útil en diversos problemas, su importancia radica en que tareas que anteriormente se realizaban con capacidades humanas ahora pueden ser realizadas mediante el uso de software (MMC Ventures & Barclays UK Ventures, 2019). Esta área abarca varios subcampos como son el aprendizaje maquina y el aprendizaje profundo (Ver Figura 6).

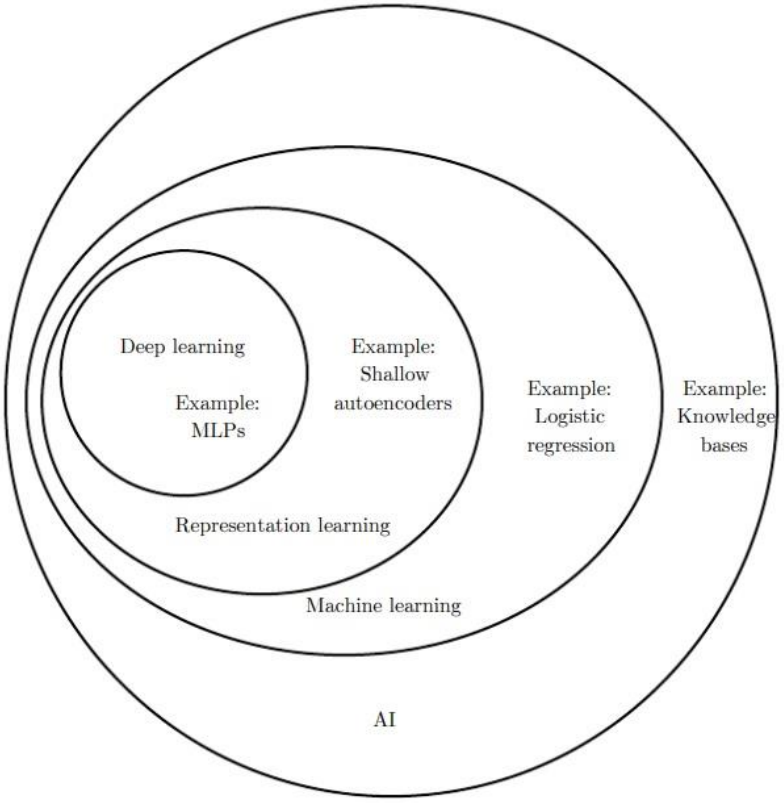

Figura 6. Diagrama de Venn de la inteligencia artificial y las ramas que incluye (Goodfellow et al., 2016).

3.5.1 Aprendizaje automático

Uno de los subcampos más utilizados dentro de IA es el aprendizaje automático, este método está basado en un enfoque basado en datos, ya que los modelos son desarrollados con base a los patrones observados en los datos (Kent & du Boulay, 2022). Los algoritmos de aprendizaje automático cuentan con diferentes componentes, como es el caso de los hiperparámetros los cuales son parámetros que deben ser definidos de manera externa (antes del entrenamiento), estos hiperparámetros ayudan al algoritmo a aproximar la función que se busca.

Básicamente este subcampo es una forma de estadística aplicada utilizada para estimar funciones complicadas y un decremento en proveer intervalos con confidencia alrededor de estos intervalos, algunas de estas implementaciones: inferencia frecuentista e inferencia bayesianas (Goodfellow et al., 2016).

Máquinas de vectores de soporte (SVM)

La máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje automático supervisado muy utilizado puesto que generalmente obtiene buenos resultados y puede ser aplicado para clasificación lineal o no lineal, regresión e incluso detección de valores atípicos. Sus principales características se basan en realizar un trazado del espacio n-dimensional para encontrar el hiperplano ideal que separe el conjunto de datos en clases.

Como se mencionó anteriormente este modelo clasifica datos linealmente separables y no lanarmente separables, el algoritmo se inicia especificando el tipo de kernel a utilizar (lineal, polinomial, función de base radial, entre otros) y sus parámetros, después se calcula el núcleo de las distancias entre los puntos de datos. En esta parte, lo principal es el cálculo de la matriz kernel o matriz de Gram $K = XX^T$, la cual es el producto escalar de los vectores originales. Para el kernel lineal se regresa K, para el polinomial de grado d se devuelve $\frac{1}{\sigma}K^d$ y para base radial (RBF) $K = \exp\left(-(x - x')^2/2\sigma^2\right)$.

El siguiente paso es el entrenamiento, el algoritmo ensambla el conjunto de restricciones como matrices a resolver y las pasa al solucionador, posteriormente identifica los vectores de soporte que son aquellos que se encuentran dentro de una distancia específica del punto más cercano y descarta el resto de datos, después calcular b^* (el cual denota el valor optimo del peso del bias).

La etapa final que es la clasificación, para un conjunto de datos Z utilizar los vectores de soporte para clasificar los datos de prueba, primero calcular el producto interno de los datos y los vectores de soporte, finalmente realizar la clasificación como $\sum_{i=1}^{n} \lambda_i t_i \boldsymbol{K}(x_i, \boldsymbol{z}) + b^*$ de esta forma devolviendo la etiqueta o el valor de la clase a la que pertenecen (Marsland, 2014).

Random forest

Es un algoritmo de aprendizaje automático su concepto está basado en que, si un árbol es bueno, múltiples arboles son mejores. Este método utiliza embolsado (bagging) y aleatoriedad a partir de un conjunto de datos, para crear un bosque se hacen distintos árboles con datos diversos. Su aleatoriedad principalmente se basa en la limitación del tamaño que el árbol de decisiones puede hacer, donde en cada nodo que se hace, un subconjunto aleatorio es dado. Lo cual ocasiona que solo se puedan elegir características del subconjunto y no de todo el conjunto.

Aunque existen múltiples variantes, de forma básica del algoritmo crea para cada de N árboles una nueva muestra Bootstrap para el conjunto de datos de entrenamiento, después utiliza esta muestra para entrenar el árbol de decisión, en cada nodo del árbol de decisión se seleccionan m características de forma aleatoria, se calcula la ganancia (índice de Gini) y se selecciona el más óptimo, este proceso se repite hasta completar el bosque. Los árboles creados después son usados para clasificar nuevos datos (Marsland, 2014).

3.5.2 Aprendizaje profundo

El aprendizaje profundo es un subcampo del aprendizaje automático el cual logra crear representaciones de los datos para su clasificación. Como se mencionaron anteriormente las técnicas tradicionales están limitadas la capacidad de procesar los datos, sin embargo, el aprendizaje profundo logra crear representaciones internas conocidas como vector de características los cuales son utilizados por los modelos para aprender y así poder identificar o clasificar patrones en los datos de entrada (LeCun et al., 2015).

Chollet (2017) define al aprendizaje profundo como una nueva forma de hacer representaciones a partir de los datos de entrada, en este método el aprendizaje sucede en las capas sucesivas que realizan representaciones cada vez más significativas. Por su parte, Krohn et al. (2019) lo define como la superposición de algoritmos simples (neuronas artificiales) en redes de varias capas.

Las redes neuronales artificiales estándar constan de muchas neuronas conectadas, cada una de las cuales produce una secuencia de activación, las neuronas de entrada se activan cuando reciben los valores de entrada, las neuronas siguientes se activan a través de las conexiones de las neuronas previas. Este tipo de aprendizaje trata de encontrar pesos que logren que las redes neuronales alcancen un comportamiento deseado.

Este campo ha logrado obtener buenos resultados en diversas tareas como lo son el reconocimiento de patrones y el PLN, especialmente en esta última área las arquitecturas de transformadores.

Transformers

La arquitectura de los *transformers* apareció por primera vez en 2017, fue propuesta por un grupo de investigadores de Google en colaboración con la universidad de Toronto. Fue creado a partir del conocimiento que los modelos con mejor rendimiento conectan un codificador y decodificador a través de un mecanismo de atención (Vaswani et al., 2017).

> **Codificador:** convierte una secuencia de entrada de tokens en una secuencia de vectores de incrustación (*embedding*). Estas incrustaciones son representaciones densas de las palabras en un espacio vectorial cuyo número de dimensiones es inferior al resultante de las palabras del vocabulario (Ekman, 2021).
>
> **Decodificador:** este módulo utiliza el estado oculto del codificador para generar una secuencia de salida en forma de tokens (Tunstall et al., 2022).

Este tipo de modelos implementan un mecanismo conocido como *autoatención* el cual permite asignar diferentes pesos (atención) a cada elemento de una secuencia de la red neuronal, los pesos indican el grado de correlación existente entre los elementos de entradas y las salidas de esta manera se el modelo se concentra en palabras específicas de una oración (Tunstall et al., 2022).

Entre las arquitecturas de *transformers* se destacan *Bidirectional Encoder Representations from Transformers* (BERT) realizado por Google y *Generative Pre-trained Transformer* (GPT) de OpenAI, la principal característica de este tipo de modelos es que ayudaron a eliminar la necesidad de entrenar arquitecturas para tareas en específico desde cero, puesto ya fueron entrenados previamente con conjuntos de datos grandes.

Anteriormente se menciona que el modelo BERT es uno de los más utilizados en tareas de PLN dado que s desempeño es bueno para distintos problemas, puesto que en un inicio BERT fue pre-entrenado con datos en inglés, para conjuntos en otros idiomas surgen otras propuestas como *Multilingual* BERT (M-BERT) que fue pre-entrenado con datos de 104 lenguajes y BETO que es específicamente para textos es español.

BERT – Multilingüe

El modelo *Multilingual* BERT (M-BERT) se basa en la arquitectura original de BERT, esta propuesta es un *transformer* de 12 capas pre-entrenado con datos obtenidos de páginas de Wikipedia con un vocabulario compartido, lo que significa que M-BERT utiliza un solo vocabulario multilingüe. La forma de transferencia sucede cuando se presentan algunas palabras durante el *fine-tuning* también se manifiestan en la evaluación (Pires et al., 2019).

Aunque se encuentra deficiencias sistemáticas entre ciertos pares de lenguajes, una posible explicación es la similitud tipológica entre los idiomas. La estructura del espacio vectorial para el modelo cambia si se aplica su parte multilingüe ya que el mapeo de una misma sentencia en dos idiomas diferentes no dependerá de la oración sino solo del par de idiomas (Pires et al., 2019).

Si bien M-BERT permite realizar clasificación de texto con conjuntos de datos multilenguaje, también se ha utilizado para la clasificación de textos de lenguajes en específico como Español, Marathi, Alemán, Swahili, combinaciones entre estos, entre otros (Velankar et al., 2022; Pàmies,2020; Dowlagar & Mamidi, 2021; Gati et al., 2021).

BETO

El modelo BETO presentado por Cañete et al. (2020) basado en el BERT aparece para cubrir la necesidad de contar un modelo que soporte el idioma español, dicho modelo esta entrenado con datos en español de Wikipedia y de OPUS Project (el conjunto de datos para el entramiento cuenta con tres mil millones de palabras).

Al igual que el original BETO cuanta con 12 capas de autoatención con 16 cabezas de atención cada una, contando con 110M de parámetros, se crearon dos versiones con datos *case* y *uncase*. A diferencia de otros modelos, para BETO se consideraron ciertas técnicas que resultaron efectivas en otros entrenamientos, una de ellas es el enmascaramiento dinámico (uso de diferentes mascaras para una misma sentencia), enmascaramiento de palabras completas, y empleo de lotes más grandes (Cañete et al., 2020).

DistilBETO

Es una variante más ligera BERT, los autores señalan que el modelo es más pequeño, rápido y barato computacionalmente puesto que reduce el tamaño del modelo en un 40% manteniendo la capacidad de comprensión del lenguaje. Introduce un nuevo método de pre-entrenado de representación del leguaje con un propósito general y utiliza pérdida triple que combina el modelado del lenguaje, destilación y perdidas por distancia coseno.

DistilBETO se basa en DistilBERT el cual utiliza una arquitectura llamada el estudiante, esta tiene tres componentes principales; es la misma arquitectura que BERT, pero las incrustaciones de tipo token y agrupado se eliminan y el número de capas se reduce en un factor de dos. Como su nombre lo indica se aplica destilación, que es destilar lotes grandes aprovechando la acumulación del gradiente al aplicar un enmascaramiento dinámico sin usar predicción de la siguiente oración (Sanh et al., 2019).

Por su parte DistilBETO es la versión pre-entrenada con un corpus en español de DistilBERT, este modelo tiene seis capas y fue entrenado en 90 mil pasos utilizando un GPU NVIDIA RTX 3090. Su comportamiento, aunque menor para algunas tareas en comparación con ALBETO y BETO obtiene buenos resultados en pruebas de rendimiento (GLUES *benchmark*) (Cañete et al., 2022).

ALBETO

ALBETO es la versión en español de ALBERT propuesto por Lan et al. (2020), trata de resolver el problema de memoria y hardware que ocurre al querer pre-entrenar modelos que tienen miles o millones de parámetros, por ello que crean ALBERT el cual proviene de *A lite*

BERT, una arquitectura más pequeña de BERT que incorpora dos técnicas de reducción de parámetros: parametrización de incrustación factorizada e intercambio de parámetros entre capas. También utilizan pérdida auto supervisada para la predicción del orden de oraciones (SOP) en lugar de predicción de la siguiente oración (NSP). Estas modificaciones permiten tener menor parámetros aun cuando se escala el tamaño del modelo.

Asimismo, Cañete et al. (2022). pre-entrenan cinco modelos de diferentes tamaños usando la arquitectura de ALBERT: ALBETO *tiny, base, large, xlarge y xxlarge*, de los cuáles el último obtiene mejor desempeño para las tareas de MLDoc, PAWS-X y XNLI que sus contrapartes BETO y DistilBETO.

3.6 Métricas de evaluación

La evaluación de los modelos de aprendizaje automático es de suma importancia, ya que se utilizan para evalúa si un modelo es bueno o malo, como se está comportando con el conjunto de datos y los hiperparámetros dados. Existen diversas formas de realizar esta evaluación siendo las más usadas: exactitud, precisión, exhaustividad y medida F1, las cuales se calculan a partir de la matriz de confusión.

Matriz de confusión

La matriz de confusión presenta de manera concisa los resultados de las predicciones de los datos de prueba en, verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativo. La Figura 7 muestra cómo se presentan los resultados en un ejemplo de clasificación binaria, los valores de color verde indican los correctos y los de color rojo los incorrectos (Marsland, 2014).

Figura 7. Formato de matriz de confusión.

Exactitud

Esta métrica se define como la suma de los verdaderos positivos y negativos divididos entre el total de datos y mide el porcentaje de predicciones correctas que realizó el modelo entrenado, a diferencia de otras métricas no se aconseja su uso de datos desbalanceados debido a que este cálculo puede no representar correctamente el desempeño de un modelo (si una clase obtiene varios positivos en comparación a otra clase con menos datos clase al sumar los positivos el puntaje será alto, lo cual no refleja el comportamiento correcto del modelo) (Marsland, 2014). La Ecuación 1 presenta la representación matemática de la exactitud:

$$Exactitud = \frac{VP + FP}{VP + FP + VN + FN}. \qquad \text{(Ecuación 1)}$$

3.6.3 Exactitud balanceada

La exactitud balanceada resuelve el problema que presenta la exactitud frente a datos desbalanceados, ya que es la media entre de la tasa de verdaderos positivos y verdaderos negativos (Guyon et al., 2015). La Ecuación 2 muestra la representación matemática de la exactitud balanceada:

$$Exactitud\ balanceada = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right). \qquad \text{(Ecuación 2)}$$

Precisión

Precisión es la relación entre los datos correctamente clasificados y la cantidad de datos positivos reales, muestra la proporción de muestras que el modelo identifico como positivos correctamente, ver Ecuación 3:

$$Precisión = \frac{VP}{VP + FP}. \qquad \text{(Ecuación 3)}$$

Exhaustividad

La exhaustividad también conocida como recuperación o sensibilidad es la relación entre el número de positivos correctos que fueron clasificados como positivos. En la Ecuación 4, se presenta la representación matemática de la exactitud:

$$Exhaustividad = \frac{VP}{VP + FN}. \qquad \text{(Ecuación 4)}$$

Medida F1

La medida F1 llama también valor-F combina la precisión y exhaustividad en una sola métrica al calcular la su media armónica. Se utiliza para comparar el desempeño entre clasificadores y datos desbalanceados ya que funciona bien para este tipo de conjuntos, ver la Ecuación 5:

$$F1 = 2\,\frac{precisión\ \times exhaustividad}{\text{precisión} + \text{Exhaustividad}}. \qquad \text{(Ecuación 5)}$$

3.7 Sistema de Información Geoespacial

Los Sistemas de Información Geoespacial (SIG) es una herramienta que permite utilizar y manejar información espacial, son esenciales en el proceso de análisis para diversos sectores como son: gobierno, logística, agricultura, salud, entre otros. Los SIG se componen de hardware, software, datos, humanos y un conjunto de protocolos organizacionales, donde todos estos componentes deben trabajar en conjunto para obtener buenos resultados. En la selección del hardware se debe tomar en cuenta la cantidad de datos que se utilizarán para realizar el análisis espacial, puesto que estos procesos suelen llevar mucho tiempo. Por su parte, el software ofrece las herramientas para administrar, manejar, procesar, analizar y visualizar información espacial (Bolstad, 2019).

Pese a que existen diferentes SIG, los más conocidos y utilizados son ArcGIS y QGIS ya que son muy capaces, Wise (2013) indica que para usar los sistemas se necesita tener conciencia de que es necesario repensar un problema de tal forma que pueda ser resuelto usando SIG.

ArcGIS

Es un Sistema de Información Geográfica lanzado en 1999 por Esri la cual es una empresa dedicada a la creación de software, mapeo y localización inteligente. ArcGIS es uno de los SIG más usados gracias a su gran documentación y soporte técnico disponible, no obstante, su desventaja es que solo esta para el sistema operativo Windows y debido a fue desarrollado bajo una licencia privada el sistema requiere del pago de una licencia para poder ser utilizado (ArcGIS Pro: GIS Professional Basic, GIS Professional Standard y GIS Professional Advanced). Entre sus capacidades se destaca: análisis basado en ubicación, herramientas para visualizar y analizar datos, crear mapas inteligentes, resolver problemas de análisis espacial, entre otros (Esri, 2022).

QGIS

Es un SIG libre fundado en 2002 por Gary Sherman bajo licencia de código abierto, el sistema fue escrito en C++ y es apto para diferentes sistemas operativos (Windows, macOS, Linux, FreeBSD, así como también Android y IOS para dispositivos móviles) (Hugentobler, 2008), su principal ventaja es que no requiere el pago de una licencia, el soporte que existe por la comunidad de voluntarios y las herramientas disponibles. Entre las funciones principales se encuentran: visualizar, gestionar, editar, explorar archivos espaciales, analizar datos, entre otros (QGIS, 2022).

3.7.1 Geovisualización

La visualización de datos geoespaciales permite crear un elemento visual ya sea en forma de mapas, gráficas, entre otros. La geovisualización también incluye la parte interactiva que se tiene al realizar pasos de análisis espacial (Çöltekin et al., 2018). Este tipo de representación facilita la interpretación de los datos a los usuarios, ya que combina diferentes áreas para crear los mapas como: cartografía, geoinformática, técnicas de visualización, entre otras (CONABIO, 2022).

Las ventajas de usar visualización geográfica son el soporte para toma de decisiones, el reconocimiento de patrones y relaciones, la visión global de un problema, la capacidad de proceso y percepción de datos, interacción entre el usuario y la información (Laurini, 2017). Existen diferentes representaciones algunos tipos de visualización son: 2D, 3D, espacio-temporal, animación, técnicas geométricas e interfaces interactivas (Nöllenburg, 2006).

3.7.2 Geocodificación

El análisis de geovisualización requiere que los datos utilizados tengan una forma de ser georeferenciados ya sea que los datos contengan coordenadas geográficas, se indique en el texto alguna dirección o pertenezcan a un área en particular que puede ser definida como un cuadro delimitador *(bounding box*). Para realizar la visualización de los datos en un mapa se quieren tener las coordenadas en formato latitud, longitud, existen nuevas técnicas que posibilitan la extracción de coordenadas a partir de un texto como lo es la geocodificación.

La técnica de geocodificación es de gran ayuda a la hora de preprocesar datos para asignar a un registro sus coordenadas, éstas después se utilizan en un SIG para visualizar los datos en

el mapa. Este proceso básicamente traduce texto (dirección, colonia, calle, código postal, localidades, países) en coordenadas geográficas numéricas (Attard et al., 2015). Existe su versión contraria el cual obtiene a partir de coordenadas una dirección, se denomina geocodificación inversa.

3.8 Herramientas de aprendizaje automático

Las herramientas son un recurso que permiten realizar tareas de forma rápida y sencilla, particularmente para el área de aprendizaje automático el uso de instrumentos que ayuden a llevar a cabo análisis e implementación de algoritmos es esencial puesto que de implementar estos procesos de forma manual puede llegar a ser costoso tanto en tiempo como recursos.

3.8.1 Python

Es un lenguaje de programación utilizado para diferentes tareas como preprocesamiento de datos, minería de texto, aprendizaje automático, entre otras. La curva de aprendizaje es baja puesto que en pocas líneas de código se logra escribir programas funcionales. Al igual que otros lenguajes Python permite hacer pruebas, depuración y reusó de código, aunado a la gran documentación disponible y el soporte de la comunidad ayudan a su fácil implementación.

Existen múltiples librerías que facilitan el desarrollo de modelos, creación de gráficas, procesamiento de datos, entre otros. Entre las más usadas se encuentran: scikit-learn, TensorFlow, Keras, Matplotlib, seaborn, Pandas, NumPy, NLTK y spaCy.

Para hacer uso de Python es necesario contar con un entorno desde el cual se pueda ejecutar el código actualmente para Windows existe la plataforma de Anaconda; en 2014 Google en colaboración con Jupyter lanzo el Colab (también conocido como Colaboratory), en donde Python se programa y ejecuta desde el navegador sin necesidad de una configuración o instalación local, además cuenta con uso de GPU y TPU.

3.8.2 Scikit-learn

Una de las librerías de Python más usadas para aplicar aprendizaje automático es Scikit-learn fue en 2007 por David Cournapeau como un proyecto de Google Summer of Code Project. La librería es de código abierto y su crecimiento depende de ayuda de la comunidad, aunque también existen instituciones públicas y privadas que ayudan a su financiamiento como Microsoft, Fujitsu, data iku, Hugging Face, entre otras (scikit-learn, 2022).

Cuenta con más de 100 algoritmos que se dividen en aprendizaje supervisado y aprendizaje no supervisado, tiene diferentes técnicas y métricas de evaluación que permiten entrenar los modelos según el método más conveniente para los datos e implementa el uso de otras librerías (NumPy, SciPy y Matplotlib) para tratar y visualizar los datos y resultados obtenidos al entrenar los modelos.

Otra parte importante es la transformación del conjunto de datos, el preprocesamiento, selección de características, reducción de dimensionalidad, entre otras. Scikit-learn permite realizar este proceso tanto con datos disponibles (conjuntos de datos juguete y mundo real) como generados.

3.8.3 Tensorflow 2

Es una plataforma de código abierto desarrollada por el equipo de Google Brain que opera en gran escala y en entornos heterogéneos, permite la creación e implementación de modelos de aprendizaje automático que debido a simplicidad ayuda a el proceso de experimentación para investigación (Pang et al., 2019), no solo cuenta con funciones de creación y ejecución de modelos si no también con herramienta de visualización TensorBoard la cual muestra graficas con evaluación obtenidas durante el entrenamiento como exactitud y perdida, también presenta grafos del modelo, histogramas de peso, entre otras.

En comparación con otras plataformas, TensorFlow es más flexible dado que realiza una abstracción de programación basada en flujo de datos permitiendo a los usuarios implementar aplicaciones en clústeres divididos, dispositivos móviles y estaciones locales. Debido a que utiliza un único grafico de flujo de datos los cálculos realizados hace que la experimentación con funciones sea más. rápida. El grafico de flujo de datos representa el cálculo y estado del algoritmo (Abadi et al., 2016), su beneficio radica en que al contener conjuntos de objetos que representan unidades de cálculo y tensores que son unidades de datos se crean grafos que son estructuras de datos que se pueden guardar, ejecutar y restaurar (TensorFlow, 2022).

3.8.4 Keras

Librería para aprendizaje profundo escrita en Python por el equipo de Keras de Google que posibilita la ejecución de modelos para realizar experimentación, su diseño este basado en torno al usuario, minimizando los pasos a realizar y reflejando de manera precisa errores que puedan ocurrir al ejecutar una acción.

Keras se describe como simple, flexible y poderosa que funciona como una API de alto nivel de TensorFlow 2 en conjunto hacen que sea adaptable tanto para el uso en industria como investigación ya que se puede implementar en una gran variedad de plataformas como servidores, navegadores, móviles (Android y IOS) y sistemas embebidos (Keras Team, 2022). Adicionalmente la librería es de código abierto, cuenta con soporte, documentación y compatibilidad con diferentes entornos (Python 3, Ubuntu, Windows y macOS). Es compatible con CPU, GPU y TPU. X|.

3.8.5 Ktrain

Es una librería de Python que ayuda a implementar de forma sencilla y fácil aprendizaje automático ya que funciona como *wrapper* para TensorFlow. Fue desarrollada por Maiya (2020) con el propósito de simplificar la construcción, entrenamiento, inspección y aplicación de modelos sin importar el nivel de aprendizaje del usuario.

Dentro de los tipos de tareas que soporta se encuentran: clasificación y regresión de texto, etiquetado de secuencias, similitud y recomendación de documentos, preguntas y respuestas, clasificación y regresión de imágenes, entre otras.

Para entrenar un modelo Ktrain solo necesita unas pocas líneas de código, primero carga y procesa los datos, después crea el modelo (este puede cargarse de los ya establecidos y soportados dentro de la librería o llamar de forma externa como lo son los modelos de Hugging Face), enseguida se utiliza una función de estimación de tasa de aprendizaje que determina el valor más optimo dado el modelo y los datos, finalmente se entrena el modelo.

Las métricas de evaluación están disponibles por medio de un reporte que incluye la precisión, exactitud, *recall*, valor F1, adicionalmente se pueden consultar los ejemplos con peor y mejor puntaje de validación.

3.8.6 Hugging Face

Es una infraestructura para inteligencia artificial establecida en 2016 por Clement Delangue y Julien Chaumond (PitchBook, 2022), usa código abierto para construir, entrenar e implementar aprendizaje máquina para JAX, PyTorch y TensorFlow.

Su principal ventaja recae en la implementación de *transformers* ya que varios modelos pre-entrenados son publicados para su posterior uso como BERT y BETO, de igual manera por medio de la API existen modelos entrenados para usar con nuevos ejemplos o aplicar en proyectos como un chatbox o traducción. Dentro de la plataforma se encuentran tutoriales y demos de como llamar un modelo, realizar la etapa de fine-tuning, compartir modelos y tokenizador, conjuntos de datos para entrenar y librerías para tareas de clasificación de texto, audio e imágenes, detección de objetos, Q&A, resumen, traducción, entre otras (Hugging Face, 2022).

Capítulo 4. Metodología

Este capítulo presenta la metodología propuesta para realizar la detección de reportes delictivos de tal manera que el proceso pueda ser repetido posteriormente, ya sea para entrenar un nuevo modelo o generar un corpus georreferenciado. El proceso comienza den la etapa de recolección y selección de datos los cuales provienen de redes sociales o datos abiertos que cuenten con un campo de georreferenciación. Al final se generan tres productos, dos conjuntos de datos uno de redes sociales etiquetado en reportes y no reportes y otro de los datos abiertos, un modelo entrenado y mapas que permitan la visualización de los resultados y sirvan para comparar las zonas delictivas e identificar si los reportes detectados en las redes sociales reflejan lo ocurrido en una zona de forma oficial (reportes realizados por el gobierno) (Ver Figura 8).

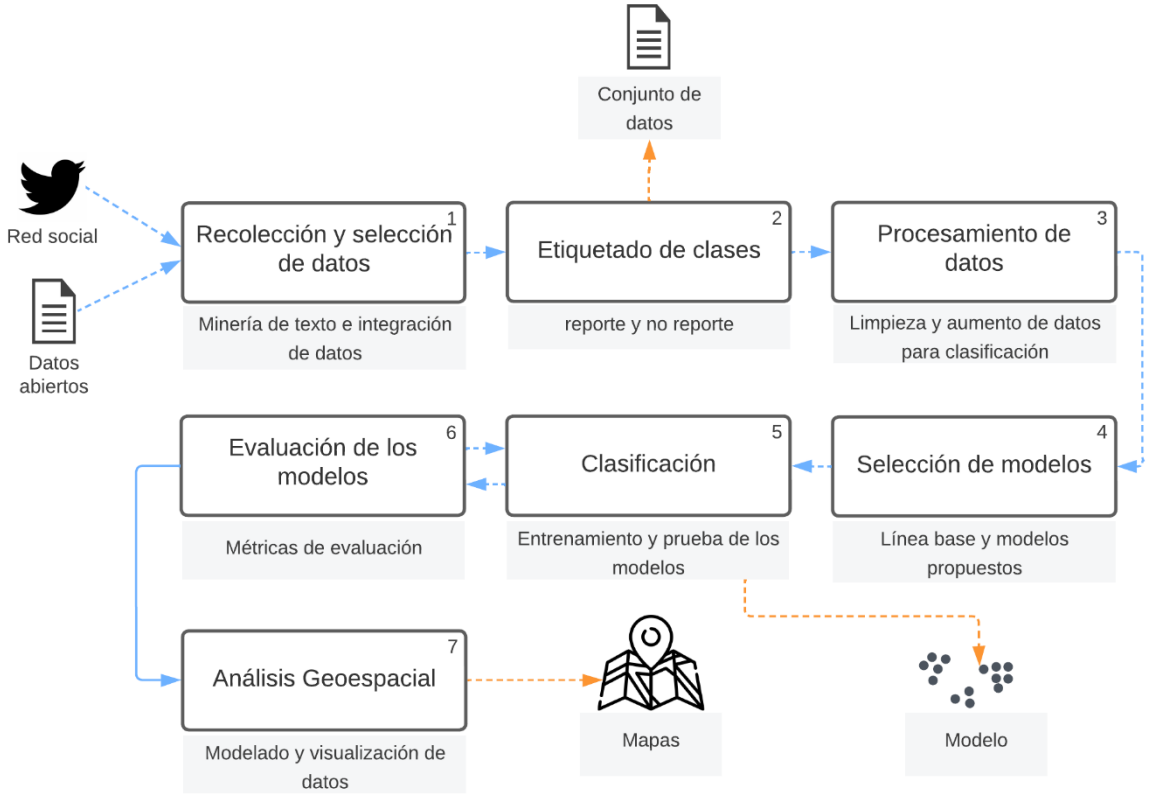

Figura 8. Metodología propuesta para la tarea de clasificación de reportes delictivos georeferenciados.

4.1 Recolección y selección de reportes delictivos

Para la recolección de datos es necesario delimitar la zona de estudio, puesto que al necesitar datos georeferenciados es importante especificar los filtros de búsqueda. También es preciso encontrar si existen conjuntos de datos abiertos disponibles de la zona seleccionada, de no existir no se podrá llevar a cabo un análisis completo.

4.1.1 Recolección de datos de redes sociales

Al igual que la zona de estudio es indispensable escoger las redes sociales de las cuales se recaudarán los datos, teniendo en cuenta cuales son las más usadas en dicho lugar y si permiten el acceso a las publicaciones de los usuarios, para este paso se recomienda el uso de Twitter.

En este proceso también se determinan los filtros y forma de consulta, puede ser de dos tipos por diccionario de palabras y hashtag o por cuentas oficiales de gobierno (ejemplo: la Ciudad de México cuenta con la Fiscalía General de Justicia y acepta reportes por medio de la red social de Twitter). Para el caso de Twitter implementar un filtro para no extraer retuits ya que los textos podrían repetirse y causar duplicados en la etapa de etiquetado.

4.1.2 Selección de datos abiertos

Los portales de datos abiertos generados por el gobierno publican bases de datos con información de investigaciones realizadas, para poder hacer un análisis comparativo entre ambos conjuntos estos datos deben ser del mismo año que los reportes recopilados de las redes sociales.

4.1.3. Procesamiento de datos

Los datos recolectados y seleccionados necesitan tener una forma de georreferenciación ya sea textualmente, en forma de coordenadas geográficas o cuadro delimitador. Este campo debe ser filtrado para contener datos que ocurrieron en el área seleccionada de no ser así se eliminan. De igual modo se eliminan datos repetidos o con poca información como tuits que contienen únicamente emojis, enlaces, menciones de otros usuarios, exclamaciones, números, entre otros.

4.2 Etiquetado de clases

Los datos recabados en la etapa uno puede o no ser reportes delictivos, es necesario identificar la clase a la que pertenece cada publicación. Para este trabajo se consideran dos etiquetas reporte y no reporte, el conjunto de datos se crea de tal forma que si un texto no es etiquetado como reporte entonces es no reporte. Solo se aplica para el conjunto de datos de redes sociales (los datos abiertos suelen ser información categorizada de investigaciones previas que ya fueron clasificados en tipos de delitos).

4.3 Procesamiento de datos

La etapa de procesamiento de datos es necesaria para depurar el conjunto de datos, con el fin de mejorar las características y prepararlo para el entrenamiento del modelo. Los datos provenientes del gobierno no necesitas pasar por todas las etapas solo se aplica limpieza de datos e identificación de direcciones.

4.3.1 Limpieza de datos

Este proceso contempla varios pasos, primero se debe cambiar el tipo de formato de codificación de caracteres al idioma del conjunto de datos (para español el más común es UTF-8), después de manera generar se eliminan valores duplicados, vacíos, enlaces, emojis,

espacios dobles, saltos de línea, caracteres especiales y signos de puntuación. En ocasiones también las palabras auxiliares, aunque depende del algoritmo a utilizar. Los *transformers* no requieren de este último paso, puesto que cuentan con su propia función de preprocesamiento.

Para el caso en específico de datos provenientes de Twitter es necesario eliminar las menciones de usuarios (ejemplo: @usuario). Los datos gubernamentales requieren de la selección de atributos de interés, eliminando así columnas redundantes, vacías o con pocos registros.

4.3.2 Geocodificación

Después de la etapa de etiquetado se georreferencian los datos de la clase de reporte (datos provenientes de Twitter), este paso únicamente se efectúa con los tuits que no cuentan con un campo de georreferenciación, es decir tuits que solo tienen el campo texto luego de minar los datos.

El campo se genera a partir del texto, detectando la alcaldía, seguida de la dirección del lugar (calle, colonia, código postal o nombre de negocio). Todos los datos contienen al menos el nombre de la alcaldía gracias al filtro aplicado en la parte de recolección.

4.3.3 Aumento de datos

El aumento de datos se aplica cuanto un conjunto de datos es reducido o esta desbalanceado, en conjuntos de datos procedentes de Twitter es común encontrar más registros de una clase que de otra, se plantean dos técnicas de desbalanceo para resolver la descompensación:

- Traducción de ida-vuelta
- Submuestreo

4.3.4 Representación del texto

Las tareas de procesamiento de lenguaje natural como son las tareas de clasificación de texto es necesario crear representaciones numéricas del texto que puedan ser procesadas computacionalmente, puesto que en su forma original no es posible realizar cálculos matemáticos. Existen diferentes técnicas entre las cuales se consideran:

- N-gramas
- Análisis semántico
- Incrustación de palabras (Word embedding)

4.3.5 Análisis de datos

Analizar la estructura de los datos permite encontrar patrones interesantes que ocurren en el lenguaje, detectar secuencias, uso y significado de las palabras utilizadas se conoce como semántica. Este proceso es importante porque ayuda a descubrir diferencias entre los textos de cada clase e identificar si las estructuras del texto interfieren con la forma en que el modelo aprende. Esta subetapa considera análisis de longitud de palabras, frecuencia de palabras, n-gramas y polaridad.

4.4 Selección de modelo

En esta etapa se determinan los modelos de aprendizaje automático supervisado, se establece la línea base y se proponen modelos actuales que han mejorado en rendimiento y desempeño tareas de PLN.

4.4.1 Línea base

Para determinar línea base se consultan los algoritmos utilizados en el estado del arte puesto que brinda fundamento para usar los modelos en la tarea presente. Estos suelen pertenecer al área de aprendizaje máquina ya que funciona bien para una tarea en específico sin usar muchos recursos o un método muy robusto. Aunque existen diferentes clasificadores se eligen los dos algoritmos que han logrado mejores resultados en la clasificación de texto:

- Máquinas de vectores de soporte
- Random forest

4.4.2 Modelos propuestos

Las tareas de clasificación de texto han tenido buenos resultados al usar el paradigma de los *transformers*, se plantear el uso de modelos basados en BERT para detectar si los resultados alcanzados en otras tareas de clasificación pueden reflejarse en la detección de reportes delictivos:

- BERT Multilingüe (M-BERT)
- BETO
- DistilBETO
- ALBETO

4.5 Clasificación

La etapa de clasificación considera el entrenamiento y prueba de los modelos seleccionados, se realiza la experimentación de los parámetros para encontrar los valores más óptimos y de esta forma hallar y guardar el mejor modelo.

4.6 Evaluación de los modelos

El desempeño de cada modelo se mide utilizado métricas de evaluación como los son precisión, sensibilidad y puntaje F1. Esta etapa va en conjunto con la etapa anterior puesto que, la clasificación depende del puntaje obtenido en la prueba para verificar si los parámetros usados son correctos o necesitan ajustes. Una vez se tengan los modelos finales, se hace una comparación entre ellos para determinar el mejor tratamiento de datos y modelo entre las combinaciones exploradas.

4.7 Análisis geoespacial

Como parte final se realiza el análisis geoespacial, es fundamental que los registros de ambos conjuntos de datos estén georeferenciados, si bien los datos gubernamentales ya cuentan con

coordenadas geográficas (latitud y longitud), el atributo en los datos de Twitter puede venir de tres formas: cuadro delimitador, punto (latitud y longitud) y texto.

Primero se normalizan los tipos de campos para que puedan ser importados al SIG. Se aplica geocodificación a los datos con direcciones textuales y se obtienen las coordenadas geográficas (latitud y longitud). Después se convierte el formato del cuadro delimitador a Well Known Text (WKT) y se calculara el centroide para asignarlo como el punto en que ocurrió el delito. Al terminar se cargan todos los registros que pertenezcan a reportes delictivos para realizar el análisis espacial.

Antes de iniciar con el análisis de los reportes es preciso efectuar un análisis del lugar de estudio, delimitar las demarcaciones territoriales, zonas pobladas y si existen áreas protegidas o algún tipo de zona restringida donde el manejo de los reportes delictivos cambie. Finalmente, se cargar todas las capas al SIG y se seleccionan los procesos necesarios para crear los mapas de calor o mapas temáticos que ayudaran a presentar de manera visual las áreas con mayor índice delictivo y así detectar lugares de alto conflicto.

Capítulo 5. Experimentación y resultados

En este capítulo se desarrolla la experimentación según la metodología propuesta, se explican los procesos, técnicas y herramientas aplicadas para realizar los análisis necesarios de tal manera que después puedan ser replicados. También se presentan los resultados obtenidos, los conjuntos de datos estructurados, los mejores clasificadores entrenados y el análisis espacial realizado. Al final se muestra una comparativa del índice delictivo ocurrido en la Ciudad de México entre los realizados conjuntos.

5.1 Recolección y selección de reportes delictivos

La primera etapa consiste en la recolección y selección de datos, dado que se busca analizar datos provenientes tanto de fuentes abiertas como de redes sociales y ambas implican procesos diferentes, se divide en dos subetapas: recolectar datos de redes sociales (Twitter) y hacer una selección de datos del repositorio gubernamental de datos abiertos (donde los datos ya fueron previamente categorizados por delitos y no es necesario recolectarlos, el conjunto ya está condensado en un solo archivo).

5.1.1 Recolección de datos de redes sociales

La recolección de datos de redes sociales depende de una API que hagan posible realizar este paso, con las nuevas medidas de privacidad de los datos algunas no cuentan con esta opción es por ello que se elige Twitter.

Twitter cuenta un portal de desarrollador el cual permite generar credenciales, más la basta documentación y fácil aplicación la hacen una buena elección. Existen diferentes niveles de acceso y versiones de la API de Twitter (Twitter API v1 y Twitter API v2), donde según los privilegios son los tipos y cantidades de consultas que se pueden realizar.

Aunque están disponibles cuatro tipos de niveles (Essential, Elevated, Elevated+ y Academic Research) solo dos son gratuitas y de uso no comercial (Essential y Academic Research), cabe mencionar que para obtener Academic Research es necesario ser trabajador o estudiante de un centro de investigación (Twitter, Inc., 2021), algunos de los beneficios que brinda este tipo de acceso son descritos en la Tabla 1.

Tabla 1. Beneficios disponibles para investigación académica.

Acceso	
Beneficios	Acceder a los datos históricos y en tiempo real con características y funcionalidades adicionales
Límite de tuits	10 millones de tuits al mes
Reglas de consulta	1024 caracteres, 1000 reglas de transmisión
Tasa de transmisión	50 peticiones/ 15 minutos por aplicación
Soporte técnico	Documentación, tutoriales, soporte y foros de la comunidad

El proceso de registro puede llegar a tardar un par de días, este consiste en llenar una serie de formularios con información referente al uso que se dará a los datos recabados más datos personales, al final se tendrá acceso al portal de desarrollador y se crearan las credenciales

necesarias para realizar las consultas, dos llaves del consumidor (usuario): API Key and Secret y dos tokens de autenticación: Bearer Token y Access Token and Secret (Ver Figura 9).

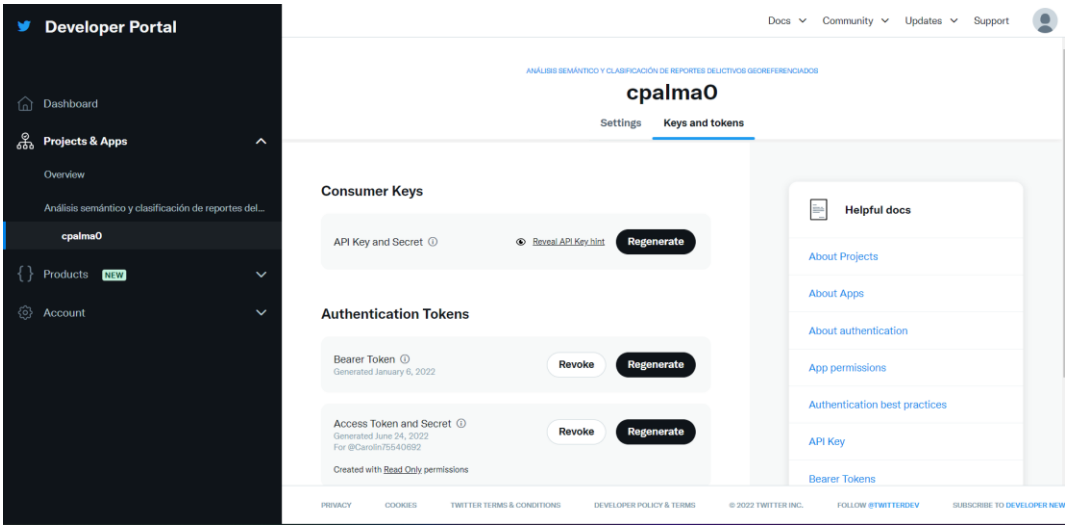

Figura 9. Portal de desarrollador de Twitter donde se muestran las claves y tokens usados para extraer información.

A pesar de que existen múltiples plataformas y librerías en diferentes lenguajes de programación para extraer los tuits como: Postman, ctw, twittered, Tweepy, Twarc, entre otras (Twitter, Inc., 2021), se elige Twarc2 por su fácil implementación y rápida ejecución.

Una vez obtenidos los requisitos para realizar las consultas se definen los filtros y parámetros a considerar. Este paso se dividió en dos formas de recolección, por diccionario de datos y por cuentas oficiales del gobierno de la Ciudad de México.

Diccionario de datos, Se creo una lista de palabras relacionadas a crimines como el trabajo realizado por Hernández et al. (2020), y se buscaron sinónimos para extender la búsqueda. Algunas palabras incluidas son: abuso, maltrato, asalto, amenaza, delito, robo, crimen, denuncia, entre otras. El diccionario cuenta con 108 palabras, la lista completa se encuentra en Anexo 1.

Cuentas oficiales, El gobierno tiene diferentes cuentas oficiales y son de interés las relacionadas a seguridad pública, específicamente la investigación y persecución de delitos. Se eligieron dos, la fiscalía general de justicia de la Ciudad de México (@fiscaliaCDMX) y la Secretaría de Seguridad Ciudadana de la Ciudad de México (@SSC_CDMX). Cabe mencionar que los tuits recolectados fueron los realizados por y hacia las cuentas.

Una vez seleccionada las palabras y cuentas de usuario, es necesario establecer filtros de búsqueda. El primer filtro establece recuperar texto solo en español, el segundo es la fecha (el periodo de estudio es el año 2021), dado que los textos requieren ser georeferenciados se establece una zona: México (MX) que después se depurara para la Ciudad de México. Finalmente, el objeto del tuit recuperado contiene varios atributos desde la fecha de creación, identificador del tuit, texto, usuario, hasta coordenadas, lenguaje, entre otros, de todos los atributos se eligen los 6 más relevantes para el estudio realizado (Ver Tabla 2).

Tabla 2. Atributos elegidos de la API de Twitter.

Campo	Contenido
id_tuit	Identificador único del tuit
id_autor	Identificador único del usuario
tuit	Texto
fecha_tuit	Fecha de publicación (dd/mm/aaaa)
lugar	Indica que el tuit está asociado, pero no necesariamente se origina en una ciudad
coordenadas	Representa la ubicación geográfica del tuit según el informe del usuario o de la aplicación (longitud, latitud)

Al completar la etapa de recolección se obtuvieron 249,788 tuits del conjunto del diccionario de palabras y 258,919 de las cuentas oficiales de gobierno.

5.1.2 Selección de datos abiertos

El portal de datos abiertos de la Ciudad de México pone a disposición de los ciudadanos información y análisis referente a la ciudad, se encuentran 424 conjuntos de datos de 31 instituciones divididos en 17 categorías (Gobierno de la Ciudad de México, 2022).

En el sector de Justicia y Seguridad se encuentran 21 conjuntos de datos, y están publicadas las carpetas de investigación de la FGJ (Gobierno de la Ciudad de México, 2021), al momento de consulta cuenta con 721,609 registros de investigaciones realizadas en múltiples años. Para establecer un parámetro de comparación con los datos de Twitter se eligieron los reportes delictivos ocurridos en 2021, teniendo así un total de 227,640 datos. Los datos de la carpeta de víctimas contienen 27 tipos de atributos, 16 categorías y 305 tipos de delitos (Ver Anexo 2).

5.1.3 Preprocesamiento

La recolección de tuits aun cuando se aplican filtros trae textos que llegan a no ser útiles y es necesario quitarlos antes de realizar el proceso de etiquetado, de ser tratados más adelante en la etapa de procesamiento el tiempo invertido en identificarlos en la etapa de etiquetado seria perdido. Se eliminaron los tuits que cumplieran con las siguientes condiciones:

- Tuits con información relacionada al COVID: se encontraron publicaciones sobre cuidados e información de salubridad sobre la enfermedad, así como mensajes de inconformidad a las medidas tomadas por el gobierno (Ver Figura 10 y 11).
- Tuits cortos con solo menciones de usuarios o enlaces (Ver Figura 12).
- Tuis que no se encuentren dentro de la Ciudad de México: los tuits cuentan con diferentes atributos que indican el lugar donde fue escrito como: geo.id, geo.coordinates (coordenadas geográficas en forma de latitud y longitud) y geo.bbox (cuadro delimitador, cuando no se conoce el lugar exacto Twitter calcula y regresa un área del lugar, este proceso requiere un tratado diferente a geo.coordinates, para visualizar los pasos realizados consultar Anexo 3). La Figura 13 muestra de forma visual la identificación de los datos, entre los que se encuentran fuera de la Ciudad de México de los que no, el proceso de extraer los tuits por se realizó con ayuda de QGIS.

Figura 10. Tuit con texto relacionado al COVID, inconformidad de un usuario a las medidas empleadas por las autoridades.

Figura 11. Tuit con texto relacionado al COVID, aviso oficial del gobierno.

Figura 12. Tuit corto que no contiene texto relevante.

Figura 13. Mapa de México con los tuits recolectados, se identifican los publicados dentro de la Ciudad de México (azul marino) y se descartan los demás (azul claro).

5.2 Etiquetado de clases

El entrenamiento de modelos de aprendizaje supervisado requiere de un conjunto de datos que este etiquetado en clases, donde a partir del conjunto de datos de entrenamiento y prueba aprenden para así poder predecir salidas.

Para la tarea de clasificación que busca identificar si un texto es o no reporte delictivo, se eligen dos clases (reporte y no reporte). Esta tarea se realizó en un periodo de cuatro meses y conto con la ayuda de un equipo de tres personas.

Como su nombre lo indica los tuits etiquetados como reportes son los que hacen referencia a un reporte o hecho delictivo, de otra forma pertenecen a la clase no reporte. En algunos textos se encuentran palabras que suponen un hecho relacionado a crimen como la Figura 14, pero no lo son. Las Figuras 14 y 15 de color azul indican ejemplos de tuits etiquetados como no reportes y las Figuras 16 y 17 de color rojo como reportes.

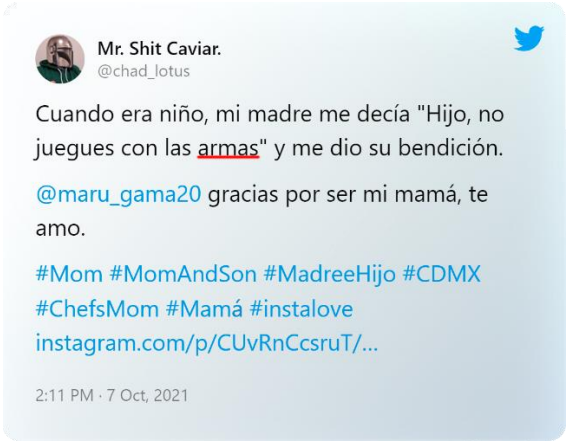

Figura 14. Tuit que contiene la palabra armas encontrada en el diccionario de palabras, pero es no reporte.

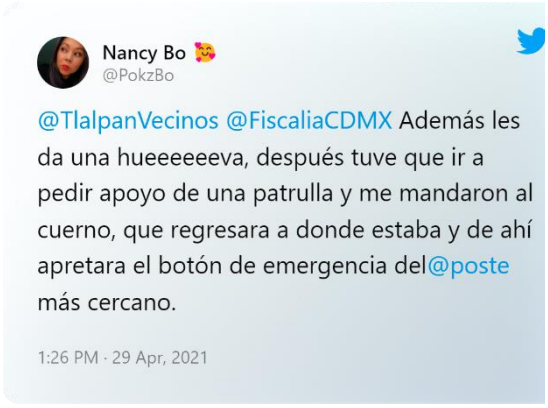

Figura 15. Tuit de inconformidad sobre la atención dada por la policia, es no reporte.

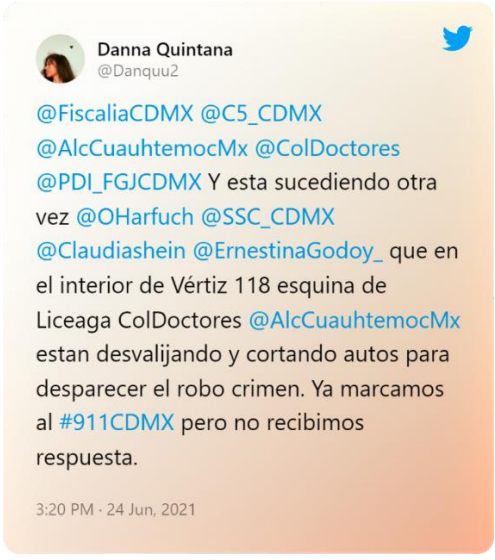

Figura 16. Tuit reporte de robo de auto, contiene información sobre la dirección donde ocurrió.

Figura 17. Tuit reporte de balacera, incluye información de donde ocurrió.

A pesar de que se encontraron más datos de la clase reporte algunos no fueron considerados porque indicaban una fecha anterior a 2021 o delitos ocurridos fuera de la Ciudad de México (tuits que serán georeferenciados según el texto). Si bien se descartaron algunos de estos tuits en la parte del preprocesamiento, este fue una limpieza rápida para eliminar tuits que contuvieran errores más obvios (duplicados, fuera del área de estudio y relacionados a COVID). Como se expresa en esta sección algunos textos pueden contener información que solo al ser revisados de forma manual logran ser verificados, el preprocesamiento anterior solo ayudo a que fueran menos textos (Ver Figura 18 y 19).

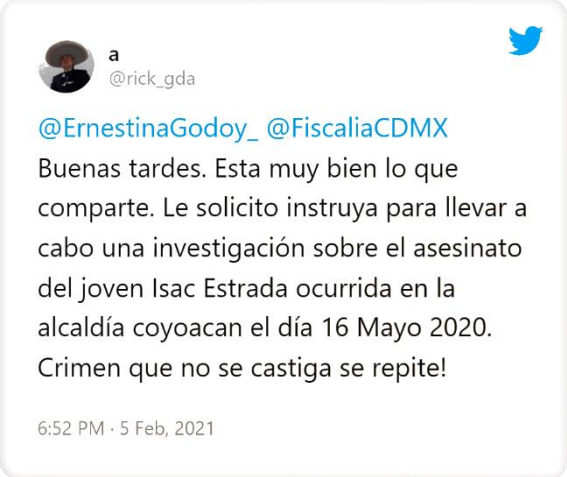

Figura 18. Reporte sobre un reporte pasado que no se encuentra dentro del rango de estudio.

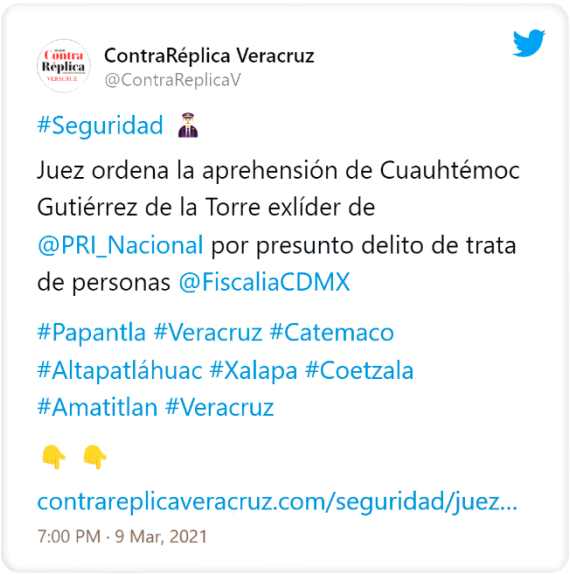

Figura 19. Reporte sobre el arresto de un individuo fuera de la Ciudad de México.

Del conjunto total recolectado, se generó un conjunto para etiquetar con 100,485 registros de los cuales se lograron etiquetar 2,740 tuits como reporte y 30,505 no reporte, siendo el 33.08% de datos etiquetados del conjunto total. Pese a que, se inició con un total de 508,707 tuits, después de preprocesar y seleccionar los datos propicios a etiquetar se redujo considerablemente a un 19.75% de los datos. Debido al poco tiempo y el pequeño grupo de etiquetadores disponibles para terminar esta etapa solo se alcanzaron a etiquetar 33,324 registros (Ver Tabla 3). Para consultar la cantidad de datos etiquetados por participante ver Anexo 4.

Tabla 3. Conjunto de datos recolectado, preprocesado y etiquetado.

Conjunto	Recolectados	Etiquetar	Etiquetados	Reporte	No reporte
Palabras clave	249,788	62,658	18,473	734	17,739
Cuentas oficiales	258,919	37,827	14,772	2,006	12,766

5.3 Procesamiento de datos

En esta etapa se procesan los datos para crear el conjunto de datos final que después será dividido en conjunto de entrenamiento y prueba, los cuales se utilizaran para entrenaran los modelos de aprendizaje supervisado.

Esta fase incluye el paso de limpieza que ayuda a separar los datos e identificar los valores importantes de los redundantes o datos con ruido, de igual forma se incluye la subetapa de geocodificación la cual es un elemento importante pues permite identificar donde ocurrieron los delitos reportados.

5.3.1 Limpieza de datos

La limpieza de datos implica identificar errores en los datos, ya sea porque están incompletos, duplicados, o existen faltantes. En ocasiones también involucra aplicar técnicas que permitan recuperar, sustituir o modificar los datos para que estos sean corregidos. En este caso, se consideró eliminar diferentes partes del texto que no aportan a la estructura, elegir los atributos más relevantes y finalmente aplicar una técnica de recuperación de coordenadas geográficas (geocodificación).

Datos de Twitter

Aun cuando se realizó un preproceso después de la recolección de datos, este no fue para eliminar información poco relevante del texto sino para evitar que el proceso de etiquetado fuera más extenso de lo necesario. Dentro de los pasos a considerar para limpiar los tuits del conjunto de datos se eliminan:

- Menciones de usuarios
- Hashtag (#)
- Saltos de línea
- Enlaces
- Caracteres especiales, puntuación, entre otros.
- Emojis
- Espacios dobles o extras
- Campos vacíos o nulos
- Duplicados
- Texto de una sola palabra

Hasta este punto se guardó un archivo con el conjunto de datos para el entrenamiento de los *transformers* manteniendo el texto con mayúsculas y minúsculas. Por otro lado, para los algoritmos de línea base se cambió el texto a minúsculas y se guardó para su posterior manejo. Debido a la extensión de código, se crean cuadernos de Jupyter los cuales manejan un formato. ipynb al ser creados en Google Colab, pero pueden ser abiertos en cualquier entorno de Python. Estos cuadernos se comparten por medio de un link público como también por el repositorio creado para el proyecto (Anexo 5).

La Figura 20 presenta un ejemplo de un tuit original, aunque no todos los textos contienen los mismos elementos a procesar este caso presenta la mayoría de componentes a eliminar como las menciones, emojis, enlaces, signos de exclamación y números.

Figura 20. Ejemplo del texto original de un tuit.

En contraste las Figuras 21 y 22 muestran el texto después de ser procesado según el tipo de modelo a emplear, la única diferencia es el manejo de mayúsculas y minúsculas.

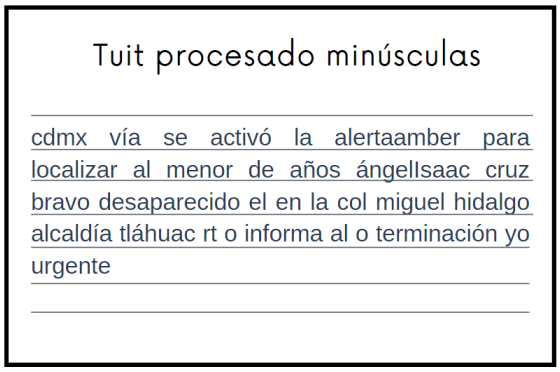

Figura 21. Tuit procesado en minúsculas para su uso en línea base.

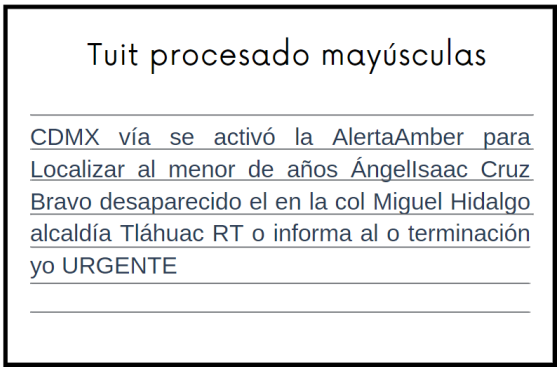

Figura 22. Tuit procesado manteniendo mayúsculas para su uso en *transformers*.

Como se mencionó anteriormente, el archivo del texto en minúsculas guardado para el entrenamiento de los modelos de línea base que pertenecen a métodos tradicionales de aprendizaje máquina, requieren de un procesamiento más exhaustivo, el cual debe considerar la eliminación de palabras auxiliares (*stop words*), estas son palabras que no brindan contexto o significado a las oraciones como artículos, pronombres, preposiciones, entre otros. Un ejemplo de un texto sin estas palabras se muestra en la Figura 23. Algunas abreviaturas o contracciones también fueron eliminadas.

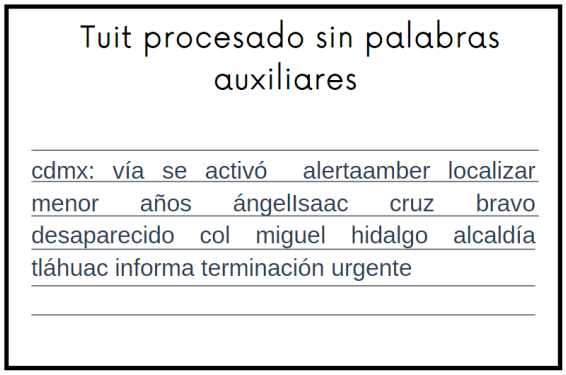

Figura 23. Ejemplo de tuit procesado sin palabras auxiliares.

Datos abiertos

El conjunto de datos abiertos cuenta con 22 atributos que indican el id, año, mes, fecha, delito, latitud, longitud, entre otras. Debido a la extensión que lleva presentar todos los atributos en una tabla, esta puede ser consultada en Anexo 6 (Se dividió en dos partes para tener una mejor visualización). La Tabla 4 muestra la existencia de varios campos redundantes como la fecha que está dividida en año y mes o la columna que tiene el valor completo (dd/mm/aaaa).

Por otro lado, se incluyen diferentes tipos de fecha, una se refiere a la fecha de inicio (cuando se registró la denuncia e inicio la investigación) y la fecha del hecho (cuando ocurrió el delito). Puesto que nos interesa realizar un análisis sobre los crímenes ocurridos y compararlos con los reportes realizados en Twitter, se elige la fecha de hecho como base para comparar el índice delictivo de un lugar. De momento la hora no es relevante al estudio realizado por lo tanto se elimina.

Tabla 4. Atributos redundantes relacionados a fecha de los datos abiertos.

Añoinicio	Mesinicio	FechaInicio	Añohecho	Meshecho	FechaHecho	HoraHecho	HoraInicio
2019	Enero	04/01/2019	2018	Agosto	**29/08/2018**	12:00:00	12:19:00
2019	Enero	04/01/2019	2018	Diciembre	**15/12/2018**	03:00:00	12:20:00
2019	Enero	04/01/2019	2018	Diciembre	**22/12/2018**	03:30:00	12:23:00

Existen seis columnas relacionadas al lugar del hecho (Ver Tabla 5). Las coordenadas geográficas (latitud y longitud) indican el punto exacto donde ocurrió el delito y los atributos referentes a colonia y calle contiene la dirección en formato de texto, varios de estos registros cuentan con errores, están vacíos o NA (no aplica), corregirlos llevaría tiempo es por ello que se eliminan los tres campos ColoniaHecho, Calle_hechos y Calle_hechos2. Aunque se quitaron los campos del conjunto final, si un registro no cuenta con latitud y longitud, pero si con valores en colonia o calle se utilizan para recuperar las coordenadas por medio de geocodificación.

Tabla 5. Atributos relacionados al lugar del delito de los datos abiertos.

AlcaldiaHechos	ColoniaHechos	Calle_hechos	Calle_hechos2	latitud	longitud
Álvaro obregón	Guadalupe inn	Insurgentes sur	NA	**19.36125**	**-99.18314**
Azcapotzalco	Victoria de las democracias	Av. Cuitláhuac	NA	**19.47181**	**-99.16458**
Coyoacán	Copilco universidad issste	Copilco	NA	**19.33797**	**-99.18611**

Por último, entre los campos restantes están el tipo de delito, categoría, tipo de persona afectada, calidad jurídica entre otros, de los cuales se mantienen idCarpeta, delito y Categoría, ya que describen que delito ocurrió y a que está relacionado (Ver Tabla 6).

Tabla 6. Atributos relacionados al tipo de delito de los datos abiertos.

idCarpeta	Delito	Categoria	CalidadJuridica	competencia
8324429	**Fraude**	**Delito de bajo impacto**	Ofendido	Fuero común
8324430	**Producción, impresión,**	**Delito de bajo impacto**	Víctima y denunciante	Fuero común
8324431	**Robo a transeúnte**	**Robo a cuentahabiente**	Víctima y denunciante	Fuero común

Finalmente, de los 22 atributos se conservan siete los cuales fueron elegidos por ser relevantes al estudio realizado (Ver Tabla 7). Una vez elegidas las columnas más importantes, se inicia con el proceso de limpieza de datos.

Tabla 7. Atributos seleccionados de los datos abiertos.

Campo	Contenido
idCarpeta	Identificador de la carpeta de investigación
Fecha_hecho	dd/mm/aaaa en que se cometió el delito
Delito	Conducta, acción u omisión de la ley
Categoría	Se clasifican en 16 tipos de delititos
Alcaldia_hecho	Alcaldía en que se cometió el delito
Latitud	Longitud de la geolocalización
Longitud	Latitud de la geolocalización

En los campos de latitud y longitud se encontraron 8,717 datos faltantes. Para recuperar las coordenadas se utilizaron los valores alcaldía, colonia y calle, buscando así completar dicho campo con el proceso de geocodificación, de no lograr recuperar un registro se elimina (Ver Tabla 8).

Tabla 8. Atributos de georreferenciación de los datos abiertos.

idCarpeta	Alcaldia	Colonia	Calle_hechos	Calle_hechos2	latitud	longitud
828794	Gustavo a madero	NA	Cometa	No scince notificacion medico lgal	NA	NA
8828900	Coyoacan	NA	No precisa	NA	NA	NA
8829174	Gustavo a madero	NA	Rio bamba	NA	NA	NA
8829699	Iztapalapa	NA	Calzada igncio zaragoza num. 1711	NA	NA	NA
8829716	Alvaro obregon	NA	Av. universidad	Notificacion de caso medico legal hospital general lopez mateos	NA	NA
8886565	Azcapotzalco	Tierra nueva	Av el rosario	NA	NA	NA
8936254	Iztapalapa	El retoño	Zacahuizco	Plutarco elias calles	NA	NA

5.3.2 Geocodificación

Existen diferentes librerías en Python para geocodificar datos que no cuentan con coordenadas geográficas, pero están georreferenciadas con un valor en forma de texto como

lo es una dirección (calle, colonia, código postal, ciudad). El proceso fue realizado con la ayuda de ArcGIS geocoding API, si bien se valoró GeoPy la plataforma de ArcGIS Developers obtuvo mejor resultados. Ambas herramientas son de uso gratuito.

La Figura 24 representa en forma de ejemplo el proceso que conlleva la geocodificación y su funcionamiento, del texto se identifica la dirección para después cambiarla a un formato estándar y finalmente por medio de una API obtener la latitud y longitud.

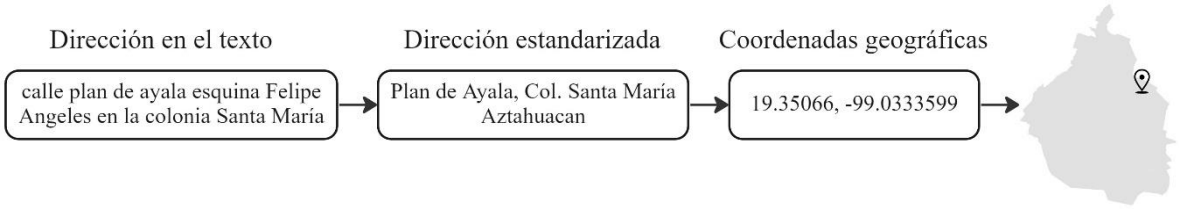

Figura 24. Ejemplo del proceso de geocodificación donde a partir del texto de un tuit se obtienen las coordenadas geográficas.

Datos abiertos

Para geocodificar los datos abiertos se utilizan las columnas de colonia, calle y calle2 (el valor de calle tiene prioridad sobre calle2, de tal forma que si un registro contiene el primero se salta el valor del segundo, por lo contario si calle este vacío se usa calle 2). La Figura 25 muestra los 8,717 registros georeferenciados, varios de los puntos se encuentran fuera de la Ciudad de México lo cual puede significar que en algunos casos los datos no fueron los suficientemente descriptivos para recuperar con éxito el lugar del reporte. Se completaron las coordenadas de 8,557 registros (160 registros no se recuperaron correctamente).

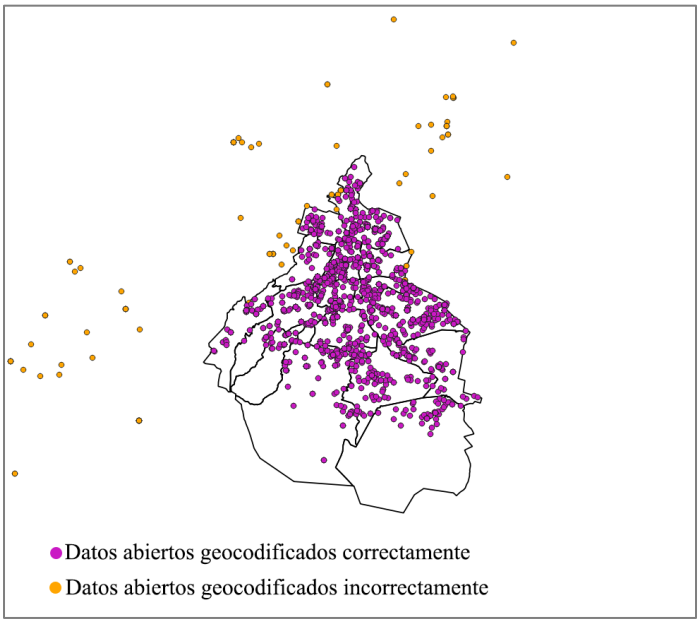

Figura 25. Registros georeferenciados por medio de geocodificación.

Twitter

Asimismo, los datos de Twitter fueron geocodificados con la diferencia de que al no existir un atributo como tal que indique el lugar del reporte este se obtiene directamente del texto,

si bien se intentó usar reconocimiento de entidades nombras (LexNLP, spaCy y Geotext) para automatizar el proceso los resultados alcanzados no fueron favorables, y al ser pocos registros etiquetados como reportes se decidió realizar de forma manual la identificación de direcciones en total se geocodificaron 2,740 tuits. La Figura 26 muestra un tuit reportando sujetos escandalizando la vía pública con la dirección Benito Juárez, Tlaltenco, Tláhuac, C. P. 13400, Ciudad de México. Aplicando geocodificación se obtienen las coordenadas [-99.0182, 19.2961].

Figura 26. Tuit de reporte con georreferencia en el texto.

Parecido al proceso de los datos abiertos, después de recuperar las coordenadas se revisan que estas se encuentren dentro de la Ciudad de México, de no ser el caso se busca la dirección escrita y se comprueba manualmente la ubicación (21 registros se corrigieron manualmente ya que al realizar la geocodificación algunos puntos de los tuits se ubicaron fuera de la Ciudad de México, pero al ser buscados manualmente se recuperaron correctamente).

5.3.3 Conjunto de datos

La creación de los conjuntos de datos permitió establecer un periodo de referencia, en este caso de un año (2021), para así comparar los delitos reportados oficialmente de los reportados en una red social (Twitter). Anteriormente se presentó el proceso de recolección, selección y limpieza, en este apartado se describen los conjuntos de datos finales y un análisis semántico del texto.

Datos abiertos

Como se mencionó los reportes de las carpetas de investigación realizados por la FGJ hasta marzo del 2022 fueron 780,440 de los cuales 227,640 se seleccionaron para crear el conjunto de datos abiertos que fue procesado. El archivo original contenía registros desde 1917 y reportes de zonas fuera de la Ciudad de México, es por ello que se procesaron para identificar los ocurriditos únicamente en las 16 alcaldías de la ciudad.

La Tabla 9 muestra el conjunto de datos final que después de la etapa de procesamiento contiene 227,478 registros. La columna de Geocodificados se refiere a los datos que se pudieron recuperar correctamente (los dos registros faltantes entre el conjunto de Zona Ciudad de México y Geocodificados son debido a que al seleccionar el lugar del reporte se utilizó el campo alcaldía, pero al ser visualizados en QGIS no se encontraban dentro del área de la ciudad y fueron eliminados).

Tabla 9. Conjunto de datos abiertos por filtro.

Conjunto	Completo	Año 2021	Zona Ciudad de México	Geocodificados	Conjunto final
Víctimas de carpetas de investigación FGJ	780,440	230,515	227,640	8,557	227,478

Twitter

A diferencia de los datos abiertos donde se realizó una selección, los datos de Twitter fueron recolectados especialmente para la tarea establecida. Se usaron 104 palabras como filtro para de recopilar los datos, las cuales se buscaron como palabras dentro del texto como también en forma de hashtag, aunado a la recolección de tuits de cuentas oficiales. Del total de tuits recuperados 508,707, se etiquetaron 33,245 donde 2,740 se determinaron como reporte y 30,505 como no reporte (Ver Tabla 10).

Tabla 10. Conjunto de datos de Twitter desde la recolección hasta el etiquetado y procesado.

Conjunto Twitter	Recolectados	Etiquetar	Etiquetados	Reporte	No reporte
Palabras clave	249,788	62,658	18,473	734	17,739
Cuentas oficiales	258,919	37,827	14,772	2,006	12,766
Total	508,707	100,485	33,245	2,740	30,505

De los 2,740 datos etiquetados como reportes 740 reportes se eliminaron en el proceso de limpieza, si bien los tuits podían estar escritos de diferente forma al agregar uno o varios emojis, anexar una imagen o video y escribir con diferentes hashtags, cuando se eliminaron algunos tuits se duplicaron al utilizar solo el texto. Similarmente para la clase no reporte 769 tuits fueron eliminados, quedando 29,736 registros para el conjunto de entrenamiento y prueba.

Aunque los tuits repetidos al tomar en cuenta el texto fueron eliminados del conjunto utilizado para los modelos, los 740 reportes se mantuvieron para el análisis geoespacial, ya que pueden ser reportes realizados por diferentes personas o múltiples reportes de una persona que intenta brindar atención a una situación en específico.

A continuación, se presenta un ejemplo del caso donde un tuit aparece como duplicado después del procesamiento de datos:

Tuit 1 (original): @GobiernoMX ?? n CAMIONES D CARGA usan Tlalpan deAUTOPISTA aTODA VELOCIDAD ponen en riesgo Y @GobCDMX @SSC_CDMX @SSPCMexico @OVIALCDMX @locatel_mx @UCS_GCDMX NO haceN nada calz Tlalpan NATIVITAS BJ sur-norte

Tuit 2 (original): @mileniotv @obritob @liliana_sosa ??n CAMIONES D CARGA usan Tlalpan deAUTOPISTA aTODA VELOCIDAD ponen en riesgo Y @GobCDMX @SSC_CDMX @SSPCMexico @OVIALCDMX @locatel_mx @UCS_GCDMX NO haceN nada calz Tlalpan NATIVITAS BJ sur-norte

El ejemplo Tuit 1 y Tuit 2 contienen diferentes elementos a primera vista, pero su diferencia está en las menciones de usuarios realizadas como @GobiernoMX y @mileniotv,

respectivamente. Después del procesamiento que implica eliminar la mención de usuarios de Twitter y caracteres especiales texto estas se vuelven iguales, por lo que se eliminan del conjunto que se usara para los modelos de aprendizaje automático.

Tuit 1 (procesado): CAMIONES D CARGA usan Tlalpan deAUTOPISTA aTODA VELOCIDAD ponen en riesgo Y NO haceN nada calz Tlalpan NATIVITAS BJ surnorte

Tuit 2 (procesado): CAMIONES D CARGA usan Tlalpan deAUTOPISTA aTODA VELOCIDAD ponen en riesgo Y NO haceN nada calz Tlalpan NATIVITAS BJ surnorte

La Tabla 11 muestra como el conjunto final fue divido para realizar el entrenamiento de los modelos, se usó la técnica de división (*train-test split*) también conocida como dejar uno afuera (*leave-one-out*) la cual consiste en dividir un conjunto en dos sub conjuntos, uno de entrenamiento y otro de prueba. Se aplicó una distribución de 80:20 ya que se ha demostrado que esta partición y la distribución 70:30 son las que llegan a obtener mejores resultados, aunado a su uso debido al principio de Pareto (Joseph, 2022), aunque dependiendo de los datos esta relación puede cambiar. No se probó 70/30 debido a los pocos datos de una clase que podría dejar menos ejemplos para la etapa de entrenamiento y no ser lo suficientemente representativos.

Tabla 11. División del conjunto de datos para entrenamiento y prueba utilizando una distribución 80:20.

Conjunto	**Porcentaje**	**Total de datos**	**Reporte**	**No reporte**
Entrenamiento	80	25,388	1,600	23,788
Prueba	20	6,348	400	5,948

5.3.4 Aumento de datos

Existen diferentes técnicas de aumento de datos, una de las más comunes siendo el sobremuestreo, submuestreo y traducción de ida-vuelta, así como también combinaciones de diferentes procesos como Easy Data Aumentation (EDA), Synthetic Minority Over-sampling Technique (SMOTE), entre otros. En muchos casos como el proceso como sobremuestreo empleados en EDA y SMOTE consisten en crear datos extra de entrenamiento, lo que significa crear textos sintéticos a partir de las muestras originales.

Para la experimentación de aumento de datos se propone la experimentación de cuatro técnicas: EDA, traducción ida-vuelta y submuestreo. EDA por su parte fue modificado debido a que utiliza cambio de sinónimos y aplica una librería en inglés para realizar este paso, por lo cual se cambia a una en español (wordnet).

Primero se obtiene el conjunto de la clase con menor muestras para realizar el aumento de datos, como se observa en la Tabla 3 la clase no reporte tienen 30,505 textos y reporte 2,740, siento esta ultima el 8.24% del total de datos, por lo cual reporte es la clase minoría.

Kumar et al. (2022) muestran una forma de calcular la relación de desequilibrio (*Imbalance Ratio,* IR), ver Ecuación 6:

$$IR = \frac{num.\ de\ muestras\ de\ la\ clase\ mayoritaria}{num.\ de\ muestras\ de\ la\ clase\ minoritaria}. \qquad \text{(Ecuación 6)}$$

Siguiendo esta fórmula el IR del conjunto de datos recolectado a etiquetar es de 11.13% (Ver Ecuación 7), cabe destacar que tanto el porcentaje del total de datos como el IR no reflejan un cálculo del total de datos recabados ya que como se mostró en el procesamiento de datos algunos fueron eliminados:

$$IR = \frac{30{,}505}{2{,}740} = 11.13. \qquad \text{(Ecuación 7)}$$

Al utilizar el IR calculado y la Tabla 12 como referencia se encuentra que la proporción de la clase minoritaria tiene un grado de desbalance de clase moderado. Aunque el IR no pertenece al grado extremo se prueban técnicas de aumento de datos para ver si los resultados obtenidos pueden mejorar en comparación con los obtenidos con el conjunto de datos original.

Tabla 12. Clasificación del grado de desbalance en el conjunto de datos (Kumar et al., 2022).

Grado de desbalance en clase	Proporción de la clase minoría
Extrema	$< 1\%$ del conjunto de datos
Moderado	$1 - 20\%$ del conjunto de datos
Poco	$20 - 40\%$ del conjunto de datos

Como se mencionó anteriormente la traducción ida-vuelta y submuestreo serán empleadas, el código puede ser consultado en el Anexo 5. Para cada método utilizado se duplico la cantidad de datos perteneciente a la clase reporte, ya que en pruebas previas se encontró que entre más datos sintéticos se usaran peor era el rendimiento de los modelos.

Para la técnica de traducción ida-vuelta se tradujo un texto cinco veces (español → francés → inglés → alemán → portugués → español), algunos textos se repitieron por lo que estos fueron eliminados (6 registros). La técnica de submuestreo elige de manera aleatoria elementos de la clase con mayor número de registros de tal forma que los datos obtenidos no necesitan de ningún proceso y mantienen los 1,600 datos seleccionados (Ver Tabla 13).

Tabla 13. Cantidad de datos por método de desbalanceo.

Técnica	Total de datos	Datos no duplicados	Conjunto final
Traducción	1,600	1,594	3,194
Submuestreo	1,600	1,600	3,200

5.3.5 Análisis de datos - Twitter

El análisis de los datos recolectados y etiquetados por clase se realiza para identificar si existen cambios en la forma en que fue redactado un tuit relacionado o no a un reporte delictivo. Las Figuras 27 y 28 presentan la cantidad de palabras por tuit por tipo de clase.

La Figura 27 muestra que los tuits que pertenecen a la clase reporte suelen ser más extensos, la mayoría de ellos tienen entre 30-40 palabras. Por el contrario, en la clase no reporte rondan entre las 2-20 palabras (Ver Figura 28). El tuit con más palabras para reporte fue de 54 y para no reporte de 59, en ambos casos el mínimo de palabras fue 2 (después del procesamiento algunos tuits contenían solo una palabra, estos fueron eliminados).

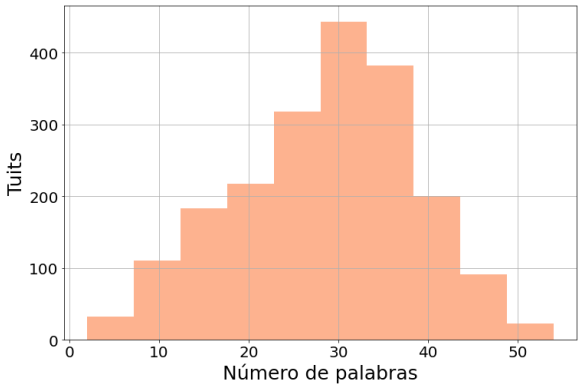

Figura 27. Conteo de palabras en tuits clasificados como reportes.

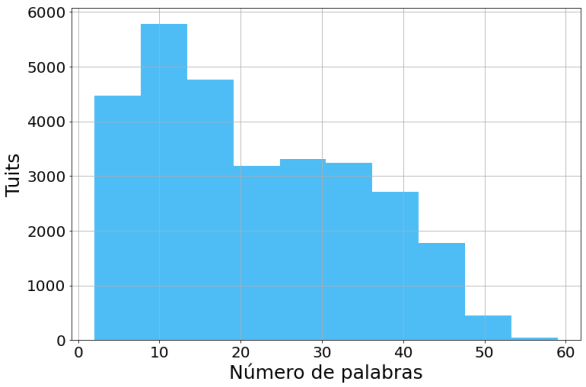

Figura 28. Conteo de palabras en tuits clasificados como no reportes.

Aunque las palabras auxiliares se eliminaron del conjunto de datos utilizado para el entrenamiento de los modelos de línea base, estas se mantuvieron para el entrenamiento de los *transformers*. Por tal motivo se presenta el análisis obtenido para los dos casos.

La Figura 29 presenta las palabras más frecuentas en la clase reporte, como es de esperar la mayor cantidad de palabras son palabras auxiliares como: *de*, *la*, *en*, *a*, *y*, *e*l, entre otras. Aun así, alcanzan a aparecer palabras relacionadas al lugar y fecha como: *alcaldía*, *colonia*, *año* y *cdmx*.

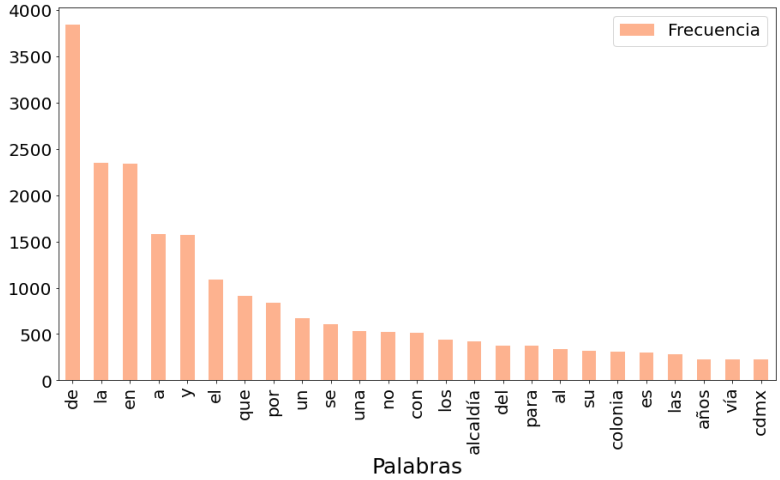

Figura 29. Frecuencia de palabras del conjunto completo para la clase reporte.

Por otro lado, la Figura 30 muestra las palabras más frecuentes para la clase no reporte que, al igual que el análisis anterior, las palabras más comunes son las auxiliares (*de*, *la*, *que*, *y*, *a*, entre otras). Entre las 25 palabras más repetidas solo se encontró una palabra no auxiliar (*corrupción*).

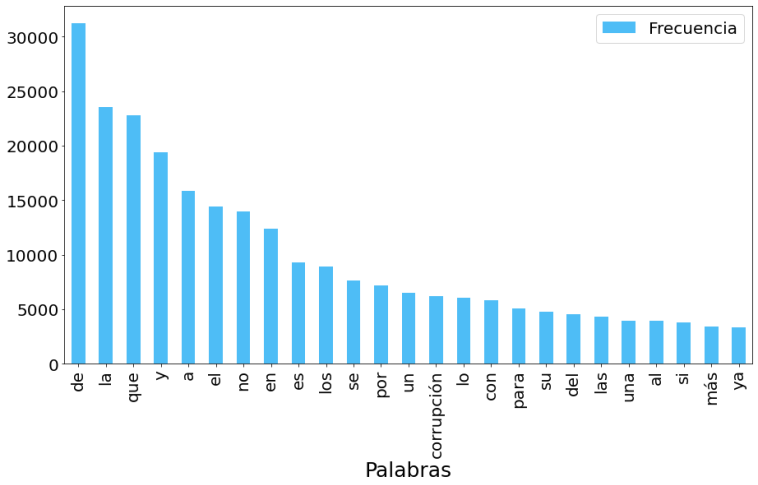

Figura 30. Frecuencia de palabras del conjunto completo para la clase no reporte.

N-gramas

De la misma forma, se realiza un análisis de n-gramas, específicamente bigramas y trigramas con el fin de observar la estructura que presentan los datos. La presencia de los bigramas obtenidos para la clase reporte de la Figura 23 encuentra que, la mayoría de los bigramas se relacionan a un mismo tipo de reporte, por el contexto provienen de reportes notificando la desaparición de un menor de edad (*de edad*, *vista por*, *alertaamber para*, *localizar a*, *última vez*, entre otros).

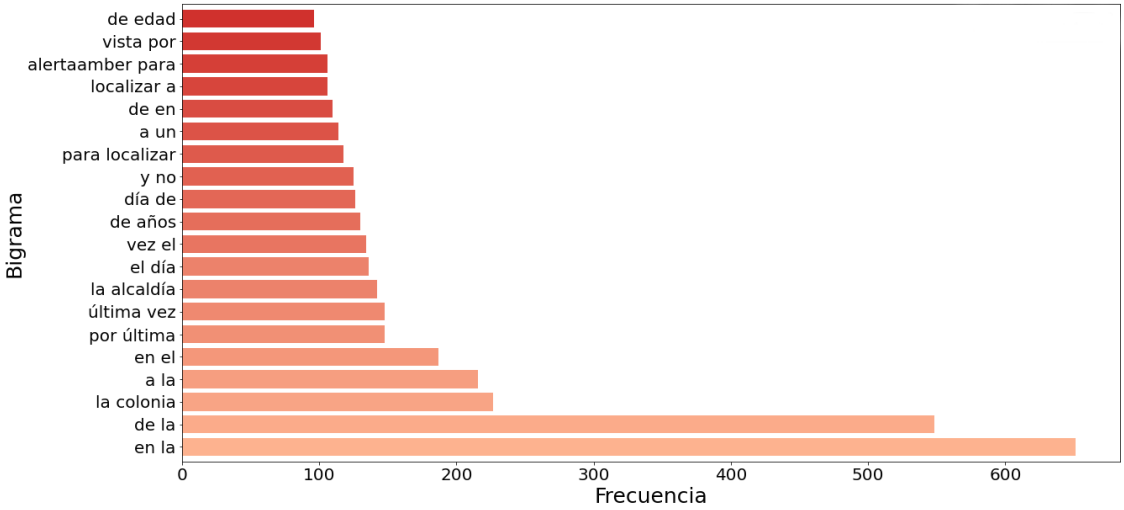

Figura 31. Frecuencia de bigramas en la clase reporte.

Por otro lado, la Figura 32 presenta los bigramas de la clase no reporte. Varios casos hacen mención del sustantivo corrupción acompañado de una preposición, (*corrupción y*, *de corrupción* y *la corrupción*).

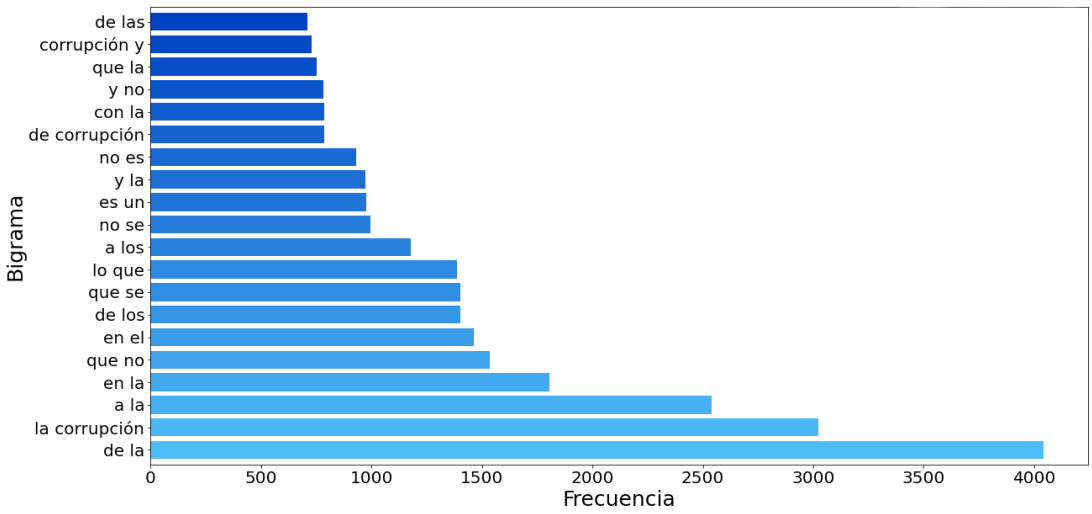

Figura 32. Frecuencia de bigramas en la clase no reporte.

Otro tipo de n-gramas son los trigramas el cual en lugar de utilizar una palabra subsecuente usa dos. Con este tipo de análisis se encuentran otro tipo de combinaciones de palabras y con ello tipos de reportes en los textos, el trigrama más frecuente (*ponen en riesgo*) se refiere a reportes por peligro de exceso de velocidad en las carretas, comportamiento indebido, entre otros. El segundo más frecuente (*años de edad*) al igual que en los bigramas son de alerta amber. Por último, el tercer más frecuente (*rt o informa*) alude a que la persona retuiteé o informe a un teléfono oficial un delito percatado (Ver Figura 33).

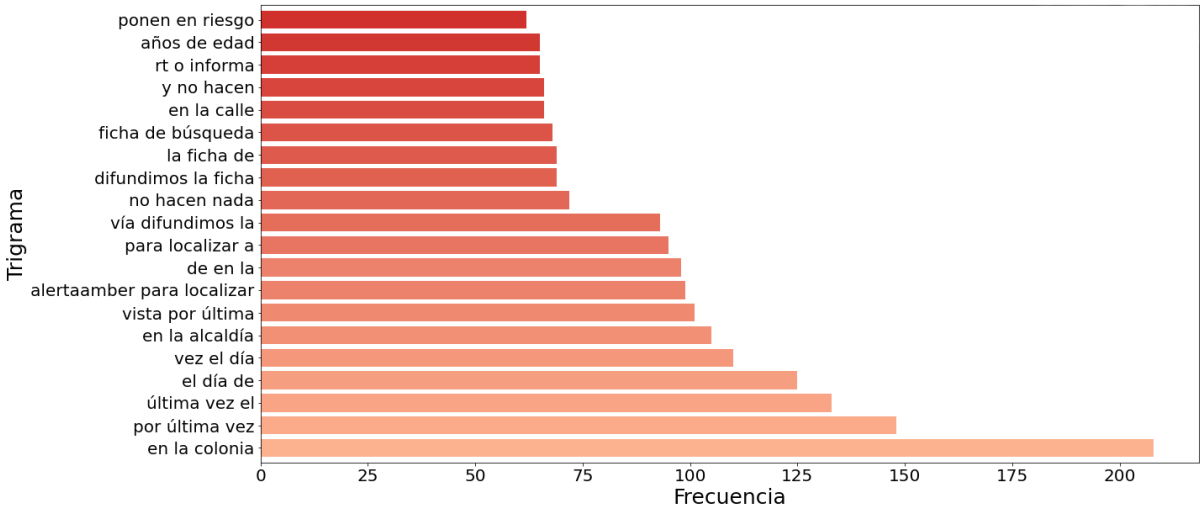

Figura 33. Frecuencia de trigramas en la clase reporte.

Finalmente, la Figura 34 muestra los trigramas de la clase no reporte donde existen diferentes estructuras particulares como el uso de la letra x para describir por en $si\ x\ la$ y $de\ si\ x$, en diez de los mas frecuentes se menciona la palabra corrupción y en dos crimen organizado y alianza.

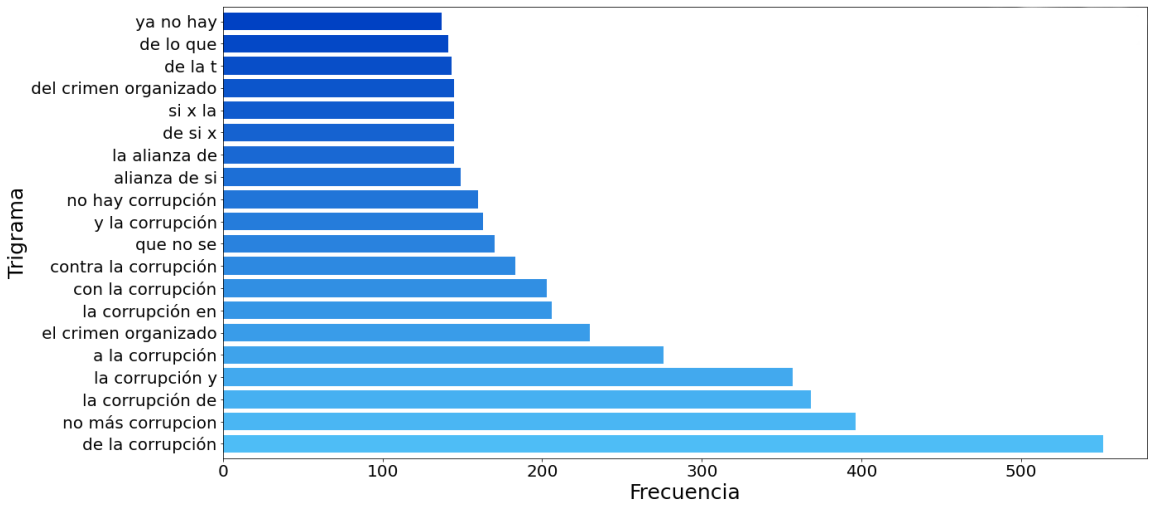

Figura 34. Frecuencia de trigramas en la clase no reporte.

Nube de palabras

Se aplico la representación de nube de palabras para mostrar de forma visual las 150 palabras más frecuentes de cada clase. A diferencia de las figuras anteriores, para la generación de la nube de palabras se lematizo el vocabulario con la intención de agrupar palabras, con la intención de reducir las variantes morfológicas (ejemplo: personas - persona).

La Figura 35 muestra que la palabra más frecuente en los reportes es calle, seguida de *persona*, *hacer*, *último vez*, entre otras, mientras que las menos frecuentes son: *ahora*, *tipo*, *esq*, *edad nombre*, *victima*, *urgir* y varios puesto que solo se repiten 31 veces. La mayoría de palabras tratan sobre un lugar (*calle*, *cdmx*, *alcaldía iztapalapa*, *azcapotzalco*, *esquina*, *tlalpan nativitas*, *álvaro obregón*), aunque también aparece palabras relacionadas con delitos (*denunciar*, *detener*, *amenaza*, *detenido*, *asaltar*, *vista último*, entre otras).

Figura 35. Nube de palabras más frecuentes en tuits etiquetados como reportes.

Figura 36. Nube de palabras más frecuentes en tuits etiquetados como no reportes.

Por otro lado, la Figura 36 presenta las palabras frecuentes escritas en tutis de la clase no reporte, donde corrupción se repitió 3,538 veces, decir y hacer más de 2,000 veces, en general se encontraron verbos (*decir*, *hacer*, *poder*, *saber seguir*). La menos repetidas en la nube

rondan en un conteo de 240 y son mano, trabajo y muerte. Se encontraron palabra relacionadas a lugares (*calle* y *mundo*) y términos negativos (*muerte*, *terminar*, *agresión*, *acoso*, *matar* y *problema*).

Análisis de polaridad – sentimiento

Finalmente, se aplicó un análisis de sentimiento al conjunto de datos final donde se cuantifica este sentimiento en un valor positivo, negativo o neutral al cual se le conoce como polaridad el cual nos ayuda a ver la tendencia de los tuits y como es que se expresan las personas. Al recolectar y etiquetar los datos se encontró que varios textos de la case no reporte tienen inclinación política, en su mayoría mencionando inconformidad hacia el gobierno. Este análisis se realizó con un pipeline de análisis de sentimientos para textos en español basado en BERT de John Snow Labs el cual clasifica si un texto es negativo, positivo o neutro (*Sentiment Analysis Pipeline for Spanish Texts*, 2021).

La Figura 37 muestra la polaridad de los tuits etiquetados como reporte, más de la mitad se identificaron como neutros (1,024), seguida de negativos (888) y muy pocos positivos (88), lo cual indica que las reacciones de las personas al reportan un delito tienen una connotación que tiende a lo negativo en lugar de positivo, hecho que es entendible debido a la naturaleza de los textos, ya que están reportando un delito. La neutralidad presentada puede ser explicada que, como la mayoría de reportes provienen de cuentas gubernamentales oficiales las cuales reportan algún hecho delictivo ocurrido tienden a ser más informativas o de alerta la forma en la que se escriben es neutral y calmada.

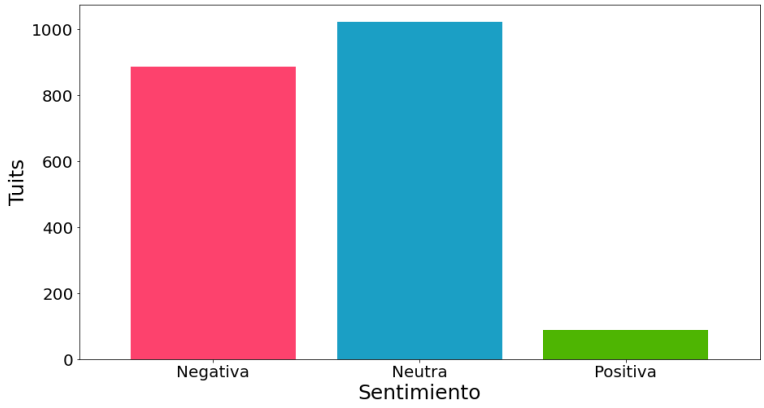

Figura 37. Análisis de polaridad para los tuits de la clase reporte.

Por otra parte, la Figura 38 que presenta la gráfica con la clasificación de los tuits para la clase no reporte indica que más de 20,000 textos tienden a ser negativos lo cual es más de dos tercios del total de datos encontrados en la clase (21,775 tuits). Los datos restantes se distribuyen en sentimiento neutro y positivo en cantidades parecidas (4,496 y 3,465, respectivamente). Dentro del análisis realizado se encontró que, opiniones políticas e inconformidad con el gobierno y el entorno conforman gran parte de los textos recabados, esto se ve reflejado y concuerda con la negatividad encontrada para la clase.

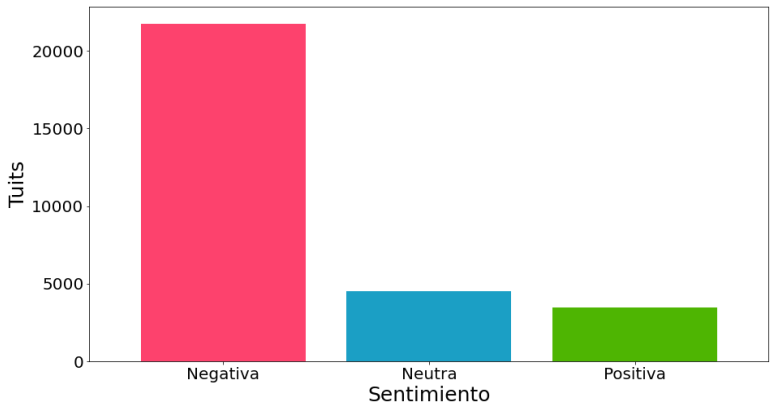

Figura 38. Análisis de polaridad para los tuits de la clase no reporte.

5.4 Selección de modelos

La selección de modelos se basa en los métodos que han obtenido mejores resultados tanto para la tarea a resolver como parecidas en este caso tareas de procesamiento de lenguaje natural. Asimismo, el uso de diferentes técnicas de representación de texto es evaluadas.

5.4.1 Línea base

La línea base es seleccionada conforme al estado del arte, pese a que existen varios trabajos relacionados al problema de clasificación de reportes delictivos, se eligieron los más parecidos. El primer trabajo aplica máquinas de vectores de soporte con trigramas, los autores recolectaron 8,000 tuis en árabe de noticias, y aplicaron diferentes técnicas para extraer las características entre ellas: Stemming, n-gramas y bolsa de palabras AL-Saif y Al-Dossari (2018).

Por su parte, Lal et al. (2019) aplicaron diferentes algoritmos de aprendizaje automático con un conjunto de datos de 369 (crimen y no-crimen), obteniendo los mejores resultados con Random forest. Como se puede observar los resultados obtenidos en tareas de clasificación de crimines rondan desde el 91% al 98% de exactitud, teniendo mejores resultados con el conjunto de datos en inglés (Ver Tabla 14).

Tabla 14. Algoritmos utilizados en el estado del arte para la detección de reportes delictivos.

Autores	**Algoritmos**	**Idioma**	**Precisión**	**Recall**	**Exactitud**
AL-Saif y Al-Dossari (2018)	SVM con trigramas	Árabe	-	-	91.55%
Lal *et al*. (2019)	Random forest	Inglés	98.2%	98.1	98.1%

Los algoritmos a evaluar para la tarea descrita en este trabajo son SVM y Random forest con extracción de características en n-gramas y *embeddings* (Network Language Model y Universal Sentence Encoder Multilingüe).

5.4.2 Modelos propuestos

Las tareas de PLN han tenido buenos resultados utilizando el paradigma de los t*ransformers* volviéndose los modelos de última generación para clasificación de texto, preguntas-respuestas, resumen de texto, traducción, entre otras.

Su mayor ventaja es que al estar pre-entrenados en diferentes lenguajes y solo necesitar realizar el paso de fine tuning el cual se acomoda específicamente a la tarea deseada, aunado a que pueden ser aplicados con conjuntos de datos pequeños y aun así obtener buenos resultados.

Se proponen tres modelos basados en BERT a evaluar, su versión multilingüe y dos modelos pre-entrenados para el lenguaje español: BETO, DistilBETO y ALBETO. En un inicio se planteó el uso de *transformers* pre-entrenados en conjuntos de datos de tuits como TWilBert presentado por González et al. (2021), pero debido a la falta de soporte y compleja implementación se deja para trabajo futuro.

La Tabla 15 muestra los modelos propuestos, los idiomas en los que fueron pre-entrenados y por quien, y cuando fueron creados, se puede apreciar que estos modelos son nuevos ya que los más recientes se publicaron en 2022, siendo el que tiene más tiempo M-BERT por ser uno de los primeros en ser entrenado por los autores de BERT.

Tabla 15. Modelos propuestos basados en BERT que fueron pre-entrenados para el lenguaje español.

Autores	Modelos	Idiomas
Devlin et al. (2019)	M-BERT	102 lenguajes
Cañete et al. (2020)	BETO	Español
Cañete et al. (2022)	DistilBETO	Español
Cañete et al. (2022)	ALBETO	Español

Aunque las variantes de los modelos basados en BERT en su mayoría usan la arquitectura original empleada por los autores. En algunos casos fue modificada para obtener mejores resultados, ya sea empleando menor o mayor número de capas, diferentes parámetros entre otros.

La configuración de los modelos propuestos se muestra en la Tabla 16, algunos valores no fueron publicados por lo que se dejan vacíos. Los dos primeros modelos utilizan una arquitectura muy parecida ya que se basan en BERT, el tercer modelo (DistilBETO) tiene parámetro parecidos pero menores debido a que se basa en una destilación de BERT y finalmente los dos últimos modelos que son variantes de ALBETO uno ligero y otro grande emplean métodos que difieren del BERT estándar.

Tabla 16. Configuración de los *transformers* basados en BERT

Modelo	Capas	Ocultas	Cabezas	Parámetros
M-BERT	12	768	12	110M
BETO	12	1024	16	110M
DistilBETO	6	-	-	67M
ALBETO *tiny*	4	312	-	5M
ALBETO *xxlarge*	12	4096	-	223M

5.5 Clasificación

La etapa de clasificación consiste en entrenar y probar los modelos tanto de línea base como los propuestos, este proceso radica en probar diferentes hiperparámetros para encontrar los valores óptimos y así obtener los mejores modelos posibles.

5.5.1 Línea base

Para el entrenamiento de línea base, se emplean diferentes formas de caracterización como n-gramas con valores Tf-idf y *embeddings* de palabras. Los hiperparámetros dependen del algoritmo a utilizar para SVM es el kernel, mientras que para el Random forest es el número de estimadores (número de árboles), aunque ambos cuentan con más hiperparámetros estos son los más importantes. La Tabla 17 muestras los tipos de kernel y número de estimadores utilizados según el algoritmo.

Tabla 17. Hiperparámetros de los algoritmos de línea base.

Algoritmo	Hiperparámetros
SVM	Kernel: lineal, polinomial, función de base radial (rbf), sigmoide
Random Forest	Estimadores: 100, 200, 250, 300, 350

Los algoritmos fueron implementados con la ayuda de la librería Scikit-learn creando un *pipeline* para cada algoritmo aplicado (RandomForestClassifier y SVC). La matriz Tf-idf se calculó con la función TfidfVectorizer, designando el valor $ngram\_range = (2,2)$ para bigramas y $ngram\_range = (3,3)$ para trigramas.

Finalmente, los *embeddings* aplicados fueron NNLM y USE Multilingüe entrenados para el lenguaje español, para su implementación se obtuvieron de TensorFlow Hub, el cual es un repositorio que pone a disposición modelos entrenados. La Tabla 18 muestra la información de cada uno de los *embeddings* utilizados, somo su nombre lo indica USE Multilingüe tiene soporte para diferentes lenguajes entre ellos: español, inglés, alemán, ruso, portugués, japones, entre otros.

Tabla 18. Información de los embeddings aplicados.

Arquitectura	Nombre	Conjunto de datos	Lenguaje
NNLM	nnlm-es-dim128	Google News	Español
CNN	universal-sentence-encoder-multilingual	-	16 lenguajes (9 familias de lenguaje)

El proceso de entrenamiento y prueba para la línea base consistió en tres pasos, no se considera la división de los conjuntos ya que este paso fue realizado después del procesamiento de datos (se dividió en 80% de datos para el conjunto de entrenamiento y 20% para prueba). El primero es caracterizar los conjuntos de datos tanto de entrenamiento como de prueba según el método a utilizar, después declarar la función del algoritmo a implementar (SVM o Random forest) y finalmente predecir los valores del conjunto de prueba para evaluar el modelo.

Se itero entre cada método y algoritmo, cambiando los hiperparámetros de cada uno para así determinar el mejor modelo, por ejemplo: unigramas con SVM, unigramas con Random forest, bigramas con SVM y bigramas con Random forest, así sucesivamente.

5.5.2 Modelos propuestos

El entrenamiento de los *transformers* implica un proceso parecido al aplicado en la línea base, pero sin necesitar funciones externas de extracción de características de los datos. Al llamar los *transformers* estos cuentan con una función propia de preprocesamiento la cual calcula las representaciones del texto según el contexto, es decir aplica su propio *embedding*.

Los modelos propuestos (M-BERT, BETO, DistilBETO, ALBETO *tiny*, ALBETO *xxlarge*) no están disponibles en todas las bibliotecas de aprendizaje profundo, de los cinco modelos solo M-BERT está directamente en TensorFlow los demás si bien esta soportados por TensorFlow y PyTorch los modelos se encuentran en Hugging Face.

El proceso que conlleva entrenar este tipo de modelos es laborioso puesto que implica varios pasos que no siempre son claros, es por ello que el utilizar un contenedor o envoltorio (wrapper) para librerías de aprendizaje automático ayuda a que la implementación sea más rápida y sencilla.

Ktrain es *wrapper* que ha sido usado en diferentes investigaciones como detección de discursos de odio (Das et al., 2021), clasificación de juego de palabras (Palma Preciado et al., 2022) y análisis de sentimiento (Othman & Yaakub, 2022), en este caso se utiliza para cargar, entrenar y guardar los modelos.

La Tabla 19 presenta los hiperparámetros probados en los modelos de aprendizaje automático, aunque existen otras opciones a elegir se resumió a cuatro variables: taza de aprendizaje, épocas, tamaño de lote y longitud máxima (se contó cual era el tuit más largo y se tomó su longitud para definir el valor). Se itero entre las diferentes variantes para encontrar los valores más óptimos.

Tabla 19. Hiperparámetros probados en los modelos de aprendizaje profundo.

Taza de aprendizaje	2e-5, 5e-3
Épocas	2
Tamaño de lote	8, 10, 12, 14, 16, 32
Longitud máxima	60

5.6 Evaluación de los modelos

Esta sección presenta los resultados obtenidos con los diferentes tratamientos y modelos aplicados y se comparan las métricas calculadas para determinar el mejor modelo para la tarea de clasificación de reportes delictivos.

5.6.1 Línea base

Los resultados obtenidos para la línea base se presentan en la Tabla 20, entre los dos algoritmos empleados (SVM y Random forest) las máquinas de vectores de soporte alcanzaron mejores resultados en cada método de caracterización, aunque en trigramas fueron muy parecidos ya que el valor F1 tiene una diferencia de una décima.

El método convencional de unigramas o bolsa de palabras alcanzo un valor F1 de 0.60 para SVM, por arriba de los resultados tanto para bigramas y trigramas que tienen un valor F1 de 0.56 y 0.46 para SVM respectivamente. Aun cuando se probó usar una combinación entre bigramas y trigramas este resultado no supero lo obtenido por si solos.

Por último, los mejores resultados se lograron con el uso de *embeddings*, específicamente USE Multilingüe ya que su valor F1 fue de 0.78 y un recall de 0.70, muy por arriba de los resultados de Random Forest (valor F1 de 0.47).

Se encontró que SVM se comporta mejor con USE Multilingüe a diferencia de Random Forest que obtiene mejores resultados con el *embedding* NNLM.

Tabla 20. Resultados de los algoritmos de línea base.

Método		Algoritmos	Precisión	Recall	Valor F1
N-gramas	**Unigramas**	SVM	0.89	0.46	0.60
		Random forest	0.99	0.38	0.55
	Bigramas	SVM	0.98	0.40	0.56
		Random forest	0.98	0.34	0.50
	Trigramas	SVM	**1.0**	0.30	0.46
		Random forest	0.97	0.29	0.45
	Bigramas/ Trigramas	SVM	**1.0**	0.34	0.51
		Random forest	0.96	0.31	0.47
Embedding	**NNLM**	SVM	0.87	0.64	0.73
		Random forest	0.97	0.38	0.54
	USE Multilingüe	SVM	0.89	**0.70**	**0.78**
		Random forest	0.99	0.31	0.47

Los resultados del entrenamiento de los algoritmos utilizando traducción ida-vuelta se presentan en la Tabla 22, de forma general el recall mejoro superando los resultados reportados en la Tabla 20. USE Multilingüe y NNLM con SVM alcanzan los mejores puntajes con un valor F1 de 0.78 y 0.77, y un recall de 0.78 y 0.77, respectivamente. Al igual que en la tabla anterior, la precisión más alta se presenta con los métodos de trigramas, bigramas/trigramas y USE Multilingüe.

Tabla 21. Resultados de los algoritmos de línea base utilizando aumento de datos (Traducción ida-vuelta).

	Método	Algoritmos	Precisión	Recall	Valor F1
Aumento de datos Traducción ida-vuelta	**Unigramas**	SVM	0.85	0.56	0.68
		Random forest	0.92	0.46	0.61
	Bigramas	SVM	0.95	0.47	0.63
		Random forest	0.92	0.43	0.59
	Trigramas	SVM	**0.99**	0.34	0.50
		Random forest	0.87	0.39	0.53
	Bigramas/ Trigramas	SVM	**0.99**	0.38	0.55
		Random forest	0.89	0.40	0.55
	NNLM	SVM	0.85	0.70	0.77
		Random forest	0.91	0.49	0.64
	USE Multilingüe	SVM	0.85	**0.73**	**0.78**
		Random forest	0.95	0.45	0.61

Por otro lado, la técnica de submuestreo para balancear datos logra mejorar el desempeño de los algoritmos con un rango significativo, las máquinas de soporte de vectores con USE Multilingüe y con NNLM alcanzan un recall de 0.92 y 0.91, respectivamente (Ver Tabla 22).

A diferencia de los resultados anteriores la precisión y el valor F1 baja. Aunque el recall sube, el valor F1 depende de ambas métricas. Por tal motivo al realizar el cálculo entre ambas este valor no refleja un mejor comportamiento que los tratamientos anteriores. El valor F1 más alto fue SVM con bigramas/trigramas y SVM con NNLM, 0.63 y 0.60 respectivamente. Al contrario, SVM con unigramas obtuvo el valor F1 más bajo con 0.47.

Tabla 22. Resultados de los algoritmos de línea base aplicando submuestreo de datos.

	Método	Algoritmos	Precisión	Recall	Valor F1
Aumento de datos Submuestreo	**Unigramas**	SVM	0.33	0.83	0.47
		Random forest	0.45	0.83	0.58
	Bigramas	SVM	0.44	0.77	0.56
		Random forest	0.46	0.71	0.56
	Trigramas	SVM	**0.89**	0.39	0.54
		Random forest	0.58	0.48	0.53
	Bigramas/ Trigramas	SVM	0.60	0.67	**0.63**
		Random forest	0.55	0.66	0.60
	NNLM	SVM	0.45	0.91	0.60
		Random forest	0.41	0.85	0.55
	USE Multilingüe	SVM	0.43	**0.92**	0.59
		Random forest	0.39	0.88	0.54

5.6.2 Modelos propuestos

Los resultados de los *transformer* propuestos se presentan en las siguientes tablas, se reporta solo el desempeño de los modelos en su versión *uncase*, ya que al experimentar con los modelos *case* los resultados estuvieron por debajo o parecidos a los obtenidos con *uncase*. Para realizar el entrenamiento se utilizó Google Colab con un entorno de ejecución GPU para el cual se asignó un Centro de datos de NVIDIA: NVIDIA-SMI 460.32.03 versión de controlador: 460.32.03 y RAM de 27.3 GB.

La Tabla 23 muestra los resultados del entrenamiento de los modelos con el conjunto de datos completo utilizando el método de Split 80:20. Los resultados obtenidos no difieren mucho entre sí, BETO alcanza mejores valores en la mayoría de las métricas de evaluación (recall 0.84, valor F1 0.86) con excepción de la precisión donde ALBETO xxlarge alcanza 0.91. Sin embargo, su versión más ligera ALBETO tiny obtiene los resultados más bajos.

Aunque parecidos, el comportamiento de los modelos difiere en las variantes implementadas dentro de la arquitectura de cada uno, es por ello que los resultados pueden llegar a variar entre sí.

Tabla 23. Resultados de los modelos propuestos utilizando el conjunto de datos completo.

Modelo	Precisión	Recall	Valor F1
M-BERT	0.87	0.79	0.83
BETO	0.87	**0.84**	**0.86**
DistilBETO	0.88	0.81	0.84
ALBETO tiny	0.89	0.76	0.82
ALBETO xxlarge	**0.91**	0.79	0.85

Al aplicar el método de aumento de datos traducción ida-vuelta el desempeño de los modelos cambia, ALBETO xxlarge alcanza un valor F1 de 0.86 igual al reportado por BETO en la Tabla 23 (Ver Tabla 24). BETO mantuvo un recall de 0.84. Sin embargo, aunque el margen es pequeño la precisión bajo para todos los modelos.

Tabla 24. Resultados de los modelos propuestos utilizando aumento de datos (Traducción ida-vuelta).

Modelo	**Precisión**	**Recall**	**Valor F1**
M-BERT	0.84	0.82	0.83
BETO	0.83	**0.84**	0.84
DistilBETO	0.86	0.82	0.84
ALBETO tiny	0.87	0.79	0.83
ALBETO xxlarge	**0.88**	0.83	**0.86**

Por último, la Tabla 25 muestra los resultados del entrenamiento de los modelos con el conjunto balanceado por medio del método de submuestreo. Los resultados reportados con esta técnica superan a los métodos anteriores en recall, BETO y ALBETO tiny obtiene un puntaje de 0.94. Al igual que el caso de la Tabla 24 el recall mejora, pero la precisión y el valor F1 baja considerablemente, ya que este método reporta los peores puntajes entre todos los modelos entrenados.

Tabla 25. Resultados de los modelos propuestos utilizando submuestreo.

Modelo	**Precisión**	**Recall**	**Valor F1**
M-BERT	0.45	0.91	0.60
BETO	0.49	**0.94**	0.65
DistilBETO	0.47	**0.94**	0.63
ALBETO tiny	0.48	0.91	0.63
ALBETO xxlarge	**0.54**	0.91	**0.67**

De los resultados anteriores se encuentra que si bien la línea base es una opción para la clasificación de textos en reporte y no reporte con un valor F1 0.78 para SVM usando *embeddings*, esta es superada por los modelos propuestos del paradigma de los *transformers* ya que BETO obtuvo un valor F1 de 0.86 con el conjunto de datos original y ALBETO xxlarge alcanzo el mismo puntaje, pero aplicado aumento de datos.

Análisis de los resultados

Después de realizar la etapa de evaluación para identificar el modelo con mejor desempeño, se realiza un análisis de los valores de las probabilidades obtenidas al predecir la clase a la que pertenece un texto. En este caso debido a que BETO y ALBETO xxlarge alcanzaron el mismo valor F1 de 0.86 se muestran los tres mejores y peores puntajes obtenidos por cada uno de los modelos para la clase reporte (etiqueta de interés). Este paso se realiza con la intención de visualizar el comportamiento de los modelos y como es que clasifican los textos.

BETO

Como se menciona anteriormente, BETO alcanzo el mejor puntaje sin considerar técnicas de balanceo de datos, es por ello que se elige como el modelo que alcanzo mejor desempeño ya que no necesito una cantidad mayor de datos para predecir correctamente reportes delictivos.

La Tabla 26, muestra los tres mejores puntajes de la clase reporte. Los ejemplos alcanzan una probabilidad de 0.996 de ser reportes, lo cual es correcto. El conjunto de datos cuenta con varios textos de Alerta AMBER, esto pudo haber ayudado a que el modelo las detecte con mayor facilidad, ya que dos de los textos mostrados pertenecen a este tipo de caso.

Tabla 26. Mejores puntaciones de BERT para la clase reporte.

Texto	**Probabilidad**
Agentes de cumplimentaron orden de aprehensión en contra de Vanessa Ballar Fallas alias La Gera por probable participación en homicidio de 2 hombres israelíes asesinados con armas de fuego en un restaurante en Plaza Artz	0.99670
Vía difundimos la ficha de búsqueda en favor de la menor de 12 años LeylaniJacqueline Rodríguez Montes de Oca vista por Última vez el da 24 de marzo 2021 en la colonia Centro alcaldía Cuauhtémoc Comparte AmberAlerta	0.99668
Vía difundimos la ficha de búsqueda para localizar a la menor de 13 años de nombre Vania Victoria Bautista Camarillo vista por Última vez el día 24 de marzo 2021 en la colonia Centro alcaldía Cuauhtémoc Comparte AmberAlerta	0.99664

A diferencia de la tabla anterior, la Tabla 27 muestra los tuits con los peores puntajes que, aunque pertenecen a la clase reporte el modelo los clasifico como no reporte (la probabilidad de ser reporte tiende a 0). Esto puede ocurrir debido a que los textos son poco descriptivos y cortos.

Tabla 27. Peores puntaciones de BERT para la clase reporte.

Texto	**Probabilidad**
Al tiro con la rata y con el articulado La Esquina del Atraco in Coyoacán Mexico City	0.00121
Muerte vial cdmx atropellado cdmx	0.00155
URGENTE terminar con la corrupción en la UNAM	0.00189

ALBETO xxlarge

El modelo ALBETO en su versión grande y usando la técnica de aumento de datos traducción ida-vuelta alcanzó el mismo puntaje de BETO con un valor F1 de 0.86, por tal motivo también se presentan los mejores y peores resultados para comparar si los textos son parecidos o difieren en la forma de clasificar.

A diferencia de BETO, ALBETO obtienen mejores puntuaciones con reportes de inconformidad, sucedidos en un área o hacia una persona. Estos textos obtienen una probabilidad alta de 0.999 de ser reportes (Ver Tabla 28).

Tabla 28. Mejores puntaciones de ALBETO para la clase reporte.

Texto	Probabilidad
Apoyo para remover objetos en vía pública en la calle Quetzal col Rosedal CP 04330 que si uno pasa por ahí los vecinos agreden si uno toca los mencionados en especial el de la casa con el 75 y	0.99968
agentes adscritos a la dirección general de la amenazan a comerciantes para que les paguen las famosas rentas es el grupo del comandante juan carlos almaraz ortiz mismo que hace días fue detenido por agredir a sus vecinos en la alcaldía	0.99963
Cc más impunidad no más corrupción en la Alcadia Venustiano Carranza apoyo jurídico	0.99956

La Tabla 29 muestra los puntajes más bajos, donde ALBETO clasifico un texto como no reporte cuando si lo era. Entre ellos se encuentran Alertas AMBER, denuncias de trajo y denuncias de construcción. Utilizando estos tuits el modelo no muestra un patrón aparente de la forma en que aprende del conjunto.

Tabla 29. Peores puntaciones de ALBETO para la clase reporte.

Texto	Probabilidad
Denuncian agresión contra trabajador de Notimex en huelga Un empleado de la agencia sin razón alguna agrede al reportero y fue acusado de ser enviado por La Prensa Noticias policiacas locales nacionales	0.00106
Vía difundimos la ficha de búsqueda por Cristofer Alexander Ramírez González Cristofer Alexander de 13 años visto por Última vez el día 7 de marzo de 2021 en la colonia Mixcóatl alcaldía Iztapalapa Tuparticipacionesvital	0.00110
como retirar estas estructuras fijas son un abuso no deben de ir Y el las noches este resta es un bar clandestino calle Altadena entre Insurgentes y Dakota	0.00154

5.7 Análisis geoespacial

El análisis geoespacial permite encontrar patrones en datos con componentes geoespaciales, una de las formas más común es mediante mapas temáticos que pueden ser coropléticos, mapas de calor, distribución de puntos, cartograma, entre otros. Esta etapa se divide en dos secciones, el modelado geoespacial que muestra el proceso llevado a cabo para procesar el conjunto de datos y generar los mapas pertinentes y el análisis de índice delictivo donde se presentan los mapas realizados durante el modelado y los patrones encontrados.

5.7.1 Modelado geoespacial

El modelado geoespacial se realiza con los conjuntos de datos finales de ambas fuentes, tanto de datos abierto como de Twitter, con el fin de analizar y comparar los resultados de cada uno. Antes de iniciar es necesario realizar una evaluación del área de estudio, en este caso la Ciudad de México, para identificar zonas pobladas, delimitación territorial, áreas protegidas, entre otras.

Como primera capa se obtiene el archivo shapefile con las delimitaciones de las alcaldías de la Ciudad de México (Agencia Digital de Innovación Pública (ADIP), 2022), después las colonias de la ciudad (Instituto Electoral de la Ciudad de México, 2022) y por último al consultar diferentes mapas del área se encuentra que la ciudad cuenta con 25 áreas naturales protegidas (ANP) que abarcan más del 14% del territorio (SEDEMA, 2022). Finalmente se genera un mapa con las tres capas para utilizar en los pasos posteriores donde se generan mapas categorizados y de calor sobre los reportes delictivos realizados. La Figura 39 muestra el proceso y resultado de las capas base.

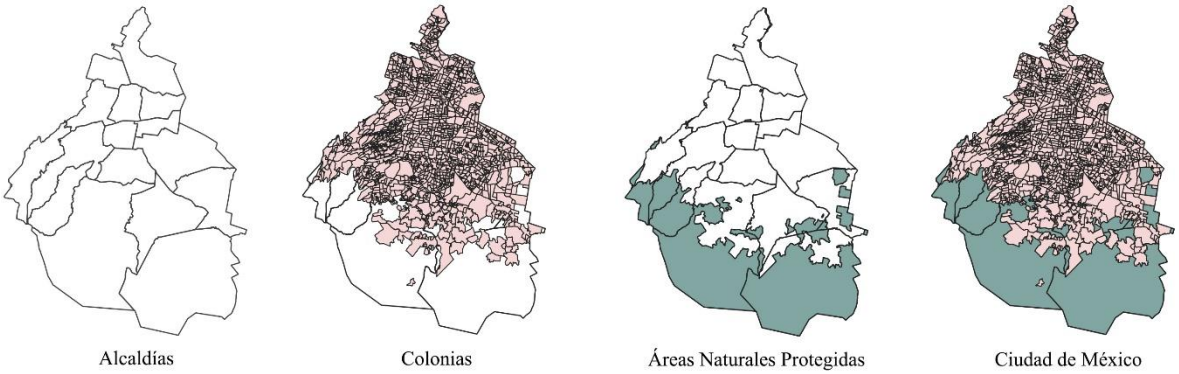

Figura 39. Análisis del área de estudio y creación de capas base.

El análisis principal consiste en contar los delitos reportados por colonias y alcaldías para determinar en qué zona existe mayor índice delictivo. Primeramente, se cargar las capas base que se mostraron en la Figura 39 y después importan los datos limpios de las carpetas de investigación. Debido a que se tienen dos archivos de los datos abiertos, uno con los datos que estaban completos desde un inicio y el otro con los datos abiertos georeferenciados. Es necesario juntar estos dos subconjuntos, para este paso se utiliza el proceso de vector general → unir capas vectoriales.

Posteriormente el para calcular la cantidad de delitos que ocurren una zona (colonia o delegación) se utiliza el análisis de vector → contar puntos de un polígono este algoritmo permite como su nombre lo indica contar los puntos que se encuentran dentro de un polígono siendo los puntos cada delito reportado y los polígonos las colonias.

Para visualizar los resultados se categoriza de forma continua los valores obtenidos, teniendo así un mapa con diferentes colores donde el color más claro indica un menor índice delictivo y por el contrario los obscuros mayor cantidad de delitos reportados (Ver Figura 40).

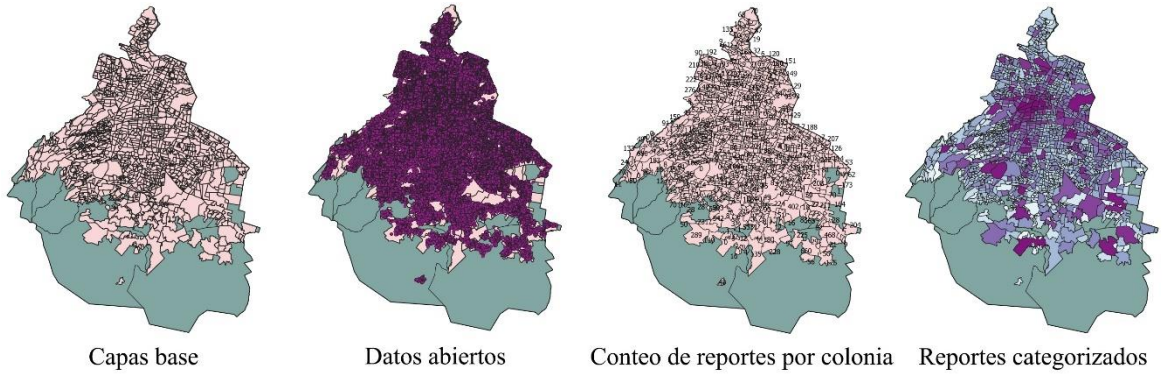

Figura 40. Conteo de puntos en un área geográfica.

Los mapas de calor son una representación gráfica de los datos que permite visualizar la concentración de los datos. La Figura 41 muestra el proceso a seguir para desde una capa vectorial de puntos generar un mapa de calor en QGIS. Primero se importan los datos al SIG, después se selecciona Interpolación → Mapa de calor (Estimación de Densidad de Núcleo) lo cual creara una capa ráster (se eligió un radio de 0.01 grados y un valor de 500 para las filas con el fin de cambiar las dimensiones de la capa de salida), esta capa ráster puede ser editada para usar colores más representativos (rampa de colores).

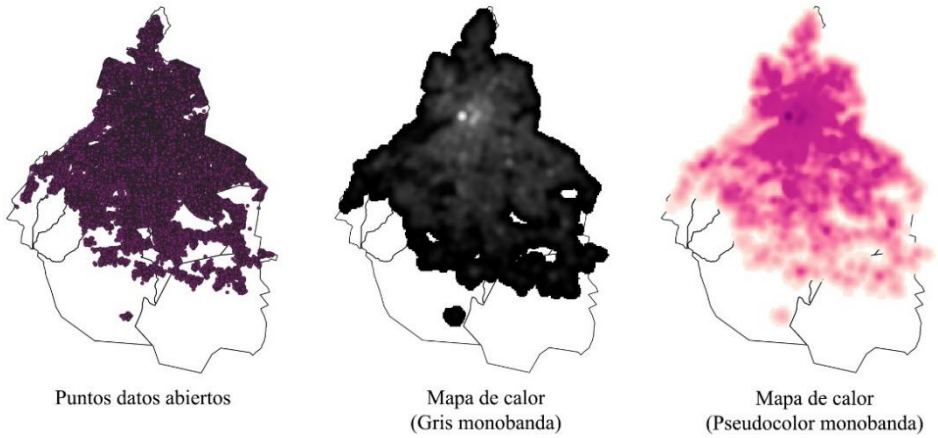

Figura 41. Proceso para la creación de mapas de calor usando puntos.

Los pasos antes mencionados para generar los mapas temáticos fueron aplicados para ambos conjuntos de datos (datos abiertos y Twitter). Los datos de Twitter a diferencia de los las investigaciones oficiales por parte de la FGJ que solo contenían atributos en forma de dirección y coordenadas, estas por su parte podían ser georreferenciadas de tres formas coordenadas, cuadro delimitador y direcciones, por lo que se generaron tres archivos diferentes que se unieron al convertirse en capas vectoriales con el proceso antes descrito (la distribución de los datos puede ser consultada en el Anexo 7).

5.7.2 Análisis de índice delictivo

El análisis geoespacial permite a través de mapas mostrar diferentes capas de información sobre un tema en específico lo cual ayuda en la toma de decisiones. A continuación, se presentan los mapas realizados en la etapa de modelado geoespacial.

Las Figuras 42 y 43 muestra los mapas temáticos generales del índice delictivo de las alcaldías de la Ciudad de México. La Figura 42 representa los reportes de las carpetas de investigación de la FGJ, en el cual se puede observar la alcaldía Iztapalapa (34,424 reportes) como la más delictiva, seguida de Cuauhtémoc (32,572) y Gustavo A. Madero (24,125). Por el contrario, la alcaldía Milpa Alta tiene menos reportes (2,476).

El mapa de la Figura 43, muestra la incidencia delictiva del conjunto de datos de Twitter. La alcaldía con mayor cantidad de reportes es Benito Juárez (590 tuits) seguida de Iztapalapa con 421 y Cuauhtémoc con 332 reportes. En contraste las alcaldías Milpa Alta y Álvaro Obregón con 8 y 52 reportes respectivamente. En ambos casos se encuentra a la alcaldía Cuauhtémoc entre las más conflictivas, y Milpa como la más segura.

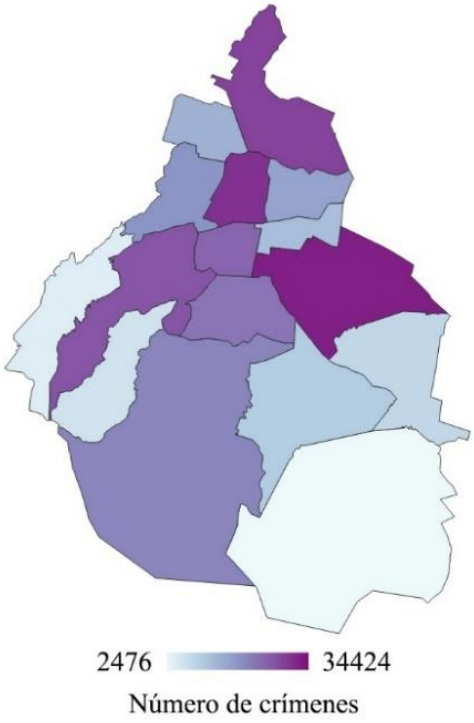

Figura 42. Mapa coroplético de incidencia delictiva por alcaldías reportada en los datos abiertos.

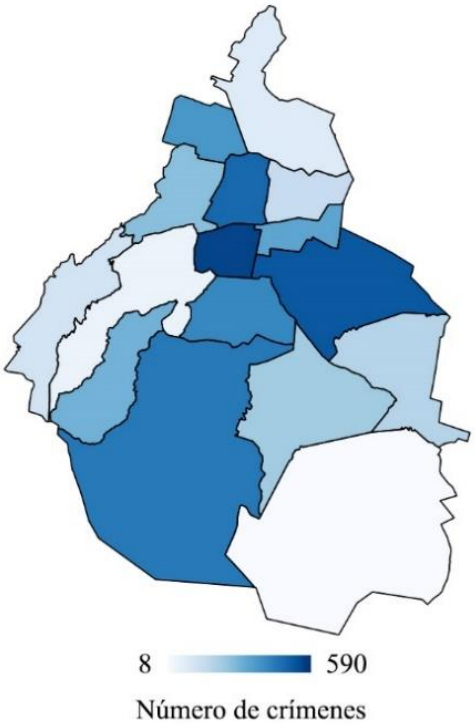

Figura 43. Mapa coroplético de incidencia delictiva por alcaldías reportada en Twitter.

Como parte complementaria también se realiza una gráfica de los valores delictivos obtenidos por cada colonia. La Figura 44 muestra que los delitos reportados en ambos conjuntos suelen seguir la misma tendencia, si bien el conjunto de datos de Twitter contiene menos registros que el conjunto de datos abiertos, se encuentran puntos similares como en las alcaldías Azcapotzalco, Coyoacán, Iztacalco, Iztapalapa, Milpa Alta, Tláhuac, Tlalpan y Xochimilco.

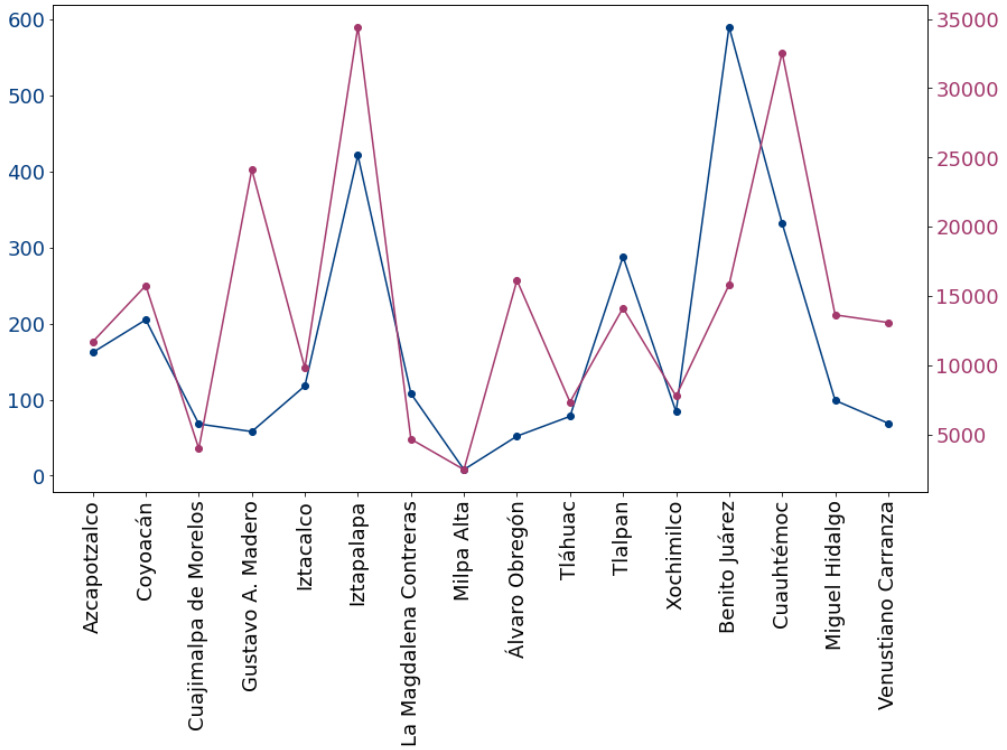

Figura 44. Delitos cometidos por alcaldía.

Las Figuras 45 y 46 presentan el mismo análisis realizado en las figuras anteriores, pero desde una perspectiva más específica puesto que se analizan los delitos por colonias. La Figura 45 muestra los delitos reportados oficialmente (las zonas de color más obscuro indican mayor índice delictivo y los colores más claros menos cantidad de crímenes). Utilizando el análisis realizado del territorio de la ciudad se encuentran que existen reportes fuera del plano de las colonias, lo que significa que estos ocurrieron en un ANP. Al realizar el conteo 873 registros pertenecen a este suceso, aunque no se descarta la posibilidad de que alguno de estos datos pueda estar mal georreferenciado.

Esto mismo ocurre en los reportes de Twitter, aunque en menor cantidad (66 registros). A diferencia de los datos abiertos los puntos de los tuits que se encuentran fuera de una colonia se encuentran justo a la mitad de las alcaldías en las que ocurrió el suceso, lo que indica que durante el proceso de geocodificación la dirección mencionada en el texto no fue lo suficientemente descriptiva para posicionar correctamente el tuit (Ver Figura 46).

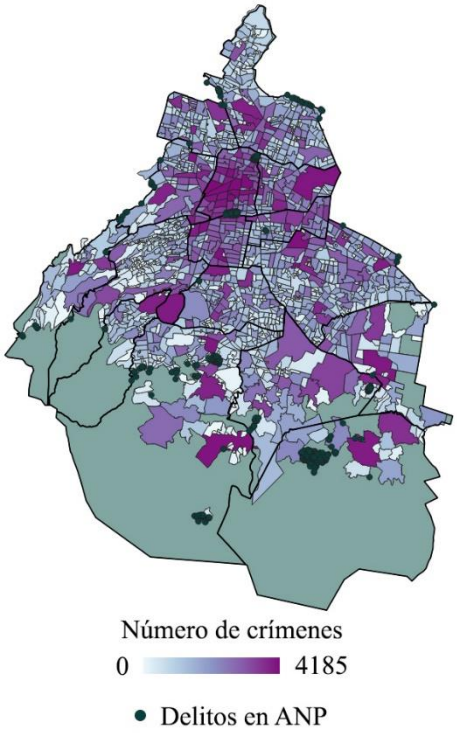

Figura 45. Mapa coroplético de incidencia delictiva reportada por colonias del conjunto de datos abiertos.

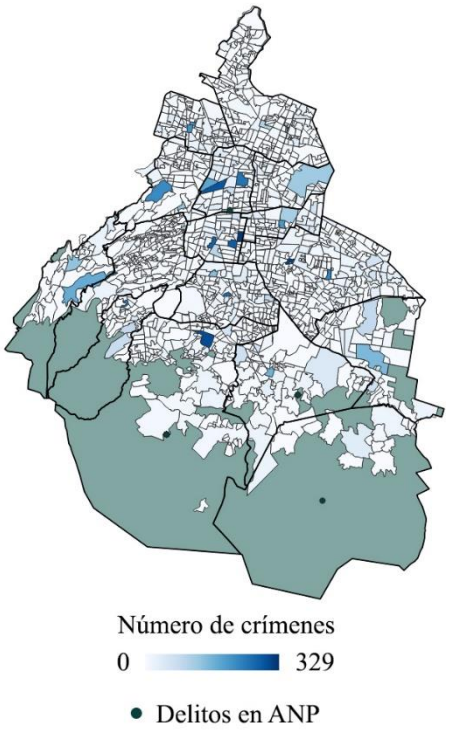

Figura 46. Mapa coroplético de incidencia delictiva reportada por colonias del conjunto de datos de Twitter.

De igual manera las Figuras 47 y 48 muestran las colonias con más crímenes, a pesar de que la cantidad de datos no es proporcional entre los conjuntos de datos se hacer una comparación entre la cantidad de reportes que realiza la ciudadanía.

La figura 47 muestra que los delitos se concentran en el centro de la ciudad destacando las colonias Cuauhtémoc, Centro, Doctores y Juárez. En contraparte, la zona más segura es La venta en Cuajimalpa de Morelos con cero reportes.

La Figura 48 muestra las colonias con mayor actividad delictiva que son Benito Juárez colonia Nativitas (329 tuis) e Iztapalapa colonia San Pablo (177 tuis). Al comparar ambos mapas se repiten zonas de alta incidencia como Cuauhtémoc, Juárez, Lomas de Chapultepec, San Pedro Cuajimalpa, Tlalpan centro, Peñon de los Baños, Ramos Millán Bramadero I y San Pedro Tláhuac.

Si bien gran parte de las colonias no cuentan con reportes para el caso de conjunto de Twitter, esto podría ocurrir debido a la poca población que utiliza la plataforma o baja delincuencia. Aun así, el uso de redes sociales es una forma viable de analizar la inseguridad y seguridad que existe en un área determinada.

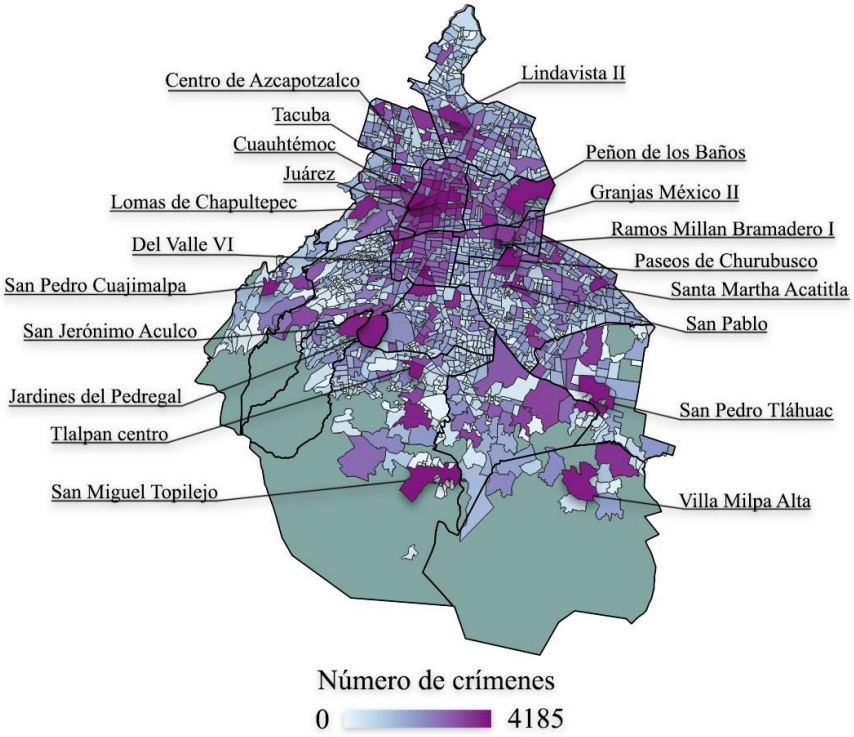

Figura 47. Mapa coroplético en el cual se resaltan las colonias con mayor cantidad de reportes del conjunto de datos abierto.

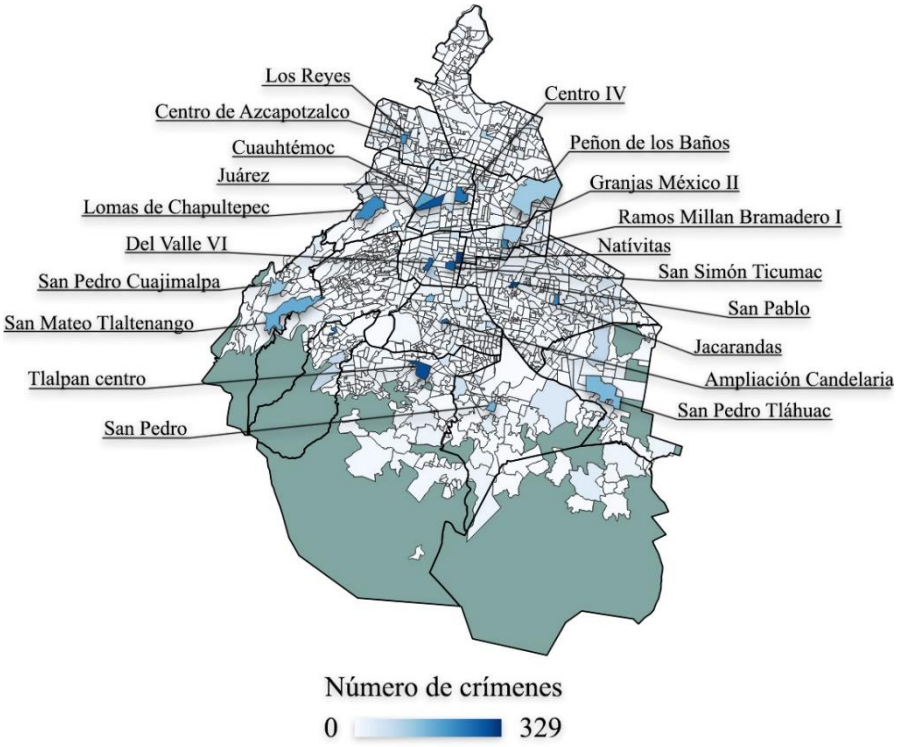

Figura 48. Mapa coroplético en el cual se resaltan las colonias con mayor cantidad de reportes del conjunto de datos de Twitter.

Finalmente se crearon mapas de calor con cada conjunto de datos para la Ciudad de México, para ambos conjuntos se encuentra que los delitos suceden en zona parecidas ya que su distribución en el mapa es similar, y que generalmente los delitos tienden a ocurrir en los centros de las alcaldías, disminuyendo conforme se acercan al límite territorial de la ciudad (Ver Figura 49 y 50).

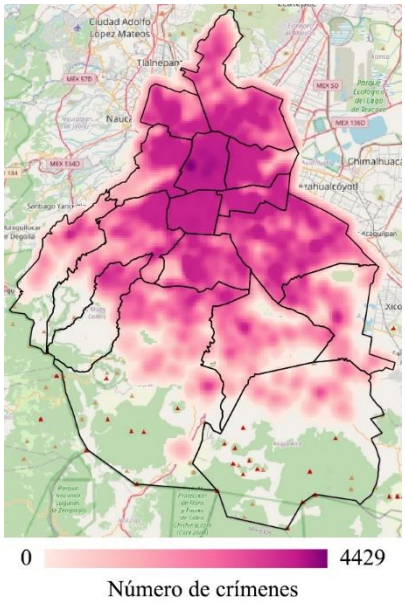

Figura 49. Mapa de calor de los delitos provenientes de conjunto de datos abierto.

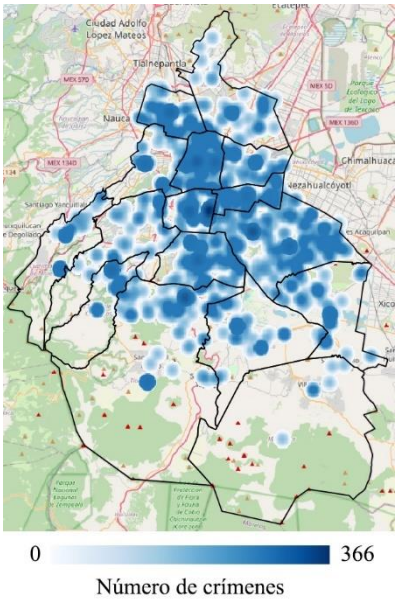

Figura 50. Mapa de calor de los delitos provenientes del conjunto de datos de Twitter.

Capítulo 6. Conclusión y trabajo a futuro

En este capítulo se presentan las conclusiones del trabajo realizado, así como también se resaltan los hallazgos y los puntos relevantes encontrados durante la investigación.

6.1 Conclusión

El uso de diferentes fuentes de datos como datos abiertos y redes sociales permiten realizar un análisis y clasificación de reportes delictivos realizados en un área de estudio. Se encontró que la extracción de datos desde Twitter es viable ya que, aunque el conjunto es pequeño se lograron identificar zonas de riesgo, que al ser comparadas con las investigaciones de fuentes oficiales estas reportan resultados parecidos.

Las técnicas de procesamiento de lenguaje natural ayudaron a procesar los datos de tal manera que después de la limpieza y manejo de datos se creó un conjunto de datos. Este después fue dividido dos subconjuntos en entrenamiento y prueba con una distribución 80:20 para entrenar los modelos de línea base (SVM y Random Forest) y los modelos de aprendizaje profundo basados en BERT.

Se entreno con éxito un clasificador que puede predecir si un texto es reporte delictivo o no. BETO alcanzo un valor F1 de 0.86, una precisión de 0.87 y 0.84 en recall, esto indica que para la tarea de clasificación de reportes delictivos el uso de aprendizaje automático es una buena solución.

Aunque el conjunto de datos tenía un desbalance moderado con un índice de 11.13 y se aplicaron técnicas de balanceo de datos los resultados no mejoraron en comparación con los modelos entrados con el conjunto de datos original.

Finalmente, el modelado geoespacial ayudo a crear representaciones visuales en forma de mapas sobre la situación actual que sucede en la ciudad, estos permitieron detectar las zonas de mayor riesgo que después pueden ser usadas para la toma de ediciones en el ámbito de seguridad ciudadana.

Para concluir, la metodología propuesta logro implementar métodos de PLN y modelado geoespacial para clasificar y encontrar patrones en zonas con mayor inseguridad. Si bien se eligió como área de estudio la Ciudad de México este proceso puede ser replicado en otras entidades.

6.2 Discusión

El desbalanceo de datos y conjuntos de datos pequeños suelen encontrar problemas de desempeño y aunque existen diferentes métodos para contrarrestar esta situación, encontrar el método más favorable no es tarea fácil. Durante el periodo de experimentación se probaron varias técnicas, entre ellas aumentos de datos como traducción ida-vuelta, pero también el método EDA modificado para el idioma español, este no fue reportado en los resultados por los bajos puntajes alcanzados tanto para línea base como para los *transformers*.

Este suceso puedo haber corrido debido a la composición de los tuits que, al ser textos cortos y EDA al implementar remplazo de sinónimos y eliminación de palabras, entre otras pudiera intervenir con el significado de los textos, aunque es necesario realizar un análisis más exhaustivo.

Por otro lado, los *transformers* al ser pre-entrenados en una gran cantidad de datos se comportan bien con conjuntos de datos pequeños después de realizar *fine tuning*. Aunque el corpus creado supera la muestra mínima necesaria, no cuenta con tantos registros de la clase de interés (reporte), por lo que el problema de la creación del conjunto de datos fue un inconveniente que se pudo resolver.

Otro aspecto a destacar es la falta de conjuntos de datos en español, al realizar un estudio de los trabajos realizados se encontró poco avance en este aspecto, si bien existen diferentes investigaciones relacionadas que aplican métodos de vanguarda existe la necesidad de aplicar y experimentar con estos métodos en el lenguaje español.

Durante la recolección de tuits se usaron dos técnicas, una que dependía de un diccionario de palabras y otra de tuits realizados por y para cuentas oficiales gubernamentales relacionadas con la seguridad pública. Para el enfoque usado se descubrió que los textos recolectados con los diccionarios de palabras contuvieron menos reportes en comparación con los tuis recolectados de las cuentas oficiales, esto puede ocurrir debido a la atención que brindan las cuentas a tutis que reportan un crimen, al contrario de solo escribir y publicar un hecho que es difícil de encontrar y realizar un seguimiento.

6.3 Contribuciones

En este trabajo se recopilo, proceso y generó un conjunto de datos en español a partir de publicaciones de Twitter para la clasificación de reportes delictivos, si bien existen algunos trabajos relacionados a esta tarea los datos en español no ha sido un tema muy explorado.

Por otro lado, se desarrolló una metodología que implementa el uso de PLN y análisis geoespacial la cual permite tanto manejar la selección y recolección de datos como su análisis de forma visual por medio de mapas para identificar la ocurrencia de eventos delictivos.

Se entrenaron modelos de aprendizaje automático supervisado para clasificar los datos, entre los cuales se realizó el fine tuning de un transformer BETO, se encontró que este tipo de clasificadores se desempeñan bien en la tarea de detección de reportes delictivos.

6.4 Trabajo futuro

El conjunto de datos creado se compone del 33% de los datos recolectados, de tal forma que aún existe una muestra considerable por revisar y etiquetar y así crear un conjunto más extenso. También considerar extraer más datos y agregar otras cuentas oficiales para tener más variedad de reportes delictivos.

La clasificación realizada fue binaria donde se identificó si un texto era un reporte delictivo o no, una mejora a esta tarea es la clasificación del tipo de delito convirtiéndose en una tarea multi clase.

Ambos conjuntos de datos cuentan con más atributos que podrían ser relevantes en otros tipos de estudios como predicción de sucesos, identificación de zonas seguras, entre otras.

Finalmente, aunque existen modelos pre-entrenados con datos de Twitter estos no pudieron ser implementados por la falta de tiempo y bajo soporte, como trabajo futuro se sugiere el uso de este tipo de modelos para comparar su desempeño con los modelos entrenados en este trabajo.

Referencias

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. *In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). USENIX Association*, 265–283. https://www.tensorflow.org/

Agencia Digital de Innovación Pública (ADIP). (2022). *Alcaldias* [Dataset]. https://datos.cdmx.gob.mx/dataset/alcaldias

Aggarwal, C. C. (2015). *Data Mining: The Textbook* (2015 ed.). Springer. https://doi.org/10.1007/978-3-319-14142-8

AL-Saif, H., & Al-Dossari, H. (2018). Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques. International Journal of Advanced Computer Science and Applications, 9(10). https://doi.org/10.14569/ijacsa.2018.091046

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*. https://doi.org/10.1145/3544558

Bolstad, P. (2019). *GIS Fundamentals: A First Text on Geographic Information Systems* (Vol. 6). XanEdu.

Boukabous, M. & Azizi, M. (2022). Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers. *Indonesian Journal of Electrical Engineering and Computer Science*, 25(2), 1131. https://doi.org/10.11591/ijeecs.v25.i2.pp1131-1139

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA. *workshop paper at PML4DC, ICLR 2020*. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

Cañete, J., Donoso, S., Bravo-Marquez, F., Carvallo, A., & Araujo, V. (2022). ALBETO and DistilBETO: Lightweight Spanish Language Models. *Proceedings of the 13th Language Resources and Evaluation Conference*. https://arxiv.org/abs/2204.09145

Carrillo-Brenes, F., Vilches-Blázquez, L. M., & Mata, F. (2020). A Proposal for Semantic Integration of Crime Data in Mexico City28-30, 2020, Proceedings: 1276. En GIS Latam: First Conference, GIS Latam 2020, Mexico City, Mexico, September (2020 ed., pp. 30–48). Springer. https://doi.org/10.1007/978-3-030-59872-3

Chakraborty, G., Pagolu, M., & Garla, S. (2013). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS (English Edition)* (Illustrated ed.). SAS Institute.

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications Company.

Çöltekin, A., Janetzko, H., & Fabrikant, S. (2018). Geovisualization. *Geographic Information Science & Technology Body of Knowledge*, *2018*(Q2). https://doi.org/10.22224/gistbok/2018.2.6

CONABIO. (2022, 27 mayo). *Geovisualización*. Biodiversidad Mexicana. Recuperado 22 de agosto de 2022, de https://www.biodiversidad.gob.mx/region/geoviz.html

Das, S., Mandal, P., & Chatterji, S. (2021). Probabilistic Impact Score Generation using Ktrain-BERT to Identify Hate Words from Twitter Discussions. *ArXiv: Computation and Language*. http://arxiv.org/pdf/2111.12939.pdf

Dowlagar, S., & Mamidi, R. (2021). HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection. *arXiv*. https://doi.org/10.48550/arXiv.2101.09007

Ekman, M. (2021). *Learning Deep Learning*. Addison Wesley Professional.

Esri. (2022). *¿Qué es ArcGIS? | ArcGIS Resource Center*. ArcGIS. Recuperado 15 de agosto de 2022, de https://resources.arcgis.com/es/help/getting-started/articles/026n00000014000000.htm

Fenner, M. (2019). *Machine Learning with Python for Everyone*. Addison-Wesley Professional.

Gati, M., Medard, M., & Young-Seob, J. (2021). Sentiment Classification in Swahili Language Using Multilingual BERT. *arXiv*. https://doi.org/10.48550/arXiv.2104.09006

Gemasih, H., Rayuwati, R., SN, A., & Mursalin, M. (2019). Classification of Criminal Crimes From Data Twitter Using Class Association Rules Mining. *Proceedings of the Proceedings of the 1st Workshop on Multidisciplinary and Its Applications Part 1, WMA-01 2018, 19–20 January 2018, Aceh, Indonesia*. https://doi.org/10.4108/eai.20-1-2018.2281925

Giridhara, P., Mishra, C., Venkataramana, R., Bukhari, S., & Dengel, A. (2019). A Study of Various Text Augmentation Techniques for Relation Classification in Free Text. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. https://doi.org/10.5220/0007311003600367

Gobierno de la Ciudad de México. (2021, January 12). *Portal de Datos Abiertos de la CDMX*. https://datos.cdmx.gob.mx/dataset/victimas-en-carpetas-de-investigacion-fgj

Gobierno de la Ciudad de México. (2022). *Portal de Datos Abiertos de la CDMX*. https://datos.cdmx.gob.mx/+

González, J. N., Hurtado, L. F., & Pla, F. (2021). TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing*, 426, 58–69. https://doi.org/10.1016/j.neucom.2020.09.078

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Tin Kam Ho, Macia, N., Ray, B., Saeed, M., Statnikov, A. & Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. *2015 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn.2015.7280767

Hansen, D. L., Shneiderman, B., & Smith, M. A. (2011). Social Media. *Analyzing Social Media Networks with NodeXL*, 11–29. https://doi.org/10.1016/b978-0-12-382229-1.00002-3

Hernández, R. G., Rosas, S. D., & Bernábe, M. B. (2020). Análisis de la incidencia delictiva del fuero común mediante la red social Twitter. Research in Computing Science, 149(8), 451–463. https://www.rcs.cic.ipn.mx/rcs/2020_149_8/Analisis%20de%20la%20incidencia%20delictiva%20del%20fuero%20comun%20mediante%20la%20red%20social%20Twitter.pdfLal,

Hugentobler, M. (2008). Quantum GIS. *Encyclopedia of GIS*, 935–939. https://doi.org/10.1007/978-0-387-35973-1_1064

Hugging Face. (2022). *Introduction - Hugging Face Course*. Recuperado 24 de agosto de 2022, de https://huggingface.co/course/chapter1/1#introduction

Instituto Electoral de la Ciudad de México. (2022). *Colonias del IECM* - 2019 [Dataset]. https://datos.cdmx.gob.mx/dataset/coloniascdmx

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Karimi, A., Rossi, L., & Prati, A. (2021). AEDA: An Easier Data Augmentation Technique for Text Classification. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2748–2754. https://doi.org/10.18653/v1/2021.findings-emnlp.234

Kent, C., & du Boulay, B. (2022). *AI for Learning* (1.a ed.). CRC Press. https://doi.org/10.1201/9781003194545

Keras Team. (2022). *Keras documentation: Why choose Keras?* Keras. Recuperado 22 de agosto de 2022, de https://keras.io/why_keras/

Krohn, J., Beyleveld, G., & Bassens, A. (2019). *Deep Learning Illustrated*. Addison-Wesley Professional.

Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., Shorfuzzaman, M., Alsufyani, A., & Uddin, M. (2022). Addressing Binary Classification

over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare*, 10(7), 1293. https://doi.org/10.3390/healthcare10071293

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Cornell University - arXiv*. https://doi.org/10.48550/arxiv.1909.11942

Laurini, R. (2017). Geovisualization and Chorems. *Geographic Knowledge Infrastructure*, 223–246. https://doi.org/10.1016/b978-1-78548-243-4.50011-6

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Ltd, T. S. S. (2022, 16 junio). *22 Essential Twitter Statistics You Need to Know in 2022*. The Social Shepherd. Recuperado 25 de julio de 2022, de https://thesocialshepherd.com/blog/twitter-statistics

Maiya, A. S. (2020). ktrain: A Low-Code Library for Augmented Machine Learning. *Journal of Machine Learning Research (JMLR)*. https://doi.org/10.48550/arXiv.2004.10703

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (Illustrated ed.). Cambridge University Press.

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective, Second Edition (Chapman & Hall/Crc Machine Learning & Pattern Recognition) (English Edition)* (2.a ed.). Chapman and Hall/CRC.

MMC Ventures & Barclays UK Ventures. (2019). *The State of AI 2019: Divergence*. https://iec2021.aaru-confs.org/The-State-of-AI-2019-Divergence.pdf

Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*. https://doi.org/10.1109/icics49469.2020.239556

Murty, M. N., & Raghava, R. (2016). *Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks* (2016 ed.). Springer. https://doi.org/10.1007/978-3-319-41063-0

Nizamani, S., Memon, N., Shah, A. A., Nizamani, S., Nizamani, S., & Ismaili, I. A. (2015). Crime Analysis using Open Source Information. Sindh University Research Journal (Science Series), 47(4), 677–682. https://arxiv.org/ftp/arxiv/papers/1902/1902.05684.pdf

Nöllenburg, M. (2006). Geographic Visualization. *Human-Centered Visualization Environments*, 257–294. https://doi.org/10.1007/978-3-540-71949-6_6

Osorio, J., & Beltran, A. (2020). Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP. 2020 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn48605.2020.9207039

Othman, H., & Yaakub, M. R. (2022). Implementing BERT With Ktrain Library For Sentiment Analysis. *Journal of Visual Language and Computing*, 2022(2), 26–34. https://doi.org/10.18293/jvlc2022-n2

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding Fundamentals and Associated Challenges. *Geospatial Data Science Techniques and Applications*, 41–62. https://doi.org/10.1201/9781315228396-3

Palma Preciado, V. M., Sidorov, G., & Palma Preciado, C. (2022). Assessing Wordplay-Pun classification from JOKER dataset with pretrained BERT humorous models. *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, 1828–1833. http://ceur-ws.org/Vol-3180/paper-142.pdf

Pàmies, M. (2020). *Multilingual identification of offensive content in social media* (id: diva2:1451543) [Tesis maestría, Linköping University].

Pang, B., Nijkamp, E., & Wu, Y. N. (2019). Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics*, *45*(2), 227–248. https://doi.org/10.3102/1076998619872761

Piña-García, C. A., & Ramírez-Ramírez, L. (2019). Exploring crime patterns in Mexico City. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0228-x

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001. https://doi.org/10.18653/v1/p19-1493

PitchBook. (2022). *Hugging Face Company Profile: Valuation & Investors | PitchBook*. Recuperado 24 de agosto de 2022, de https://pitchbook.com/profiles/company/168527-08#overview

QGIS. (2022). *Descubre QGIS*. Recuperado 15 de agosto de 2022, de https://www.qgis.org/es/site/about/index.html#:%7E:text=QGIS%20proporciona%20una%20creciente%20gama,lista%20m%C3%A1s%20detallada%20de%20caracter%C3%ADsticas.

Quarati, A., de Martino, M., & Rosim, S. (2021). Geospatial Open Data Usage and Metadata Quality. *ISPRS International Journal of Geo-Information*, *10*(1), 30. https://doi.org/10.3390/ijgi10010030

Rouder, J. N. (2015). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069. https://doi.org/10.3758/s13428-015-0630-z

S., Tiwari, L., Ranjan, R., Verma, A., Sardana, N., & Mourya, R. (2020). Analysis and Classification of Crime Tweets. Procedia Computer Science, 167, 1911–1919. https://doi.org/10.1016/j.procs.2020.03.211

SEDEMA. (2022). *Áreas Naturales Protegidas*. http://www.sadsma.cdmx.gob.mx:9000/rally/pex/assets/pages/anp.php

Sentiment Analysis Pipeline for Spanish texts. (2021, November 29). John Snow Labs. https://nlp.johnsnowlabs.com/2021/11/29/classifierdl_bert_sentiment_pipeline_es.html

Sharma, N., Dhamne, A., More, N., Rathi, V., & Supekar, A. (2018). Detection of crime and non-crime tweets using Twitter. International Journal of Advanced Research in Computer and Communication Engineering, 7(3), 229–232. https://doi.org/10.17148/IJARCCE.2018.7344

Sidorov, G. (2019). *Syntactic n-grams in Computational Linguistics*. Springer Publishing. https://doi.org/10.1007/978-3-030-14771-6

Statista. (2022, 13 mayo). *México: porcentaje de usuarios por red social 2022*. Recuperado 25 de julio de 2022, de https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/

TensorFlow. (2022, 19 enero). *Introduction to graphs and tf.function | TensorFlow Core*. Recuperado 21 de agosto de 2022, de https://www.tensorflow.org/guide/intro_to_graphs

Tidke, P. (2022, 26 febrero). *Text Data Augmentation in Natural Language Processing with Texattack*. Analytics Vidhya. Recuperado 28 de agosto de 2022, de https://www.analyticsvidhya.com/blog/2022/02/text-data-augmentation-in-natural-language-processing-with-texattack/

Tunstall, L., von Werra, L., Wolf, T., & von Werra, L. (2022). *Natural Language Processing with Transformers*. Van Duuren Media.

Twitter, Inc. (2021a). *Twitter API for Academic Research | Products*. Twitter Developer Platform. https://developer.twitter.com/en/products/twitter-api/academic-research

Twitter, Inc. (2021b). *Twitter API v2 tools & libraries*. Docs | Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries/v2

Vajjala, S., Majumder, B., Surana, H., & Gupta, A. (2020). *Practical Natural Language Processing*. Van Duuren Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Velankar, A., Patil, H., & Joshi, R. (2022, 19 abril). Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi. *arXiv*. Recuperado 1 de agosto de 2022, de https://doi.org/10.48550/arXiv.2204.08669

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), 6382–6388. https://doi.org/10.18653/v1/d19-1670

Wise, S. (2013). *GIS Fundamentals* (2.a ed.). Amsterdam University Press.

Wong, Carolyn. 2021. Analyzing Easy Data Augmentation Techniques for Text Classification. Bachelor's thesis, Harvard College.

Anexos

Esta sección contiene información complementaria al documento.

Anexo 1. Diccionario de palabras

El diccionario de palabras cuenta con 108 palabras las cuales se usaron para extraer tuis:

abuso	balacera	extraviada	perpetrador
acoso	brutalidad	extraviado	persecución
agarraron	bullying	falsificación	pillaron
agravio	captura	fechoría	portación de arma
agresión	cateo	feminicidio	posesión
agresividad	corrupción	fraude	rapto
agresor	crimen	golpes	ratear
allanamiento	criminal	golpear	ratero
amenaza	crueldad	homicida	redad
amenazas	custodia	homicidio	reporte
armas	daño	hostigarían	robar
arrestaron	delictiva	hostigamiento	robaron
arresto	delictivo	hurtar	robo
asaltante	delincuencia	hurto	secuestrar
asaltantes	delincuente	inseguridad	secuestro
asaltar	delincuentes	intimidación	soborno
asaltaron	delito	intruso	suicidio
asalto	demanda	ladrón	tiroteo
asesinato	denuncia	ladrones	transgresión
asesino	desaparecida	maleante	transgresor
atacante	desaparecido	malhechor	violación
ataque	desaparecidos	maltrato	violada
atentado	detención	masacre	violado
atracar	disparo	matanza	violador
atraco	disparos	mato	violar
atrapar	entrada forzada	muerte	violencia
atraparon	estafa	ofensa	vulneración

Anexo 2: Atributos de los datos abiertos

Las carpetas de investigación de la FGJ es un conjunto de datos extenso que está en constante cambio, ya sea porque se actualizan los registros o porque el formato de publicación cambia.

A la fecha marzo 2022 los datos cuentan 27 atributos (Ver Tabla 30), 16 categorías de delitos (Ver Tabla 31) y 305 tipos de delitos (Ver Tabla 32), el diccionario de datos también puede ser consultado en el portal de datos abiertos de la Ciudad de México.

Tabla 30. Diccionario de atributos de los datos abiertos.

Nombre de variable	Definición
idCarpeta	Número entero que representa el identificador único usado por PGJ asociado a cada carpeta de investigación dentro de su sistema.
Delito	Es la conducta, acción u omisión típica (descrita por la ley), antijurídica (contraria a la ley) y culpable, a la que le corresponde una sanción.
Categoría	Las carpetas de investigación se clasifican en función del tipo de delito cometido, los cuales se dividen en 16 tipos.
FechaHecho	Día y hora en que se cometió el delito, según el reporte de la víctima.
HoraHecho	Hora y minuto en que se cometió el delito, según el reporte de la víctima.
FechaInicio	Día y hora en que se hizo la denuncia para iniciar la carpeta de investigación.
HoraInicio	Hora y minuto en que se hizo la denuncia para iniciar la carpeta de investigación.
Año_hecho	Año en que se cometió el delito, según el reporte de la víctima.
Mes_hecho	Mes en que se cometió el delito, según el reporte de la víctima.
Año_inicio	Año en que se hizo la denuncia para iniciar la carpeta de investigación.
Mes_inicio	Mes en que se hizo la denuncia para iniciar la carpeta de investigación.
Sexo	Sexo de la víctima del delito reportado.
Edad	Edad de la víctima del delito reportado en la carpeta de investigación.
TipoPersona	Cómo se reconoce a él o los denunciantes al ser sujetos de obligaciones y derechos salvaguardados legislación penal vigente.
CalidadJuridica	Título con el que se registra y se identifica a la persona, física o moral, en la carpeta de investigación.
Competencia	Variable categórica a través de la cual se clasifican los hechos según su naturaleza. Hechos no delictivos corresponde a aquellos que son denunciados a la PGJ, pero no constituyen un delito en sí mismos como por ejemplo un suicidio. Incompetencias son aquellos hechos delictivos que suceden fuera de la Ciudad de México y son denunciados a la PGJ de la Ciudad de México, por lo que no se deben tomar en cuenta como incidencia delictiva propia de la ciudad. Fuero

	común son los delitos que ocurren y se denuncian dentro de la Ciudad de México.
lon	Longitud de la geolocalización, uno de dos elementos que componen la referencia angular que permite localizar el lugar donde se cometió el delito. WGS84
lat	Latitud de la geolocalización, uno de dos elementos que componen la referencia angular que permite localizar el lugar donde se cometió el delito. WGS84
AlcaldiaHechos	Alcaldía en que se cometió el delito, según el reporte de la víctima. Notar que puede ser fuera de la CDMX.
ColoniaHechos	Colonia en que se cometió el delito, según el reporte de la víctima. Notar que puede ser fuera de la CDMX.

Tabla 31. Categorías de los tipos de delitos en los datos abiertos.

Categorías
Delito de bajo impacto
Hecho no delictivo
Homicidio doloso
Lesiones dolosas por disparo de arma de fuego
Robo a casa habitación con violencia
Robo a cuentahabiente saliendo del cajero con violencia
Robo a negocio con violencia
Robo a pasajero a bordo de microbús con y sin violencia
Robo a pasajero a bordo de taxi con violencia
Robo a pasajero a bordo del metro con y sin violencia
Robo a repartidor con y sin violencia
Robo a transeúnte en vía pública con y sin violencia
Robo a transportista con y sin violencia
Robo de vehículo con y sin violencia
Secuestro
Violación

Tabla 32. Tipos de delitos en el conjunto de datos abierto.

Delito		
Aborto	intimidación	Robo a repartidor con violencia
Abuso de autoridad y uso ilegal de la fuerza publica	Intimidación (evitar denuncia, aporte información o pruebas)	Robo a repartidor sin violencia
Abuso de confianza	La administración de justicia	Robo a repartidor y vehículo con violencia
Abuso sexual	Lenocinio	Robo a repartidor y vehículo sin violencia
Acoso sexual	Lesiones culposas	Robo a sucursal bancaria (supermercado) sin violencia

Acoso sexual agravado en contra de menores	Lesiones culposas accidente laboral	Robo a sucursal bancaria con violencia
Allanamiento de morada, despacho, oficina o establecimiento mercantil	Lesiones culposas con excluyentes de responsabilidad	Robo a sucursal bancaria dentro de tiendas de autoservicio con violencia
Amenazas	Lesiones culposas por caída	Robo a sucursal bancaria dentro de tiendas de autoservicio s/v
Ataque a las vías de comunicación (daño a vías o medios de transporte)	Lesiones culposas por caída de vehículo en movimiento	Robo a sucursal bancaria sin violencia
Ataque a las vías generales de comunicación	Lesiones culposas por quemaduras	Robo a transeúnte a bordo de taxi público y privado con violencia
Ataques a la paz publica	Lesiones culposas por tránsito vehicular	Robo a transeúnte a bordo de taxi público y privado sin violencia
Bigamia	Lesiones culposas por tránsito vehicular en colisión	Robo a transeúnte conductor de taxi público y privado con violencia
Cambio de uso de suelo	Lesiones dolosas por quemaduras	Robo a transeúnte de celular con violencia
Coalición de servidores públicos	Lesiones intencionales	Robo a transeúnte de celular sin violencia
Cobranza ilegitima	Lesiones intencionales por arma blanca	Robo a transeúnte en hotel con violencia
Cohecho	Lesiones intencionales por arma de fuego	Robo a transeúnte en negocio con violencia
Concusión	Lesiones intencionales por golpes	Robo a transeúnte en parques y mercados con violencia
Contagio venéreo	Lesiones intencionales y robo de vehículo	Robo a transeúnte en restaurant con violencia
Contaminación o residuos	Ley federal de armas de fuego y explosivos	Robo a transeúnte en terminal de pasajeros con violencia
Contra el cumplimiento de la obligación alimentaria	Maltrato animal	Robo a transeúnte en vía pública (nomina) con violencia
Contra el estado civil	Narcomenudeo posesión con fines de venta, comercio y suministro	Robo a transeúnte en vía pública (nomina) sin violencia
Contra funcionarios públicos	Narcomenudeo posesión simple	Robo a transeúnte en vía pública con violencia
Contra la intimidad sexual	Negación del servicio publico	Robo a transeúnte en vía pública sin violencia
Contra la ley federal de población	Omisión de auxilio o de cuidado	Robo a transeúnte saliendo del banco con violencia
Corrupción de menores e incapaces	Operaciones con recursos de procedencia ilegal	Robo a transeúnte saliendo del cajero con violencia
Corrupción de personas menores de edad o personas que no tengan capacidad para comprender el significado del hecho o de	Operaciones con recursos de procedencia ilícita	Robo a transeúnte y vehículo con violencia

personas que no tengan capacidad de resistir la conducta		
Daño en propiedad ajena culposa	Oposición a que se ejecute alguna obra o trabajo públicos	Robo a transportista y vehículo pesado con violencia
Daño en propiedad ajena culposa por tránsito vehicular a automóvil	Otros ambientales	Robo a transportista y vehículo pesado sin violencia
Daño en propiedad ajena culposa por tránsito vehicular a bienes inmuebles	Otros delitos	Robo de accesorios de auto
Daño en propiedad ajena culposa por tránsito vehicular a vías de comunicación	Pandilla, asociación delictuosa y delincuencia organizada	Robo de alhajas
Daño en propiedad ajena intencional	Peculado	Robo de animales
Daño en propiedad ajena intencional a automóvil	Peligro de contagio	Robo de arma
Daño en propiedad ajena intencional a bienes inmuebles	Perdida de la vida asfixia por alimentos/ líquidos	Robo de contenedores de tráileres s/v
Daño en propiedad ajena intencional a casa habitación	Perdida de la vida por accidente laboral	Robo de dinero
Daño en propiedad ajena intencional a negocio	Perdida de la vida por ahogamiento	Robo de documentos
Daño en propiedad ajena intencional a vías de comunicación	Perdida de la vida por asfixia	Robo de fluidos
Daño suelo (actividad, invasión o extracción)	Perdida de la vida por caída	Robo de infante
DDH anónimas	Perdida de la vida por congestión alcohólica	Robo de maquinaria con violencia
DDH cerezo	Perdida de la vida por derrumbe	Robo de maquinaria sin violencia
DDH fes	Perdida de la vida por enfermedad	Robo de mercancía a transportista c/v
DDH frvt	Perdida de la vida por envenenamiento	Robo de mercancía en contenedores en áreas federales
DDH incompetencia	Perdida de la vida por intoxicación	Robo de motocicleta con violencia
DDH oficio colaboración	Perdida de la vida por otras causas	Robo de motocicleta sin violencia
DDH otras materias	Perdida de la vida por paro cardiaco	Robo de objetos
DDH redes	Perdida de la vida por precipitación	Robo de objetos a escuela
DDH relacionadas	Perdida de la vida por quemadura	Robo de objetos del interior de un vehículo
DDH sin datos	Perdida de la vida por suicidio	Robo de placa de automóvil
Delitos ambientales	Perdida de la vida por suicidio en el metro	Robo de vehículo de pedales

Delitos contra la salud	Personas extraviadas y ausentes	Robo de vehículo de servicio de transporte con violencia
Delitos de abogados, patronos, litigantes y asesores jurídicos	Plagio o secuestro	Robo de vehículo de servicio de transporte sin violencia
Delitos electorales	Pornografía	Robo de vehículo de servicio oficial con violencia
Denuncia de hechos	Portación de arma de fuego	Robo de vehículo de servicio oficial sin violencia
Denuncia de hechos por robo de celular	Portación, fabricación e importación de objetos aptos para agredir	Robo de vehículo de servicio particular con violencia
Desaparición forzada de personas	Posesión de vehículo robado	Robo de vehículo de servicio particular sin violencia
Desobediencia de particulares	Priv. ilegal de la lib. y robo de vehículo	Robo de vehículo de servicio público con violencia
Desobediencia y resistencia de particulares	Privación de la libertad personal	Robo de vehículo de servicio público sin violencia
Despojo	Privación de la libertad personal (realizar acto sexual)	Robo de vehículo eléctrico monopatín
Difamación	Procreación asistida, inseminación artificial y esterilización forzada	Robo de vehículo en pensión, taller y agencias c/v
Discriminación	Producción, impresión, enajenación, distribución, alteración o falsificación de títulos al portador, documentos de crédito públicos o vales de canje	Robo de vehículo en pensión, taller y agencias s/v
Disparos de arma de fuego	Quebrantamiento de sellos	Robo durante traslado de valores (nomina) con violencia
Ejercicio abusivo de funciones	Regulación urbana	Robo durante traslado de valores (nomina) sin violencia
Ejercicio ilegal y abandono del servicio publico	Responsabilidad profesional y técnica	Robo en eventos masivos (deportivos, culturales, religiosos y artísticos) s/v
Ejercicio indebido del propio der.	Retención de menores	Robo en interior de empresa (nomina) con violencia
Encubrimiento	Retención o sustracción de menores incapaces	Robo en interior de empresa (nomina) sin violencia
Encubrimiento por favorecimiento	Revelación de secretos	Robo s/v dentro de negocios, autoservicios, conveniencia
Encubrimiento por favorecimiento y receptación	Robo a casa habitación con violencia	Sabotaje
Enriquecimiento ilícito	Robo a casa habitación sin violencia	Secuestro express (para cometer robo o extorsión)
Entrega ilegitima de un menor	Robo a casa habitación y vehículo con violencia	Sustracción de menores
Estupro	Robo a casa habitación y	Tala

	vehículo sin violencia	
Evasión de presos	Robo a locales semifijos (puestos de alimentos, bebidas, enseres, periódicos, lotería, otros)	Tentativa de extorsión
Exhortos	Robo a negocio con violencia	Tentativa de feminicidio
Explotación laboral de menores, personas con discapacidad física o mental y adultos mayores	Robo a negocio con violencia por farderos (tiendas de autoservicio)	Tentativa de fraude
Exposición de menores	Robo a negocio con violencia por farderos (tiendas de conveniencia)	Tentativa de homicidio
Extorsión	Robo a negocio sin violencia	Tentativa de robo
Fabricación, comercialización y uso indebido de insignias y uniformes	Robo a negocio sin violencia por farderos	Tentativa de robo de vehículo
Falsedad ante autoridades	Robo a negocio sin violencia por farderos (tiendas de autoservicio)	Tentativa de suicidio
Falsificación de sellos, marcas, llaves, cuños, troqueles, contraseñas y otros	Robo a negocio sin violencia por farderos (tiendas de conveniencia)	Tentativa de violación
Falsificación o alteración y uso indebido de documentos	Robo a negocio y vehículo con violencia	Tortura
Feminicidio	Robo a negocio y vehículo sin violencia	Trafico de infantes
Feminicidio por arma blanca	Robo a oficina pública con violencia	Tráfico de influencia
Feminicidio por disparo de arma de fuego	Robo a oficina pública sin violencia	Trata de personas
Feminicidio por golpes	Robo a pasajero / conductor de vehículo con violencia	Uso de documento falso
Fraude	Robo a pasajero a bordo de cablebus sin violencia	Uso indebido de atribuciones y facultades
Gestión ambiental	Robo a pasajero a bordo de metro con violencia	Usurpación de funciones publicas
Homicidio culposo	Robo a pasajero a bordo de metro sin violencia	Usurpación de identidad
Homicidio culposo con excluyentes de responsabilidad	Robo a pasajero a bordo de Metrobús con violencia	Usurpación de profesión
Homicidio culposo fuera del D.F (atropellado)	Robo a pasajero a bordo de Metrobús sin violencia	Utilización indebida de la vía publica
Homicidio culposo fuera del D.F (colisión)	Robo a pasajero a bordo de pesero colectivo con violencia	Variación de nombre o domicilio
Homicidio culposo por arma de fuego	Robo a pasajero a bordo de pesero colectivo sin violencia	Violación
Homicidio culposo por instrumento punzo cortante	Robo a pasajero a bordo de pesero y vehículo con violencia	Violación a los derechos humanos
Homicidio culposo por tránsito	Robo a pasajero a bordo de	Violación de correspondencia

vehicular	transporte público con violencia	
Homicidio culposo por tránsito vehicular (atropellado)	Robo a pasajero a bordo de transporte público sin violencia	Violación de la intimidad
Homicidio culposo por tránsito vehicular (caída)	Robo a pasajero en autobús foráneo con violencia	Violación equiparada
Homicidio culposo por tránsito vehicular (colisión)	Robo a pasajero en autobús foráneo sin violencia	Violación equiparada por conocido
Homicidio por ahorcamiento	Robo a pasajero en ecobus con violencia	Violación equiparada y robo de vehículo
Homicidio por arma blanca	Robo a pasajero en ecobus sin violencia	Violación tumultuaria
Homicidio por arma de fuego	Robo a pasajero en rtp con violencia	Violación tumultuaria equiparada
Homicidio por golpes	Robo a pasajero en rtp sin violencia	Violación tumultuaria equiparada por conocido
Homicidios intencionales (otros)	Robo a pasajero en tren ligero con violencia	Violación y robo de vehículo
Incesto	Robo a pasajero en tren ligero sin violencia	Violencia familiar
Inhumación, exhumación y respeto a los cadáveres o restos humanos	Robo a pasajero en tren suburbano con violencia	
Injurias	Robo a pasajero en tren suburbano sin violencia	

Anexo 2. Cuadro delimitador

Algunos tuits recolectados contienen un campo de georreferenciación conocido como área delimitadora (*bounding box*) compuesta por cuatro coordenadas que forman el área. Un ejemplo de esto es la Figura 51 que muestra el cuadro delimitador de la Ciudad de México con las coordenadas geográficas [-99.396765,19.035002,-98.923237,19.613237] (realizado con la ayuda de BoundingBox).

Figura 51. Ejemplo de cuadro delimitador en la Ciudad de México.

En la tabla de los datos se identifican los tuits que tienen el atributo bbox y no cuentan con valores en coord, se separan y se procesan para detectar cuales fueron escritos dentro de la Ciudad de México.

El formato de bbox no es reconocido por QGIS al leer un archivo .csv por lo que es necesario convertir estos valores a un polígono en formato de texto conocido (Well Known Text, WKT). El valor de bbox se divide en cuatro columnas (minx, miny, maxx, maxy) con estos valores se aplica un formato y se convierte a WKT (Ver Tabla 33). Al terminar el proceso de conversión entre formatos, los datos se cargan a QGIS.

$$POLYGON((minx\ miny, maxx\ miny, maxx\ maxy, minx\ maxy, minx\ miny))$$

Tabla 33. Formato WKT para generar polígonos.

bbox	minx	miny	maxx	maxy	wkt
[-87.3227, 20.7342, -86.7406, 21.3632]	-87.3227	20.7342	-86.7406	21.3632	POLYGON ((-87.3227 20.7342, -86.7406 20.7342, -86.7406 21.3632, -87.3227 21.3632, -87.3227 20.7342))

La Figura 52 muestra los cuadros delimitadores de todos los datos recolectados como se puede ver muchos de estos cuadros representan áreas generales, es decir, abarcan todo el país y algunas más específicas están fuera de la Ciudad de México. Para encontrar los tuits que fueron publicados dentro del área de interés se calcula el centroide de cada uno de los recuadros.

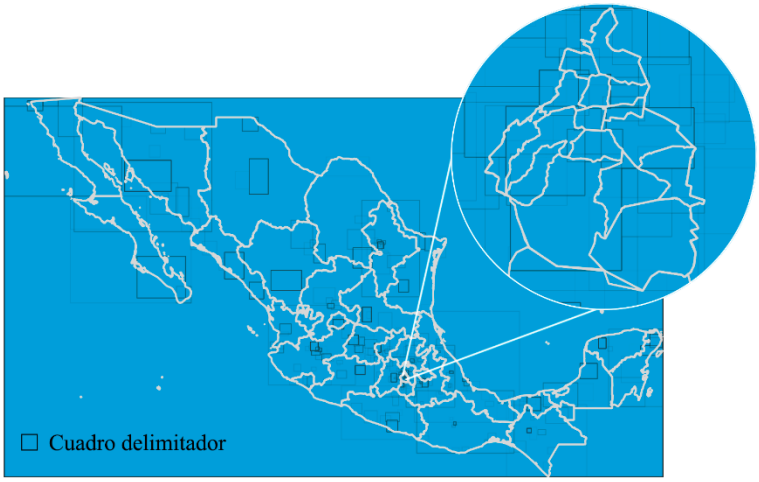

Figura 52. Mapa con los cuadros delimitadores de los tuits antes de ser limpiados.

La Figura 53 muestra los puntos obtenidos después calcular el centroide, para seleccionar los tuis que se encuentran dentro de la ciudad se utiliza la herramienta selección por localización logrando así encontrar 718 registros.

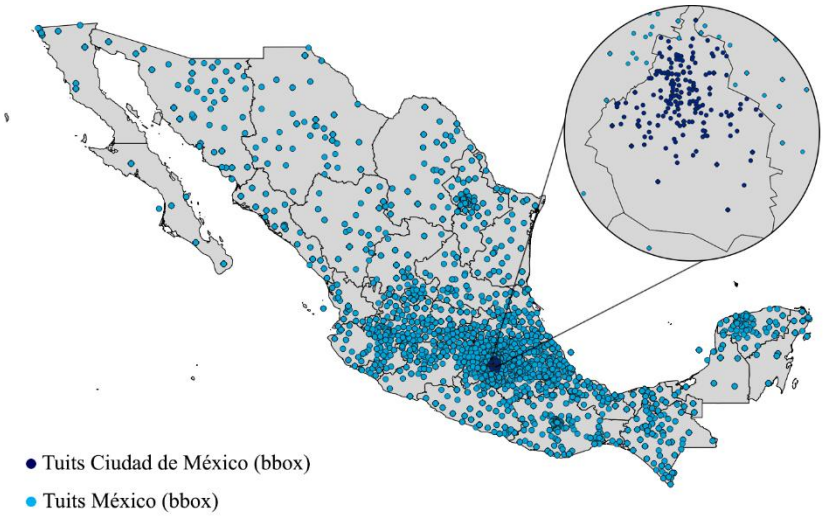

Figura 53. Mapa de México con los puntos donde se publicó un tuit.

Cabe mencionar que después de la selección de los puntos que se encuentran dentro de la Ciudad de México, se calculó el área del cuadro delimitador perteneciente a cada centroide con la intención de validar el área en la que se encuentra y verificar si la zona no es muy extensa y con esto la precisión del punto obtenido.

La Figura 54 presenta el agrupamiento de las áreas de los cuadros delimitadores, de los 718 registros, 258 tienen un área menor a 50 m^2, seguido de 202 registros con un área entre 51-100 metros. Como se puede ver entre mayor es el área menos registros tienen este valor, por lo que pocos datos tienen un margen de error alto. Estos resultados ayudan a corroborar que el uso del centroide representar favorablemente el lugar desde el cual se publicó un tuit.

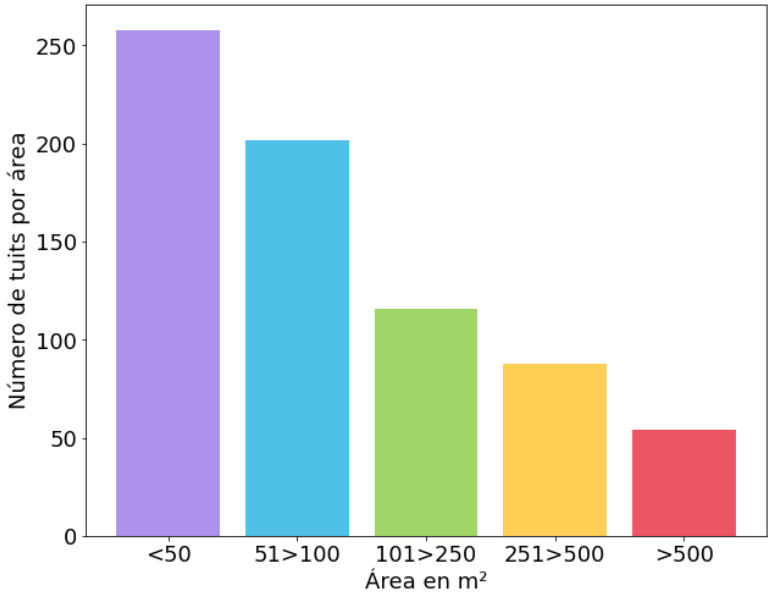

Figura 54. Área de los cuadros delimitadores pertenecientes a los reportes obtenidos de Twitter

Anexo 4: Etiquetadores

Como se menciona en la etapa de etiquetado de datos, los tuis del conjunto de datos de Twitter fueron etiquetado manualmente con la ayuda de tres participantes durante el periodo de cuatro meses contando el tiempo trabajado. Debido a que cada persona inicio en diferentes periodos la cantidad de datos varia, aunque no por mucho. La Figura 55 muestra la cantidad de datos realizados por cada participante, el etiquetador uno reviso 10,095 datos, el etiquetador dos 9,748 y el etiquetador tres 13,402 tuits.

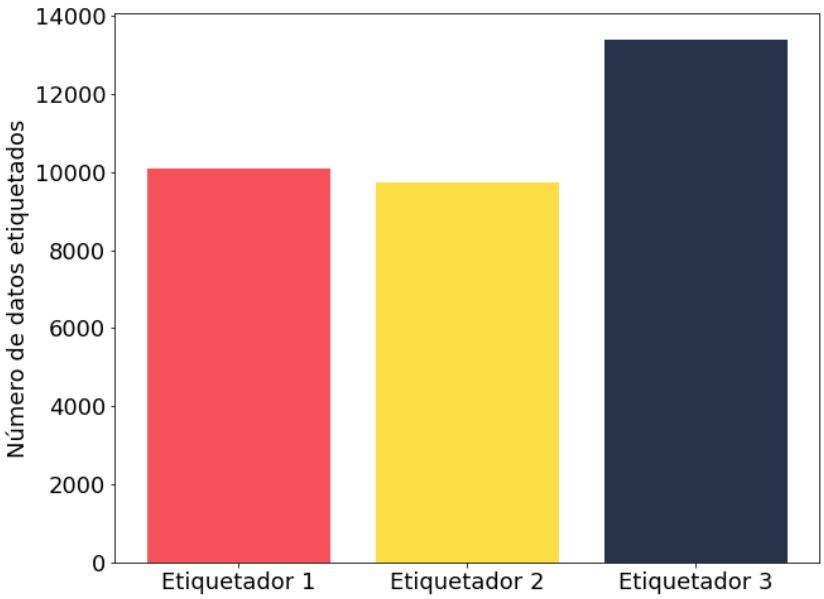

Figura 55. Datos etiquetados por persona (conjunto de datos Twitter).

Anexo 5: Cuadernos de Python

Los cuadernos de Jupyter sirven para escribir y ejecutar código en Python, son una forma de escribir código más interactivo y visual puesto que permite agregar texto e imágenes. A continuación, los siguientes enlaces contiene los cuadernos desarrollados durante el proceso de experimentación.

Extracción de datos utilizando Twitter API

https://colab.research.google.com/drive/1h_9WAMq2sYs9GflnK2PZ7KsGKBuIgTQg?usp=sharing

Procesamiento de datos

https://colab.research.google.com/drive/1kDAZTs_hu1r8DbhUa-GtolSlSsP45r5f?usp=sharing

Geocodificación

https://colab.research.google.com/drive/1HfeBY7Z191480VsG8PmuqOJwJHh2nG4i?usp=sharing

Análisis de los datos

https://colab.research.google.com/drive/1T8by7JmAaI_xqbUYe2WWueqW2PusAtlQ?usp=sharing

Aumento de datos

https://colab.research.google.com/drive/1t_5KW0u1T1kK4FuOZoHpMrWc_IkjuJD7?usp=sharing

Algoritmos de línea base

https://colab.research.google.com/drive/1ep3o5pIMbHkmRz0xYO2qunv4rP6XwaJY?usp=sharing

Modelos propuestos

https://github.com/cpalma0/Analisis-clasificacion-reporte-delictivos/tree/main/models

Conjunto de datos

https://github.com/cpalma0/Analisis-clasificacion-reporte-delictivos/tree/main/dataset

Repositorio del trabajo realizado

https://github.com/cpalma0/Analisis-clasificacion-reporte-delictivos

Anexo 6: Muestra de los atributos del conjunto de datos abiertos

idCarpeta	Año_inicio	Mes_inicio	FechaInicio	Delito	Categoria	Sexo	Edad	TipoPersona	CalidadJuridica
8324429	2019	Enero	04/01/2019	FRAUDE	DELITO DE BAJO IMPACTO	Masculino	62	FISICA	OFENDIDO
8324430	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	38	FISICA	VICTIMA Y DENUNCIA
8324431	2019	Enero	04/01/2019	ROBO A TRANSEUNTE SALIENDO DEL BAN	ROBO A CUENTAHABIENTE	Masculino	42	FISICA	VICTIMA Y DENUNCIA
8324435	2019	Enero	04/01/2019	ROBO DE VEHICULO DE SERVICIO PARTICUI	ROBO DE VEHÍCULO CON Y S	Masculino	35	FISICA	VICTIMA Y DENUNCIA
8324438	2019	Enero	04/01/2019	ROBO DE MOTOCICLETA SIN VIOLENCIA	ROBO DE VEHÍCULO CON Y S	Masculino	NA	FISICA	VICTIMA
8324442	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	42	FISICA	OFENDIDO
8324444	2019	Enero	04/01/2019	ROBO A TRANSEUNTE DE CELULAR SIN VIC	DELITO DE BAJO IMPACTO	Femenino	55	FISICA	VICTIMA Y DENUNCIA
8324451	2019	Enero	04/01/2019	VIOLACION	VIOLACIÓN	Masculino	13	FISICA	VICTIMA
8324454	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	83	FISICA	VICTIMA Y DENUNCIA
8324455	2019	Enero	04/01/2019	OMISION DE AUXILIO O DE CUIDADO	DELITO DE BAJO IMPACTO	Masculino	15	FISICA	VICTIMA
8324457	2019	Enero	04/01/2019	DESPOJO	DELITO DE BAJO IMPACTO	Masculino	45	FISICA	VICTIMA Y DENUNCIA
8324459	2019	Enero	04/01/2019	ROBO A PASAJERO A BORDO DE PESERO C	ROBO A PASAJERO A BORD	Masculino	27	FISICA	VICTIMA Y DENUNCIA
8324465	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	58	FISICA	VICTIMA Y DENUNCIA
8324469	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	31	FISICA	VICTIMA Y DENUNCIA
8324477	2019	Enero	04/01/2019	ROBO DE ACCESORIOS DE AUTO	DELITO DE BAJO IMPACTO	Masculino	61	FISICA	VICTIMA Y DENUNCIA
8324479	2019	Enero	04/01/2019	ROBO A TRANSEUNTE A BORDO DE TAXI P	DELITO DE BAJO IMPACTO	Masculino	39	FISICA	VICTIMA Y DENUNCIA

idCarpeta	Año_inicio	Mes_inicio	FechaInicio	Delito	Categoria	Sexo	Edad	TipoPersona	CalidadJuridica
8324429	2019	Enero	04/01/2019	FRAUDE	DELITO DE BAJO IMPACTO	Masculino	62	FISICA	OFENDIDO
8324430	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	38	FISICA	VICTIMA Y DENUNCIA
8324431	2019	Enero	04/01/2019	ROBO A TRANSEUNTE SALIENDO DEL BAN	ROBO A CUENTAHABIENTE	Masculino	42	FISICA	VICTIMA Y DENUNCIA
8324435	2019	Enero	04/01/2019	ROBO DE VEHICULO DE SERVICIO PARTICUI	ROBO DE VEHÍCULO CON Y S	Masculino	35	FISICA	VICTIMA Y DENUNCIA
8324438	2019	Enero	04/01/2019	ROBO DE MOTOCICLETA SIN VIOLENCIA	ROBO DE VEHÍCULO CON Y S	Masculino	NA	FISICA	VICTIMA
8324442	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	42	FISICA	OFENDIDO
8324444	2019	Enero	04/01/2019	ROBO A TRANSEUNTE DE CELULAR SIN VIC	DELITO DE BAJO IMPACTO	Femenino	55	FISICA	VICTIMA Y DENUNCIA
8324451	2019	Enero	04/01/2019	VIOLACION	VIOLACIÓN	Masculino	13	FISICA	VICTIMA
8324454	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	83	FISICA	VICTIMA Y DENUNCIA
8324455	2019	Enero	04/01/2019	OMISION DE AUXILIO O DE CUIDADO	DELITO DE BAJO IMPACTO	Masculino	15	FISICA	VICTIMA
8324457	2019	Enero	04/01/2019	DESPOJO	DELITO DE BAJO IMPACTO	Masculino	45	FISICA	VICTIMA Y DENUNCIA
8324459	2019	Enero	04/01/2019	ROBO A PASAJERO A BORDO DE PESERO C	ROBO A PASAJERO A BORD	Masculino	27	FISICA	VICTIMA Y DENUNCIA
8324465	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	58	FISICA	VICTIMA Y DENUNCIA
8324469	2019	Enero	04/01/2019	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, I	DELITO DE BAJO IMPACTO	Femenino	31	FISICA	VICTIMA Y DENUNCIA
8324477	2019	Enero	04/01/2019	ROBO DE ACCESORIOS DE AUTO	DELITO DE BAJO IMPACTO	Masculino	61	FISICA	VICTIMA Y DENUNCIA
8324479	2019	Enero	04/01/2019	ROBO A TRANSEUNTE A BORDO DE TAXI P	DELITO DE BAJO IMPACTO	Masculino	39	FISICA	VICTIMA Y DENUNCIA

Anexo 7: Tuits asociados a un lugar

El conjunto de datos final de reportes que se utiliza para realizar el análisis geoespacial, contiene tuits georreferenciados en tres diferentes formas: coordenadas geográficas, cuadro delimitador y dirección mencionada en la publicación por el usuario.

Al realizar la extracción de los datos se especificó que los tuits recolectados utilizando el diccionario de palabras necesitaban contener el atributo de coordenadas por lo que estos contenían un valor especifico de donde ocurrió (734 tuits). sin embargo, los tuits recolectados utilizando cuentas oficiales no contenían este valor por lo que los 2,006 tuits clasificados como reportes ocuparon geocodificación (Ver Tabla 34).

Tabla 34. Tuits por atributos de georreferenciación.

Atributo	Tuits
Coordenadas	16
Cuadro delimitador	718
Dirección en texto	2,006